Polynomial Based RKHS

With Applications to Data-Driven Modeling

He ZHANG (hqz5159@psu.edu)

Penn State, Math Department

Poster Session, Workshop on Mathematical Machine Learning and Application

Abstract

We focus on a nonparametric density estimator formulated by the kernel embedding of distributions. In particular, we consider the "Mercer-type" kernels constructed based on the classical orthogonal bases defined on non-compact domains, such as the Hermite and Laguerre polynomials. While the resulting representation is analogous to Polynomial Chaos Expansion (PCE), by studying the orthogonal polynomial approximation in the reproducing kernel Hilbert space (RKHS) setting, we establish the uniform convergence of the estimator. More importantly, the RKHS formulation allows one to systematically address a practical question of identifying the PCE basis for a consistent estimation through the decay property of the target functions quantified using the available data. Numerically, we apply our density estimator to data-driven modeling in recovering the linear response statistics of an unknown underlying dynamics. The poster is based on the recent works [4, 5] with my advisor Prof. John Harlim and co-advisor Prof. Xiantao Li.

The universality of RKHS

Definition 1. *Let X be a non-empty set and $\mathcal{H}$ be a $\mathbb{R}$-Hilbert function space over X, i.e., a $\mathbb{R}$-Hilbert space of functions that maps X to $\mathbb{R}$. Then $\mathcal{H}$ is called an RKHS with kernel k, if $k(\cdot, x) \in \mathcal{H}$, $\forall x \in X$, and we have the reproducing property*

$$f(x) = \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} \tag{1}$$

holds for all $f \in \mathcal{H}$ and all $x \in X$. In particular, we call such $k(\cdot,\cdot)$ a reproducing kernel of $\mathcal{H}$.

There is a one-to-one correspondence between the RKHS and kernel. The RKHS has the remarkable property that the norm convergence implies the pointwise convergence. More precisely, consider $f_n \to f$ in $\mathcal{H}$, that is, $\|f_n - f\|_{\mathcal{H}} \to 0$ as $n \to \infty$. Then, $\forall x \in X$, we have

$$|(f_n - f)(x)| = |\langle f_n - f, k(\cdot, x)\rangle_{\mathcal{H}}| \le \|f_n - f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}} \to 0,\ n \to \infty. \tag{2}$$

Eq. (2) suggests that if $\|k(\cdot, x)\|_{\mathcal{H}}$ is bounded uniformly in x, we will have the uniform convergence of f_n.

Lemma 2. *Let X be a topological space and k be a kernel on X with RKHS $\mathcal{H}$. If k is bounded in the sense that*

$$\|k\|_\infty := \sup_{x\in X} \sqrt{k(x, x)} < \infty.$$

and $k(\cdot, x) : X \to \mathbb{R}$ is continuous $\forall x \in X$, then $\mathcal{H} \subset C_b(X)$ (space of bounded and continuous functions on X), and the inclusion $\mathrm{id} : \mathcal{H} \to C_b(X)$ is continuous with $\|\mathrm{id} : \mathcal{H} \to C_b(X)\| = \|k\|_\infty$.

As a subspace of $C_b(X)$, it is natural to ask whether the RKHS $\mathcal{H}$ is dense in the Banach space $C_b(X)$ equipped with the uniform norm. In this poster, we are interested in the case where X is non-compact, e.g., $X = \mathbb{R}^d$, and the target f is a continuous density function which vanishes at infinity. For a locally compact Hausdorff (LCH) space X, let $C_0(X)$ denote the space of all continuous functions on X that vanish at infinity.

Definition 3. *(c_0-universal) Let X be an LCH space and let k be a bounded kernel on $X \times X$ and $k(\cdot, x) \in C_0(X)$, $\forall x \in X$. The kernel k is said to be c_0-universal if the RKHS, $\mathcal{H}$, induced by k is dense in $C_0(X)$ with respect to the uniform norm.*

Notice $\forall f \in \mathcal{H}$,

$$|f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| \le \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} k^{\frac{1}{2}}(x, x), \quad \forall x \in X, \tag{3}$$

that is, functions in the RKHS have the same decay rate as $k^{\frac{1}{2}}(x, x)$. For general C_0-function, we may use a weight function q to characterize its decay rate. For example, take $X = \mathbb{R}^d$, and $q \propto \exp(-\theta\|\boldsymbol{x}\|^2)$ $(\theta > 0)$, then the functions in $C_0(\mathbb{R}^d, q^{-1})$ are continuous with a Gaussian decay rate.

