1 EMPIRICAL INFERENCE

1.1 Research Overview

The problems studied in the department can be subsumed under the heading of *empirical inference*, i.e., inference performed on the basis of empirical data. This includes statistical learning, but also the inference of causal structures from statistical data, leading to models that provide insight into the underlying mechanisms, and make predictions about the effect of interventions. Likewise, the type of empirical data can vary, ranging from biological measurements (e.g., in neuroscience) to astronomical observations. We are conducting theoretical, algorithmic, and experimental studies to try and understand the problem of empirical inference.

The department was started around statistical learning theory and kernel methods. It has since broadened its set of inference tools to include a stronger component of Bayesian methods, including graphical models with a strong focus on issues of causality. In terms of the inference tasks being studied, we have moved towards tasks that go beyond the relatively well-studied problem of supervised learning, such as semi-supervised learning or transfer learning. Finally, we have continuously striven to analyze challenging datasets from biology, astronomy, and other domains, leading to the inclusion of several application areas in our portfolio.

The most competitive publication venues in empirical inference are NeurIPS (Neural Information Processing Systems), ICML (International Conference on Machine Learning), UAI (Uncertainty in Artificial Intelligence), and for theoretical work, COLT (Conference on Learning Theory). The presence at these conferences makes us one of the top international machine learning labs. In addition, we sometimes submit our work to the leading application oriented conferences in neighboring fields including computer vision (ICCV, ECCV, CVPR) and data mining (KDD, ICDM, SDM), as well as to specialized journals.

Our work has earned us a number of awards, including best paper prizes at the major conferences in the field (NeurIPS, ICML, UAI, COLT, ALT, CVPR, ECCV, ISMB, IROS, KDD). Recent awards include IEEE SMC 2016, ECML-PKDD 2016, a honorable mention at ICML 2017, and the test-of-time award for Olivier Bousquet at NeurIPS 2018, received for work he started while he was still member of our department (at MLSS 2003 in Tübingen with Leon Bottou).

Theoretical studies, algorithms, and applications often go hand in hand. For instance, it may be the case that someone working on a specific application will develop a customized algorithm that turns out to be of independent theoretical interest. Such serendipity is a desired side effect caused by interaction across groups and research areas, for instance during our frequent departmental talks. It concerns cross-fertilization of methodology (e.g., kernel independence measures used in causal inference), the transfer of algorithmic developments or theoretical insights to application domains (e.g., causal inference in neuroscience), or the combination of different application areas (e.g., using methods of computational photography for magnetic resonance imaging).

The linear organization of the text does not permit an adequate representation of all these connections. Below, we have opted for an organization of the material that devotes individual sections to our main application areas (computational imaging, robot learning, and neuroscience), and that comprises four methodological sections, on learning algorithms, causal inference, probabilistic inference, and statistical learning theory. We begin with the latter.

Statistical Learning Theory A machine learning algorithm is given training data and tries to learn a model that is well-suited to describe the data and that can be used to make predictions. The goal of statistical learning theory is to assess to which extent such algorithms can be successful in principle. The general approach is to assume that the training data have been generated by an unknown random source, and to develop mathematical tools to analyze the performance of a learning algorithm in statistical terms: for example, by bounding prediction errors ("generalization bounds") or by analyzing large sample behavior and convergence of algorithms on random input ("consistency").

The department has made various contributions to this area, especially in areas of machine learning where statistical learning theory is less well developed. These include settings like active and transfer learning, privacy preserving machine learning as well as unsupervised generative modeling. Our goal is to contribute statistical foundations to these exciting new challenges where pioneering work can be done.

Active learning exploits structure and information in unlabeled data to reduce label supervision by requesting labels only for a small set of points from a large pool. Developing a novel active query procedure that takes in an unlabeled data and constructs a compressed version of the underlying labeled sample, we showed that active learning can provide label savings even in non-parametric learning settings. 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 was then conducted in [237].

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, and many more. Inspired by the James-Stein estimator, in [100] we introduced a new type of KME estimators called kernel mean shrinkage estimators (KMSEs) and showed that it can converge faster than the empirical KME estimator. We have studied the optimality of KME estimators in the minimax sense in [62] and shown that the convergence rate for the KME, and many other methods published in the literature, is optimal and can not be improved. The advances and characterizations for kernel mean embeddings also play a role for privacy preserving machine learning.

