RESEARCH STATEMENT

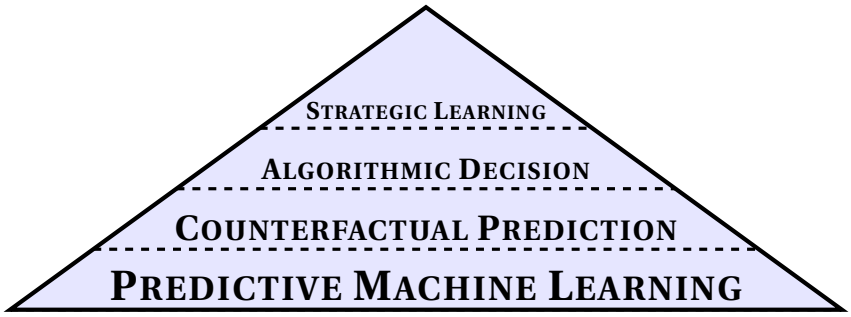
KRIKAMOL **MUANDET**

Max Planck Institute for Intelligent Systems • Max-Planck-Ring 4 • Tübingen 72076 • Germany
http://krikamol.org • krikamol@tuebingen.mpg.de • ☎ +49 (0)7071 601 559

Summary

I have a broad interest in **machine learning** and the synergy between this exciting field and other disciplines including economics, social science, and mechanism design. The emergence of machine learning as a new problem solver for governmental organizations, private sectors, and scientific institutions has not only increased its value considerably, but has also raised pressing concerns and challenges that require immediate attention such as algorithmic bias, discrimination, and unfairness. *The goal of my research is to understand—both theoretically and practically—different ingredients of machine learning (e.g., data collection, algorithm design, and deployment) and how they interact, which can then be used to develop novel learning algorithms that can help solve the aforementioned problems.* I believe that my research will not only increase the practical merit of machine learning in improving the welfare of our society, but will also provide insights into the foundation of learning that paves the way for building a truly intelligent machine, which is an ultimate goal of artificial intelligence.

The following diagram depicts my current research interest starting from predictive machine learning as a foundation to a complex form of strategic learning. A summary of my current research is also given below.



- **Modern Kernel Methods, Mean Embeddings, and Beyond**—The backbone of kernel methods is a positive definite kernel $k(x,x')$ which defines a data representation $\phi(x),\phi(x')$ *implicitly* in a so-called *reproducing kernel Hilbert space* (RKHS) $\mathscr{H}$ such that $k(x,x') = \langle \phi(x),\phi(x')\rangle_{\mathscr{H}}$. Besides their mathematical elegance, kernel methods are among the most popular and powerful techniques for predictive machine learning because (i) they are flexible and can work with a variety of data types (*e.g.*, strings, graphs, and groups) in an integrated framework, (ii) we can incorporate prior knowledge about the learning problem through diverse choices of kernel functions, and (iii) there exist abundant learning algorithms that support kernel functions. Recently, the kernel function becomes an important ingredient of *neural tangent kernel* (NTK) which offers theoretical insights into the training dynamics of deep neural networks [1]. Further, a *kernel mean embedding* (KME) $\mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[k(x,\cdot)] =: \mu_{\mathbb{P}}$ extends the whole arsenal of kernel methods to probability measures [2, 3, 4], leading to new research opportunities in deep learning [5, 6], causality [7, 8, 9], econometrics [10], among others. I am constantly pushing for a wide range of new and exciting applications of kernel methods.

- **Learning from Probability Measures**—In traditional machine learning, data are assumed i.i.d. points $z_1,\ldots,z_n \in \mathscr{Z}$ that are distributed according to some unknown distribution $\mathbb{P}$. In many scenarios, however, data may arrive in the form of probability distributions $\mathbb{P}_1,\ldots,\mathbb{P}_n$ over the input space $\mathscr{Z}$, which are assumed i.i.d according to some unknown *meta-distribution* $\mathscr{M}$ over a space of probability measures $\mathscr{P}$. The meta-distribution $\mathscr{M}$ governs a data generating process at a higher level. For example, learning with uncertain inputs [11], group anomaly detection [12], and causal inference [8] can be viewed as learning problems with probability measures as input data. The kernel mean embeddings $\mu_{\mathbb{P}_1},\mu_{\mathbb{P}_2},\ldots,\mu_{\mathbb{P}_n}$ offer a theoretically elegant, computationally efficient, and flexible representation for predictive machine learning on these distributions. My interest lies in developing novel learning algorithms for distributions with applications in out-of-distribution generalization [13], meta-learning [6], and counterfactual prediction [7]. I believe that my research in this direction will contribute to the quest for machines that generalize better across different environments.

