A Fourier Representation of Kernel Stein Discrepancy with Application to Goodness-of-Fit Tests for Measures on Infinite Dimensional Hilbert Spaces

George Wynne, Mikołaj Kasprzak, Andrew Duncan

The presentation contains results obtained in project Stein-ML that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 01024264.

Motivation

- *Functional data analysis* (FDA) is a growing field and has found numerous applications, for instance in finance, biomedicine or weather forecasting.
- Among the most common statistical tasks performed by practitioners using functional data is goodness-of-fit testing.
- However, dealing with infinite-dimensional data poses many challenges - for instance it prohibits the use of density-based methods.
- This issue is often side-stepped through the *project first* approach to FDA.
- But choosing the right projection is challenging - it might not capture the variability of the random functions and might not yield close form expressions.
- A convenient way of fully characterizing probability distributions (including examples of infinite-dimensional distributions) is offered by Stein's method.
- Stein's method has been used in the past to construct goodness-of-fit tests via kernel Stein discrepancies (KSD) but only in finite dimensions.
- Our work:
  - Formulates KSD for measures on general separable Hilbert spaces.
  - Identifies conditions which ensure that such KSD separates measures.
  - Formulates a KSD goodness-of-fit test for measures absolutely continuous wrt Gaussians, directly on separable Hilbert spaces, without projections.
- Along the way we derive a new Fourier representation which gives insight on the behaviour of KSD in finite and infinite dimensions.

Overview

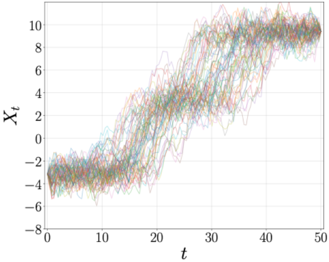
Formulation of the problem

Formulation of the problem

- $\mathcal{X}$ is a separable Hilbert space, e.g. $\mathbb{R}^d$ or $L^2([0,1]^d)$
- Observed samples $\{X_n\}_{n=1}^N \sim Q$ for a prob. measure Q on $\mathcal{X}$.
- How far is Q from P such that $\frac{dP}{dN_C} \propto e^{-U}$ for N_C denoting a Gaussian measure with mean zero and covariance operator C?

Example (Conditioned SDE)

- $\mathcal{X} = L^2([0,50])$
- $dX_t = 0.7\sin(X_t)dt + dW_t$
- Condition on $X_0 = -\pi$, $X_{50} = 3\pi$.
- N_C - Brownian bridge, U - from Girsanov theorem



Example from Bierkens et al, Stat. Comput., 2021

Stein's method and Stein discrepancies

Stein's method - original motivation

Original aim: find a bound on an integral probability metric $\sup_{h\in\mathcal{H}} |\mathbb{E}_Q h - \mathbb{E}_P h|$, where P is the target (known) distribution, Q is the approximating law and $\mathcal{H}$ is a suitable class of real-valued functions.

- Step 1: Find an operator $\mathcal{A}$ acting on a class of real-valued functions, such that:

$$\forall f \in \mathrm{Domain}(\mathcal{A}) \quad \mathbb{E}_P \mathcal{A} f = 0.$$

- Step 2: For a given function $h \in \mathcal{H}$, find $f = f_h$, such that:

$$\mathcal{A} f = h - \mathbb{E}_P h.$$

- Step 3: Study the properties of f_h and bound $\sup_{h\in\mathcal{H}} |\mathbb{E}_Q \mathcal{A} f_h|$, using various mathematical techniques (exchangeable pairs, Taylor expansions, Malliavin calculus...).

