Krikamol **Muandet**, *Dr. Rer. Nat.*

Mahidol University • 272 Rama VI Road • Ratchathewi District • Bangkok 10400 • Thailand
http://krikamol.org • krikamol@gmail.com • +66 0-2201-5344

Research Statement

I have a broad interest in machine learning. The emergence of data as a new natural resource for governmental organizations, private sectors, and scientific institutions has increased its value considerably. The goal of my research are to understand—both theoretically and practically—different ingredients of machine learning and how they interact, and then to leverage this knowledge in developing better learning algorithms that will contribute to this movement. I believe it will not only increase the practical merit of machine learning in our society, but 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 current themes of my research can be summarized as follows.

- **Kernel methods, mean embeddings, and beyond**—Kernel methods are among the most popular and powerful machine learning techniques because (i) they are mathematically elegant and flexible methods which enable us to work with a variety of data types (*e.g.*, strings, graphs, and semi-groups) in a coherent framework, (ii) we can incorporate prior knowledge about the learning problem through the diverse choices of kernel functions, and (iii) there exist abundant learning algorithms with which kernel functions can directly be used. Furthermore, recent advances in kernel mean embedding $\mathbb{P} \mapsto \mathbb{E}_{x\sim\mathbb{P}}[k(x,\cdot)] =: \mu_{\mathbb{P}}$ extend the whole arsenal of kernel methods to probability measures [1, 2, 3]. I am particularly interested in a wide range of novel applications—*e.g.*, large-scale learning, statistics, and causality—resulted from such a development.

- **Learning from probability measures**—Traditionally, it is assumed that data are i.i.d. points $x_1,\ldots,x_n \in \mathscr{X}$ drawn from some unknown distribution $\mathbb{P}$, whereas—in many scenarios—representing these data as distributions $\mathbb{P}_1,\ldots,\mathbb{P}_n$ over $\mathscr{X}$ may be preferable. 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 to work with these distributions. Many learning algorithms in domain adaptation, anomaly detection, and causal inference can be extended to distributions, leading to new methodologies in astronomy, high-energy physics, and statistics. My particular interest lies in developing novel algorithms for distributions and employing them in statistics and causality.

- **Causal learning and counterfactuals**—Cause-effect inference is a grand challenge in science because such a relationship would provide insights into the effect of an intervention. The causal understanding of diseases like cancer could ultimately help save million of lives around the world. In many situations, *experimental data* can be unethical, expensive, or even impossible to obtain, so we must rely on *observational data*. 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. Prior knowledge about cause-effect relations may also prove useful in some domains of machine learning.

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 recently worked on the *kernel mean embedding of distributions* [1, 2, 3] 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 [4], 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 [5] and Bayesian non-parametric [6]. 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 $\mathbb{P}$.

The kernel mean $\mu_{\mathbb{P}}$ is central to kernel methods in that it is used by many classical algorithms such as kernel PCA, and it also forms 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 [7, 8], 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. 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" [9]—*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 [10]. 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 [11]. 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. Last but not least, there are several theoretical questions—such as oracle inequalities and minimax lower bounds [12]—which remain to be addressed. 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.

The success of machine learning algorithms generally depends on data representation as is evident from the recent breakthrough in the neural network community [13]. Unlike deep learning algorithms, kernel methods often rely on a fixed representation—*i.e.*, a canonical feature map $k(x,\cdot)$—defined implicitly by the kernel. Hence, the success of kernel methods is heavily dependent on the choice of kernel k. Unfortunately, how to choose the "right" kernel remains an ultimate open question. The promise of the "big data" revolution is that by using these data we may find an answer to this fundamental question. That is, a more expressive representation can be learned directly from such data. Many well-known algorithms such as multiple kernel learning (MKL) permit efficient learning of such a representation at the price of very restrictive class of representations which may be insufficient for complex tasks such as image and speech recognition. My interest is thus to design a mathematically elegant framework that can efficiently leverages huge datasets for learning expressive representations [14]. From a theoretical perspective, it is instructive to understand what characterizes a good representation in the context of learning theory.

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 statistics. 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 $\mathscr{H}$ [15, 16]. 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 [17] 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 [15]. Moreover, we developed one-class SMM (OCSMM) [16] with connections to variable kernel density estimation (VKDE) and novel applications in group anomaly detection on astronomical data and high-energy physics data. In [18], we investigated 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). I believe that this 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. Secondly, large-scale learning has become increasingly important. While a common approach is to scale up learning algorithms, another effective approach is to reduce the amount of data, while preserving most of the necessary information. For instance, representing a set of data points by a distribution can capture most of the information while reducing the amount of computation required. Distributions also help to conceal sensitive information about individual samples, *i.e.*, privacy-preserving. Lastly, on a theoretical side, I am also investigating if injecting noise to data points to form distributions corresponds to a more flexible form of regularization.

Causal Learning and Counterfactuals

Relying on the learning framework for distributions described above, 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 $\mathbb{P}(X,Y)$ [19]. 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 very 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 some 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 causal effect $\text{ACE} = \mathbb{E}[Y(0) - Y(1)] = \mathbb{E}[Y(0)] - \mathbb{E}[Y(1)]$ 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 ACE has been used only for real-valued outcomes, restricting the potential of this framework. A research direction I am pursuing is to generalize the ACE to multivariate and structured outcomes by means of a kernel mean embedding of counterfactual distributions. I believe that this research could open up a new frontier in this framework.

Future Directions

I am excited that kernel methods have recently enabled a transition from data points to distributions as well as from correlation to causation, prompting exciting research opportunities, new challenges, and novel applications. Besides the aforementioned open questions, I am also keen to explore the following research directions:

- *Scalable learning for highly structured data*—A recent surge in data science has not only expanded the research frontiers in machine learning, but also created attractive business opportunities. It is relatively easier to obtain an unprecedented amount of data having complex structures. Hence, we need ML algorithms that are scalable and are capable of coping with such an increasing complexity.
- *Learning meaningful representation*—Good representation can improve learning significantly, as is evident from the recent breakthrough in deep learning. Nevertheless, deep learning models are often designed to have millions, if not billions, of parameters, and thereby require a huge amount of data to train. But does it imply we cannot learn good, compact representation from small amount of data?
- *Reinforcement learning*—RL allows machines to learn from a direct interaction with their environments, or at least via a realistic simulation. In many applications like medical diagnosis, collecting data from the environment can be expensive, prohibited, or sometimes unethical. Building RL systems that can learn successfully from data-scarce domains remains a challenging problem.
- *High dimensional setting*—Learning from high dimensional data is one of the most challenging problems in machine learning. For instance, high-dimensional generative models, *e.g.*, generative adversarial network (GAN) and variational autoencoder (VAE), has received attention recently due to its ability to generate realistic samples. However, training such models is considered to be difficult, especially for large models.
- *Applications*—Recommendation systems, probabilistic programming, medical data analysis, and natural language processing.

References

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

[2] 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.

[3] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, 2017. Accepted.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 05 2015.

[14] R. Babbar, K. Muandet, and B. Schölkopf, "A scalable mixed-norm approach for learning lightweight models in large-scale classification," in *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 234–242, 2016.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.