Empirical Inference

Report for the Scientific Advisory Board
April 18-20, 2016

Max Planck Institute for Intelligent Systems, Stuttgart • Tübingen
Contents

1 Empirical Inference 3
  1.1 Research Overview ............................................. 3
  1.2 Research Projects ............................................... 11
  1.3 Equipment ....................................................... 33
  1.4 Awards & Honors ............................................... 35
  1.5 Bernhard Schölkopf ............................................. 39
  1.6 Publications ..................................................... 41
1 EMPIRICAL INFERENCE

1.1 Research Overview

The problems studied in the department\(^1\) 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 recent 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. No matter whether the applications are done in the department or in collaboration with external partners, considering a whole range of applications helps us *study principles and methods of inference*, rather than inference applied to one specific problem domain.

The most competitive publication venues in empirical inference are NIPS (Neural Inform-\(^1\)The department of EI previously belonged to the MPI for biological Cybernetics. The present extended evaluation covers the period since the last departmental evaluation as part of that institute, i.e., since 2009.

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, neuroscience, and bioinformatics), 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”). Over the last decade, statistical learning theory has made significant progress in the analysis of supervised learning problems such as classification and regression.

The department has made various contributions to this area, including two recent NIPS orals (2011, 2012), where we prove that regressors can escape the curse of dimensionality if the intrinsic dimensionality of the data is low [546], and provide methods to increase performance by introducing gradient weights [481]. We have, however, gradually shifted our attention to areas of machine learning where statistical learning theory is less well developed. These include settings that incorporate both labeled and unlabeled data, such as semi-supervised or transductive learning, active learning and transfer learning, as well as unsupervised learning tasks such as clustering, manifold learning, graph-based learning, and the emerging field of complex network science. 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. We provided novel statistical guarantees for cluster-based active learning procedures [340] and a first formal analysis of the use of active learning for learning under covariate shift [341].

We studied theoretical aspects of transductive learning [342], introduced a new complexity measure of function classes called Permutational Rademacher Complexity, which is better suited for transductive problems than other popular analogues, and used it to derive data-dependent generalization bounds for transductive algorithms.
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 non-parametric 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 [487].

In recent years, we have developed kernel mappings of probability distributions to RKHSes [142]. For sufficiently rich feature spaces, these maps can be shown to be injective [253], i.e., they represent probability distributions without any loss of information. This property is profoundly linked to questions of optical imaging, by viewing a Fraunhofer imaging process as a particular case of kernel mapping [423].

Using such mappings, we can compare samples from two distributions, to determine whether they are distinguishable to a statistically significant degree [142]. This comprises the problem of independence testing as a special case, where we determine the difference between the embedding of a joint distribution and that of the product of the marginals.

They can be generalized to yield embeddings of conditional distributions, which are a core building block of a number of learning algorithms. Conditional distribution embeddings can be used to perform Bayesian inference on graphical models [556, 558, 602]. A second application is in reinforcement learning, where conditional expectations are used to update the state of an agent given its observations and actions [462]. A final application of conditional mean embeddings is the challenging task of testing for conditional dependence [390, 519]. This is of interest for finding which covariates contain redundant information, but also when performing causal inference, by providing a generalization of the well-known PC algorithm by Glymour and Spirtes that can incorporate nonlinear interactions.

Although kernel methods were originally associated with statistical learning theory, our work has helped move them towards probabilistic approaches. This includes our generalization of SVMs to the case where the inputs are not point observations, but nontrivial probability measures [415, 478]. This allows us to model uncertainty of the inputs. It also provides a generalization of SVMs towards multi-scale kernel expansions that naturally arise as a consequence of nontrivial input measures, as well as a generalization of SVMs towards multi-layer systems.

Kernel means are used throughout the field nowadays, and it is thus of significant interest to study their estimation properties. This led us to generalizations of classical results on Stein shrinkage estimation [35, 395].

