Prediction, Causation, and Regulation in Machine Learning

Krikamol Muandet

***Abstract*— In this note, I summarize my research across the "prediction-causation-regulation" spectrum in machine learning and illustrate my vision for future research. Recent breakthroughs in algorithmic predictions have not only led to widespread use of predictive models in critical domains, but also expedited scientific discoveries. Nevertheless, important problems in machine learning, statistics, and economics, require inferences and learning over entire probability distributions. To this end, I have worked extensively on a kernel mean embedding (KME) of distributions which offers powerful mathematical tools to accomplish these tasks. In particular, the framework also provides a succinct representation of an interventional distribution which is the key ingredient in causal reasoning and decision making. However, an algorithmic decision remains unattainable as it demands additional knowledge that can only be acquired from real-world experiments. As collecting experimental data can typically be expensive, time-consuming, or even unethical, most algorithmic decisions are made on the basis of non-experimental data alone. To improve the reliability of these decisions, I am working on projects that address important challenges such as endogeneity and heterogeneity by combining ideas from machine learning and econometrics. Finally, the omnipresence of the data-driven systems and scarcity of resources in socioeconomic systems call for proper mechanism and oversight which will ensure long-term sustainability. This challenge has drawn my attention towards the applicability of concepts in game theory and mechanism design to machine learning and vice versa.**

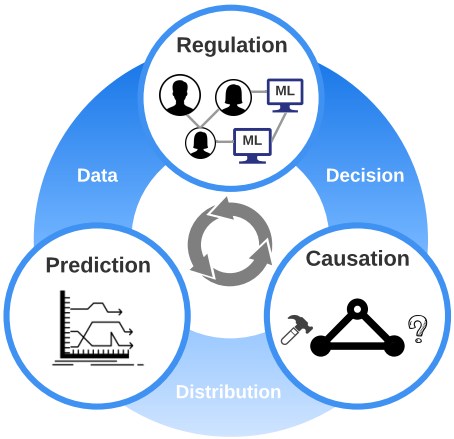

Fig. 1: My **Prediction-Causation-Regulation (PCR)** research agenda. (**P**) A good predictive modeling requires an accurate estimate of the underlying probability distribution. (**C**) The interventional distribution informs the consequences of algorithmic decisions in the real world. (**R**) Due to the scarcity, these decisions, especially personalized ones, may introduce incentives for individuals to change their behaviour. This in turn creates a feedback loop in the data collection process.

I. Distributional Prediction

An algorithmic prediction relies on a good estimate of the conditional mean $\mathbb{E}[Y \mid X = x]$. Many machine learning (ML) algorithms have been proposed for this task including k-nearest neighbor (KNN), support vector machine (SVM), ensemble methods, and deep learning (DL) [1–3]. However, important problems in machine learning and statistics such as two-sample testing, causal inference, and transfer learning involve inferences over entire probability distributions.

A. Kernel Mean Embedding of Distributions

I have worked extensively on a **kernel mean embedding** (KME) of distributions [4, 5] and have published a highly-cited book entitled "Kernel Mean Embedding of Distributions: A Review and Beyond" [6]. The idea is to represent a distribution $P(X)$ over some random variable $X \in \mathcal{X}$ as a function in a high-dimensional feature space known as a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Formally, the embedding of $P(X)$ is defined by $\mu_P := \mathbb{E}_X[k(X, \cdot)]$ where $X \sim P(X)$. If k is bounded, *i.e.*, $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$, then μ_P is well-defined as a function in $\mathscr{H}$ (see Fig. 2 for an illustration). The KME has two important properties. First, there exists a class of kernels k, known as *characteristic kernels*, for which μ_P captures all information about P, *i.e.*, for distributions P and Q, $\mu_P = \mu_Q$ if and only if $P = Q$ [7, 8]. Examples of characteristic kernels on $\mathbb{R}^d$ include Gaussian RBF kernel $k(x, x') = \exp(-\|x - x'\|_2^2 / 2\sigma^2)$ and Laplacian kernel $k(x, x') = \exp(-\|x - x'\|_1 / \sigma)$ with a bandwidth parameter $\sigma > 0$. Second, it follows from the *reproducing property*[1] of $\mathscr{H}$ that $\mathbb{E}_X[f(X)] = \mathbb{E}_X[\langle f, k(x, \cdot)\rangle_{\mathscr{H}}] = \langle f, \mu_P \rangle_{\mathscr{H}}$ for any $f \in \mathscr{H}$ [5]. That is, μ_P provides not only a complete representation of $P(X)$, but also basic operations on $P(X)$. For example, it allows us to compare distributions using a so-called maximum mean discrepancy (MMD) [9, 10], i.e., $\mathrm{MMD}(P, Q, \mathscr{H}) := \|\mu_P - \mu_Q\|_{\mathscr{H}}$. With this framework, my colleagues and I have contributed in areas such as hypothesis testing [11–13], causality [14–18], and deep learning [19, 20] among others.