Stein discrepancies

- For a suitable class of test functions $\mathcal{F}$, let

$$SD_{\mathcal{A},\mathcal{F}}(Q,P) := \sup_{f\in\mathcal{F}} \left|\mathbb{E}_Q\left[\mathcal{A}f\right]\right|.$$

- A convenient choice for $\mathcal{F}$ is the unit ball of a RKHS: $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_k \leq 1\}$, giving rise to the Kernel Stein Discrepancy:

$$KSD_{\mathcal{A},k}(Q,P) := \sup_{\|f\|_k\leq 1} \left|\mathbb{E}_Q\left[\mathcal{A}f\right]\right|.$$

- KSD on $\mathbb{R}^d$ has a convenient representation that allows it to be easily estimated with a U-statistic, given samples from Q:

$$\begin{aligned} KSD_{\mathcal{A},k}(Q,P)^2 =& \mathbb{E}_{(X,X')\sim Q\times Q}\left[\left(\mathcal{A}\otimes\mathcal{A}\right)k(X,X')\right] \\ =:& \mathbb{E}_{(X,X')\sim Q\times Q}\left[h_{\mathcal{A},k}(X,X')\right], \end{aligned}$$

where $h_{\mathcal{A},k}$ is called the Stein kernel.

How to find a Stein operator?

Remember: We want to find an operator $\mathcal{A}$ acting on a class of real-valued functions, such that:

$$\forall f \in \text{Domain}(\mathcal{A}) \quad \mathbb{E}_P \mathcal{A} f = 0.$$

One way of doing this is to:

- construct a Markov process (X_t) whose stationary distribution is P;
- take $\mathcal{A}$ to be its infinitesimal generator:

$$\mathcal{A} f(x) = \lim_{t \downarrow 0} \frac{\mathbb{E}[f(X_t)|X_0 = x] - f(x)}{t}.$$

Example (Langevin Stein operator on $\mathbb{R}^d$)

Suppose that P is a density over $\mathbb{R}^d$.

- Markov process: $dX_t = \nabla \log P(X_t) dt + \sqrt{2} dW_t$.
- Infinitesimal generator:

$$\mathcal{A} f(x) = \Delta f(x) + \langle \nabla \log P(x), \nabla f(x) \rangle_{\mathbb{R}^d}.$$

Note that $\nabla \log P$ kills the normalising constant of P.

Stein discrepancy in infinite dimensions

- Suppose we work on a separable Hilbert space $\mathcal{X}$.
- Our target measure P is such that $\frac{dP}{dN_C} \propto e^{-U}$.
- Consider the pre-conditioned Langevin diffusion on $\mathcal{X}$:

$$dX_t = -(X_t + CDU(X_t))dt + \sqrt{2}dW_t,$$

  Take its infinitesimal generator:

$$\mathcal{A}f(x) = \mathrm{Tr}(CD^2 f(x)) - \langle Df(x), x + CDU(x)\rangle_{\mathcal{X}}$$

  as our Stein operator.

Some questions about the resulting KSD:

- When is $KSD_{\mathcal{A},k}(Q,P) := \sup_{\|f\|_k \leq 1} |\mathbb{E}_Q[\mathcal{A}f]|$ well defined?
- When does it separate measures?
- When does it metrize weak convergence?

Our results about (infinite-dimensional) KSD

Our assumptions

- $\mathcal{X}$ is a separable Hilbert space.
- The target P is such that $\frac{dP}{dN_C} \propto e^{-U}$.
- The Stein operator is the pre-conditioned Langevin generator: $\mathcal{A}f(x) = \text{Tr}(CD^2 f(x)) - \langle Df(x), x + CDU(x)\rangle_{\mathcal{X}}$.
- The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has bounded and continuous second derivatives: $k \in C_b^{(2,2)}(\mathcal{X} \times \mathcal{X})$.
- Additional mild integrability and differentiability conditions on the potential U hold:
  - $\mathbb{E}_{X\sim Q}\left[\|CDU(X)\|_{\mathcal{X}}\right] < \infty$;
  - $e^{-U(\cdot)/2} \in W_C^{1,2}(\mathcal{X})$ $\left(\text{i.e. } \mathbb{E}_{X\sim N_C}\left\|e^{-U(X)/2}(X)\right\|_{\mathcal{X}}^2 < \infty \text{ and } \mathbb{E}_{X\sim N_C}\left\|C^{1/2}D\left(e^{-U(\cdot)/2}\right)(X)\right\|_{\mathcal{X}}^2 < \infty\right)$;
  - $\mathbb{E}_{X\sim N_C}\left[\|C^{1/2}DU(X)\|_{\mathcal{X}}^2\right] < \infty$.
- Second moments of candidate Q exist.