Recently, we have started exploiting kernel means in the context of probabilistic programming, proposing an elegant way of representing the distribution of arbitrary arithmetic functions of random variables. Being a kernel method, this approach is applicable also to non-vectorial data types and accordingly holds promise for generalizing standard computing operations to distributions over data types [43]. Further details on our work on learning algorithms, including for...
the new field of network science, is included in the project reports below.

**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 aspects:

1. We usually 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). Our work in this domain includes the use of restricted function classes such as additive noise models, which solve the previously open problems of cause-effect inference [84, 624] and allow the detection of confounders in certain cases [625]. It also includes functional models where the functions are implemented by Turing machines and the implied dependences are not statistical ones, but characterized in terms of Kolmogorov complexity. We have developed a theory of graphical models for such models, analogous to the known probabilistic approach. It uses algorithmic mutual information rather than the usual statistical notion of mutual information, and it can be derived from a computing model (using Turing machines) just like graphical models can be derived from FCMs. Since it is based on algorithmic complexity of bit strings, it is in principle also applicable to individual (non-statistical) observations [234, 590].

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 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 mechanisms are autonomous entities of the world that (in the generic case) do not depend on each other, and changing (or intervening on) one of them often leaves the remaining ones invariant.

In a first study, we were able to show that if the effect is a deterministic invertible function of the cause which satisfies a suitable independence condition relative to the cause distribution, then this independence condition is provably violated between the effect and the inverse function [134, 562]. We have used the general independence assumption to connect the problems of covariate shift and transfer learning to causality, arguing that the causal conditionals (or mechanisms) should often be invariant under domain changes [360, 470]. We have also exploited it to develop methods for inferring causal graphs based on this kind of invariance [20].

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. We have shed light on the conditions under which semi-supervised learning (SSL) works, showing that knowledge of the underlying causal structure allows us to predict whether SSL can possibly work [24, 470]. Likewise, we believe that machine learning approaches are useful in advancing the field of causality. Some examples include the kernel-based tests for (conditional) independence developed in our lab [390, 519], our work on consistency of cause-effect inference [400], and the application of kernel mean embeddings to learn and analyze cause-effect-inference as a classification problem [356].

Our lab has played a major role in putting causal inference on the agenda of the machine learning community, and we expect that causal inference will have practical implications for many scientific inference problems, especially where interventions are not feasible (e.g., in astronomy [349] and neuroscience [14, 341]). It touches statistics, econometrics, and philosophy, and it constitutes one of the most exciting field for conceptual basic research in machine learning today.

**Probabilistic Inference** The probabilistic formulation of inference—conditioning probabil- ity measures encoding prior assumptions on observed data by multiplying with a likelihood des- cribing the data’s assumed generative process—remains 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 [240, 371, 543]. We also produced widely used software for Gaussian process regression [249] and inference in generalized linear models [130].

The central computational bottleneck in Bayesian models is the marginalization of latent variables. This can be computationally demanding, so approximate inference routines reducing computational complexity are a major research theme. This includes work on variational models, expectation propagation [498], and custom-built approximations for specific applications [232, 452, 651], as well as general integration routines [396] constructed de novo.

Members of the department have also been interested in uses of probabilistic inference for reinforcement learning and control [16, 404].

A broad new research theme has emerged around *probabilistic numerics*, the description of computation itself as probabilistic inference. Although this work was started in the department and several group members are still department members, currently we are not discussing it in further detail. More on this direction can be found in the project section on probabilistic numerics, and the separate section on the Emmy Noether Research Group that has arisen from it.

**Computational Imaging** Our work in this domain focuses on how to recover an image from corrupted measurements by modeling the distortion process and solving the resulting ill-posed inverse problem using image priors to restore the most likely image explaining our observations. We studied two major types of distortions: blur and noise. The latter included learning approaches using neural networks, some of them beating the state of the art BM3D algorithm [450, 539], as well as a collaboration with a department of our Stuttgart site [52].