Given an i.i.d. sample $x_1, \ldots, x_n$ from $P(X)$, μ_P can be approximated by an empirical kernel mean $\hat{\mu}_P := (1/n) \sum_{i=1}^n k(x_i, \cdot)$, which was shown to converges to μ_P

[1] For any $f \in \mathscr{H}$ and $x \in \mathcal{X}$, $f(x) = \langle f, k(x, \cdot)\rangle_{\mathscr{H}}$.

at the rate of $O_P(1/\sqrt{n})$ as n goes to infinity [5, 15, 21, 22]. Importantly, this result holds without any distributional assumption on P, which makes this framework ideal for applications in economics where such an assumption is undesirable. In [22], my colleagues and I showed that this rate of convergence is **minimax optimal** in $\|\cdot\|_{\mathscr{H}}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$-norms over the class of discrete measures and the class of measures that have an infinitely differentiable density, with k being a continuous translation-invariant kernel on $\mathbb{R}^d$. Surprisingly, my colleagues and I showed that the kernel mean estimation can still be improved in practice thanks to the well-known Stein phenomenon [23]. In [24, 25], we proposed a family of estimators called **kernel mean shrinkage estimators** (KMSE): $\hat{\mu}_\alpha = \alpha f^* + (1-\alpha)\hat{\mu}_P$ for an arbitrary data-independent $f^* \in \mathscr{H}$ and a shrinkage parameter $\alpha \in (0,1)$. I developed an efficient leave-one-out cross validation procedure to select an optimal value of α that attains the right bias-variance tradeoff. In [26], I further developed a non-linear extension of $\hat{\mu}_\alpha$ based on spectral filtering algorithms [27]. Our KMSE have been shown to improve the estimation of (cross-) covariance operators and tensors of higher order in RKHS as demonstrated in [28] for increasing the power of kernel independence test.

The KME has been generalized to represent a conditional distribution $P(Y|X)$. Let Y be another random variable taking values in a measurable space $\mathcal{Y}$. A **conditional mean embedding** (CME) of $P(Y|X=x)$ for some $x \in \mathcal{X}$ can be defined in a similar way as $\mu_{Y|x} := \mathbb{E}_{Y|x}[\ell(Y,\cdot)] \in \mathscr{F}$ where $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ is a reproducing kernel on $\mathcal{Y}$ with the RKHS $\mathscr{F}$ [29]. Under the assumption that $\mathbb{E}_{Y|X}[g(Y)|X=\cdot] \in \mathscr{H}$, it was shown in [29] that there exists an operator $\mathcal{U}_{Y|X} : \mathscr{H} \to \mathscr{F}$ for which the following two essential properties associated with $P(Y|X)$ hold: (i) $\mu_{Y|x} = \mathcal{U}_{Y|X}k(x,\cdot)$. (ii) $\mathbb{E}_{Y|x}[g(Y)|X=x] = \langle g, \mu_{Y|x}\rangle_{\mathscr{F}}$ for all $g \in \mathscr{F}$. This conditional mean operator (CMO) can be expressed in terms of covariance operators $\mathcal{C}_{YX}$ and $\mathcal{C}_{XX}$ as $\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ [7, 29].[2] Given a joint sample $(x_1,y_1),\ldots,(x_n,y_n)$ from $P(X,Y)$, the empirical estimate of $\mu_{Y|x}$ is $\hat{\mu}_{Y|x} = \sum_{i=1}^n \alpha_i \ell(y_i,\cdot)$ where $\alpha := (K + n\lambda I)^{-1}\mathbf{k}_x \in \mathbb{R}^n$ with $K_{ij} := k(x_i,x_j)$ and $\mathbf{k}_x := (k(x_1,x),\ldots,k(x_n,x))^\top$. In [29, Theorem 6], the rate of convergence $O_P((n\lambda)^{-1/2} + \lambda^{1/2})$ was established, suggesting that the CME is harder to estimate than the KME. Under the condition that the eigenvalues $(\gamma_m)_{m=1}^\infty$ of $\mathcal{C}_{XX}$ decay as $\gamma_m \le \beta m^{-b}$ for some $\beta > 0$, the rate of $O_P(n^{-b/(4b+1)})$ was established in [30]. Combined with KME, the CME allows us to perform important operations such as sum rule, product rule, and Bayes rule by means of the RKHS embeddings which are basic building blocks for probabilistic inference [30–33]. For example, my colleagues and I proposed in [34] a novel conditional density estimation model termed the **conditional density operator** (CDO) that is competitive with recent neural conditional density models and Gaussian processes.

The prevalent definition of CME relies on two stringent assumptions that hinder the theoretical analyses, namely, that $\mathcal{C}_{XX}^{-1}$ exists and that $\mathbb{E}_{Y|X}[g(Y)|X=x]$, as a function in x, lives in $\mathscr{H}$. In [35], my colleagues and I overcame these long-standing limitations via an alternative operator-free definition based on a measure-theoretic perspective, i.e., **we define $\mu_{Y|X}$ as a X-measurable random variable taking values in $\mathscr{F}$**. As natural by-products, we extended MMD and HSIC to the conditional setting, which we call the maximum conditional mean discrepancy (MCMD) and the Hilbert-Schmidt conditional independence criterion (HSCIC). This novel definition not only eases theoretical analyses associated with the CME, but also inspires its future applications.

