Kernel Mean Embedding for Dynamical Systems

A New Distribution-Free Concept for Representing, Comparing, and Propagating Uncertainty in Dynamical Systems

with Kernel Probabilistic Programming

Based on joint work with Krikamol Muandet,

Moritz Diehl, and Bernhard Schölkopf

Jia-Jie Zhu
Max Planck Institute for Intelligent Systems
Tübingen, Germany

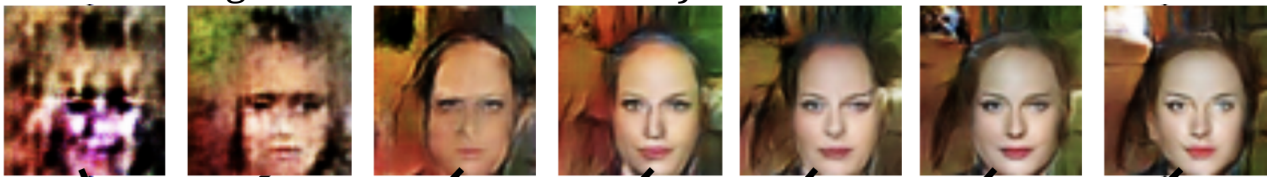
Learning with kernels

- Consider the inner product in finite dimensions

$$\langle x, x' \rangle$$

  This is a similarity measure.

- How do we generalize this similarity measure to more use cases?



Learning with kernels

- To scale up the similarity measures, consider applying a transformation to the space of interest

$$x \mapsto \phi(x)$$

- Evaluate the inner product in the mapped space,

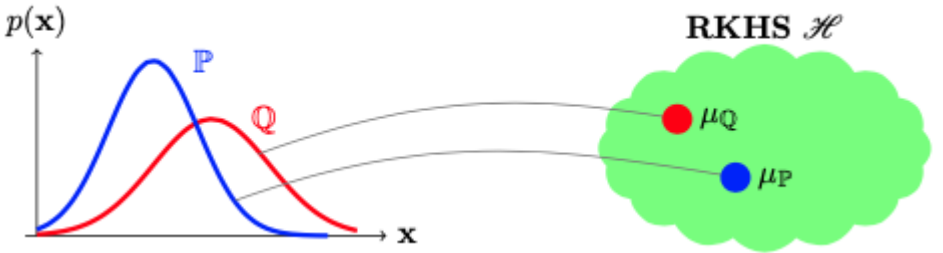
$$k(x, x') := \langle \phi(x), \phi(x') \rangle$$

  We refer to ϕ as a feature map and k as a kernel function.

Kernel mean embedding

- Recall a kernel is a symmetric, positive semi-definite bivariate function, e.g., $k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|_2^2\right)$.
- Kernel mean embedding (KME) maps probability distributions to functions in a Hilbert space, called the reproducing kernel Hilbert space (RKHS).

$$\mu : P \mapsto \int k(x, \cdot)\, dP(x), \quad \hat{\mu} : P \mapsto \sum_{i=1}^{N} \alpha_i k(x_i, \cdot),\ x_i \sim P$$



- μ can be thought of a generalized moment vector

Example: Second-order polynomial kernel embedding

- Polynomial kernel of order two $k(x, x') = (x^\top x' + 1)^2$
- The kernel mean embedding is given by

$$\mu_X = \int k(x, \cdot) dP(x) = \int \left(x^\top (\cdot) + 1\right)^2 dP(x)$$
$$= (\cdot)^\top \mathbb{E} x x^\top (\cdot) + 2\mathbb{E} x^\top (\cdot) + 1$$

- The embedding keeps track of the mean and variance of X.
- Universal kernel (e.g., Gaussian, $k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|_2^2\right)$.) keeps track of infinite-order moments

Embedding dynamical system

In a nutshell, we represent the distribution of x_t, the state of the dynamical systems (continuous or discrete time), by its KME.

$$\mu_{x_t} = \int k(x_t, \cdot) dP(\xi), \quad \hat{\mu}_{x_t} = \sum_{i=1}^{N} \alpha_i k(x_t^{(i)}, \cdot)$$

for some weights α_i.

