

A Dictionary of Closed-Form Kernel Mean Embeddings

François-Xavier Briol

Alexandra Gessner

Toni Karvonen

Maren Mahsereci

University College London, United Kingdom

University of Tübingen, Germany¹

Lappeenranta–Lahti University of Technology LUT, Lappeenranta, Finland

Yahoo Research, Berlin, Germany

Abstract

Kernel mean embeddings – integrals of a kernel with respect to a probability distribution – are essential in Bayesian quadrature, but also widely used in other computational tools for numerical integration or for statistical inference based on the maximum mean discrepancy. These methods often require, or are enhanced by, the availability of a closed-form expression for the kernel mean embedding. However, deriving such expressions can be challenging, limiting the applicability of kernel-based techniques when practitioners do not have access to a closed-form embedding. This paper addresses this limitation by providing a comprehensive dictionary of known kernel mean embeddings, along with practical tools for deriving new embeddings from known ones. We also provide a Python library that includes minimal implementations of the embeddings.

Proceedings of the 1st International Conference on Probabilistic Numerics (ProbNum) 2025, Sophia Antipolis, France. PMLR Volume 271. Copyright 2025 with the author(s)

1 Introduction

Let P be a probability measure on a subset Ω of \mathbb{R}^d , and $K : \Omega \times \Omega \rightarrow \mathbb{R}$ be a reproducing kernel (Berlinet and Thomas-Agnan, 2004) satisfying $\int_{\Omega} K(x, x) dP(x) < \infty$. This paper considers two important quantities for kernel-based methods: the *kernel mean embedding* K_P and its integral K_{PP} , given by

$$K_P(x) = \int_{\Omega} K(x, y) dP(y), \quad (1)$$

$$K_{PP} = \int_{\Omega} \int_{\Omega} K(x, y) dP(x) dP(y). \quad (2)$$

These quantities are ubiquitous in kernel methods, but are also needed for implementing Bayesian quadrature, a probabilistic numerical method for integration which motivates this paper (Hennig et al., 2022). They are also needed for many other kernel-based numerical integration techniques, and for statistical inference. Unfortunately, K_P and K_{PP} are often tedious, difficult, or even impossible to compute in closed form, limiting the applicability of these algorithms. The primary purpose of this article is therefore to collect K_P and K_{PP} for a number of common pairs of K and P . We have implemented the embeddings in a Python library (https://github.com/mmahsereci/kernel_embedding_dictionary).

The paper is structured as follows. In Section 2, we introduce the broad range of kernel-based techniques which rely on closed-form expressions for kernel mean embeddings. Section 3 then provides a *dictionary of kernel mean embeddings* for common pairs of kernel and distribution. Section 4 discusses common techniques for deriving new closed-form expressions for kernel mean embeddings. Section 5 introduces our Python library. Most

of the embeddings in Section 3 are not new (e.g., Briol et al., 2019b, Table 1). However, they have never been collated in one place and are currently found in disparate sources throughout the machine learning, statistics, signal processing, and numerical analysis literatures. As a result many a hapless researcher has had to rederive these embeddings over the years.

2 Uses of Kernel Embeddings

The following algorithms use kernel mean embeddings.

Bayesian quadrature. Numerical integration is the computational task of approximating an integral

$$I(f) = \int_{\Omega} f(x) dP(x) \quad (3)$$

using evaluations of f at points $X = \{x_1, \dots, x_n\} \subset \Omega$. A natural approach in this context is to use a quadrature rule $\hat{I}(f) = \sum_{i=1}^n w_i f(x_i)$, where each function evaluation $f(x_i)$ is assigned a weight $w_i \in \mathbb{R}$.

Bayesian quadrature (O’Hagan, 1991; Briol et al., 2019b) is a probabilistic numerical method for integration. It typically models f as a zero-mean Gaussian process $GP(0, K)$ with covariance kernel K that encodes prior knowledge such as differentiability, periodicity, or sparsity. Conditioning the prior on data $\mathcal{D} = \{X, Y\}$ consisting of evaluations $Y = (f(x_1), \dots, f(x_n))^T$ of f at some pairwise distinct nodes X yields a posterior for the integral:

$$I(f) \mid \mathcal{D} \sim \mathcal{N}(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}^2), \quad \begin{aligned} \mu_{\mathcal{D}} &= m^T C^{-1} Y, \\ \sigma_{\mathcal{D}}^2 &= K_{PP} - m^T C^{-1} m. \end{aligned} \quad (4)$$

Here, $C = (K(x_i, x_j))_{i,j=1}^n$ is an $n \times n$ positive-definite covariance matrix, and $m = (K_P(x_1), \dots, K_P(x_n))^T$ a vector of kernel mean evaluations. The name Bayesian

¹At the time of publication, AG is an employee at the Centre for Artificial Intelligence, AstraZeneca, Barcelona, Spain. However, this work has been carried out while employed at the University of Tübingen.

quadrature is justified by the posterior mean being a quadrature rule with $w = (w_1, \dots, w_n)^\top = m^\top C^{-1}$. The computation of K_P and K_{PP} is a major challenge in the implementation of Bayesian quadrature.

Integration in Kernel Spaces. The quantities K_P and K_{PP} also play a key role in other quadrature rules. Given a reproducing kernel K , denote by \mathcal{H} the reproducing kernel Hilbert space (RKHS) associated with K and by $\|\cdot\|_{\mathcal{H}}$ the norm of \mathcal{H} . Assuming that $f \in \mathcal{H}$, we can bound the error of an arbitrary quadrature rule $\hat{I}(f) = \sum_{i=1}^n w_i f(x_i)$ as

$$|I(f) - \hat{I}(f)| \leq \|f\|_{\mathcal{H}} \sup_{\|f\|_{\mathcal{H}} \leq 1} |I(f) - \hat{I}(f)|. \quad (5)$$

The second term $\text{WCE} := \sup_{\|f\|_{\mathcal{H}} \leq 1} |I(f) - \hat{I}(f)|$ is called the *worst-case (integration) error* (WCE) and has the straightforward expression (e.g., Briol et al., 2019b)

$$\sqrt{K_{PP} - 2 \sum_{i=1}^n w_i K_P(x_i) + \sum_{i=1}^n \sum_{j=1}^n w_i w_j K(x_i, x_j)}. \quad (6)$$

The WCE can be computed when K_P and K_{PP} have known expressions. The weights minimising the WCE can also be derived and are identical to the Bayesian quadrature weights. This non-Bayesian construction is typically called kernel quadrature (Sommariva and Vianello, 2006; Fuselier et al., 2014; Belhadji et al., 2019; Kanagawa et al., 2020; Epperly and Moreno, 2023).

Many other quadrature rules are constructed or analysed using the WCE and therefore require closed-form expressions for K_P or K_{PP} . For example, quasi-Monte Carlo (Niederreiter, 1992; Dick and Pillichshammer, 2010; Dick et al., 2013) is a set of quadrature rules for P being a uniform measure. It uses equal weights and point sets selected to guarantee that the WCE decreases at a fast rate in n , and evaluations of the WCE are often returned as computable guarantees on the performance of the method. Relatedly, kernel herding (Chen et al., 2010; Bach et al., 2012; Lacoste-Julien et al., 2015) also uses equal weights, but selects points by directly minimising the WCE by repeatedly evaluating it.