B. Distributional Learning and Generalization

The idea described in Section I-A enables learning at the level of probability distributions [36–41]. This allows us to reason about the aggregate behaviours that could arise from a collection of data which might represent individuals, institutions, or even countries (see the middle of Fig. 2). To this end, let $\mathscr{P}$ be a space of probability distributions and $\mathscr{M}$ a probability distribution over $\mathscr{P}$ called a **meta-distribution**, *i.e.*, a distribution over distributions. In supervised learning on distributional data, we are interested in learning a function $F : \mathscr{P} \to \mathcal{Y}$ from the labeled sample $(P_1,y_1),\ldots,(P_n,y_n) \sim \mathscr{M}(\mathcal{P},\mathcal{Y})$ where $\mathcal{Y}$ denotes an output space. In an unsupervised setting, a collection of $P_1,P_2,\ldots,P_n \sim \mathscr{M}(\mathcal{P})$ contains some useful information about the underlying process $\mathscr{M}$. Unlike in the standard setting, P_i can only be observed through the sample $x_1^i,x_2^i,\ldots,x_{n_i}^i$ from P_i for $i=1,\ldots,n$; in other words, empirical distributions $\hat{P}_i := (1/n)\sum_{j=1}^{n_i}\delta_{x_j^i}$.

Based on the KME, I developed a framework called **distributional risk minimization** (DRM) that generalizes the conventional empirical risk minimization (ERM) to a space of distributions [36, 39]. I showed that any $f \in \mathscr{H}$ minimizing the regularized risk functional $L(P_1,y_1,\mathbb{E}_{x\sim P_1}[f(x)],\ldots,P_n,y_n,\mathbb{E}_{x\sim P_n}[f(x)]) + \Omega(\|f\|_{\mathscr{H}})$ admits a form $f = \sum_{i=1}^n \alpha_i \mu_{P_i}$ for some $\alpha \in \mathbb{R}^n$ [36, Theorem 1]. It generalizes the well-known **representer theorem** [42] to a space of distributions. Based on this framework, I proposed learning algorithms on distributions called **support measure machine** (SMM) and **one-class support measure machine** (OCSMM) [36, 39]. These algorithms also generalize the support vector machine (SVM) algorithm with extended **kernel functions on probability distributions**, *e.g.*, $K(P,Q) = \langle \mu_P, \mu_Q\rangle_{\mathscr{H}}$ and $K(P,Q) = \exp(-\|\mu_P - \mu_Q\|_{\mathscr{H}}^2/2\sigma^2)$. Because we have no direct access to $P_1,\ldots,P_n$, learning from them involves a two-stage estimation [37].

[2] A covariance operator $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$ is a unique bounded operator satisfying $\langle g, \mathcal{C}_{YX} f\rangle_{\mathscr{F}} = \mathrm{Cov}[g(Y), f(X)]$ for all $f \in \mathscr{H}$, $g \in \mathscr{F}$ [6, Sec. 3.2].

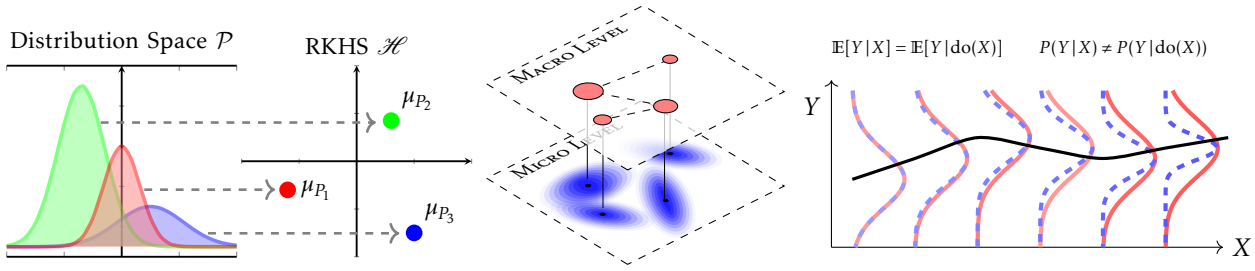

Fig. 2: (**Left**) The kernel mean embedding (KME) $P \mapsto \mu_P := \mathbb{E}_{X\sim P}[k(X,\cdot)]$. (**Middle**) The distributional learning with KME allows us to reason about the aggregate behaviors that arise from multiple collections of data. (**Right**) A comparison between the observational distribution $P(Y|X)$ and interventional distribution $P(Y|\mathrm{do}(X))$ with the same conditional means, but different higher-order moments, *i.e.*, conditional variances. Ignoring higher-order effects can lead to unreliable decisions.