Our results

Given the assumptions from the previous slide we show the following:

- The kernel Stein discrepancy $KSD_{\mathcal{A},k}(Q,P)$ is well-defined.
- The double expectation representation of KSD is well-defined in the desired generality of (potentially) infinite-dimensional $\mathcal{X}$:

$$KSD_{\mathcal{A},k}(Q,P)^2 = \mathbb{E}_{(X,X')\sim Q\times Q}\left[(\mathcal{A}\otimes\mathcal{A})\, k(X,X')\right].$$

- Let μ be a Borel measure on $\mathcal{X}$ and $\hat{\mu}$ be its Fourier transform. If $k(x,y)=\hat{\mu}(x-y)$ then the following Fourier representation holds:

$$KSD(Q,P)^2_{\mathcal{A},k} = \int_{\mathcal{X}} \left|\mathbb{E}_{X\sim Q}\left[\mathcal{A}\left(e^{i\langle s,\cdot\rangle_{\mathcal{X}}}\right)(X)\right]\right|^2_{\mathbb{C}} d\mu(s).$$

- Suppose that $k(x,y)=\hat{\mu}(x-y)$ for a Borel measure μ with full support. Then the KSD separates measures:

$$KSD(Q,P)_{\mathcal{A},k} = 0 \iff Q = P.$$

A closer look at the Fourier representation

If $k(x,y) = \hat{\mu}(x-y)$ then:

$$KSD_{\mathcal{A},k}(Q,P) := \sup_{\|f\|_k \leq 1} |\mathbb{E}_Q[\mathcal{A}f]| = \int_{\mathcal{X}} \left| \mathbb{E}_{X\sim Q}\left[\mathcal{A}\left(e^{i\langle s,\cdot\rangle_{\mathcal{X}}}\right)(X)\right]\right|_{\mathbb{C}}^2 d\mu(s).$$

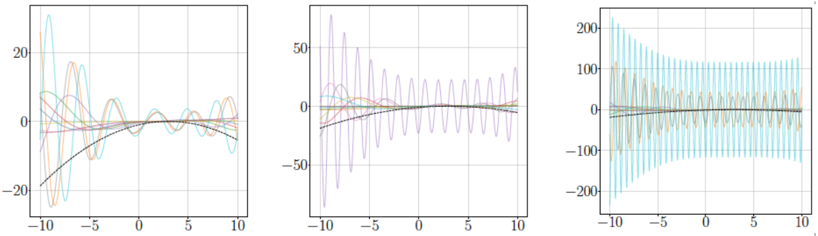
- Applies to more general Stein operators
- Related to the following expression for MMD:

  $$\mathrm{MMD}_k(Q,P) = \int_{\mathcal{X}} \left|\hat{Q}(s) - \hat{P}(s)\right|_{\mathbb{C}}^2 d\mu(s) = \int_{\mathcal{X}} \left|\mathbb{E}_{X\sim Q}\left[\Theta\left(e^{i\langle s,\cdot\rangle_{\mathcal{X}}}\right)(X)\right]\right|_{\mathbb{C}}^2 d\mu(s)$$

  for $\Theta f(x) = f(x) - \mathbb{E}_P f$.
- Relates KSD to L^2 - based tests (Ebner, Henze 2020).
- Supremum over RKHS = Average over μ.
- The kernel choice only influences the integrating measure μ. The integrand is determined by the Stein operator $\mathcal{A}$.
- The heavier the tails of μ, the more weight is placed upon the test functions $\mathcal{A}\left(e^{i\langle s,\cdot\rangle_{\mathcal{X}}}\right)(\cdot)$ for values of s with large norm and KSD becomes more discerning between P and Q.

