Kernel methods

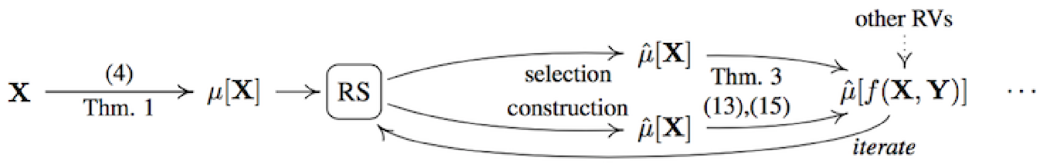

Figure 1.2: Overview of the application of kernel mean embedding in [43]. Together with the reduced set (RS) techniques to limit the complexity of the RKHS expansion, the kernel mean embedding is used to approximate the embedding of the functional of random variables $Z = f(X,Y)$.

A Hilbert space embedding of distributions (KME)—which generalizes the feature map of individual points to probability measures—has emerged as a powerful machinery for probabilistic modeling, machine learning, and causal discovery. The idea behind this framework is to map distributions into a reproducing kernel Hilbert space (RKHS) endowed with a kernel k. It enables us to apply RKHS methods to probability measures and has given rise to a great deal of research and novel applications of kernel methods.

Given an i.i.d. sample $x_1, x_2, \dots, x_n$ from $\mathbb{P}$, the most natural estimate of the embedding $\mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X,\cdot)]$ is an empirical average $\hat{\mu}_{\mathbb{P}} = (1/n)\sum_{i=1}^{n} k(x_i,\cdot)$. In [35, 370], we showed that this estimator is not optimal in a certain sense. Inspired by James-Stein estimator, we proposed the so-called kernel mean shrinkage estimators (KMSEs) which improves upon the standard estimator. A suitable explanation for the improvement is a bias-variance tradeoff: the shrinkage estimator reduces variance substantially at the expense of a small bias. In addition, we presented a class of estimators called spectral shrinkage estimators in [395] which also incorporates the RKHS structure via the eigenspectrum of the empirical covariance operator. Our empirical studies suggest that the proposed estimators are very useful for "large p, small n" situations (e.g. medical data, gene expression analysis, and text documents).

A natural application of KME is in testing for similarities between samples from distributions. We refer to the distance between two distribution embeddings as the maximum mean discrepancy (MMD). We have formulated a two-sample test [142] (of whether two distributions are the same), and showed that the independence test (of whether two random variables observed together are statistically independent) is a special case. A further application of the MMD as independence criterion is in feature selection, where we maximize dependence between features and labels [143]. We have further developed alternative independence tests based on space partitioning approaches and classical divergence measures (such as the ℓ_1 distance and KL-divergence) [268]. Lastly, we also constructed the test for non-i.i.d. data such as time-series in [441].

Given that the MMD depends on the particular kernel that is chosen, we proposed two kernel selection strategies [494], the earlier one relying on a classification interpretation of the MMD, and the later one explicitly minimizing the probability of Type II error of the associated two-sample test (that is, the probability of wrongly accepting that two unlike distributions are the same, given samples from each).

We have also used the KME to develop a variant of an SVM which operates on distributions rather than points [478], permitting modeling of input uncertainties. One can prove a generalized representer theorem for this case, and in the special case of Gaussian input uncertainties and Gaussian kernel SVMs, it leads to a multiscale SVM, akin to an RBF network with variable widths, which is still trained by solving a quadratic optimization problem. In [356], we applied this framework to perform bivariate causal inference between X and Y as a classification problem on joint distribution $\mathbb{P}(X,Y)$. Another interesting application is in domain adaptation [407, 676]. This idea has also been extended to develop a variant of One-class SVM that operates on distributions, leading to applications in group anomaly detection [415].

A recent application uses kernel means in visualization. When using a power-of-cosine kernel for distributions on the projective sphere, the kernel mean can be represented as a symmetric tensor. In the context of diffusion MRI, this permits an efficient visual and quantitative analysis of the uncertainty in nerve fiber estimates, which can inform the choice of MR acquisition schemes and mathematical models [110, 388].

A natural question to consider is whether the MMD constitutes a metric on distributions, and is zero if and only if the distributions are the same. When this holds, the RKHS is said to be characteristic. We have determined necessary and sufficient conditions on translation invariant kernels for injectivity, for distributions on compact and non-compact subsets of $\mathbb{R}^d$ [253]: specifically, the Fourier transform of the kernel should be supported on all of $\mathbb{R}^d$. Gaussian, Laplace, and B-spline kernels satisfy this requirement. The MMD is a member of a larger class of metrics on distributions, known as the integral probability metrics (IPMs). In [16, 4], we provide estimates of IPMs on $\mathbb{R}^d$ which are taken over function classes that are not RKHSs, namely the Wasserstein distance (functions in the unit Lipschitz semi-norm ball) and the Dudley metric (functions in the unit bounded Lipschitz norm ball), and establish strong consistency of our estimators. Comparing the MMD and these two distances, the MMD converges fastest, and at a rate independent of the dimensionality d of the random variables – by contrast, rates for the classical Wasserstein and Dudley metrics worsen when d grows.

Embeddings of distributions can be generalized to yield embeddings of conditional distributions. The first application is to Bayesian inference on graphical models. We have developed two approaches: in the first [556, 602], the messages are conditional density functions, subject to smoothness constraints; these were orders of magnitude faster than competing nonparametric BP approaches, yet more accurate, on problems including depth reconstruction from 2-D images and robot orientation recovery. In the second approach [558], conditional distributions $P(Y|X = x)$ are represented directly as embeddings in the RKHS, allowing greater generality (for instance, one can define distributions over structured objects such as strings or graphs, for which probability densities may not exist). We showed the conditional mean embedding to be a solution to a vector valued regression problem [492], which allows us to formulate sparse estimates. The second application is to reinforcement learning. In [491], we estimate the optimal value function for a Markov decision process using conditional distribution embeddings, and the associated policy. This work was generalized to partially observable Markov decision processes in [462], where the kernel Bayes' rule was used to integrate over distributions of the hidden states.

Another important application of conditional mean embeddings is in testing for conditional independence (CI). We proposed a Kernel-based Conditional Independence test (KCI-test) [519] which avoids the classical drawbacks of CI testing. Most importantly, we further derived its asymptotic distribution under the null hypothesis, and provided ways to estimate such a distribution. Our method is computationally appealing and is less sensitive to the dimensionality of Z compared to other methods. This is the first time that the null distribution of the kernel-based statistic for CI testing has been derived. Recently, we proposed a new permutation-based CI test [390] that easily allows the incorporation of prior knowledge during the permutation step, has power competitive with state-of-the-art kernel CI tests, and accurately estimates the null distribution of the test statistic, even as the dimensionality of the conditioning variable grows.

Lastly, we have recently leveraged the KME in computing functionals of random variables $Z = f(X_1, X_2, \ldots, X_n)$ [43], which is ubiquitous in various applications such as probabilistic programming. Our approach allows us to obtain the distribution embedding of Z directly from the embeddings of $X_1, X_2, \ldots, X_n$ without resorting to density estimation. It is in principle applicable to all functional operations and data types, thank to the generality of kernel methods. Based on the proposed framework, we showed how it can be applied to non-parametric structural equation models, with an application to causal inference. As an aside, we have also developed algorithms based on distribution embedding for identifying confounders [422], which is one of the most fundamental problems in causal inference**.**

More information: https://ei.is.tuebingen.mpg.de/project/kernel-distribution-embeddings