For the purpose of deblurring, we developed an online method for *multi-frame blind decon- volution* [620] and generalized it to be able to super-resolve and to be robust against saturated pixels. To extend the applicability of deconvolution algorithms to spatially varying blur, we proposed the *Efficient Filter Flow* (EFF) model [570], and combined it successfully with our online blind deconvolution method [620]. This paved the way to consider *camera shake*, a major source of non-uniform blur in photography. While it had previously been modeled as spatially invariant, real-world shake leads to spatially variant blur that we modeled using EFF, improving upon the state-of-the-art [597]. By incorporating physical constraints into the shake model, we significantly decreased running time and improved reconstruction quality [537, 718].

Non-uniform blur also results from *lens aberrations*, usually minimized through sophisticated lens design. Nonetheless, even professional lenses exhibit residual aberrations, resulting in smoothly varying blur that tends to get worse in the image corners. We developed meth-

---

3These methods have led to a start-up involving several department members.
ods that automatically reduce lens aberrations from images, with or without a priori knowledge of the lens blur [472, 535]. More recently, we cast this into a data-driven regression framework, leading to what is probably the current state of the art [320]. This has the potential to change future lens designs by shifting the focus from expensive hardware solutions towards a combination of hardware and sophisticated software.

Two relevant application areas of computational imaging are astronomy and medical imaging, and we briefly mention some of our work in that field, starting with the latter. Motivated by our work on camera shake removal, we assayed whether similar ideas could be applied to magnetic resonance imaging (MRI) motion correction. The setup is different from regular camera shake: MR images are taken by recording trajectories in Fourier space (“k-space”), and a moving object will blur the different frequencies in a way that depends on the order in which the k-space is scanned. By carefully modeling the distortion process and imposing a sharpness prior on the resulting image, we were able to recover excellent images from sequences that would be rendered useless for standard MR reconstruction algorithms. This work applies not only for translational movements but also for rotational ones [98]. A second line of research, pursued in collaboration with the University of Tübingen’s Laboratory for Preclinical Imaging, applies machine learning and image processing methods in combined PET-MR systems. We have developed a solution to the problem of PET attenuation correction, which was licensed to Siemens and won a best paper award of the field’s flagship journal [103, 205].

In astronomy, our work initially focused on high-resolution imaging using deconvolution [184], but we applied our expertise also to other astronomical data analysis problems. Recent highlights include an approach to combine heterogeneous imaging data by voting methods [361], as well as causal modeling methods for light curves that have led to the discovery of a number of previously unknown exoplanets [38, 349].

**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 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. To overcome such difficulties, we use imitation learning to teach robots new motor skills [238]. A human demonstrator first provides one or several examples of the skill. Information recorded through motion capture or physical interaction is used by the robot to automatically generate a controller that can replicate the movements. This step can be accomplished using modern machine learning techniques. Imitation learning allows robots to learn the observed behavior and have a good starting point for self-improvement. This so called self-improvement of the task can help the robot adapt the learned movement to the characteristics of its own body or the requirements of the current context. Hence, even if the examples presented by humans are not optimal, the robot can still use them to bootstrap its behavior. In our lab, imitation learning has already been used to teach complex motor skills to various kinds of robots. This includes skills such as grasping of novel objects [239, 574, 575], robot table tennis [48, 581], ball-in-the-cup [178], throwing & catching [144, 467, 499], and tether ball [455, 465]. Our work on grasping won the ICINCO 2010 Best Paper Award. New machine learning methods that can acquire motor skills faster are developed. The goal of this research is to have intelligent robots that can autonomously enlarge their repertoire of skills by observing or interacting with human teachers.

However, supervised learning is not always sufficient for motor learning problems, partly because an expert teacher or idealized version of the behavior is often not available. Accordingly, one of our goals is the development of reinforcement learning methods which can handle the dimensionality of humanoid robots and generate actions for seven or more degrees of freedom. Efficient reinforcement learning for continuous states and actions is essential for robotics and control [683]. For high-dimensional state and action spaces, it is often easier to directly learn policies without estimating accurate system models. The resulting algorithms are parametric pol-
Machinen Lernen in Neurowissenschaften 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 30). This research aims to construct systems that replace and rehabilitate lost communication and motor skills in patient populations. The contribution of machine learning to BCI is in developing, refining, and applying algorithms to improve the accuracy with which neural signals are decoded. Our algorithmic developments ([13, 186, 187, 283, 560]) as well as improvements in experimental methodology (for example, as described in [137, 223, 298]) are significant contributions towards making clinical BCI systems a reality.