Beyond these, many algorithms aim to minimise the WCE in \mathcal{H} but rely on sample-based approximations of K_P and/or K_{PP} . These algorithms are typically designed this way so as to make them more widely applicable, but they would most likely benefit from access to closed-form expressions. Examples include gradient flows (Arbel et al., 2019; Hertrich et al., 2024; Chen et al., 2024a; see also Xu et al., 2022; Chen et al., 2025 and Belhadji et al., 2025 for the benefits of closed-forms), thinning algorithms (Dwivedi and Mackey, 2024), and quadrature rules based on leverage scores (Chatalic et al., 2023).

Statistical Inference with Kernel Embeddings. The maximum mean discrepancy (MMD; Gretton et al., 2012) is a probability metric under mild condition on the kernel (Sriperumbudur et al., 2010). It compares two distribution by the magnitude of the difference in their mean embedding measured in the RKHS norm.

The MMD admits a straightforward expression:

$$\begin{aligned} \text{MMD}^2(P, Q) &:= \|K_P - K_Q\|_{\mathcal{H}}^2 \\ &= K_{PP} - 2K_{PQ} + K_{QQ} \end{aligned} \quad (7)$$

where $K_{PQ} = \int_{\Omega} \int_{\Omega} K(x, y) dP(x) dQ(y)$. Note that when Q is the empirical measure $Q_n := \sum_{i=1}^n w_n \delta_{x_i}$, the MMD becomes exactly the WCE from Equation (6).

The convenient expression of the MMD has led to multitudes of applications of kernel mean embeddings (Muandet et al., 2016a). For example, goodness-of-fit testing can be performed using $\text{MMD}^2(P, Q_n)$ as a test statistic (Lloyd and Ghahramani, 2015; Kellner and Celisse, 2019), where P is the model under the null hypothesis and Q_n the observed data. Another example is parametric estimation through minimum distance estimation (Briol et al., 2019a; Chérif-Abdellatif and Alquier, 2022; Alquier and Gerber, 2023), approximate Bayesian computation, generalised Bayesian inference (Chérif-Abdellatif and Alquier, 2020; Dellaporta et al., 2022; Pacchiardi et al., 2024), or variational inference (Huang et al., 2023), which typically make use of evaluations of $\text{MMD}^2(P_{\theta}, Q_n)$, where P_{θ} is the parametric model and Q_n the observed data.

Many other algorithms in statistics and machine learning either have limited applicability due to the lack of known expressions for K_P and K_{PP} , or require that these quantities are approximated through samples, which introduces additional error (Chamakh and Szabó, 2024). This includes two-sample testing (Gretton et al., 2012), causal inference (Singh et al., 2019, 2024; Muandet et al., 2021; Sejdinovic, 2024), density estimation (Song et al., 2008), the analysis of variance (Durrande et al., 2013), linked emulation (Ming and Guillas, 2021), kernel Bayes (Fukumizu et al., 2013), and the construction of auto-encoders (Tolstikhin et al., 2018; Rustamov, 2021) and generative adversarial networks (Dziugaite et al., 2015).

3 Dictionary of Embeddings

Now that we have highlighted the importance of K_P and K_{PP} , this section collates expressions for some commonly used kernels and distributions. We focus primarily on the uniform and Gaussian, though other formulae exist in the literature (e.g., Nishiyama and Fukumizu, 2016; Nishiyama et al., 2020). Our expressions are summarised in Table 1. In absence of references we only sketch the derivations, which are in most cases tedious and uninteresting. Our Python library and its tests can be thought of as “numerical proofs” of the identities here.

Uniform Distribution. Let $a_i < b_i$ for $i = 1, \dots, d$ and define $r_i = b_i - a_i > 0$. The *uniform measure* on $\Omega = [a_1, b_1] \times \dots \times [a_d, b_d] \subset \mathbb{R}^d$ has density

$$p(x) = \prod_{i=1}^d (b_i - a_i)^{-1} = \prod_{i=1}^d r_i^{-1}. \quad (8)$$

Table 1: References to the equations for kernel mean embeddings of kernel-measure pairs.

Kernel	Degree	Measure P	Domain Ω	K_P	K_{PP}
Gaussian		Uniform	$\prod_{i=1}^d [a_i, b_i]$	(11)	(12)
		Gaussian	\mathbb{R}^d	(13)	(14)
Matérn	$n + 1/2$	Uniform	$[a, b]$	(23)	(26)
	$1/2, 3/2, 5/2, 7/2$	Uniform	$[a, b]$	(28), (29), (30), (31)	(34), (35)
	$1/2, 3/2, 5/2$	Gaussian	\mathbb{R}	(36), (37), (38)	?, ?, ?
Wendland	0	Uniform	$[a, b]$	(42)	(43)
	0, 2	Gaussian	\mathbb{R}	(44), (45)	?, ?
Brownian Motion	H	Uniform	$[a, b]$	(47)	(48)
Power series		Uniform	$\prod_{i=1}^d [a_i, b_i]$	(50)	(51)
		Gaussian	\mathbb{R}^d	(52)	(53)
Sobolev	$3/2$	Uniform	\mathbb{S}^2	(55)	(55)
	∞	Uniform	\mathbb{S}^2	(57)	(57)
Periodic Sobolev	$2r$	Uniform	\mathbb{S}^1 or $[0, 1]$	(60)	(60)
Langevin Stein		Unnormalised	\mathbb{R}^d	(71)	(71)

Gaussian Distribution. Let $\Sigma \in \mathbb{R}^{d \times d}$ be a positive-definite matrix and $\mu \in \mathbb{R}^d$. A *Gaussian measure* on $\Omega = \mathbb{R}^d$ has the density

$$p(x) = C(d, \Sigma) \cdot \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right), \quad (9)$$

where $C(d, \Sigma) = (2\pi)^{-d/2}(\det \Sigma)^{-1/2}$. The *centered* Gaussian distribution has $\mu = 0$. The *isotropic Gaussian distribution* has a diagonal covariance $\Sigma = \sigma^2 I$ for $\sigma > 0$.

3.1 Gaussian Kernel

Let $\Lambda \in \mathbb{R}^{d \times d}$ be a positive-definite length-scale matrix. The *Gaussian kernel* is defined for any $x, y \in \mathbb{R}^d$ as

$$K(x, y) = \exp\left(-\frac{1}{2}(x - y)^\top \Lambda^{-1}(x - y)\right). \quad (10)$$

Uniform Distribution. Consider the uniform distribution in (8) and suppose that the length-scale matrix is $\Lambda = \text{diag}(\ell_1^2, \dots, \ell_d^2)$ for $\ell_1, \dots, \ell_d > 0$. Then

$$K_P(x) = \left(\frac{\pi}{2}\right)^{\frac{d}{2}} \prod_{i=1}^d \frac{\ell_i}{r_i} \left[\text{erf}\left(\frac{b_i - x_i}{\ell_i \sqrt{2}}\right) - \text{erf}\left(\frac{a_i - x_i}{\ell_i \sqrt{2}}\right) \right], \quad (11)$$

$$K_{PP} = (2\pi)^{\frac{d}{2}} \prod_{i=1}^d \frac{\ell_i}{r_i^2} \left[\frac{\ell_i \sqrt{2}}{\sqrt{\pi}} \left(\exp\left(-\frac{r_i^2}{2\ell_i^2}\right) - 1 \right) + r_i \text{erf}\left(\frac{r_i}{\ell_i \sqrt{2}}\right) \right], \quad (12)$$

where $\text{erf}(x) = (\pi)^{-1/2} \int_{-x}^x \exp(-t^2) dt$ is the error function. It is straightforward to derive (11) and (12) from the definition of the error function, but we are not aware of closed-form expressions for non-diagonal Λ .