My work [43] also laid the foundation for subsequent works in the out-of-distribution (OOD) generalization. In [43], my colleagues and I studied the **domain generalization** (DG) problem [44]: *Given data sets from multiple domains* $P_1(X,Y),\ldots,P_n(X,Y) \sim \mathscr{M}(\mathcal{P}_{XY})$, *learn a predictor* $f : \mathcal{P}_X \times \mathcal{X} \to \mathcal{Y}$ *that generalizes well to any previously unseen domains* $P_*(X)$.[3] In health care, domains, i.e., $P_1,\ldots,P_n$, may correspond to different patients from which data were collected and the goal is to learn a predictive diagnosis model that will work well on data collected from the new patients, i.e., P_*. To achieve this goal, I proposed a simple algorithm called **domain-invariant component analysis** (DICA) to extracts feature representation $\phi(X)$ that renders the marginals invariant, *i.e.*, $P_1(\phi(X)) \approx \cdots \approx P_n(\phi(X))$, while preserving functional relationship between X and Y. The invariance property was achieved via the notion of a **distributional variance** $\mathbb{V}_{\mathscr{H}}(\mathscr{M})$ which quantifies the dispersion of the meta-distribution $\mathscr{M}$ based on the empirical mean embeddings $\hat{\mu}_{P_1},\ldots,\hat{\mu}_{P_n}$ [43, Sec. 2.1]. Our learning-theoretic analysis showed that **this representation leads to improved out-of-distribution generalization bound** [43, Theorem 5]. To allow for better generalization, my intern and I later combined this idea with deep neural network and applied it to a few-shot learning problem [20]. A novel task representation called **model-aware task embedding** (MATE) incorporates not only the data distributions of different tasks, but also the complexity of the tasks.

C. Real-world Applications

Using the aforementioned ideas, I have contributed a number of real-world applications, especially in science, health care, and causality (see [6] for a review).

Science. Applications of machine learning to scientific data such as micro-array and astronomical data can be challenging due to a measurement error. To account for this uncertainty, I suggested to **represent each data point as a distribution** and adopted our distributional learning framework [36, 39]. Secondly, learning aggregate information from distributions of high-throughput data may reveal interesting phenomena that would have been impossible to discover from individual data. In [39], for example, I applied the OC-SMM algorithm to detect "**anomalous group**" of galaxies from the Sloan Digital Sky Survey (SDSS) dataset and to identify "**anomalous collision events**" in high-energy particle physics experiments that deviate from known Standard Model, e.g., Higgs boson. Thirdly, data in complex dynamical systems like molecular dynamics, fluid dynamics, atmospheric sciences, and control theory can be described by "transfer operators" such as the Perron-Frobenius or Koopman operator. In molecular dynamics, the eigenfunctions of these operators can help detect meta-stable sets, to project the dynamics onto the dominant slow processes, or to separate superimposed signals. Based on the CME, my colleagues and I proposed in [45] a **kernel transfer operator** (KTO) which extends transfer operator theory in complex dynamical systems to RKHS. Lastly, together with my colleagues, we proposed in [46] a **quantum mean embedding** (QME) to represent a pure quantum state of a system that is described by an infinite dimensional Hilbert space.

Health care. Medical data such as electronic health record (EHR) are often collected from multiple heterogeneous sources (patients, cohorts, or hospitals) with different data distributions. In [43], I applied our domain generalization (DG) algorithms to **gating of graft-versus-host disease (GvHD) data and to Parkinson's telemonitoring data**. The GvHD dataset consists of weekly peripheral blood samples obtained from 31 patients following allogenic blood and marrow transplant. The goal of gating is to identify $CD3^+CD4^+CD8\beta^+$ cells, which were found to have a high correlation with the development of GvHD. In the latter, the aim is to predict the clinician's motor and total UPDRS scoring of Parkinson's disease symptoms from 16 voice measures. There are around 200 recordings per patient. Our algorithms could improve

[3] This problem was first studied in [44], but the term "domain generalization (DG)" was coined for the first time in our paper [43].

generalization of the learned diagnosis to new patients [43, Sec. 3.2 & 3.3].

Causality. Conceptually, causal inference involves reasoning about the entire distributions rather than the individual data. In this light, my colleagues and I casted **a bivariate causal inference, *i.e.*, deciding if X causes Y or vice versa, as a classification problem on the joint distribution** $P(X,Y)$ [15]. Given a training sample $\{(\hat{P}_i(X,Y), l_i)\}_{i=1}^n$ where $\hat{P}_i(X,Y)$ denotes an empirical distribution and $l \in \{-1,+1\}$ is a ground-truth label indicating whether X causes Y, we learned a classifier that would allow us to predict the causal direction on the unseen datasets. Our approach outperforms classical causal inference algorithms, demonstrating the benefit of machine learning in causal inference [15, Sec. 5]. In [14], we showed how to use KME to model functional relationship $Z = f(X,Y)$, *e.g.*, a structural equation model (SEM) $Y = f(X)+\varepsilon$, also with applications in bivariate causal inference.