Lemma 4. *(weighted c_0-universal) If the kernel k satisfies $k(\cdot, x) \in C_0(X, q^{-1})$, $\forall x \in X$. Then, $\tilde{k}(x, y) := q^{-1}(x)k(x, y)q^{-1}(y)$ defines a kernel on X, and the RKHS $\mathcal{H}$ induced by k is dense in $C_0(X, q^{-1})$ if and only if the kernel $\tilde{k}$ is c_0-universal.*

From orthogonal polynomials to RKHS

Let $\{p_{\vec{m}}(\boldsymbol{x})\}$ be the orthonormal polynomial in $L^2(\mathbb{R}^d, \boldsymbol{W})$, we define the "Mercer-type" kernel,

$$k_\beta(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\vec{m}\ge 0} \lambda_{\vec{m}} p_{\vec{m}}(\boldsymbol{x}) p_{\vec{m}}(\boldsymbol{y}) \boldsymbol{W}^\beta(\boldsymbol{x}) \boldsymbol{W}^\beta(\boldsymbol{y}), \quad \lambda_{\vec{m}} := \prod_{i=1}^d \lambda_{m_i}, \tag{4}$$

for $\beta \ge \frac{1}{2}$. Here, λ_n is a monotonically decreasing sequence, which can be interpreted as the eigenvalues with eigenfunctions $\left\{\Psi_{\beta,\vec{m}} := p_{\vec{m}}\boldsymbol{W}^\beta\right\}$.

Proposition 5. *For any fixed $\beta \ge \frac{1}{2}$, we have the following results.*

1. For any sequence $\left\{\hat{f}_{\vec{m}}\right\} \in \ell_2$ satisfying

$$\sum_{\vec{m}\ge 0} \frac{\hat{f}^2_{\vec{m}}}{\lambda_{\vec{m}}} < \infty, \tag{5}$$

where $\lambda_{\vec{m}}$ is defined in (4)*, the sequence of functions*

$$f_n := \sum_{\|\vec{m}\|_1 \le n} \hat{f}_{\vec{m}} \Psi_{\beta,\vec{m}}, \quad n \ge 0,$$

converge uniformly in $C_0(\mathbb{R}^d)$. Moreover, the limit, denoted as f^*, satisfies $f^* \in L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta}) \cap C_0(\mathbb{R}^d)$.

2. The function space

$$\mathcal{H}_\beta := \left\{ f = \sum_{\vec{m}\ge 0} \hat{f}_{\vec{m}} \Psi_{\beta,\vec{m}} \;\middle|\; \sum_{\vec{m}\ge 0} \frac{\hat{f}^2_{\vec{m}}}{\lambda_{\vec{m}}} < \infty \right\}, \tag{6}$$

is a well-defined subspace of $L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta}) \cap C_0(\mathbb{R}^d)$. Further, define the map $\langle\cdot,\cdot\rangle : \mathcal{H}_\beta \times \mathcal{H}_\beta \to \mathbb{R}$ as

$$\langle f, g\rangle := \sum_{\vec{m}\ge 0} \frac{\hat{f}_{\vec{m}}\hat{g}_{\vec{m}}}{\lambda_{\vec{m}}}, \quad f = \sum_{\vec{m}\ge 0} \hat{f}_{\vec{m}}\Psi_{\beta,\vec{m}},\ g = \sum_{\vec{m}\ge 0} \hat{g}_{\vec{m}}\Psi_{\beta,\vec{m}} \in \mathcal{H}_\beta.$$

Then $\langle\cdot,\cdot\rangle$ defines an inner product, and $\mathcal{H}_\beta$, equipped with the inner product $\langle\cdot,\cdot\rangle$, is a Hilbert space.