Recently significant progress has been made in the field of *unsupervised (deep) 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 samples, especially in the domain of natural images. Traditionally the training objectives in VAEs and GANs have been based on f-divergences. We showed, starting from Kantorovich's primal formulation of the optimal transport problem, that this can be equivalently written in terms of probabilistic encoders, which are constrained to match the latent posterior and prior distributions.[1] This leads to a new training procedure of latent variable models, so called Wasserstain auto-eoncoders (WAEs) as described in [149]. While WAEs share many of the nice properties of VAEs, the generated samples often exhibit better quality and leads to interesting properties of the learned latent representations as described in [114] and [113]. Another theoretical study of generative modeling led us to propose the *AdaGAN*, a boosting approach to greedily build mixtures of generative models (e.g., GANs or VAEs) by solving, at each step, an optimization problem that results in the best additional model to reduce the discrepancy between the current mixture model and the target [200].

Learning Algorithms Learning algorithms based on kernel methods have enjoyed considerable success in a wide range of supervised learning tasks such as regression and classification. One reason for the popularity of these approaches is that they solve difficult nonparametric problems by mapping data points into high dimensional spaces of features, specifically reproducing kernel Hilbert spaces (RKHSes), in which linear algorithms can be brought to bear, leading to solutions taking the form of kernel expansions [59].

Based on foundational work from our department, kernel methods underlie methods determining the goodness of fit of a model and more recently of differentially private learning. In [126], we address the problem of measuring the relative goodness of fit of two models using kernel mean embeddings. Given two candidate models, and a set of target observations, the goal is to produce a set of so-called informative features, which indicate the regions in the data domain where one model fits better than the other. The task is formulated as a statistical test whose runtime complexity is linear in the sample size. Privacy-preserving machine learning algorithms aim to come up with database release mechanisms that allow third-parties to construct consistent estimators of population statistics while ensuring that the privacy of each individual contributing to the database is protected. In [155], we develop privacy-preserving algorithms based on the kernel mean embedding, allowing us to release a database while guaranteeing the privacy of the database records.

Optimization lies at the heart of most machine learning algorithms and our department aims to understand the convergence property of coordinate descent as well as Frank-Wolfe optimization algorithms under different sampling schemes and constraints. In [188], we provide a theoretical understanding of greedy coordinate descent for smooth functions. Similarly, in [172] we propose an adaptive recursive sampling scheme based on the min-max optimal solution of the variance reduction problem to achieve faster convergence for coordinate descent, which can also be applied to stochastic gradient descent. A connection between matching pursuit and Frank-Wolfe optimization is explored in [144]

Causal inference The detection and use of statistical dependences form the core of statistics and machine learning. In recent years, machine learning methods have enabled us to perform rather accurate prediction, often based on large training sets, for complex nonlinear problems that not long ago would have appeared completely random. However, in many situations we would actually prefer a causal model to a purely predictive one; i.e., a model that might tell us that a specific variable (say, whether or not a person smokes) is not just statistically associated with a disease, but causal for the disease.

Pearl's graphical approach to causal modeling generalizes Reichenbach's common cause principle and characterizes the observable statistical (conditional) independences that a causal structure should entail. Many causal inference methods build on these independences to infer causal graphs from data. This "graphical models" approach to causal inference has several weaknesses that we try to address in our work: it only can infer causal graphs up to Markov equivalence, it does not address the hardness of conditional independence testing, and it usually does not worry about the complexity of the underlying functional regularities that generate statistical dependences in the first place. Our work in this field is characterized by the following three

[1] 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.

aspects:

1. We often work in terms of structural equation models (SEMs) or functional causal models (FCMs), i.e., we do not take statistical dependences as primary, but rather study mechanistic models which give rise to such dependences. In FCMs, each variable is modeled as a deterministic function of its direct causes and some noise variable N, e.g., $Y = f(X, Z, N)$; all noise variables are assumed to be jointly independent. FCMs do not only allow us to model observational distributions; one can also use them in order to model what happens under interventions (e.g., gene knockouts or randomized studies).

Under suitable model assumptions like additive independent noise, causal knowledge and the framework of SCMs admits novel techniques of noise removal via so-called half-sibling regression [89, 102], or revealed previously unknown aspects of the arrow of time [86, 234].

2. Viewed from an FCM perspective, the crucial assumption of the graphical approach to causality is statistical independence of all noise terms. Intuitively, it is clear that as the noises propagate through the graph, they pick up dependences due to the graph structure, hence the assumption of initial independence of the noise terms allows us to tease out properties of that structure. We believe, however, that much can be gained by considering a more *general independence assumption* related to notions of invariance and autonomy of causal mechanisms. Here, the idea is that causal mechanism are autonomous entities of the world that (in the generic case) do not depend on or influence each other, and changing (or intervening on) one of them often leaves the remaining ones invariant.

