Statistical Learning Theory

Machine learning algorithms are designed to generalize from past observations across different problem settings. The goal of learning theory is to analyze statistical and computational properties of learning algorithms and to provide guarantees on their performance. To do so, it poses these tasks in a rigorous mathematical framework and deals with them under various assumptions on the data-generating process.

In [237] we initiate a formal analysis of **compressing a data sample** so as to encode a set of functions consistent with (or of minimal error on) the data. We propose several formal requirements (exact versus approximate recovery and worst case versus statistical data generation) for such compression and identify parameters of function classes that characterize the resulting compression sizes. In [233] we provide a novel analysis for a **life-long learning** setup where performance guarantees are required for every encountered task. Such a setup had previously been analyzed only under rather restrictive assumptions on the data-generating process. Our study generalizes a natural lifelong learning (where at every step, if possible, a predictor is created as an ensemble over previously learned ones using only little data) and identifies conditions of task relatedness that render such a scheme data efficient. In [232] we show that **active learning** can provide label savings in non-parametric learning settings. Previously this had mostly been done in parametric learning of a classifier from a fixed class of bounded capacity. We develop a novel active query procedure that takes in unlabeled data and constructs a compressed version of the underlying labeled sample while automatically adapting a number of label queries. We then show that this procedure maintains performance guarantees of nearest neighbor classification.

In recent years the **kernel mean embedding (KME)** of distributions started to play an important role in various machine learning tasks, including independence testing, density estimation, implicit generative models, and more. Given a reproducing kernel and its corresponding reproducing kernel Hilbert space (RKHS), KME maps a distribution P over the input domain to the element $\mu_P := \int k(X,\cdot)dP(X)$ in the RKHS. An important step in many KME-based learning methods is to estimate the distribution embedding μ_P using observations $x_1,\ldots,x_n$ sampled from P. Inspired by the James-Stein estimator, in [100] we introduced a new type of KME estimators called **kernel mean shrinkage estimators (KMSEs)** and proved that it can converge faster than the empirical KME estimator $\hat{\mu}_P := \sum_{i=1}^n k(x_i,\cdot)/n$. This improvement is due to the bias-variance tradeoff: the shrinkage estimator reduces variance substantially at the expense of a small bias. We also empirically showed that KMSE is particularly useful when the sample size n is small compared to the input space dimensionality.

We have studied the optimality of KME estimators in the **minimax sense**. In [62] we show that the rate $O(n^{-1/2})$ achieved by $\hat{\mu}_P$, KMSE, and many other methods published in the literature is optimal and cannot be improved. This holds for any continuous translation-invariant kernel and for various classes of distributions, including both discrete and smooth distributions with infinitely differentiable densities. In [226] we also study the minimax optimal estimation of the **maximum mean discrepancy (MMD)** between two probability distributions, which is defined as the RKHS distance between their KMEs: $\mathrm{MMD}(P,Q) := \|\mu_P - \mu_Q\|$. We show that for any radial universal kernel the rate $O(n^{-1/2} + m^{-1/2})$ achieved by existing estimators is minimax optimal.

The properties of MMD are known to depend on the underlying kernel and have been linked to three fundamental concepts: **universal, characteristic, and strictly positive definite kernels**. In [37] we show that these concepts are essentially equivalent and give the first complete characterization of those kernels whose associated MMD metrizes the weak convergence of probability measures. Finally, we show that KME can be extended to Schwartz-distributions and analyze properties of these distribution embeddings.

While MMDs are known to metrize convergence in distribution, the underlying conditions are too stringent when one only aims to metrize convergence to a fixed distribution, which is the case for instance in goodness-of-fit tests. To address this, we derive necessary and sufficient conditions for MMD to **metrize tight convergence**

to a fixed target distribution.[2] We use our characterizations to analyze the convergence properties of the targeted kernel Stein discrepancies (KSDs) commonly employed in the goodness-of-fit testing. The results validate the use of KSDs for a broader set of targets, kernels, and approximating distributions.

The problem of **estimating a distribution of functions of random variables** plays an important role in the field of probabilistic programming, where it can be used to generalize functional operations to distributions over data types. In [227] we proposed a non-parametric way to estimate the distribution of $f(X)$ for any continuous function f of a random variable X. The proposed KME based estimators are proven to be asymptotically consistent. We provide finite-sample guarantees under stronger assumptions.

Motivated by recent advances in **privacy-preserving machine learning** and building upon the results of [227], we have proposed a theoretical framework for a novel database release mechanism that allows third-parties to construct consistent estimators of population statistics while ensuring that the privacy of each individual contributing to the database is protected [155]. Our framework is based on newly introduced differentially private and consistent estimators of KMEs, of interest in their own right.

Visually impressive progress in machine learning has been made in the field of **unsupervised generative modeling** with generative adversarial networks (GANs), variational autoencoders (VAEs) and other deep neural network based architectures, significantly improving the state of the art in the quality of generated samples, especially in the domain of natural images.

In [200] we study the **training of mixtures of generative models** from a theoretical perspective. We find a globally optimal closed form solution for performing greedy updates while approximating an unknown distribution with mixtures in any given f-divergence. We then derive a boosting style meta-algorithm which can be combined with many modern generative models (including GANs and VAEs).

While training objectives in VAEs and GANs are based on f-divergences, it has been recently shown that other divergences, in particular, **optimal transport distances**, may be better suited to the needs of generative modeling. Starting from Kantorovich's primal formulation of the optimal transport problem, we show that it can be equivalently written in terms of probabilistic encoders, which are constrained to match the latent posterior and prior distributions.[3] We then apply this result to train latent variable generative models in [149]. When relaxed, the constrained optimization problem leads to a new **regularized autoencoder algorithm** which we call Wasserstein auto-encoders (WAEs). WAEs share many of the properties of VAEs (stable training, nice latent manifold structure) while generating samples of better quality, as measured by the Frechet Inception score across multiple datasets.

In [113, 114] we focus on **properties of the latent representations** learned by WAEs and draw several interesting conclusions based on various experiments. First, we show that there are fundamental problems when training WAEs with deterministic encoders when the intrinsic dimensionality of the data is different from the latent space dimensionality. Second, we point out that training WAEs with probabilistic encoders is a challenging problem, and propose a heuristic approach with promising results on several datasets.

Many deep neural network based architectures have been proven vulnerable to so-called **adversarial attacks**. In the case of natural image classifiers, carefully chosen but imperceptible image perturbations can lead to drastically changing predictions. We showed that adversarial vulnerability increases with the gradients of the training objective when viewed as a function of the inputs.[4] For most current network architectures, we prove that the ℓ_1-norm of these gradients grows as the square root of the input size. These nets therefore become increasingly vulnerable with growing image size.

More information: https://ei.is.mpg.de/project/statistical-learning-theory

[2] C. J. Simon-Gabriel, L. Mackey. Targeted Convergence Characteristics of Maximum Mean Discrepancies and Kernel Stein Discrepancies. In *PhD thesis: Distribution-Dissimlarities in Machine Learning (University of Tübingen)*, 2018.

[3] O. Bousquet, S. Gelly, I. Tolstikhin, C. J. Simon-Gabriel, B. Schölkopf. From Optimal Transport to Generative Modeling: the VEGAN cookbook. *CoRR* **abs/1705.07642**, 2017.

[4] C. J. Simon-Gabriel, Y. Ollivier, B. Schölkopf, L. Bottou, D. Lopez-Paz. Adversarial Vulnerability of Neural Networks Increases with Input Dimension. *CoRR* **abs/1802.01421**, 2018.