3. $\mathcal{H}_\beta$ is the RKHS with reproducing kernel k_β in (4)*.*

Example. Let $\boldsymbol{W}$ be the d-dimensional standard Gaussian distribution, following (4), we define the kernel

$$k_{\beta,\rho}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\vec{m}\ge 0} \rho^{\|\vec{m}\|_1} \Psi_{\beta,\vec{m}}(\boldsymbol{x}) \Psi_{\beta,\vec{m}}(\boldsymbol{y}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,\ \Psi_{\beta,\vec{m}} = \psi_{\vec{m}}\boldsymbol{W}^\beta, \tag{7}$$

where $\{\psi_n\}$ are normalized Hermite polynomials with eigenvalues $\lambda_n = \rho^n$, $\rho \in (0, 1)$. For this special case, we do have an explicit expression for $k_{\beta,\rho}$. For example, when $d = \beta = 1$, the kernel $k_{1,\rho}$ in (7) is known as the Mehler kernel with

$$k_{1,\rho}(x, y) = \sum_{m=0}^\infty \rho^m \Psi_m(x)\Psi_m(y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right). \tag{8}$$

For $\beta \in [\frac{1}{2}, \infty)$ and $\rho \in (0, 1)$, we will call the kernel $k_{\beta,\rho}$ in (7) and the corresponding RKHS, $\mathcal{H}_{\beta,\rho}$, following Proposition 5, the d-dimensional *Mehler kernel* and *Mehler RKHS*, respectively.

Corollary 6. *(universality of the Mehler RKHS) Let*

$$q_{\beta,\rho}(\boldsymbol{x}) := \exp\left[-\left(\frac{1}{2(1+\rho)} + \frac{\beta-1}{2}\right)\|\boldsymbol{x}\|^2\right], \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{9}$$

then $\mathcal{H}_{\beta,\rho}$ is dense in $C_0\left(\mathbb{R}^d, q^{-1}_{\beta,\rho}\right)$.

This means that one can approximate any continuous function that has Gaussian (or faster) decaying rate up to any desirable accuracy using an estimator that belongs to the Mehler RKHS.

So far we have introduced a framework to construct RKHS as a subspace of $L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta}) \cap C_0(\mathbb{R}^d)$. The resulting RKHS extracts features of both $L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta})$ and $C_0(\mathbb{R}^d)$, e.g., the expansion formula

$$f = \sum_{\vec{m}\ge 0} \hat{f}_{\vec{m}} \Psi_{\beta,\vec{m}}, \quad \hat{f}_{\vec{m}} := \int_{\mathbb{R}^d} f(\boldsymbol{x}) \Psi_{\beta,\vec{m}}(\boldsymbol{x}) \boldsymbol{W}^{1-2\beta}(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x} = \int_{\mathbb{R}^d} f(\boldsymbol{x}) p_{\vec{m}}(\boldsymbol{x}) \boldsymbol{W}^{1-\beta}(\boldsymbol{x})\,\mathrm{d}\,\boldsymbol{x}, \tag{10}$$

makes sense not only in $L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta})$ but also pointwise. In practice, given data $\{X_n\}_{n=1}^N$ sampled from the target density f, we will choose a function in $\mathcal{H}_\beta$ with a finite sum, $\|\vec{m}\|_1 \le M$, where $M \ll N$, as an estimator for f. While the choice of M allows us to specify the theoretical "bias" or "approximation error", thanks to the orthogonal representation, the resulting hypothesis function is parametric and the evaluation of f on a new $\boldsymbol{x} \in \mathbb{R}^d$ amounts to evaluating $\binom{M+d}{M}$ components of $\left\{\Psi_{\beta,\vec{m}}(\boldsymbol{x}) \mid \|\vec{m}\|_1 \le M\right\}$. This is computationally much cheaper than evaluating $f(\boldsymbol{x}) = \langle f, k(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}}$, with radial-type kernels, such as $k(\boldsymbol{x}, \boldsymbol{y}) = h(\|\boldsymbol{x} - \boldsymbol{y}\|)$ for some positive function h, since the computation of the inner product requires evaluating $h(\|X_n - \boldsymbol{x}\|)$, for all $n = 1, \ldots, N$.