Gaussian Distribution. Consider the Gaussian distribution in (9). Then

$$K_P(x) = \det(I + \Sigma \Lambda^{-1})^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(x - \mu)^\top (\Lambda + \Sigma)^{-1}(x - \mu)\right), \quad (13)$$

$$K_{PP} = \det(I + \Sigma \Lambda^{-1})^{-\frac{1}{2}} \det(I + \Sigma(\Lambda + \Sigma)^{-1})^{-\frac{1}{2}} = \sqrt{\frac{\det(\Lambda)}{\det(\Lambda + 2\Sigma)}}. \quad (14)$$

If Λ and Σ are diagonal with diagonal elements ℓ_i^2 and σ_i^2 , respectively, the expressions simplify to

$$K_P(x) = \prod_{i=1}^d \sqrt{\frac{\ell_i^2}{\ell_i^2 + \sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2(\ell_i^2 + \sigma_i^2)}\right) \quad (15)$$

$$K_{PP} = \prod_{i=1}^d \sqrt{\frac{\ell_i^2}{\ell_i^2 + 2\sigma_i^2}}. \quad (16)$$

Equations (13) and (14) are derived from the usual completion of the square trick. The embedding for log-Gaussian distributions can be obtained by using a log-transformation on the Gaussian kernel as per [Chen et al. \(2024b\)](#).

3.2 Matérn Kernels

Let ℓ be a positive length-scale parameter. The *Matérn kernel* of order $\nu > 0$ is

$$K^\nu(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} \tau)^\nu K_\nu(\sqrt{2\nu} \tau), \quad (17)$$

with $\tau = \|x - y\|_2 / \ell$ and $x, y \in \mathbb{R}^d$, where K_ν is the modified Bessel function of the second kind of order ν and Γ is the gamma function.

For $\nu = n + 1/2$, $n \in \mathbb{N}_0$, we have the simpler form:

$$K^{n+\frac{1}{2}}(x, y) = \exp(-\sqrt{2n+1}\tau) \frac{n!}{(2n)!} \times \sum_{k=0}^n \frac{(n+k)!}{k!(n-k)!} (2\sqrt{2n+1}\tau)^{n-k}. \quad (18)$$

For $n \in \{0, 1, 2, 3\}$, this gives

$$K^{\frac{1}{2}}(x, y) = \exp(-\tau), \quad (19)$$

$$K^{\frac{3}{2}}(x, y) = (1 + \sqrt{3}\tau) \exp(-\sqrt{3}\tau), \quad (20)$$

$$K^{\frac{5}{2}}(x, y) = \left(1 + \sqrt{5}\tau + \frac{5}{3}\tau^2\right) \exp(-\sqrt{5}\tau), \quad (21)$$

$$K^{\frac{7}{2}}(x, y) = \left(1 + \sqrt{7}\tau + \frac{14}{5}\tau^2 + \frac{7^{\frac{3}{2}}}{15}\tau^3\right) \exp(-\sqrt{7}\tau). \quad (22)$$

Uniform Distribution ($d = 1$). Consider the uniform distribution in (8) on $\Omega = [a, b] \subset \mathbb{R}$ with $a < b$. Let $r = b - a > 0$. Then

$$K_P^{n+\frac{1}{2}}(x) = \frac{\alpha_n}{r} \cdot \frac{n!}{(2n)!} \left[2c_{n,0} - Q_n\left(\frac{x-a}{\alpha_n}\right) - Q_n\left(\frac{b-x}{\alpha_n}\right) \right], \quad (23)$$

where

$$\alpha_n = \frac{\ell}{\sqrt{2n+1}}, \quad Q_n(z) = e^{-z} \sum_{m=0}^n c_{n,m} z^m \quad (24)$$

and

$$c_{n,m} = \frac{1}{m!} \sum_{i=0}^{n-m} \frac{(n+i)!}{i!} 2^{n-i}. \quad (25)$$

This formula is obtained from a formula for $\Omega = [0, 1]$ in Section 9 of [Ginsbourger et al. \(2016\)](#) with a change of variables. From an additional change of variables and the formula $\gamma(m+1, x) = \int_0^x t^m e^{-t} dt = m!(1 - e^{-x} \sum_{i=0}^m x^i/i!)$ for the lower incomplete gamma function at $m \in \mathbb{N}_0$ it follows that

$$K_{PP}^{n+\frac{1}{2}} = \frac{2\alpha_n^2}{r^2} \cdot \frac{n!}{(2n)!} \left[\frac{r}{\alpha_n} c_{n,0} - \sum_{m=0}^n c_{n,m} \gamma_{m+1} \right], \quad (26)$$

where

$$\gamma_{m+1} = m! \left[1 - \exp\left(-\frac{r}{\alpha_n}\right) \sum_{i=0}^m \frac{1}{i!} \left(\frac{r}{\alpha_n}\right)^i \right]. \quad (27)$$

Let $d_n(x, y) = (x - y)/\alpha_n$ and $\rho_n = r/\alpha_n$. Then, for $n \in \{0, 1, 2, 3\}$ the embeddings are

$$K_P^{\frac{1}{2}}(x) = \frac{1}{\rho_0} \left[2 - e^{d_0(a,x)} - e^{d_0(x,b)} \right], \quad (28)$$

$$K_P^{\frac{3}{2}}(x) = \frac{1}{\rho_1} \left[4 - e^{d_1(x,b)} (2 - d_1(x, b)) - e^{d_1(a,x)} (2 - d_1(a, x)) \right], \quad (29)$$

$$K_P^{\frac{5}{2}}(x) = \frac{1}{3\rho_2} \left[16 - e^{d_2(x,b)} (8 - 5d_2(x, b) + d_2(x, b)^2) - e^{d_2(a,x)} (8 - 5d_2(a, x) + d_2(a, x)^2) \right], \quad (30)$$

$$K_P^{\frac{7}{2}}(x) = \frac{1}{15\rho_3} (96 - h(x, b) - h(a, x)). \quad (31)$$

where $h(x, y) = e^{d_3(x,y)} (48 - 33d_3(x, y) + 9d_3(x, y)^2 - d_3(x, y)^3)$. The integrals of these four embeddings are

$$K_{PP}^{\frac{1}{2}} = \frac{2}{\rho_0^2} [\rho_0 - 1 + e^{-\rho_0}], \quad (32)$$

$$K_{PP}^{\frac{3}{2}} = \frac{2}{\rho_1^2} [2\rho_1 - 3 + e^{-\rho_1} (\rho_1 + 3)], \quad (33)$$

$$K_{PP}^{\frac{5}{2}} = \frac{2}{3\rho_2^2} [8\rho_2 - 15 + e^{-\rho_2} (\rho_2^2 + 7\rho_2 + 15)], \quad (34)$$

$$K_{PP}^{\frac{7}{2}} = \frac{2}{15\rho_3^2} [3(16\rho_3 - 35) + e^{-\rho_3} (\rho_3^3 + 12\rho_3^2 + 57\rho_3 + 105)]. \quad (35)$$