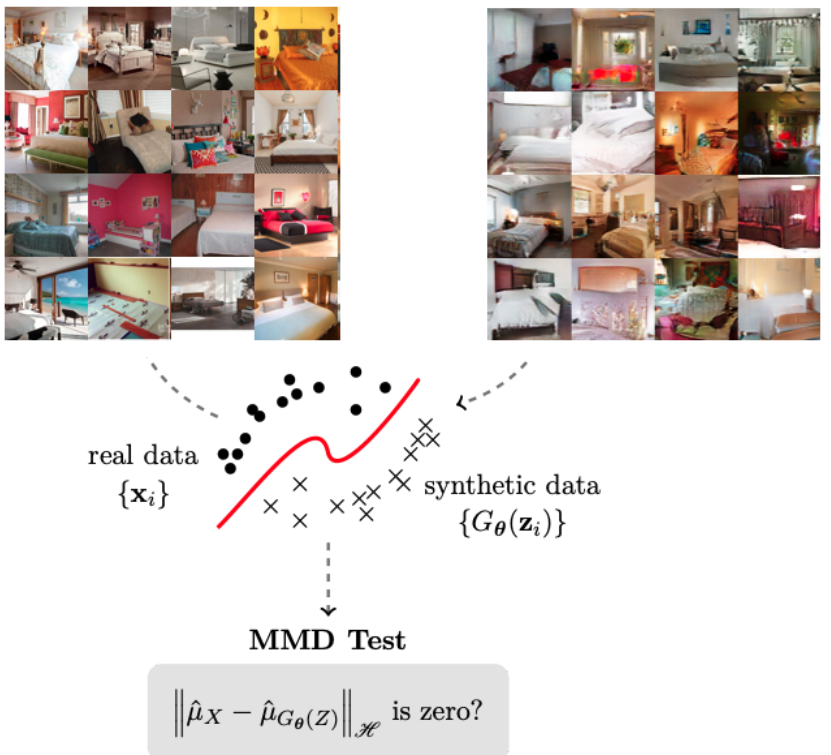
- (Statistical consistency). The embedding estimator $\hat{\mu}_{\hat{x}(t,\xi)}$ produced by a one-step numerical integration rule with step size h is consistent, i.e.,

$$\hat{\mu}_{\hat{x}(t,\xi,h)} \to \mu_{x(t,\xi)}, \forall t, \text{ as } N \to \infty, \ h \to 0$$

The maximum mean discrepancy (MMD)

Given RKHS $\mathcal{H}$,

$$MMD(\mathcal{H}, P, Q) := \sup_{f \in \mathcal{H}} \{ \int f dP - \int f dQ \} = \|\mu_P - \mu_Q\|_{\mathcal{H}}$$



The maximum mean discrepancy (MMD)

- Given two sets of samples $\{x_i\}_{i=1}^M$ and $\{y_i\}_{i=1}^N$ from simulations of two dynamical systems, a sample-based estimator for $\|\mu_{x_t} - \mu_{y_t}\|_{\mathcal{H}}$ is given by

$$\frac{1}{M^2}\sum_{i,j=1}^{M} k(x_i, x_j) - \frac{2}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} k(x_i, y_j) + \frac{1}{N^2}\sum_{i,j=1}^{N} k(y_i, y_j)$$

- In essense, MMD gives a new metric in the probability simplex. (other metrics include the Wasserstein distance, ϕ-divergence.)

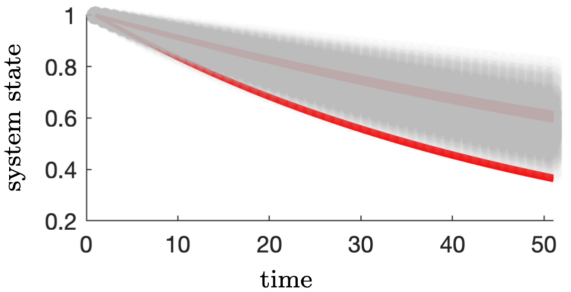
- See Z et al., (Kernel DRO), for deeper mathematical connections between MMD and robust optimization