II. Causation and Decision Making

The observational distributions $P(Y|X)$ and $P(X)$ alone do not provide a full picture of causation which lies at the heart of decision making. Inspired by concerns over the societal impact of machine learning, my current research focuses on a synergy between machine learning and economics [47, 48]. I am developing algorithms to solve challenging economic problems, while bringing the economic perspective into the design of new learning algorithms.

A. Distributional Policy Intervention

In economics, a prime objective is to predict the effect of a policy intervention or a counterfactual change in economic conditions on some outcome variables. That is, we are concerned with an **interventional distribution** $P(Y|\mathrm{do}(X))$ which describes how the distribution of Y would change (or would have changed) as a result of some intervention on X.[4] In other words, it encodes the causal effect of X on Y. Since $P(Y|\mathrm{do}(X)) \neq P(Y|X)$ in general, one needs further restrictions on the data generating process (DGP) which mostly come in the form of causal graphs [50, 51], exchangeability conditions [52], or completeness conditions [53, 54].

The policy intervention generally aims to understand non-trivial effect of change in distribution $P(X)$ of relevant covariates, *e.g.*, pregnant women who smoke, on the entire outcome distribution $P(Y)$, *e.g.*, the birth weight of the babies [55, Sec. 5.2]. However, most of existing works in the literature focus on mean effects such as average treatment effect (ATE) and conditional average treatment effect (CATE), which do not inform changes in higher-order moments such as the variance. In [16], my colleagues and I proposed a **counterfactual mean embedding** (CME) $\mu_{Y\langle 1|0\rangle}$ as a Hilbert space representation of the counterfactual distribution [56]:

$$P_{Y\langle 1|0\rangle}(y) := \int P_{Y_1|X_1}(y|x)\,\mathrm{d}P_{X_0}(x) \tag{1}$$

The counterfactual $P_{Y\langle 1|0\rangle}$ represents **the outcome of hypothetical change which by definition is not observable**, e.g., the birth weight of the babies had the mothers abstained from smoking. Without any parametric assumptions, we proposed consistent estimators $\hat{\mu}_{Y\langle 1|0\rangle}$ for the embedding $\mu_{Y\langle 1|0\rangle}$ based on samples from the observable distributions P_{X_0} and $P_{Y_1X_1}$. We established consistency and convergence rate of $\hat{\mu}_{Y\langle 1|0\rangle}$ which requires weaker assumptions than the previous work [16, Sec. 4]. One of the important applications of our work is **offline policy evaluation** [16, Section 6], which is relevant to a program evaluation in economics. In [17], my colleagues and I extended this idea to conditional distributions by proposing the **conditional distributional treatment effect** (CoDiTE), which, in contrast to the more common conditional average treatment effect (CATE), is designed to encode a treatment's distributional aspects and heterogeneity beyond the mean. We applied CoDiTE to LaLonde's well-known National Supported Work (NSW) dataset [57] to model the distribution of income. Since income distributions are known to be skewed, CoDiTE can capture the treatment effect better than CATE [17, Sec. 6.2]. Fig. 2 highlights the importance of distributional effects.

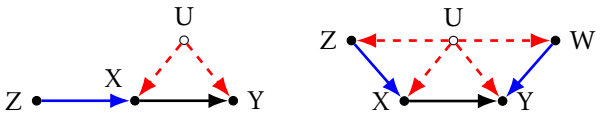

Fig. 3: Causal graphs depicting an instrumental variable Z (**left**) and proxy variables (Z,W) (**right**). The causal effect of X on Y is confounded by a hidden variable U. By utilizing Z or (Z,W), one can mitigate the effect of U.

B. Nonparametric Instrumental Variable (NPIV)

A major obstacle to the policy evaluation is an **endogeneity** that arises from hidden confounders. Suppose that X and Y denote treatment and outcome variables, *e.g.*, education and level of income, whose functional relationship can be described by $Y = f(X)+\varepsilon$, $\mathbb{E}[\varepsilon] = 0$. The presence of unobserved confounders, *e.g.*, socio-economic status, prevents us from learning f with the standard nonparametric regression because $\mathbb{E}[\varepsilon|X] \neq 0$, *i.e.*, $\mathbb{E}[Y|\mathrm{do}(X)] \neq \mathbb{E}[Y|X]$. The idea of an **instrumental variable (IV) regression** is to leverage an instrument Z that (i) induces a variation in X (*relevance*), (ii) affects Y only through X (*exclusion restriction*), (iii) is independent of the error term ϵ (*exchangeability*). Fig. 3 provides an illustration. In this case, f satisfies the

[4] The notation $\mathrm{do}(X)$ denotes a mathematical operator that simulates physical interventions on X [49, Sec. 3.2.1].