Gaussian Distribution ($d = 1$). Consider the centered univariate Gaussian distribution in (9) with $\Sigma = \sigma^2$ and let $\Phi(x) = \frac{1}{2}[1 + \text{erf}(x/\sqrt{2})]$ denote the cumulative distribution function of the standard normal distribution. [Ming and Guillas \(2021\)](#) have derived the following three mean embeddings:

$$K_P^{\frac{1}{2}}(x) = \exp\left(\frac{\sigma^2 + 2\ell(x - \mu)}{2\ell^2}\right) \Phi\left(\frac{\mu - \frac{\sigma^2}{\ell} - x}{\sigma}\right) + \exp\left(\frac{\sigma^2 - 2\ell(x - \mu)}{2\ell^2}\right) \Phi\left(\frac{x - \mu - \frac{\sigma^2}{\ell}}{\sigma}\right) \quad (36)$$

$$K_P^{\frac{3}{2}}(x) = \exp\left(\frac{3\sigma^2 + 2\sqrt{3}\ell(x - \mu)}{2\ell^2}\right) \times \left[\left(1 - \frac{\sqrt{3}(x - \mu_1)}{\ell}\right) \Phi\left(\frac{\mu_1 - x}{\sigma}\right) + \sqrt{\frac{3\sigma^2}{2\pi\ell^2}} \exp\left(-\frac{(\mu_1 - x)^2}{2\sigma^2}\right) \right] + \exp\left(\frac{3\sigma^2 - 2\sqrt{3}\ell(x - \mu)}{2\ell^2}\right) \times \left[\left(1 + \frac{\sqrt{3}(x - \mu_2)}{\ell}\right) \Phi\left(\frac{x - \mu_2}{\sigma}\right) + \sqrt{\frac{3\sigma^2}{2\pi\ell^2}} \exp\left(-\frac{(x - \mu_2)^2}{2\sigma^2}\right) \right], \quad (37)$$

where $\mu_1 = \mu - \sqrt{3}\sigma^2/\ell$ and $\mu_2 = \mu + \sqrt{3}\sigma^2/\ell$, and

$$K_P^{\frac{5}{2}}(x) = \exp\left(\frac{5\sigma^2 + 2\sqrt{5}\ell(x - \mu)}{2\ell^2}\right) \times \left[\left(1 - \frac{\sqrt{5}(x - \mu_3)}{\ell} + \frac{5((x - \mu_3)^2 + \sigma^2)}{3\ell^2}\right) \Phi\left(\frac{\mu_3 - x}{\sigma}\right) + \left(\frac{\sqrt{5}}{\ell} + \frac{5(\mu_3 - x)}{3\ell^2}\right) \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(\mu_3 - x)^2}{2\sigma^2}\right) \right] + \exp\left(\frac{5\sigma^2 - 2\sqrt{5}\ell(x - \mu)}{2\ell^2}\right) \times \left[\left(1 + \frac{\sqrt{5}(x - \mu_4)}{\ell} + \frac{5((x - \mu_4)^2 + \sigma^2)}{3\ell^2}\right) \Phi\left(\frac{x - \mu_4}{\sigma}\right) + \left(\frac{\sqrt{5}}{\ell} + \frac{5(x - \mu_4)}{3\ell^2}\right) \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_4)^2}{2\sigma^2}\right) \right], \quad (38)$$

where $\mu_3 = \mu - \sqrt{5}\sigma^2/\ell$ and $\mu_4 = \mu + \sqrt{5}\sigma^2/\ell$. See [Ming and Guillas \(2021\)](#) for more related formulae.

3.3 Wendland Kernels

Wendland kernels (Wendland, 1995) are compactly supported kernels on \mathbb{R}^d and give rise to sparse Gram matrices. Let ℓ be a positive length-scale parameter. The first three even-order Wendland kernels are given by

$$K^0(x, y) = (1 - \tau)_+, \quad (39)$$

$$K^2(x, y) = (1 - \tau)_+^3 (3\tau + 1) \quad (40)$$

$$K^4(x, y) = (1 - \tau)_+^5 (8\tau^2 + 5\tau + 1) \quad (41)$$

for $x, y \in \mathbb{R}^d$, where $\tau = \|x - y\|/\ell$ and $(a)_+ := \max(0, a)$. Because the kernel is defined piecewise, the mean embeddings are exceptionally unwieldy. We only include selected embeddings for kernels of orders zero and two. We have computed these embeddings with Mathematica.

Uniform Distribution ($d = 1$). Consider the uniform distribution in (8) on $\Omega = [a, b] \subset \mathbb{R}$ with $a < b$. Denote $r = b - a > 0$. Then

$$K_P^0(x) = \begin{cases} \frac{\ell}{r} & \text{if } b \geq x + \ell \text{ and } a + \ell < x, \\ \frac{1}{2r\ell} [2x(a + \ell) + \ell^2 - a^2 - 2a\ell - x^2] & \text{if } b \geq x + \ell \text{ and } a + \ell \geq x, \\ \frac{1}{2r\ell} [2b(\ell + x) + \ell^2 - b^2 - 2\ell x - x^2] & \text{if } b < x + \ell \text{ and } a + \ell < x, \\ \frac{1}{2r\ell} [2(b\ell + bx + ax) - a^2 - b^2 - 2(a\ell + x^2)] & \text{otherwise} \end{cases} \quad (42)$$

and

$$K_{PP}^0 = \begin{cases} \frac{1}{3r^2} \ell (3r - \ell) & \text{if } \ell < r, \\ 1 - \frac{1}{3\ell} r & \text{if } \ell \geq r. \end{cases} \quad (43)$$

Similar expressions are available for Wendland kernels of higher order, but these are omitted due to their complexity.

Gaussian Distribution ($d = 1$). Consider the centered univariate Gaussian distribution in (9) with $\Sigma = \sigma^2$ and let $\text{erf}(x) = (\pi)^{-\frac{1}{2}} \int_{-x}^x \exp(-t^2) dt$ again denote the standard error function, $\varphi(x) = \exp(-x^2/(2\sigma^2))$ the unnormalised Gaussian density, and $s = \sqrt{2}\sigma$. Then

$$K_P^0(x) = \frac{1}{2\ell} \left[(\ell - x) \text{erf}\left(\frac{\ell - x}{s}\right) + (\ell + x) \text{erf}\left(\frac{\ell + x}{s}\right) + \frac{s}{\sqrt{\pi}} [\varphi(\ell - x) + \varphi(\ell + x) - 2\varphi(x)] - 2x \text{erf}\left(\frac{x}{s}\right) \right], \quad (44)$$

$$K_P^2(x) = \frac{1}{2\ell^4}$$

$$\times \left[\frac{s}{\sqrt{\pi}} \left[\left(\varphi(x - \ell) + \varphi(x + \ell) \right) \left(\ell^3 - \ell(7\sigma^2 + 5x^2) \right) + 16\ell(2\sigma^2 + x^2)\varphi(x) - \left(\varphi(x + \ell) - \varphi(x - \ell) \right) \left(\ell^2 x + 3x(5\sigma^2 + x^2) \right) \right] + \left[\ell^4 - 6\ell^2(\sigma^2 + x^2) + 8\ell(3\sigma^2 x + x^3) \right] \right]$$

$$\begin{aligned} & - 3(3\sigma^4 + 6\sigma^2 x^2 + x^4) \left] \text{erf}\left(\frac{\ell - x}{s}\right) \right. \\ & + \left[\ell^4 - 6\ell^2(\sigma^2 + x^2) - 8\ell(3\sigma^2 x + x^3) \right. \\ & \left. - 3(3\sigma^4 + 6\sigma^2 x^2 + x^4) \left] \text{erf}\left(\frac{\ell + x}{s}\right) \right. \\ & \left. + 16\ell x(3\sigma^2 + x^2) \text{erf}\left(\frac{x}{s}\right) \right]. \end{aligned} \quad (45)$$

Again, similar but more complex expressions exist for higher-order Wendland kernels. We have not found or been able to compute the integrals of the mean embeddings.