- **Causal Learning and Counterfactual Prediction**—Cause-effect inference is a grand challenge of scientific research. Insights into causal relationships can help assess the consequences of certain interventions such as medical treatments or public policies. The causal understanding of diseases like cancer and possible treatments could ultimately help save million of lives around the world. Furthermore, causal knowledge is a foundation

of *counterfactual* prediction, which is an essential part in decision making. In practice, however, experimental data can be unethical, expensive, or even impossible to collect, so we must rely on observational data which can suffer from the confounding bias. Observational studies are ubiquitous in medical diagnosis, recommendation systems, and personalization. According to the so-called Reichenbach's principle, the dependence between two random variables implies that either one causes the other, or that they simply have a common cause. Put differently, a straightforward deployment of learning algorithms alone is insufficient for solving causal problems—but it certainly can help. My interest in this direction lies in developing novel algorithms that can achieve better counterfactual prediction from observational data [7, 14, 15].

- **Strategic and Algorithmic Decision Making**—While machines that outperform human at prediction tasks have become inexpensive to build, machines that can mimic *human judgement and decision* is still out of reach. A rapid deployment of predictive models in important sectors such as online advertising, criminal justice, education, and labor market has raised widespread concerns about the societal impact of algorithmic decisions. Complex decisions based on model predictions alone—even when they are perfect—can lead to unintended consequences such as discrimination and unfairness [16]. Unlike a prediction, making a good decision from past data can be challenging because (i) it involves makeing counterfactual prediction and reasoning about potential outcomes, (ii) there can be many, interrelated factors or alternatives to consider, e.g., self-driving cars, (iii) the impact of the decision may be significant, e.g., health care and justice system, and (iv) the decisions may be subject to a strategic behavior. My goal is to bridge the gap between machine prediction and human judgement by exploring ideas in economics, mechanism design, game theory, and offline reinforcement learning.

Modern Kernel Methods, Mean Embeddings, and Beyond

Many classical learning algorithms such as principal component analysis (PCA), support vector machines (SVMs), and Gaussian processes (GPs) can be expressed entirely in terms of the inner product $\langle x, x' \rangle$. Kernel methods enable us to construct their nonlinear counterparts simply by replacing $\langle x, x' \rangle$ with positive definite kernel $k(x, x')$ which corresponds to an inner product $\langle \phi(x), \phi(x') \rangle_{\mathscr{H}}$ in a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ of functions on $\mathscr{X}$. The success of kernel methods in practice—together with their beautiful theories—lends themselves to numerous applications and makes them one of the most popular techniques in machine learning.