Example

- Let $\mathcal{X} = \mathbb{R}$, P have density $p(x) \propto \exp\left(-\left(\frac{x-3}{3}\right)^3\right)$.
- Let $\mathcal{A}$ be the standard Langevin-Stein operator:
  $\mathcal{A}f(x) = f''(x) + (\log p)'(x)f'(x)$.



(a) $\mu = \text{Gaussian}(0, 1)$ (b) $\mu = \text{Students-}t(\nu = 2)$ (c) $\mu = \text{Cauchy}(0, 1)$

Figure 1: Plots corresponding to Example 4.1 of the real part of the test functions $\mathcal{A}(e^{is\cdot})(x)$ for 10 samples from different choices of μ, the heavier the tails of μ the larger the samples of s hence the greater the magnitude and periodicity of the test functions. In black is $(\log p)'(x)$ where $p(x) \propto \exp(-\left(\frac{x-3}{3}\right)^3)$.

Role of the assumptions on the kernel

Remember: For the Fourier representation and separation of measures we required $k(x, y) = \hat{\mu}(x - y)$.

- In finite dimensions, it's necessary and sufficient that k be translation invariant and continuous (by Bochner's theorem).
- In infinite dimensions additional strong smoothness conditions are required.
- For instance, $k(x, y) = \exp\left(-\frac{1}{2}\|x - y\|_{\mathcal{X}}^2\right)$ is not a Fourier transform of any measure.

Example

Consider the Squared Exponential (SE-T) and Inverse Multi Quadric-T (IMQ-T) kernels:

$$k_{SE-T}(x, y) = \exp\left(-\frac{1}{2}\|Tx - Ty\|_{\mathcal{X}}^2\right), \quad k_{IMQ-T}(x, y) = \left(\|Tx - Ty\|_{\mathcal{X}}^2 + 1\right)^{-1/2}.$$

Suppose that T is symmetric, positive-definite and trace class. Then the $SE - T^{1/2}$ and $IMQ - T^{1/2}$ are characteristic functions of measures with full support.

Additional result on the separation of measures

Theorem

Assume Q has bounded second moments and U satisfies the mild integrability and differentiability conditions:

- $\mathbb{E}_{X\sim Q}\left[\|CDU(X)\|_{\mathcal{X}}\right] < \infty$;
- $e^{-U(\cdot)/2} \in W_C^{1,2}(\mathcal{X})$;
- $\mathbb{E}_{X\sim N_C}\left[\|C^{1/2}DU(X)\|_{\mathcal{X}}^2\right] < \infty$.

Suppose that $T \in L(\mathcal{X})$ and T^* is surjective. If k is either SE-T or IMQ-T then

$$KSD_{\mathcal{A}_v,k}(Q,P) = 0 \Longleftrightarrow Q = P,$$

where $\mathcal{A}_v$ is a special *vectorised* version of the pre-conditioned Langevin operator instroduced before.

Goodness-of-fit testing

Framework for goodness-of-fit testing

Testing framework developed in Chwiałkowski et al. (2016) and Liu et al. (2016), adapted to our (potentially) infinite-dimensional setup:

- Given i.i.d. samples $\{X_i\}_{i=1}^n$ from Q consider the U-statistic:
$$\widehat{\mathrm{KSD}}(Q,P)^2 = \frac{1}{n(n-1)} \sum_{1 \le i \neq j \le n} h(X_i, X_j),$$
where h is the Stein kernel.
- Use a bootstrap procedure to generate samples.
- After generating bootstrap samples, reject the null if the test statistic falls outside a certain percentile of the empirical histogram.
- Computational cost: $O(n^2 BH)$, where n - number of data points, B - number of bootstrap repetitions, H - cost of evaluating h.