3.4 Fractional Brownian Motion Kernels

Let $d = 1$ and $\Omega = [a, b]$ for $b > a > 0$. The *Brownian motion kernel* is $K^{1/2}(x, y) = \min\{x, y\}$. Let $H \in (0, 1)$ be a parameter known as Hurst index. The Brownian motion kernel is obtained by setting $H = 1/2$ in the family of *fractional Brownian motion kernels*

$$K^H(x, y) = \frac{1}{2} (|x|^{2H} + |y|^{2H} - |x - y|^{2H}). \quad (46)$$

Uniform Distribution ($d = 1$). Consider the uniform distribution in (8) on $\Omega = [a, b]$. Set $h = 2H + 1$. Then

$$K_P^H(x) = \frac{b^h - a^h - (b - x)^h - (x - a)^h}{2h(b - a)} + \frac{x^{h-1}}{2}, \quad (47)$$

$$K_{PP}^H = \frac{(h + 1)(b^h - a^h) - (b - a)^h}{h(h + 1)(b - a)}. \quad (48)$$

These are obtained via straightforward integration. Note that the above corresponds to a Brownian motion with zero boundary at $x = 0$, but a version with zero boundary at $x = 1$ is sometimes also used in the QMC literature. See Section 2.4 of (Dick and Pillichshammer, 2010) for details and the expressions of K_P and K_{PP} .

3.5 Power Series Kernels

Let $c_\alpha \in \mathbb{R}$. A *power series kernel* has the form

$$K(x, y) = \sum_{\alpha \in \mathbb{N}_0^d} c_\alpha x^\alpha y^\alpha \quad \text{for } x, y \in \mathbb{R}^d, \quad (49)$$

where the sum is over d -dimensional non-negative multi-indices, operations on which are defined in the usual way. Setting $c_\alpha = 1$ for $|\alpha| = 1$ and $c_\alpha = 0$ otherwise gives the linear kernel $K(x, y) = \langle x, y \rangle_2 = x^\top y$.

Uniform Distribution. Consider the d -dimensional uniform distribution in (8). By integrating polynomials we get

$$K_P(x) = \sum_{\alpha \in \mathbb{N}_0^d} x^\alpha c_\alpha \prod_{i=1}^d \frac{b_i^{\alpha_i+1} - a_i^{\alpha_i+1}}{(\alpha_i + 1)(b_i - a_i)}, \quad (50)$$

$$K_{PP} = \sum_{\alpha \in \mathbb{N}_0^d} c_\alpha \left(\prod_{i=1}^d \frac{b_i^{\alpha_i+1} - a_i^{\alpha_i+1}}{(\alpha_i + 1)(b_i - a_i)} \right)^2. \quad (51)$$

Gaussian Distribution (Diagonal). Consider the centered Gaussian distribution in (9) with covariance

$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$. Since Σ is diagonal, the formula for central moments of a univariate Gaussian yields

$$K_P(x) = \sum_{\alpha \in 2\mathbb{N}_0^d} x^\alpha c_\alpha \prod_{i=1}^d \sigma_i^{\alpha_i} (\alpha_i - 1)!!, \quad (52)$$

$$K_{PP} = \sum_{\alpha \in 2\mathbb{N}_0^d} c_\alpha \left(\prod_{i=1}^d \sigma_i^{\alpha_i} (\alpha_i - 1)!! \right)^2, \quad (53)$$

where $2\mathbb{N}_0^d$ denotes the set of multi-indices with even elements and $n!! = 1 \cdot 3 \cdots (n-2)n$. Isserlis' theorem could be used to compute K_P and K_{PP} for general Σ .

3.6 Stationary Kernels on Spheres

Let $\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$ denote the d -dimensional unit sphere and let P be the uniform spherical measure on \mathbb{S}^d . In this case many stationary kernels have constant embeddings. For example, the kernel

$$K(x, y) = 2 - \|x - y\|_2 \quad \text{for } x, y \in \mathbb{S}^2, \quad (54)$$

whose RKHS is a Sobolev space of order $3/2$ on \mathbb{S}^2 , has

$$K_P(x) = K_{PP} = \frac{2}{3} \quad \text{for all } x \in \mathbb{S}^2. \quad (55)$$

The infinitely smooth kernel

$$K(x, y) = 48 \exp(-12\|x - y\|_2) \quad \text{for } x, y \in \mathbb{S}^2, \quad (56)$$

has

$$K_P(x) = K_{PP} = 1 - \exp(-48) \quad \text{for all } x \in \mathbb{S}^2. \quad (57)$$

See [Gräf \(2013\)](#) and [Ehler et al. \(2019\)](#) for these and other kernels on closed manifolds.

Periodic Sobolev kernels of order $2r$ ($r \in \mathbb{N}$) are another class of kernels with constant embeddings. They are

$$K^{2r}(x, y) = 1 + (-1)^{r+1} (2\pi)^{2r} \frac{B_{2r}(|x - y|)}{(2r)!} \quad (58)$$

$$= 1 + 2 \sum_{k=1}^{\infty} k^{-2r} \cos(2\pi k(x - y)). \quad (59)$$

for $x, y \in [0, 1]$, where B_{2r} is the Bernoulli polynomial of degree $2r$; see [\(Wahba, 1990, Ch. 2\)](#) for the series expansion. Since $K^{2r}(\cdot, y)$ is continuous and periodic for each $y \in [0, 1]$, K^{2r} can be viewed as a kernel on \mathbb{S}^1 . All terms in the sum (59) integrate zero. Therefore

$$K_P^{2r}(x) = K_{PP}^{2r} = 1 \quad \text{for all } x \in [0, 1], \quad (60)$$

where P can be interpreted as either the uniform distribution on $[0, 1]$ or the spherical measure on \mathbb{S}^1 . See [Rathinavel and Hickernell \(2019\)](#) and [Belhadji et al. \(2019\)](#) for uses of these kernels in Bayesian/kernel quadrature.

The identities (60) remain true if k^{-2r} in (59) are replaced with any positive coefficients for which the series converges. Additional terms that integrate to zero can be included without altering the embeddings in (60). For example, the term 1 in (58) could be replaced with $\sum_{\tau=0}^r B_\tau(x)B_\tau(y)/(\tau!)^2$ as in [Dick and Pillichshammer \(2010, Sec. 15.4\)](#) since B_τ for $\tau \geq 1$ integrate to zero. Related kernels with constant embeddings include digital shift invariant and scramble invariant kernels ([Dick and Pillichshammer, 2010, Thms. 12.7 & 13.20](#)).

4 Building Tractable Embeddings

We now discuss what to do when the pair of kernel K and distribution P of interest is not one of those above.