In addition to the advancement of methods for brain decoding, further progress in this field is contingent on a better understanding of the neural basis of BCI control. Research has largely focused on neural signals that provide information on a subject’s intent, neglecting those processes that determine whether a subject is in a state of mind suitable for operating a BCI. In collaboration with the Institute of Medical Psychology and Behavioral Neurobiology at the University of Tübingen, we have studied the neuro-physiological causes of good and poor BCI-control in healthy subjects and patient populations. This work has provided insights into the relation of attention networks and BCI-performance [149, 174], as recognized by the annual International BCI Research Award 2011, and has led to a new class of cognitive BCI paradigms for patients in late stages of amyotrophic lateral sclerosis (ALS) [81, 346].

Our interest in interfacing with the brain is not confined to communication. We are further exploring the use of BCI systems as assistive devices in stroke rehabilitation. Online decoding of movement intent in severely disabled patients can be employed to deliver congruent haptic feedback through robotic devices, thereby re-establishing the disrupted sensorimotor feedback loop. This synchronization may support cortical reorganization, hopefully resulting in enhanced recovery [183]. We have developed a BCI-controlled robotic system in collaboration with the Neurosurgery Department at the University of Tübingen, and demonstrated the feasibility of real-time haptic feedback on movement intent in healthy subjects and stroke patients [194].

We have also designed machine learning techniques to assist with the interpretation of experimental brain data (see page 31). Unsupervised learning tools were designed to automatically identify key events in large neural recordings, and were an integral part of several high-impact publications on memory-triggering events [25, 160]. 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. We have been developing causal inference tools adapted to these challenges [46, 174, 224, 353], and this work led to significant neuroscientific findings such as the role of traveling electrical waves in routing sensory information across the visual cortex [34]. Machine learning also pro-

Our work on reinforcement learning (RL) also comprises algorithmic and theoretical developments that are of independent interest. In this context, significant progress has been made regarding new information theoretic RL approaches [218, 508, 586], hierarchical RL approaches [455], as well as structured approaches [463]. We also studied theoretical foundations for Bayesian RL approaches [540] and the analysis of reinforcement learning algorithms in the bandit setting [536] and within inverse reinforcement learning [508].

Our resulting approach [465] has won the IROS 2012 Best Cognitive Paper, and was both IROS 2012 Best Student Paper Finalist and Best Paper Award Finalist.

In this context, significant progress has been made regarding new information theoretic RL approaches [218, 508, 586], hierarchical RL approaches [455], as well as structured approaches [463]. We also studied theoretical foundations for Bayesian RL approaches [540] and the analysis of reinforcement learning algorithms in the bandit setting [536] and within inverse reinforcement learning [508].
vided theoretical insights into biological learning by showing that observed properties of synaptic plasticity are optimal for neurons to learn under metabolic constraints imposed by limited energetic resources [95, 473].

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 part-time member of our department. In one line of work, we use machine learning to uncover the behaviorally relevant features observers use in complex perceptual tasks: given high-dimensional sensory input, which are the features that sensory systems base their computations on? We have applied our methods successfully to visual saliency [290], gender discrimination [242], and auditory tone-in-noise detection [148]. We have also published a discussion of the application of kernel methods in the flagship journal of cognitive science [289]. A second focus is 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 [108].

**Bioinformatics** In the past, the department was active in several branches of computational biology. While these activities have been significantly reduced, they overlapped with the reporting period and will thus briefly be included.

Genetics has witnessed an impressive increase in datasets that contain detailed information both about the genetic properties of individuals and their appearance properties, that is about their *genotype* and *phenotype*. Large genetics consortia around the world are recording single-nucleotide polymorphisms (SNPs), i.e., single genome bases that differ between individuals, for large sets of individuals. The standard technique is then to compute which of these $10^5$ to $10^7$ genome positions correlate most strongly with the disease status. Due to the enormous computational costs that an exhaustive scan for, say, $10^{14}$ candidate pairs would entail, these studies usually ignore any interactive effect between genomic loci, although these are believed to be important in many diseases. We have developed new engineering approaches and mathematical results to tackle this fundamental computational problem in Statistical Genetics [513].