I have been working on the *kernel mean embedding of distributions* [2, 3, 4] which is defined as a map $\mu : \mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[k(x,\cdot)] =: \mu_{\mathbb{P}}$ for a probability distribution $\mathbb{P}$ over some measurable space $(\mathscr{X}, \mathscr{A})$. One may think of $\mu_{\mathbb{P}}$ as a feature map of the distribution $\mathbb{P}$. When $\mathbb{P} = \delta_x$, *i.e.*, a Dirac measure at x, for some $x \in \mathscr{X}$, $\mu_{\mathbb{P}}$ reduces to the canonical feature map $k(x,\cdot)$ as a special case. This representation enables the whole arsenal of kernel methods to novel applications in two-sample testing, causal inference, probabilistic inference, and reinforcement learning. For instance, in [9], my colleagues and I showed that more general operations—such as multiplication and exponentiation—on random variables, *i.e.*, $Z = h(X, Y)$, can also be performed via kernel mean embedding. Moreover, for characteristic kernels such as Gaussian and Laplace kernels, the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective. Hence, we do not lose any information about $\mathbb{P}$ when adopting $\mu_{\mathbb{P}}$ as its representation. This is an appealing property for many applications such as conditional dependence [17], Bayesian nonparametric [18], and quantum computing [19]. In practice, the kernel mean $\mu_{\mathbb{P}}$ can be estimated simply by the empirical kernel mean $\hat{\mu}_{\mathbb{P}} = (1/n)\sum_{i=1}^{n} k(x_i, \cdot)$ where $x_1, x_2, \ldots, x_n$ is an i.i.d. sample from $\mathbb{P}$. Conditional distributions can also be embedded into an RKHS in a similar manner. In [20], we overcome one of the critical limitations of the classical definition of conditional mean embedding (CME) by taking a measure-theoretic approach to define the CME as a random variable taking values in a reproducing kernel Hilbert space. This new definition of CME does not depends on stringent assumptions that hinder its analysis of the existing operator-based approach.

The kernel mean $\mu_{\mathbb{P}}$ is central to the core inference step of modern kernel methods which rely on embedding probability distributions in RKHSs. Since most of them rely on the empirical estimate $\hat{\mu}_{\mathbb{P}}$, it is fundamental to ask if the estimation of $\hat{\mu}_{\mathbb{P}}$ can be improved. In [21, 22], my colleagues and I showed that this estimator is in a certain sense not optimal. We proposed a new class of estimators called *kernel mean shrinkage estimators* (KMSEs) which we showed to be "better" than the classical estimator, though it is shown to be minimax optimal [23]. These works are to some extent inspired by Stein's seminal work in 1955, which showed that the maximum likelihood estimator (MLE) of the mean $\boldsymbol{\theta}$ of a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ is "inadmissible" [24]—*i.e.*, there exist a better estimator—though it is minimax optimal. Our setting, however, is more general and thus differs fundamentally from Stein's work in that it is *non-parametric* and involves a *non-linear feature map* into a high-dimensional space $\mathscr{H}$. Subsequent observations and insights allowed us to construct a *nonlinear* estimator in RKHS via spectral filtering algorithms developed originally for supervised learning problems [25]. This estimator also takes into account the geometric structure of RKHS $\mathscr{H}$ encoded in the eigenspectrum of the empirical covariance operator. The theoretical aspects of the latter remain open questions which I plan to investigate in future works.

Our findings suggest some interesting research directions. Firstly, our shrinkage estimators can be used to estimate (cross-) covariance operators and tensors of higher order in RKHS as has already been used, for example, in increasing the power of kernel independence test [26]. Furthermore, I have observed that the improvement of the KMSE over the classical estimator is substantial in the "large p, small n" paradigm. In particular, the improvement

increases as the dimensionality p grows. This phenomenon is surprising as it is generally believed that the ambient dimension p of data does not significantly affects the performance of kernel methods (*i.e.*, the estimation of an empirical kernel mean). Understanding this phenomenon may shed light on the ultimate kernel choice problem. I believe that research in this direction will subsequently foster our understanding of a fundamental link between Stein estimation in statistics and Tikhonov regularization in inverse problems.

Moving beyond probability measures, the most recent line of my work is *conditional moment embedding* (CMME) [10], which is a synergy between kernel methods and econometrics. The CMME is a Hilbert space embedding of a *conditional moment restriction* (CMR): for correctly specified models, the conditional mean of certain functions of data is almost surely equal to zero. Many problems in economics can be formulated in terms of the CMR. Inspired by the maximum mean discrepancy (MMD), my colleague and I develop a measure of moment restriction called a *maximum moment restriction* (MMR). Not only can the MMR capture all information about the original CMR, but it also has a closed-form expression that enables the practical ease of implementation.