4.1 Building on Known Embeddings

A first approach is to obtain tractable expressions from transformations of known expressions. This trick has long been used in the literature, and was formalised by [Li et al. \(2021\)](#) in the context of probabilistic circuits.

Product Kernels and Product Distributions. Suppose that our distribution and kernel factorise so that

$$p(x) = \prod_j p_j(x_j), \quad K(x, y) = \prod_j K^j(x_j, y_j), \quad (61)$$

and assume that the embeddings of P_j with K^j are known in closed-form for all j . Then, the kernel embedding and its integral are given by

$$K_P(x) = \prod_j K_{P_j}^j(x), \quad K_{PP} = \prod_j K_{P_j P_j}^j. \quad (62)$$

Note that this approach can be straightforwardly generalised to the case of products of multivariate marginals.

Sum Kernels and Mixture Distributions. Suppose that we have a mixture distribution and/or a sum kernel:

$$p(x) = \sum_j w_j p_j(x), \quad (63)$$

$$K(x, y) = \sum_{j,j'} \gamma_{j,j'} K^{j,j'}(x, y), \quad (64)$$

where $K_{P_j}^{j'}$ and $K_{P_j P_j}^{j,j'}$ are known in closed-form for all j, j' . Then, the kernel mean embedding and its integral are themselves the sum of known quantities:

$$K_P(x) = \sum_{j,j'} w_j \gamma_{j,j'} K_{P_j}^{j'}(x), \quad (65)$$

$$K_{PP} = \sum_{j,j'} w_j w_{j'} \gamma_{j,j'} K_{P_j P_j}^{j,j'}. \quad (66)$$

Change of Measure. Suppose we want to compute the integral $I(f)$ of a function f against P . If K_P and K_{PP} are intractable, but we have access to closed-forms for K_Q and K_{QQ} for another distribution Q , then one approach is the “change of measure trick” (or “importance sampling trick”; e.g., [Karvonen et al., 2019, Sec. 5](#)). Suppose that P and Q have densities p and q . Then

$$\begin{aligned} I(f) &= \int_{\Omega} f(x) p(x) dx = \int_{\Omega} \left(\frac{f(x)p(x)}{q(x)} \right) q(x) dx \\ &= \int_{\Omega} g(x) q(x) dx. \end{aligned} \quad (67)$$

This trick works for Bayesian quadrature, but cannot necessarily be used more broadly since it is not an approach for computing unknown kernel mean embeddings.

Change of Variable. Suppose that Q is some distribution on Ω_Q for which we have closed-form expressions of K_Q and K_{QQ} . If we are interested in having closed-form embeddings for the distribution $P = \varphi_{\#}Q$, the push-forward of Q through the invertible map $\varphi: \Omega_Q \rightarrow \Omega$, then one approach is to use the “change of variable trick”

(also sometimes called the “*inverse transform trick*”). This consists of using a kernel of the form

$$K^\varphi(x, y) = K(\varphi(x), \varphi(y)), \quad (68)$$

since a change of variables gives us

$$\begin{aligned} K_P^\varphi(x) &= \int_{\Omega} K(\varphi(x), \varphi(y)) dP(y) \\ &= \int_{\Omega_Q} K(\varphi(x), y) dQ(y) = K_Q(\varphi(x)). \end{aligned} \quad (69)$$

One simple example is to take φ to be the inverse cumulative distribution function for P , in which case one can use a kernel K with closed-form embeddings against the uniform distribution. This trick is particularly natural when computing embeddings with respect to simulators/generative models (Bharti et al., 2023).

Matrix-Valued Kernels. In applications such as multi-output Bayesian quadrature (Xi et al., 2018; Gessner et al., 2020; Karvonen et al., 2019; Sun et al., 2023) one works with vector-valued RKHSs (Álvarez et al., 2012). This leads to matrix-valued kernels $K: \Omega \times \Omega \rightarrow \mathbb{R}^{T \times T}$, $T \in \mathbb{N}$. In these settings, it is often possible to recover embeddings through the embeddings of scalar-valued kernels. For example, a common construction is to take $K(x, y) = BK^s(x, y)$ where $K^s: \Omega \times \Omega \rightarrow \mathbb{R}$ is a scalar-valued kernel and $B \in \mathbb{R}^{T \times T}$ a positive semi-definite matrix. In this case, both the kernel mean embedding and its integral are matrices that can be directly obtained as $K_P(x) = BK_P^s(x)$ and $K_{PP} = BK_{PP}^s$.

4.2 Stein Reproducing Kernels

Since K_P and K_{PP} are known for few kernel/distribution pairs, an alternative is to *design* a reproducing kernel such that these quantities are available in closed-form. This is the idea behind Stein reproducing kernels (Oates et al., 2017; Anastasiou et al., 2023). One example is the *Langevin Stein reproducing kernel*, \tilde{K} . Suppose that $\Omega = \mathbb{R}^d$, K is a sufficiently regular and P satisfies $\int_{\Omega} \|\nabla_x \log p(x)\|_2 dP(x) < \infty$. Then

$$\begin{aligned} \tilde{K}(x, y) &= K(x, y)(\nabla_x \log p(x))^T \nabla_y \log p(y) \\ &\quad + \nabla_x K(x, y)^T \nabla_y \log p(y) \\ &\quad + \nabla_y K(x, y)^T \nabla_x \log p(x) \\ &\quad + \text{Tr}(\nabla_x \nabla_y K(x, y)). \end{aligned} \quad (70)$$

The i th entry of $\nabla_x K(x, y) \in \mathbb{R}^d$ is $\partial K(x, y)/\partial x_i$ and the (i, j) th entry of the matrix $\nabla_x \nabla_y K(x, y) \in \mathbb{R}^{d \times d}$ is $\partial^2 K(x, y)/\partial x_i \partial y_j$. The kernel \tilde{K} has the key property

$$\tilde{K}_P(x) = \tilde{K}_{PP} = 0 \quad \text{for all } x \in \Omega \quad (71)$$

by construction. Alternatively, if having a kernel mean embedding of zero is not convenient, one can use a kernel of the form $\tilde{K}^C(x, y) = \tilde{K}(x, y) + C$ for $C \in \mathbb{R}$, so that $\tilde{K}_P^C(x) = C$ and $\tilde{K}_{PP}^C = C$. This is particularly useful for Bayesian quadrature, in which case C can be viewed as a kernel hyperparameter to be optimised.

One of the main advantages of the Langevin Stein kernel is that it only requires knowledge of P through

evaluations of the score function $\nabla_x \log p$, which is available for most densities known up to normalisation constant. Indeed, suppose that $p(x) = \tilde{p}(x)/Z$ where \tilde{p} can be evaluated pointwise but $Z > 0$ is unknown. Then $\nabla_x \log p(x) = \nabla_x \log \tilde{p}(x)$. The score can hence be obtained through automatic differentiation using \tilde{p} . This allows the user to obtain closed-form embeddings for Bayesian posterior distributions, or unnormalised models such as large graphical models and deep energy models.