Kernel embedding data-driven modeling

Assume the target d-dimensional density function f lives in the Mehler RKHS $\mathcal{H}_{\beta,\rho}$. We define the order-M *kernel embedding estimates* as the order-M truncation of Eq. (10),

$$f_M := \sum_{\|\vec{m}\|_1 \le M} \hat{f}_{\vec{m}} \Psi_{\beta,\vec{m}}. \tag{11}$$

We should point out that, with this choice of basis representation, we arrive at a polynomial chaos approximation of f. But the convergence $f_M \to f$ as $M \to \infty$ is valid in both $L^2(\mathbb{R}^d, \boldsymbol{W}^{1-2\beta})$ and $C_0(\mathbb{R}^d)$. In practice, the integral in (11) can be approximated by a Monte-Carlo average, and we define the order-M *empirical kernel embedding estimate* of f as

$$f_{M,N} := \sum_{\|\vec{m}\|_1 \le M} \hat{f}_{\vec{m},N} \Psi_{\beta,\vec{m}}, \quad \hat{f}_{\vec{m},N} := \frac{1}{N}\sum_{n=1}^N \psi_{\vec{m}}(X_n)\boldsymbol{W}^{1-\beta}(X_n), \tag{12}$$

where $\{X_n\}_{n=1}^N$ are sampled from the target density function f. For general density functions, we run statistical tests to identify the tail of the marginal distribution. Subsequently, we choose an appropriate RKHS basis based on the tail information. Finally, we construct the basis using tensor product, and compute the empirical kernel embedding estimates following (12).

In [5], we propose the following "semi-parametric" extended Langevin equation as an imperfect model to recover the response statistics of an unknown underlying dynamics

$$\begin{aligned} \dot{x} &= v, \\ \dot{v} &= \Lambda\nabla_x \log\left(\hat{\rho}(x)\right) - \Gamma v + \sigma\dot{W}_t. \end{aligned} \tag{13}$$

By "semi-parametric", we refer to the combination of linear parametric equation in the right-hand-side with a "nonparametric" term that involves $\hat{\rho}(x)$ that is estimated by the kernel embedding approximation of the marginal distribution of x at equilibrium.

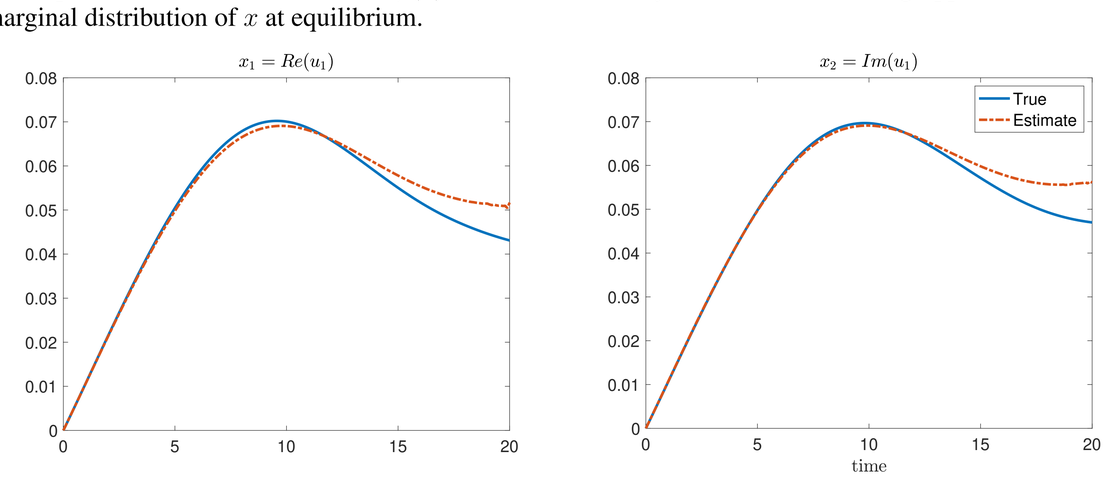

Figure 1: The prediction of the full response. The blue solid curves indicate the full response of the underlying dynamics. The red dash curves are the full response from the imperfect model (13) based on the parameter estimates.

References

[1] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008.

[2] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[3] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

[4] He Zhang, John Harlim, and Xiantao Li. Computing linear response statistics using orthogonal polynomial based estimators: An rkhs formulation. *arXiv preprint arXiv:1912.11110*, 2019.

[5] He Zhang, John Harlim, and Xiantao Li. Linear response based parameter estimation in the presence of model error. *arXiv preprint arXiv:1910.14113*, 2019.