In the context of the classical pattern recognition task with handwritten digits, [139] shows that learning causal models that contain independent mechanisms helps in transferring information across substantially different data sets. Theoretical work in [152] shows that the independence of causal mechanisms can be formalized via group symmetry.

3. This leads to the third characteristic aspect of our work on causality. Wherever possible, we attempt to establish connections to machine learning, and indeed we believe that some of the hardest problems of machine learning (such as those of domain adaptation and transfer) are best addressed using causal thinking. To this end, one may assume, for instance, that structural equations remain constant across data sets and only the noise distributions change [2], or that some of the causal conditionals in a causal Bayesian network change, while others remain constant [30] or that they change independently [181], which results in new approaches to domain adaptation [242].

Our lab has played a major role in putting causal inference on the agenda of the machine learning community, including a recent award-winning textbook [2], and we expect that causal inference will have practical implications for many inference problems (e.g., in astronomy [89] and neuroscience [98]), as well as increasingly in machine learning methods, including deep learning [139] and reinforcement learning [127]. We also expect that it will play a major role in societal aspects of AI, including fairness [173] and interpretability/accountability. Causality touches statistics, econometrics, and philosophy, and it constitutes one of the most exciting field for conceptual basic research in machine learning today. We expect that going forward, causality will play a major role in taking *representation learning* to the next level, moving beyond the mere representation of statistical dependence structures towards models that support intervention and planning (and thus Konrad Lorenz' notion of *thinking* as *acting in an imagined space*).

Probabilistic Inference The probabilistic formulation of inference is one of the main research streams within machine learning. One of our main themes in this field has been nonparametric inference on function spaces using Gaussian process models [123, 222, 223]. The approaches developed at the department allow finding the best kernel bandwidth hyperparameter efficiently and are especially well-suited for online learning.

A crucial bottleneck in Bayesian models is the marginalization of latent variables. This can be computationaly demanding, so approximate inference routines reducing computational complexity are a major research theme. In [124, 151], we study the convergence properties of this approach from a modern optimization viewpoint, establishing connections to the classic Frank-Wolfe algorithm. The analyses yield novel theoretical insights regarding the sufficient conditions for convergence, explicit rates, and algorithmic simplifications.

Members of the department have also been working on aspects of probabilistic programming and studied the problem of representing the distribution of $f(X)$ for a random variable X and a function f [227]. We use kernel mean embedding methods to construct consistent estimators of the mean embedding of $f(X)$. The method is applicable to arbitrary data types on which suitable kernels can be defined. It thus allows us to generalize (to the probabilistic case) functions in computer programming which are originally only defined on deterministic data types.

As algorithmic decision making systems are becoming ubiquitously trained from historical data collected from, as well as implemented in a wide variety of online as well as offline services; there is a growing concern that these automated decisions can lead to a lack of fairness, i.e., their outcomes can disproportionately hurt (or, benefit) particular groups of people sharing one or more sensitive attributes (e.g., race, gender). As a consequence, there has been an increase in research on computational (un)fairness. The contributions of the department on in this domain are twofold. First, we have focused on proposing new definitions and metrics of fairness, as well as on designing automatic decision systems that incorporate a fairness definition in their training step to avoid discrimination towards particular groups of people sharing certain sensitive attributes, while providing clear mechanisms to trade-off fairness and accuracy [13, 159]. Second, we have done pioneering work in connecting fairness to causality, showing that whether an algorithmic decision is fair or not should really take into account the underlying causal graph rather than just the observational distribution [173].

Computational Imaging Handheld video cameras are now commonplace and available in every smartphone, and thus images and videos are recorded in unprecented amounts. Our research focus is on digital image restoration that aims at computationally enhancing the quality of images and recovering the most likely original image by undoing the adverse effects of image degradation such as noise and blur.

To recover a high-resolution image from a single low-resolution input, we proposed [202] a novel method for automated texture synthesis in combination with a perceptual loss focusing on creating realistic textures rather than optimizing for a pixel-accurate reproduction of ground truth images during training. By using feed-forward fully convolutional neural networks in an adversarial training setting, we achieve a significant boost in image quality even at high magnification ratios.

For recovering an image from corrupted measurements, e.g., due to unwanted camera shake during recording, or moving objects in the scene, we developed methods for propogating information between multiple consecutive blurry observations to help restore the desired sharp image or video. In a number of works [134, 183, 184], we have developed efficient recurrent network architectures to deblur frames taking temporal information into account, which can efficiently handle both ego and object motion for arbitrary spatial and temporal input sizes.