Learning from Probability Measures

A distribution can represent highly structured and high-level regularity in the data which makes it suitable for problems in transfer learning, domain adaptation, and statistical inference. When the measurement is noisy, we may incorporate that uncertainty by treating the data points themselves as distributions. This is often the case for microarray data and astronomical data in which the measurement process is imprecise. In order to obtain reliable data, costly and time-consuming measurements or experiments have to be replicated. Moreover, distributions not only embody individual data points, but also contain information about their interactions which can be beneficial for structural learning in fields such as high-energy physics, cosmology, and causality. Lastly, classical problems in statistics such as statistical estimation, hypothesis testing, and causal inference may be interpreted in a decision-theoretic sense as learning a function that maps empirical distributions to the desired statistics, which is in contrast to standard estimation based on plug-in estimators. Rephrasing these problems in this way leads to novel approach for statistical estimation.

Relying on the kernel mean embedding, my colleagues and I have developed a frameworks called *distributional risk minimization* (DRM) that operates directly on a space of distributions by representing them as $\mu_{\mathbb{P}_1}, \ldots, \mu_{\mathbb{P}_n}$ in some RKHS $\mathscr{H}$ [11, 12]. Compared to the contemporary divergence method, kernel density estimation, and probabilistic models, our framework requires minimal assumptions on the distributions, can learn more efficiently, and achieves superior performance on some tasks. In the supervised setting, we showed that the representer theorem [27] for probability distributions holds and reduces to its data-point counterpart when the inputs are Dirac measures $\delta_{x_1}, \ldots, \delta_{x_n}$. Based on this framework, we proposed the *support measure machines* (SMM) which is a generalization of SVM to probability space [11]. Moreover, we developed one-class SMM (OCSMM) [12] with connections to variable kernel density estimation (VKDE) and novel applications in group anomaly detection on astronomical data and high-energy physics data.

In [13], we improve an out-of-distribution generalization by investigating the *domain generalization* problem with applications to flow cytometry analysis. In this case, we have data from m domains $\mathbb{P}_1(X,Y), \mathbb{P}_2(X,Y), \ldots, \mathbb{P}_m(X,Y)$ (*e.g.*, data from m patients) and the goal is to train a classifier that generalizes well to the previously unseen domains $\mathbb{P}^*$ (*i.e.*, new patients). Our idea is to learn a representation that is invariant across the training distributions $\mathbb{P}_1(X,Y), \mathbb{P}_2(X,Y), \ldots, \mathbb{P}_m(X,Y)$ using their kernel mean embeddings. To allow for better generalization, my colleague and I have also recently proposed a novel task representation called *model-aware task embedding* (MATE) that incorporates not only the data distributions of different tasks, but also the complexity of the tasks through the models used [6]. I believe that this line of work has a potential in medical research and healthcare as it will enable doctors to quickly transfer the diagnosis and treatment to future patients.

Several exciting open questions persist. For instance, in flow cytometry, the distributions $\mathbb{P}(X,Y)$ may correspond to different patients, whereas the domain on which X and Y are defined is essentially the same across patients. Hence, the *domain-specific knowledge* can be generalized across domains by accounting for change in distributions. In contrast, many problems in statistics involve statistical properties such as independence and causal relation which are transferable across domains. Specifically, X and Y are said to be independent if $\mathbb{P}(X,Y) = \mathbb{P}(X)\mathbb{P}(Y)$ regardless of what X and Y represent. This implies that—to be able to generalize well the *domain-general knowledge*—we need to also deal with the invariant representation of X and Y across domains. Learning the right invariance from data remains one of the most challenging problems in machine learning.

Causal Learning and Counterfactual Prediction

Counterfactual reasoning is a hallmark of human thought which enables human to imagine possible futures that could have happened; something that is contrary to what actually happened. This form of learning requires algorithms that can infer causal relationships from observational data. Relying on the learning framework for distributions, my colleagues and I have explored the possibility of solving bivariate causal inference—*i.e.*, deciding if X causes Y, or vice versa from the observation $\{(x_i, y_i)\}_{i=1}^m$—as a classification on joint distributions $\mathbb{P}(X,Y)$ [8]. In contrast to classical approaches which rely on specific causal assumptions, we assume access to a training sample $\{(\hat{\mathbb{P}}_i(X,Y), l_i)\}_{i=1}^n$ where $\hat{\mathbb{P}}_i(X,Y)$ are empirical distributions and $l_i \in \{-1,+1\}$ indicates whether X causes Y, or vice