The idea of using Stein reproducing kernels to obtain closed-form kernel mean embeddings has been used successfully in a broad range of application areas. It has been used for Bayesian/kernel quadrature and related control variate approaches (Oates et al., 2017, 2019; Barp et al., 2022; Karvonen et al., 2018; Si et al., 2022; Sun et al., 2023; South et al., 2022), but also in the context of the maximum mean discrepancy, in which case the discrepancy is called *kernel Stein discrepancy*. This has led to new kernel herding algorithms (Chen et al., 2018, 2019), gradient flows (Korba et al., 2021), goodness-of-fit tests (Chwialkowski et al., 2016; Liu et al., 2016), parameter estimators (Barp et al., 2019), and generalised posterior distributions (Matsubara et al., 2022) for which kernel embeddings are available by construction.

5 Library

Kernel mean embeddings, while useful, are cumbersome to implement and test, which raises the bar for their practical usefulness. A small number of existing software packages contain closed-form expressions. These include the `ProbNum` (Wenger et al., 2021) and `Emukit` (Paley et al., 2023) packages in Python, and the `regMMD` package in R (Alquier and Gerber, 2024). However, the main focus of these libraries is to provide the user with the final method (which uses a kernel embedding) rather than to make the closed-form expression itself accessible. Thus, code related to kernel embeddings is often “hidden away” or linked with code for other purposes. To accompany the collection of embeddings in this paper, we therefore provide an accessible Python library called `kernel_embedding_dictionary`² whose purpose is to collect and make available embeddings in one place.

The structure is simple: i) we first instantiate a kernel mean embedding object and ii) then evaluate it at x . The example below shows how to do this for the uniform distribution (Lebesgue measure) with a Gaussian kernel.

```
from kernel_embedding_dictionary import \
    get_embedding

ke = get_embedding("expquad", "lebesgue")

# evaluate kernel mean embedding
x = np.random.randn(3, 1)
ke.mean(x)

array([0.92829 , 0.844137, 0.334471])
```

²https://github.com/mmahsereci/kernel_embedding_dictionary

Parameters of the distribution and the kernel can be defined in a configuration.

```
config_measure = {
    "ndim": 2,
    "bounds": [(0, 1), (-1, 0.5)],
    "normalize": True
}

config_kernel = {
    "ndim": 2,
    "lengthscales": [1.0, 2.0],
}

ke = get_embedding("expquad", "lebesgue", \
    config_kernel, config_measure)
```

All our embeddings are unit tested and neatly listed in the form of functions in a single module. This makes it easy to find and reuse the appropriate code for the user’s own scientific project (MIT License). Given its intended use, the library does not provide any methods that use kernel embeddings and is not optimized for efficiency. The library can be thought of as a dictionary of kernel embeddings “in the form of code,” providing one of the cumbersome building blocks for writing more elaborate project code. We hope that over time `kernel_embedding_dictionary` will become a point of reference and contain a representative collection of embeddings contributed by the open source community.

6 Conclusion

This paper provides a dictionary of kernel/distribution pairs for which the kernel mean embedding and its integral have a known closed-form, and reviewed several approaches to construct new expressions. Our hope is that this will save many a researcher the time needed to derive or implement kernel embeddings from scratch.

Additional related integrals not discussed in this paper are also occasionally needed. For example, some extensions of Bayesian quadrature (Gunter et al., 2014; Deisenroth et al., 2009; Prüher and Straka, 2018) require integrating certain products of kernels not covered in this paper. Expanding our dictionary with these expressions could therefore be useful. Some algorithms also require embeddings of conditional distributions (Muandet et al., 2016a, Sec. 4); see for instance (Chen et al., 2024b) for their use in Bayesian quadrature. Several of the embeddings above can already be interpreted as embeddings of conditional distributions, but again expanding our dictionary with this focus in mind could be of interest.

Of course, the paper would be incomplete without mentioning the broad literature studying approximations of kernel mean embeddings. In principle, any quadrature rule can be used (Sommariva and Vianello, 2006). Given independent samples x_1, \dots, x_n from P , the most common approximation is obtained through a Monte Carlo estimator: $K_P(x) \approx \frac{1}{n} \sum_{i=1}^n K(x, x_i)$, which Tolstikhin et al. (2017) show to be minimax-optimal and for which finite-sample bounds are available in Wolfer and Alquier (2024). Several other estimators have also been proposed, including a shrinkage estimator (Muandet et al., 2016b), a kernel density estimation-based

estimator (Sriperumbudur, 2016), a Gaussian process-based approach (Flaxman et al., 2016), a quasi-Monte Carlo estimator (Niu et al., 2023) and even a Bayesian quadrature estimator (Bharti et al., 2023). These can typically improve the error rate or provide uncertainty quantification, but at the cost of additional regularity assumptions. In certain cases, approximating embeddings is easier than in others. For example, it may be known that the embeddings are constant (as in Section 3.6), so that only one integral needs to be approximated, or there may be symmetries that drastically reduce the number of approximations needed (Karvonen et al., 2018, 2019).

Acknowledgements

The authors would like to thank Siu Lun Chau, Zonghao Chen, Krikamol Muandet, Masha Naslidnyk, Luc Pronzato, and Anatoly Zhigljavsky for helpful discussions. Equation (26) is due to Anatoly Zhigljavsky. FXB was supported by the EPSRC grant [EP/Y022300/1]. TK was supported by the Research Council of Finland grant 359183.

References

- P. Alquier and M. Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1): 71–92, 2023.
- P. Alquier and M. Gerber. *regMMD: Robust regression and estimation through maximum mean discrepancy minimization*, 2024. URL <https://cran.r-project.org/package=regMMD>. R package version 0.0.1.
- M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.
- A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, L. Mackey, C. J. Oates, G. Reinert, and Y. Swan. Stein’s method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.
- M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, pages 1355–1362, 2012.
- A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, volume 32, pages 12964–12976, 2019.
- A. Barp, C. J. Oates, E. Porcu, and M. Girolami. A Riemannian–Stein kernel method. *Bernoulli*, 28(4): 2181–2208, 2022.

- A. Belhadji, R. Bardenet, and P. Chainais. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems*, volume 32, pages 12927–12937, 2019.
- A. Belhadji, D. Sharp, and Y. Marzouk. Weighted quantization using MMD: From mean field to mean shift via gradient flows. *arXiv:2502.10600*, 2025.
- A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.
- A. Bharti, M. Naslidnyk, O. Key, S. Kaski, and F.-X. Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, pages 2289–2312, 2023.
- F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944v1*, 2019a.
- F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? (with discussion and rejoinder). *Statistical Science*, 34(1):1–22, 2019b.
- L. Chamakh and Z. Szabó. Keep it tighter - A story on analytical mean embeddings. *arXiv:2110.09516v3*, 2024.
- A. Chatalic, N. Schreuder, E. De Vito, and L. Rosasco. Efficient numerical integration in reproducing kernel Hilbert spaces via leverage scores sampling. *arXiv:2311.13548v1*, 2023.
- W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning*, pages 843–852, 2018.
- W. Y. Chen, A. Barp, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. J. Oates. Stein point Markov chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, pages 1011–1021, 2019.
- Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010.
- Z. Chen, A. Mustafi, P. Glaser, A. Korba, A. Gretton, and B. K. Sriperumbudur. (De)-regularized maximum mean discrepancy gradient flow. *arXiv:2409.14980v1*, 2024a.
- Z. Chen, M. Naslidnyk, A. Gretton, and F.-X. Briol. Conditional Bayesian quadrature. *Uncertainty in Artificial Intelligence*, pages 648–684, 2024b.
- Z. Chen, T. Karvonen, H. Kanagawa, F.-X. Briol, and C. J. Oates. Stationary MMD points for cubature. *arXiv:2505.20754*, 2025.
- B.-E. Chérif-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference (AABI)*, pages 1–21, 2020.
- B.-E. Chérif-Abdellatif and P. Alquier. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1): 181–213, 2022.
- K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of the International Conference on Machine Learning*, pages 2606–2615, 2016.
- M. P. Deisenroth, M. F. Huber, and U. D. Hanebeck. Analytic moment-based Gaussian process filtering. In *Proceedings of the International Conference on Machine Learning*, pages 225–232, 2009.
- C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 943–970, 2022.
- J. Dick and F. Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.
- J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.
- N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.
- R. Dwivedi and L. Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25, 2024.
- G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.
- M. Ehler, M. Graef, and C. J. Oates. Optimal Monte Carlo integration on closed manifolds. *Statistics and Computing*, 29:1204–1214, 2019.
- E. N. Epperly and E. Moreno. Kernel quadrature with randomly pivoted cholesky. In *Advances in Neural Information Processing Systems*, volume 36, pages 65850–65868, 2023.
- S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi. Bayesian learning of kernel embeddings. In *Uncertainty in Artificial Intelligence*, pages 182–191, 2016.
- K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes’ rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