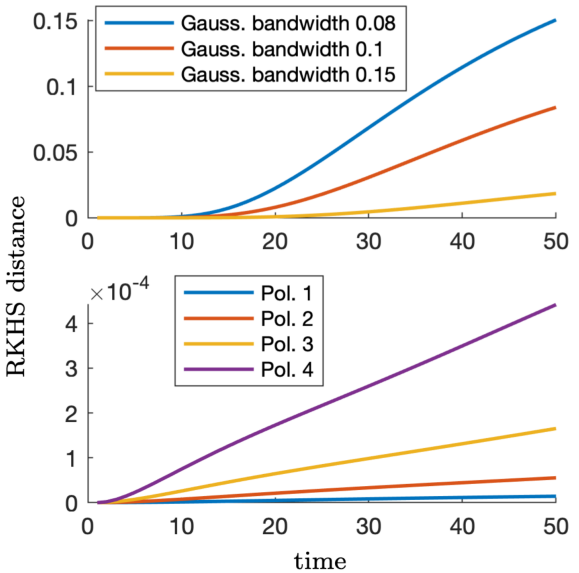
Example: comparing two dynamical system

$$\dot{x}(t) = \xi\, x, \quad x(0) = x_0$$

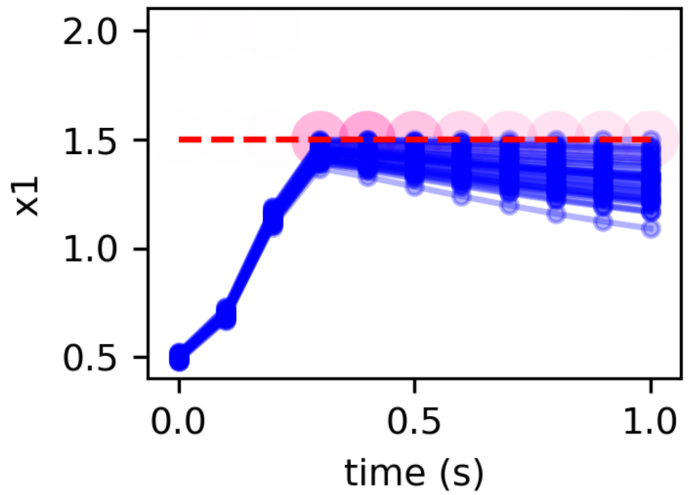
$$\xi \sim \mathrm{GMM},\ \xi' \sim \mathrm{N}(m, \sigma)$$

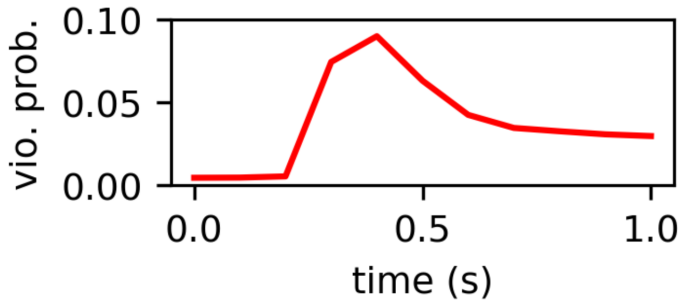
$$\mathbb{E}\xi = \mathbb{E}\xi', \mathbb{E}\xi^2 = \mathbb{E}\xi'^2.$$





Application: estimating chance constraint violation probability





What if P is uncertain? We can solve the kernel moment problem:

$$\sup_{\mathbb{P}} \mathbb{P}(c(x) > 0)$$

$$\text{subject to } \|\mu_{\mathbb{P}} - \mu_{\hat{\mathbb{P}}}\|_{\mathcal{H}} \leq \epsilon$$

Going deeper: distributionally robust optimization with kernels?

- Universal kernel (e.g., Gaussian, $k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|_2^2\right).$) keeps track of infinite-order moments.
- Mathematically, the MMD associated with a universal kernel is a proper metric in the probability simplex. Hence, one can perform DRO with it

$$\min_{\theta} \sup_{P} \left\{ \int l(\theta, \xi)\, dP(\xi) : \mathrm{MMD}(P, \hat{P}) \leq \epsilon \right\}$$

- This can be reformulated using conic duality as a tractable program. (Z. et al. 2020, Kernel DRO).

Takeaway

- This paper proposed to use kernel mean embedding as a tool for dynamical systems.
- RKHS inuduces a metric in probability measures, the MMD.
- Possible applications:
  - System identification: MMD can be viewed as a test statistic
  - Kernel distributionally robust optimization: risk-averse optimization under distributional ambiguity (K-DRO. See a subsequent work in *Z et al., 2020*.)

Thank you! This talk is based on

- **Z, Muandet, Diehl, Schölkopf, 2019. A New Distribution-Free Concept for Representing, Comparing and Propagating Uncertainty in Dynamical Systems with Kernel Probabilistic Programming. IFAC 2020**
- **Z, Jitkrittum, Diehl, Schölkopf, 2020. Kernel Distributionally Robust Optimization. Arxiv preprint**
- Z, Jitkrittum, Diehl, Schölkopf, 2020. Worst-Case Risk Quantification under Distributional Ambiguity using Kernel Mean Embedding in Moment Problem. Arxiv preprint
- Z, Diehl, Schölkopf, 2020. A Kernel Mean Embedding Approach to Reducing Conservativeness in Stochastic Programming and Control. L4DC 2020

For more information: contact me at jzhu@tuebingen.mpg.de