versa. Using the kernel mean representation of $\mathbb{P}_i(X,Y)$, we train a classifier on $\{(\hat{\mu}_{\mathbb{P}_i(X,Y)}, l_i)\}_{i=1}^n$ and then use it to infer causal direction. Our approach outperforms some classical causal inference algorithms, demonstrating the benefit of machine learning in causal inference. Moreover, since X and Y may correspond to semantically different variables, this raises a philosophical question as to what kind of knowledge is being learned. Causal inference involves the investigation of how the distribution of outcome changes as a result of some intervention, so I believe that learning framework on distributions can be quite useful in this direction.

Structural equation models and graphical models are among the most popular tools in causal inference. However, another framework that I find appealing is the *potential outcomes framework* used primarily in political science, social science, and epidemiology. In this framework, the effect of a treatment $T = 1$ (vs. a control condition $T = 0$) on an outcome Y for a subject i is expressed as a difference between two potential outcomes $Y_i(1) - Y_i(0)$ where $Y_i(1)$ represents the value of the outcome the subject would experience if exposed to the treatment, and $Y_i(0)$ represents the outcome if the subject is exposed to the control. The average treatment effect ATE $= \mathbb{E}[Y(0) - Y(1)] = \mathbb{E}[Y(0)] - \mathbb{E}[Y(1)]$ is often used to characterize the causal effect. Unfortunately, we can only observe either $Y_i(0)$ or $Y_i(1)$ for each individual i due to the *fundamental problem of causal inference* (*e.g.*, one cannot both take the pill and not take the pill at the same time). Moreover, the ATE has been used only for real-valued outcomes, restricting the potential of this framework. In our recent work [7], my colleague and I propose a *distributional treatment effect* (DTE) that allows for causal inference over the entire landscape of the counterfactual distribution. The DTE relies on a novel kernel mean embedding of counterfactual distribution that we call a *counterfactual mean embedding* (CME).

My previous works have paved the way for me to recognize one of the most challenging obstacles in causal inference, namely, *hidden confounders*, i.e., unobserved common causes between treatment and outcomes. When a randomized control trials is infeasible, an *instrumental variable* (IV) have become standard tools for economists, epidemiologists, and social scientists to uncover causal relationships from observational data when hidden confounders exist. To improve learning with instrumental variables, my colleague and I have recently developed simple kernel-based algorithms that allow practitioners to efficiently perform IV regression in a nonlinear setting [14, 15].

Strategic and Algorithmic Decision Making

An accurate counterfactual reasoning will enable agents to make better decisions as it gives them the flexibility in thinking about possible outcomes of actions that are different from the ones taken in the past. However, an adoption of such agents in the real-world setting faces several key challenges: First, in critical applications like education and health care, the agents cannot collect data by interacting directly with the environment because the cost of such decisions are high. Hence, the decisions must be learned from historical data which can potentially be biased. In [7], my colleague and I propose to use the counterfactual mean embedding (CME), a kernel mean embedding of counterfactual distributions, for off-policy evaluation task. Based on the CME, we show that under reasonable assumptions the historical data collected from a logged policy can be used to consistently evaluate the new policy without actually implementing it. This work allows the agents to effectively assess the quality of their decisions when the interaction with the environment is limited, if not impossible. The second challenge is a *feedback loop* in the data collection process. That is, the historical data are used in building the models which in turn are used to collect more data to improve the current models. In loan decisions, for example, a bank may decide whether or not to offer a loan based on learned models of the credit default. These decisions generate more data that are used to improve the models. However, my colleague and I analyze in [16] consequential decision making using imperfect predictive models, which are learned from data gathered by potentially biased historical decisions. We articulate that when starting with a nonoptimal deterministic policy, this approach fails to optimize utility in for sequential decisions. To avoid this failure mode while respecting a common fairness constraint, we suggest to directly learn the decisiobs with exploring policies. Last but not least, the impact of the decision can be significant. The widespread adoption of machine learning models in important sectors such as online advertising, education, criminal justice, and labor market, and health care raises general concerns on the extent to which they could potentially change someone's life. As a result, the decisions themselves may be subject to a strategic behavior. Understanding the strategic behavior that arises from interactions between human and machines is one of my future research directions.