Fredholm integral equation of the first kind:

$$\mathbb{E}[Y\,|\,Z] = \int_{\mathcal{X}} f(x)\,\mathrm{d}P(x|Z). \tag{2}$$

Solving (2) for f is an ill-posed inverse problem [53, 58, 59]. Existing methods in machine learning can be categorized either as a two-stage approach [60, 61] or GMM-based approach [62, 63]. In [64], I showed that learning f via (2) can be reformulated as **a two-player game with a convex-concave utility function**, *i.e.*,

$$\min_{f\in\mathcal{F}}\max_{u\in\mathcal{U}} \mathbb{E}[(Y-f(X))u(Y,Z)] - \frac{1}{2}\mathbb{E}[u^2(Y,Z)]. \tag{3}$$

When $\mathcal{F}$ and $\mathcal{U}$ are both RKHSes, **the global equilibrium can be obtained in closed-form** [64, Sec. 4]. Furthermore, this work reveals a close connection between the two-stage and GMM-based approaches as a dual problem, as pointed out in [65, Appendix F] and our work [18, Sec. 3.3]. This reformulation also elucidates the kind of problems for which a game-theoretic perspective as a search for Nash equilibrium can lead to simpler algorithms than the standard ones that solve (2) directly.

The IV has revolutionized economics [66] and epidemiology [67, 68], but the statistical tools employed for estimation are fairly rudimentary. Hence, I envision not only the development of novel algorithms for this task, but also a widespread use of IV in machine learning applications such as offline RL [69, 70].

C. Maximum Moment Restriction

Alternatively, the IV regression problem can be solved via a moment restriction: $\mathbb{E}[Y-f(X;\theta)|z]=0$ for almost all $z\in\mathcal{Z}$ [62, 63, 71]. In fact, most econometric models are often specified in terms of a **conditional moment restriction** (CMR): *for correctly specified models, the conditional mean of certain functions of data is almost surely equal to zero* [72, 73]. For the true parameter $\theta_0\in\Theta$, the CMR is expressed mathematically as

$$\mathbb{E}[\varphi(X,\theta_0)\,|\,Z] = \mathbf{0}, \quad P_Z - \text{a.s.} \tag{4}$$

where (X,Z) is a data vector and $\varphi(\cdot,\cdot)$ is a vector of problem-dependent *generalized residual function*.

The major challenge here is that (4) implies an infinite number of unconditional moment restriction (UMR): $\mathbb{E}[\varphi(X,\theta_0)^\top h(Z)] = 0$ for any measurable vector-valued function h. The function h is often referred to as an **instrument** whose optimal choice remains an open problem in econometrics. In [13], I proposed a **maximum moment restriction** (MMR):

$$\begin{aligned}\mathrm{MMR}^2(\mathcal{H},\theta) &:= \sup_{f\in\mathcal{H},\|f\|\le 1}\left|\mathbb{E}_{XZ}\left[\varphi(X,\theta)^\top h(Z)\right]\right|^2 \\ &= \mathbb{E}[\varphi(X,\theta)^\top K(Z,Z')\varphi(X',\theta)] \end{aligned}\tag{5}$$

where $\mathcal{H}$ is a **vector-valued RKHS** (vv-RKHS) with the kernel K [74, 75]. In words, the original CMR (4) is transformed to a maximum of the interaction between the generalized residual function and functions of the conditioning variables that belong to a unit ball of the vv-RKHS. We showed that, if the kernel K satisfies certain properties, not only $\mathrm{MMR}(\mathcal{H},\theta_0)=0$ if and only if the original CMR (4) is fulfilled, but there also exists a closed-form, easy-to-use solution (5). My colleagues and I have applied this framework successfully to the problems of **conditional moment (CM) test** [13], **IV regression** [71], and **proximal causal learning** [18]; see Fig. 3. Therefore, modern kernel methods clearly have potentials to help solve many problems in economics.

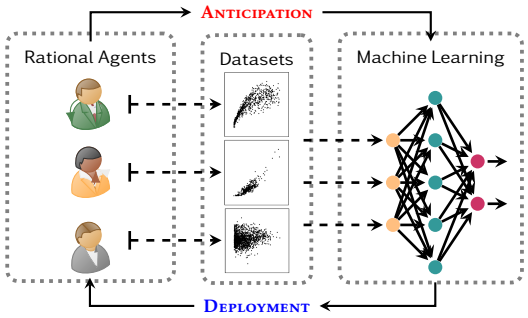

Fig. 4: **Learning as a robust mechanism design**: When data come from self-interest agents (individuals, institutions, or even countries) that can anticipate the outcome of data-driven models, the deployment of static models can be *sub-optimal*. In this setting, the problem is akin to designing a game to be played by the data against the model.

III. Model Regulation and Mechanism Design

Machine learning will soon enable accurate causal reasoning and reliable decision making. However, the deployment of learned models can affect the real world and change the environments in which they operate.

The first challenge is a **feedback loop in the data collection process**, as illustrated in Fig. 1. That is, historical data are used to build models which are then deployed to collect more data. In loan decisions, for example, a bank may decide whether or not to offer a loan based on models of the credit default. These decisions generate more data that are subsequently used to improve the models. In [76], my colleagues and I analyzed consequential decision making using imperfect predictive models, which are learned from data gathered by potentially biased historical decisions. **We articulated that when starting with a non-optimal deterministic policy, this approach fails to optimize utility for sequential decisions**. To avoid this failure mode while respecting a common fairness constraint, we suggested to directly learn the decisions with **exploring policies** [76, Sec. 3.2]. Our results highlight the need of regulation that governs the design and applications of predictive models in the real world.