Robot Learning Research in robotics and artificial intelligence has lead to the development of complex robots ranging from anthropomorphic arms to complete humanoids. In order to be meaningfully applied in human-inhabited environments, robots need to possess a variety of physical abilities and skills. However, programming such skills is a labor- and time-intensive task which often leads to brittle solutions and requires a large amount of expert knowledge. In particular, it often involves transforming intuitive concepts of motions and actions into formal mathematical descriptions and algorithms.

While kinematic optimization allows for efficient representation and online generation of hitting trajectories, e.g., in robot table tennis, learning to track such dynamic movements with inaccurate models remains an open problem. To achieve accurate tracking for such tasks in a stable and efficient way, we have proposed a series of novel adaptive Iterative Learning Control (ILC) algorithms that are implemented efficiently and enable caution during learning [6]. Moreover, we have built a muscular robot system in order to study the problem of how to accurately control musculoskeletal robots by learning control. Muscular systems are hard to control by classical methods, but offer beneficial properties to achieve human-comparable performance in complex tasks [244].

In real robot experiments on a Barrett WAM, we have studied the properties of optimal trajectory generation in robot table tennis strikes [34], and how to robustly learn such primitives from multiple demonstrations as well as adapt them to new goals [5]. We have also recently demonstrated how a table tennis serve can be captured and successfully be reproduced [4, 231].

Robot table tennis has a number of components that are representative of tasks encountered by natural intelligent systems, including perception and action, as well as various aspects of social interaction (opponent modeling, competition, collaboration). We have recently completed the move of our robotics laboratory to the new building, now operating a two-robot setup. In the long term, this will enable us to study a rich set of problems, including cooperative game play.

Machine Learning in Neuroscience The neurosciences present some of the steepest challenges to machine learning. Nearly always there is a very high-dimensional input structure — particularly relative to the number of exemplars, since each data point is usually gathered at a high cost. To avoid overfitting, inference must thus make considerable use of domain knowledge. Relevant regularities are often subtle, the rest being made up of noise that may be of much larger magnitude (often composed largely of the manifestations of other neurophysiological processes, besides the ones of interest). In finding generalizable solutions, one usually has to contend with a high degree of variability, both between individuals and across time, leading to problems of covariate shift and non-stationarity.

One specific neuroscientific application area in which we have a long-standing interest is that of brain-computer interfacing, or BCI (see page 10). BCIs hold promise in restoring communication for completely locked-in stage (CLIS) patients in late stages of amyotrophic lateral sclerosis (ALS). Despite more than two decades of research, however, late-stage ALS patients remain incapable of operating BCIs, arguably because such systems currently rely on brain processes that are impaired as a result of disease progression. In a series of studies, we have investigated how ALS affects neural- and cognitive processes [19, 60].

Building upon these novel insights, we have developed and validated a new type of cognitive BCI for late-stage ALS patients [84, 252]. To translate this system from laboratory- into home-use, we have pioneered a transfer learning approach for BCIs [61, 95]. These advances now enable us to build high-performance, cognitive BCIs with off-the-shelf hardware, thereby (in ongoing work) rendering BCI systems available for large-scale application outside of laboratory environments.

We have also designed machine learning techniques to assist with the interpretation of experimental brain data (see page 20). Unsupervised learning tools based on non-negative Matrix Factorization were designed to automatically identify activity patterns among large populations of recorded neurons [27]. Finding causal relationships between neural processes is also of particular interest to neuroscientists, but hard to address in living neural systems due to ethical and practical concerns. Combining machine learning with detailed computational biophysical network models provided theoretical insights into biological learning by identifying key neural circuits underlying the reliable replay of memorized events [29].

We conclude this section with a short summary of our contributions to cognitive science and vision research, performed in collaboration with Felix Wichmann, professor of computational neuroscience at the local University Tübingen and until recently also part-time member of our department. In one line of work, we developed methods to gain more information from psychophysical data, linking traditional methods with machine learning approaches [80, 246]. We investigate reliable supra-threshold psychophysical paradigms which are more intuitive and require less training and are thus more likely to yield reliable crowd-sourcing data [54, 63]. A second focus was the development of a predictive image-based model of spatial vision. We integrated the large psychophysical literature on simple detection and discrimination experiments and proposed a model based on maximum-likelihood decoding of a population of model neurons predicting the most important spatial vision data sets simultaneously, using a single set of parameters [56]. Third, we have explored similarities and differences of DNNs and the human visual system [55, 109].