References

[1] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 8571–8580, Curran Associates, Inc., 2018.

[2] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[3] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pp. 13–31, Springer-Verlag, 2007.

[4] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[5] Y. Zhang, S. Tang, K. Muandet, C. Jarvers, and H. Neumann, "Local temporal bilinear pooling for fine-grained action parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] X. Chen, Z. Wang, S. Tang, and K. Muandet, "MATE: Plugging in model awareness to task embedding for meta learning," in *NeurIPS*, vol. Forthcoming, 2020.

[7] K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat, "Counterfactual mean embedding," *JMLR*, vol. Accepted, 2020.

[8] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a learning theory of cause-effect inference," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, pp. 1452–1461, JMLR, 2015.

[9] B. Schölkopf, K. Muandet, K. Fukumizu, and J. Peters, "Computing functions of random variables via reproducing kernel Hilbert space representations," *Statistics and Computing*, vol. 25, no. 4, pp. 755–766, 2015.

[10] K. Muandet, W. Jitkrittum, and J. Kübler, "Kernel conditional moment test via maximum moment restriction," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, vol. 124 of *Proceedings of Machine Learning Research*, pp. 41–50, PMLR, 2020.

[11] K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf, "Learning from distributions via support measure machines," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 10–18, 2012.

[12] K. Muandet and B. Schölkopf, "One-class support measure machines for group anomaly detection," in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 449–458, AUAI Press, 2013.

[13] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 10–18, JMLR, 2013.

[14] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, "Dual instrumental variable regression," in *NeurIPS*, vol. Forthcoming, 2020.

[15] R. Zhang, M. Imaizumi, B. Schölkopf, and K. Muandet, "Maximum moment restriction for instrumental variable regression," *ArXiv Preprint*, 2020.

[16] N. Kilbertus, M. G. Rodriguez, B. Schölkopf, K. Muandet, and I. Valera, "Fair decisions despite imperfect predictions," in *AISTATS*, vol. 108, pp. 277–287, PMLR, 2020.

[17] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf, "A permutation-based kernel conditional independence test," in *30th Conference on Uncertainty in Artificial Intelligence (UAI2014)*, pp. 132–141, 2014.

[18] K. Muandet, "Hilbert space embedding for dirichlet process mixtures." NIPS 2012 Workshop on confluence between kernel methods and graphical models (oral presentation), Dec 2012.

[19] J. M. Kübler, K. Muandet, and B. Schölkopf, "Quantum mean embedding of probability distributions," *Physical Review Research*, vol. 1, no. 3, p. 033159, 2019.

[20] J. Park and K. Muandet, "A measure-theoretic approach to kernel conditional mean embeddings," in *NeurIPS*, vol. Forthcoming, 2020.

[21] K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Kernel mean estimation and Stein effect," in *Proceedings of The 31st International Conference on Machine Learning*, vol. 32, pp. 10–18, JMLR, 2014.

[22] K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf, "Kernel mean shrinkage estimators," *Journal of Machine Learning Research*, vol. 17, no. 48, pp. 1–41, 2016.

[23] I. Tolstikhin, B. Sriperumbudur, and K. Muandet, "Minimax estimation of kernel mean embeddings," *CoRR*, vol. abs/:1602.04361, 2016.

[24] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 197–206, University of California Press, 1955.

[25] K. Muandet, B. Sriperumbudur, and B. Schölkopf, "Kernel mean estimation via spectral filtering," in *Advances in Neural Information Processing Systems 27*, pp. 10–18, Curran Associates, Inc., 2014.

[26] A. Ramdas and L. Wehbe, "Nonparametric independence testing for small sample sizes," in *Proceedings of the 2015 International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3777–3783, 2015.

[27] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pp. 416–426, Springer-Verlag, 2001.