The second challenge is that **the deployment may introduce incentives for individuals to change their**

behaviour. Hence, the models can become *sub-optimal* if observed data are manipulable, e.g., in heterogeneous pricing, individualized credit offer, and target social program, as illustrated in Fig. 4. In these situations, we must design a mechanism with which the models will be governed to reach a desirable long-term social welfare, for example, by eliciting the right incentives for people [77]. This question leads to a growing interest in strategic classification [78–80] and performative prediction [81] among others. Compared to immediate impacts of machine learning such as unfairness and algorithmic biases, long-term impacts are often neglected as they are more difficult to foresee.

Hence, **my vision for future research in machine learning lies in the challenges that arise from interactions between machine learning systems and real-world environments**. In particular, to create models that generalize better to the real world it is important to understand how to incorporate the information about the data collection process and the effects of the deployment in complex environments into the design of learning algorithms. Critical challenges include market equilibrium, temporal data, heterogeneity, delayed effect, endogeneity, and strategic response. To fulfil this vision, we need the synergy between machine learning, economics, game theory, and mechanism design.

Acknowledgement

I am indebted to Isabel Valera for her comments on the first draft of this manuscript. Ruth Urner also gave me early advice that has helped improve the writing of this manuscript. I also thank all of my collaborators who made it possible for me to write this manuscript and with whom I have had the privilege to collaborate.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[2] K. P. Murphy, *Machine learning : a probabilistic perspective*. MIT Press, 2013.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

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

[5] 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)*, Springer-Verlag, 2007, pp. 13–31.

[6] 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.

[7] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[8] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *The 21st Annual Conference on Learning Theory (COLT)*, Omnipress, 2008, pp. 111–122.

[9] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, 2006.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[11] 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)*, 2014, pp. 132–141.

[12] J. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet, "Learning kernel tests without data splitting," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 6245–6255.

[13] 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*, ser. Proceedings of Machine Learning Research, vol. 124, PMLR, 2020, pp. 41–50.

[14] 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.

[15] 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, JMLR, 2015, pp. 1452–1461.

[16] K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat, "Counterfactual mean embedding," *Journal of Machine Learning Research*, vol. Accepted, 2021. [Online]. Available: https://arxiv.org/abs/1805.08845.

[17] J. Park, U. Shalit, B. Schölkopf, and K. Muandet, "Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[18] A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. J. Kusner, A. Gretton, and K. Muandet, "Proximal causal learning with kernels: Two-stage estimation and moment restriction," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, forthcoming, 2021.

[19] 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)*, 2019.

[20] X. Chen, Z. Wang, S. Tang, and K. Muandet, "MATE: Plugging in model awareness to task embedding for meta learning," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 11 865–11 877.

[21] B Sriperumbudur, K Fukumizu, A Gretton, B Schölkopf, and G Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[22] I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet, "Minimax estimation of kernel mean embeddings," *Journal of Machine Learning Research*, vol. 18, no. 86, pp. 1–47, 2017.

[23] 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, University of California Press, 1955, pp. 197–206.

[24] 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, JMLR, 2014, pp. 10–18.

[25] 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.

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

[27] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri, "Spectral algorithms for supervised learning," *Neural Computation*, vol. 20, no. 7, pp. 1873–1897, 2008.

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

[29] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, ACM, 2009, pp. 961–968.

[30] K. Fukumizu, L. Song, and A. Gretton, "Kernel Bayes' rule: Bayesian inference with positive definite kernels," *Journal of Machine Learning Research*, vol. 14, pp. 3753–3783, 2013.

[31] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf, "Tailoring density estimation via reproducing kernel moment matching," in *Proceedings of the 25th International Conference on Machine Learning (ICML)*, ACM, 2008, pp. 992–999.

[32] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin, "Kernel belief propagation," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR, 2011, pp. 707–715.

[33] L. Song, A. P. Parikh, and E. P. Xing, "Kernel embeddings of latent tree graphical models.," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2708–2716.

[34] I. Schuster, M. Mollenhauer, S. Klus, and K. Muandet, "Kernel conditional density operators," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108, PMLR, 2020, pp. 993–1004.

[35] J. Park and K. Muandet, "A measure-theoretic approach to kernel conditional mean embeddings," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 21 247–21 259.

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

[37] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton, "Learning theory for distribution regression," *Journal of Machine Learning Research*, vol. 17, no. 152, pp. 1–40, 2016.

[38] J. B. Oliva, B. Póczos, and J. G. Schneider, "Distribution to distribution regression.," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, ser. JMLR Proceedings, vol. 28, JMLR.org, 2013, pp. 1049–1057.

[39] 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)*, AUAI Press, 2013, pp. 449–458.