Numerical experiments

Experiments

- We will use the:
  - SE-$\gamma^{-1}T$ kernel $k_{\mathrm{SE}}(x,y) = \exp\left(-\frac{1}{2\gamma^2}\|Tx - Ty\|_{\mathcal{X}}^2\right)$;
  - IMQ-$\gamma^{-1}T$ kernel $k_{IMQ}(x,y) = \left(\gamma^{-2}\|Tx - Ty\|_{\mathcal{X}} + 1\right)^{-1/2}$.
- We use the median heuristic for γ and choose $\gamma = \mathrm{Med}\{\|TX_i - TX_j\|_{\mathcal{X}}, 1 \leq i \neq j \leq n\}$, where $\{X_j\}_{j=1}^n$ are i.i.d. samples from the unknown measure Q.
- We make the following choices for T:
  - $T_1 = I_{\mathcal{X}}$
  - $T_2 x = \sum_{i=1}^{\infty} \eta_i \langle x, e_i \rangle_{\mathcal{X}} e_i$, where $\eta_i = \lambda_i^{-1}$ for $1 \leq i \leq 50$ and $\eta_i = 1$ for $i > 50$ with e_i, λ_i the eigensystem of Brownian motion.

Example: Conditioned non-linear SDE

- $\mathcal{X} = L^2([0,50])$, $dX_t = 0.7\sin(X_t)dt + dW_t$, conditioned on $X_0 = X_{50} = 0$.
- N_C - Brownian bridge, U - from Girsanov theorem:

$$U(x) = \frac{1}{2}\int_0^{50} 0.49\sin(x(s))^2 + 0.7\cos(x(s))ds.$$

- Simulate samples using the piecewise-deterministic Markov process sampler from Bierkens et al. (2021).
- Consider deviations from the target by a deterministic drift $Y_t = X_t + \delta t/50$, for $\delta \in \mathbb{R}$. The null hypothesis is given by $\delta = 0$.

δ	SE-T_1	SE-T_2	IMQ-T_1	IMQ-T_2
0	0.08	**0.05**	0.07	**0.05**
0.05	0.17	0.17	**0.20**	0.18
0.1	0.43	0.4	**0.45**	0.43
0.15	**0.81**	0.77	0.79	0.77
0.2	**0.97**	0.95	0.96	0.96

Table 3. Proportion of times the null was rejected on the non-linear conditioned SDE experiment, δ denotes the parameter controlling the deviation from the null.

Example: Euler-Maruyama Discretisation Error

- Consider the same conditioned diffusion
- Use KSD to check how sensitive the SDE is to the arising discretization error arising from the Euler-Maruyama method.
- Use the IMQ-T kernel for $T \in \{T_1, T_2\}$.

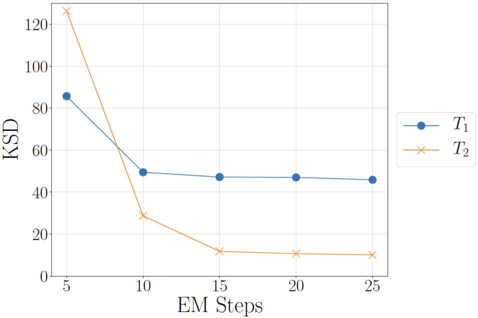


Figure 2: A plot of KSD using the IMQ kernel against the number of steps in the Euler-Maruyama simulation to simulate the target measure. The target measure is the conditioned SDE (25). The KSD value was estimated using 2000 samples of the Euler-Maruyama simulation, keeping the trajectories with $|X(50)| < 0.1$.

Conclusions

- KSD is well-defined for a wide array of kernels and infinite-dimensional targets.
- Infinite-dimensional KSD separates measures for certain commonly used kernels.
- Therefore, KSD may be used to test goodness of fit of functional data.
- The Fourier representation gives insight into the behaviour of KSD.
- There are many questions remaining!

G. Wynne, M.J. Kasprzak, A.B. Duncan: A Spectral Representation of Kernel Stein Discrepancy with Application to Goodness-of-Fit Tests for Measures on Infinite Dimensional Hilbert Spaces arXiv:2206.04552.