- E. Fuselier, T. Hangelbroek, F. J. Narcowich, J. D. Ward, and G. B. Wright. Kernel based quadratures on spheres and other homogeneous spaces. *Numerische Mathematik*, 127(1):57–92, 2014.
- A. Gessner, J. Gonzalez, and M. Mahsereci. Active multi-information source Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 712–721, 2020.
- D. Ginsbourger, O. Roustant, D. Schuhmacher, N. Durand, and N. Lenz. On ANOVA decompositions of kernels and Gaussian random field paths. In *Monte Carlo and Quasi-Monte Carlo Methods*, volume 163 of *Springer Proceedings in Mathematics & Statistics*, pages 315–330, 2016.
- M. Gräf. *Efficient Algorithms for the Computation of Optimal Quadrature Points on Riemannian Manifolds*. PhD thesis, Chemnitz University of Technology, 2013.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.
- T. Gunter, M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, volume 24, pages 2789–2797, 2014.
- P. Hennig, M. A. Osborne, and H. P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.
- J. Hertrich, C. Wald, F. Altekruiger, and P. Hagemann. Generative sliced MMD flows with Riesz kernels. In *International Conference on Learning Representations*, 2024.
- D. Huang, A. Bharti, A. Souza, L. Acerbi, and S. Kaski. Learning robust statistics for simulation-based inference under model misspecification. In *Advances in Neural Information Processing Systems*, pages 7289–7310, 2023.
- M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20:155–194, 2020.
- T. Karvonen, C. J. Oates, and S. Särkkä. A Bayes–Sard cubature method. In *Advances in Neural Information Processing Systems*, volume 31, pages 5882–5893, 2018.
- T. Karvonen, S. Särkkä, and C. J. Oates. Symmetry exploits for Bayesian cubature methods. *Statistics and Computing*, 29:1231–1248, 2019.
- J. Kellner and A. Celisse. A one-sample test for normality with kernel methods. *Bernoulli*, 25(3):1816–1837, 2019.
- A. Korba, P. C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein discrepancy descent. In *Proceedings of the International Conference on Machine Learning*, pages 5719–5730, 2021.
- S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 544–552, 2015.
- W. Li, Z. Zeng, A. Vergari, and G. van den Broeck. Tractable computation of expected kernels. In *Uncertainty in Artificial Intelligence*, pages 1163–1173, 2021.
- Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning*, pages 276–284, 2016.
- J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. *Advances in Neural Information Processing Systems*, 28:829–837, 2015.
- T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 84(3):997–1022, 2022.
- D. Ming and S. Guillas. Linked Gaussian process emulation for systems of computer models using Matérn kernels and adaptive design. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1615–1642, 2021.
- K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2016a.
- K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41, 2016b.
- K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):7322–7392, 2021.
- H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992.
- Y. Nishiyama and K. Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17(180):1–28, 2016.
- Y. Nishiyama, M. Kanagawa, A. Gretton, and K. Fukumizu. Model-based kernel sum rule: kernel Bayesian inference with probabilistic models. *Machine Learning*, 109(5):939–972, 2020. ISSN 15730565.
- Z. Niu, J. Meier, and F.-X. Briol. Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. *Electronic Journal of Statistics*, 17(1):1411–1456, 2023.
- C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(3):695–718, 2017.

- C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein’s identity. *Bernoulli*, 25(2):1141–1159, 2019.
- A. O’Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.
- L. Pacchiardi, S. Khoo, and R. Dutta. Generalized Bayesian likelihood-free inference. *Electronic Journal of Statistics*, 18:3628–3686, 2024.
- A. Paleyes, M. Mahsereci, and N. D. Lawrence. Emukit: A Python toolkit for decision making under uncertainty. *Proceedings of the Python in Science Conference*, 2023.
- J. Prüher and O. Straka. Gaussian process quadrature moment transform. *IEEE Transactions on Automatic Control*, 63(9):2844–2854, 2018.
- J. Rathinavel and F. J. Hickernell. Fast automatic Bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229, 2019.
- R. M. Rustamov. Closed-form expressions for maximum mean discrepancy with applications to Wasserstein auto-encoders. *Stat*, 10(1):1–12, 2021.
- D. Sejdinovic. An overview of causal inference using kernel embeddings. *arXiv:2410.22754*, 2024.
- S. Si, C. J. Oates, A. B. Duncan, L. Carin, and F.-X. Briol. Scalable control variates for Monte Carlo methods via stochastic optimization. In *Monte Carlo and Quasi-Monte Carlo Methods. MCQMC 2020*, pages 205–221. Springer, 2022.
- R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 4593–4605, 2019.
- R. Singh, L. Xu, and A. Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024.
- A. Sommariva and M. Vianello. Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3-4):295–310, 2006.
- L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the International Conference on Machine Learning*, pages 992–999, 2008.
- L. F. South, T. Karvonen, C. Nemeth, and C. J. Oates. Semi-exact control functionals from Sard’s method. *Biometrika*, 109(2):351–367, 2022.
- B. K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.
- B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11, 2010.
- Z. Sun, A. Barp, and F.-X. Briol. Vector-valued control variates. In *Proceedings of the International Conference on Machine Learning*, pages 32819–32846, 2023.
- I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(1):3002–3048, 2017.
- I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.
- G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.
- H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4: 389–396, 1995.
- J. Wenger, N. Krämer, M. Pförtner, J. Schmidt, N. Bosch, N. Effenberger, J. Zenn, A. Gessner, T. Karvonen, F.-X. Briol, M. Mahsereci, and P. Hennig. ProbNum: Probabilistic numerics in Python. *arXiv:2112.02100v1*, 2021.
- G. Wolfer and P. Alquier. Variance-aware estimation of kernel mean embedding. *arXiv:2210.06672v2*, 2024.
- X. Xi, F.-X. Briol, and M. Girolami. Bayesian quadrature for multiple related integrals. In *Proceedings of the International Conference on Machine Learning*, pages 8533–8564, 2018.
- L. Xu, A. Korba, and D. Slepčev. Accurate quantization of measures via interacting particle-based optimization. In *Proceedings of the International Conference of Machine Learning*, pages 24576–24595, 2022.