[40] J. B. Oliva, B. Poczos, and J. Schneider, "Fast distribution to real regression," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 33, JMLR.org, 2014, pp. 706–714.

[41] B. Póczos, A. Singh, A. Rinaldo, and L. A. Wasserman, "Distribution-free distribution regression.," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 31, JMLR.org, 2013, pp. 507–515.

[42] 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*, ser. COLT '01/EuroCOLT '01, Springer-Verlag, 2001, pp. 416–426.

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

[44] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2178–2186.

[45] S. Klus, I. Schuster, and K. Muandet, "Eigendecompositions of transfer operators in reproducing kernel hilbert spaces," *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 283–315, 2020.

[46] J. M. Kübler, K. Muandet, and B. Schölkopf, "Quantum mean embedding of probability distributions," *Physical Review Research*, vol. 1, no. 3, p. 033 159, 2019.

[47] S. Athey, "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*, ser. NBER Chapters, National Bureau of Economic Research, Inc, 2018, pp. 507–547.

[48] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," *Annual Review of Economics*, vol. 11, no. 1, pp. 685–725, 2019.

[49] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[50] ——, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[51] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference – Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.

[52] D. Rubin, "Causal inference using potential outcomes," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[53] W. K. Newey and J. L. Powell, "Instrumental variable estimation of nonparametric models," *Econometrica*, vol. 71, no. 5, pp. 1565–1578, 2003.

[54] X. D'Haultfoeuille, "On the completeness condition in nonparametric instrumental problems," *Econometric Theory*, vol. 27, no. 3, pp. 460–471, 2011.

[55] C. Rothe, "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, vol. 155, pp. 56–70, 2010.

[56] V. Chernozhukov, I. Fernández-Val, and B. Melly, "Inference on counterfactual distributions," *Econometrica*, vol. 81, no. 6, pp. 2205–2268, 2013.

[57] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, vol. 76, no. 4, pp. 604–620, 1986.

[58] R. Kress, *Linear Integral Equations*. Springer, 1989, vol. 3.

[59] J. L. Horowitz, "Applied nonparametric instrumental variables estimation," *Econometrica*, vol. 79, no. 2, pp. 347–394, 2011.

[60] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, PMLR, 2017, pp. 1414–1423.

[61] R. Singh, M. Sahani, and A. Gretton, "Kernel instrumental variable regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 4595–4607.

[62] A. Bennett, N. Kallus, and T. Schnabel, "Deep generalized method of moments for instrumental variable analysis," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 3564–3574.

[63] N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis, "Minimax estimation of conditional moment models," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 12 248–12 262.

[64] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, "Dual instrumental variable regression," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 2710–2721.

[65] L. Liao, Y.-L. Chen, Z. Yang, B. Dai, M. Kolar, and Z. Wang, "Provably efficient neural estimation of structural equation models: An adversarial approach," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 8947–8958.

[66] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[67] S. Burgess, D. S. Small, and S. G. Thompson, "A review of instrumental variable estimators for Mendelian randomization," *Statistical Methods in Medical Research*, vol. 26, no. 5, pp. 2333–2355, 2017.

[68] S. Burgess, C. N. Foley, E. Allara, J. R. Staley, and J. M. M. Howson, "A robust and efficient method for Mendelian randomization with hundreds of genetic variants," *Nature Communications*, vol. 11, no. 1, p. 376, 2020.

[69] L. Liao, Z. Fu, Z. Yang, M. Kolar, and Z. Wang, "Instrumental variable value iteration for causal offline reinforcement learning," *arXiv preprint arXiv:2102.09907*, 2021.

[70] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton, "Learning deep features in instrumental variable regression," in *International Conference on Learning Representations*, 2021.

[71] R. Zhang, M. Imaizumi, B. Schölkopf, and K. Muandet, "Maximum moment restriction for instrumental variable regression," *ArXiv Preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2010.07684.

[72] W. Newey, "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics*, vol. 11, 1993, ch. 16, pp. 419–454.

[73] C. Ai and X. Chen, "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, vol. 71, no. 6, pp. 1795–1843, 2003.

[74] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.

[75] C. Carmeli, E. De Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Analysis and Applications*, vol. 04, no. 04, pp. 377–408, 2006.

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

[77] J. Kleinberg and M. Raghavan, "Algorithmic classification and strategic effort," *SIGecom Exch.*, vol. 18, no. 2, pp. 45–52, 2020.

[78] M. Hardt, N. Megiddo, C. H. Papadimitriou, and M. Wootters, "Strategic classification," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, M. Sudan, Ed., ACM, 2016, pp. 111–122.

[79] J. Kleinberg and M. Raghavan, "How do classifiers induce agents to invest effort strategically?" In *Proceedings of the 2019 ACM Conference on Economics and Computation*, Association for Computing Machinery, 2019, pp. 825–844.

[80] J. Miller, S. Milli, and M. Hardt, "Strategic classification is causal modeling in disguise," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020, pp. 6917–6926.

[81] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020, pp. 7599–7609.