Another recent area in Computational Biology is Molecular Genomics, in which new sequencing technologies enable an unprecedented wealth of data for exploring the structure of the genome and of individual genes. We have developed several tools to detect genes within the genome, and to establish a fine-grained picture of gene structure and of the impact of sequence variation on gene expression levels [312, 313, 650].
### 1.2 Selected Research Projects

<table>
<thead>
<tr>
<th>Research Area</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning theory</td>
<td>12</td>
</tr>
<tr>
<td>Random geometric graphs</td>
<td>13</td>
</tr>
<tr>
<td>Kernel methods</td>
<td>14</td>
</tr>
<tr>
<td>Optimization and large scale learning</td>
<td>16</td>
</tr>
<tr>
<td>Structure and dynamics of diffusion networks</td>
<td>17</td>
</tr>
<tr>
<td>Causal Inference</td>
<td>18</td>
</tr>
<tr>
<td>Probabilistic control</td>
<td>20</td>
</tr>
<tr>
<td>Probabilistic numerics</td>
<td>21</td>
</tr>
<tr>
<td>Computational photography</td>
<td>23</td>
</tr>
<tr>
<td>Medical and neuroscientific imaging</td>
<td>25</td>
</tr>
<tr>
<td>Inference in Astronomy</td>
<td>26</td>
</tr>
<tr>
<td>Robot skill learning</td>
<td>27</td>
</tr>
<tr>
<td>Reinforcement learning</td>
<td>28</td>
</tr>
<tr>
<td>Brain-computer interfacing</td>
<td>30</td>
</tr>
<tr>
<td>Learning and inference for neuroscience</td>
<td>31</td>
</tr>
<tr>
<td>Psychophysics and computational models of behaviour</td>
<td>32</td>
</tr>
</tbody>
</table>
Learning theory

Machine learning algorithms are designed to yield accurate predictions on future data by generalizing from past observations. The main goal of learning theory is to analyze their statistical and computational properties, and to provide performance guarantees on the new data. To do so, it poses these tasks in a rigorous mathematical framework and deals with them under various assumptions on the data-generating process.

PAC-Bayesian analysis (Probably Approximately Correct) provides guarantees on the generalization ability of randomized predictors within the classical PAC framework. We derived general PAC-Bayesian inequalities which allow to control the concentration of weighted averages of possibly uncountably many simultaneously evolving and interdependent martingales [138]. We applied these inequalities to contextual bandits providing a new regret bound and learning algorithm [536].

We have studied the relations and differences between Vapnik-Chervonenkis (VC) and Popper’s dimensions. While both VC and Popper’s dimensions aim to quantify the process of falsification, the first focuses on passive falsification based on examples provided by nature, whereas the latter focuses on active falsification through experiment design. We have shown that concepts that are related to Popper’s notion of falsifiability occur in the domain of query-learning and derived relations between Popper’s, exclusion, and VC-dimension [677].

We have analyzed how the performance of learning algorithms adapts to local structure in the data, which often helps when the ambient dimension is huge. We established that certain regression procedures, such as kernel and k-NN regression, adapt to a locally low intrinsic dimension [439, 546]. Further, we have shown how to improve regression performance by adapting the learning process to differing variability in various coordinates of the function to be estimated [481]. For this, we proposed a novel, consistent estimator for gradient norms. We also developed new data-structures for regression in a streaming setting [440].

We showed the first statistical consistency result (under surprisingly general distributional conditions) for additive noise methods in causal inference [400]. Further, we posed the problem of cause-effect inference as binary classification of probability measures, proposed a computationally efficient algorithm based on kernel mean embeddings achieving state-of-the-art performance, and provided a risk bound for the approach [356].

In transductive learning the learner observes labeled training and unlabeled test points with the final goal of correctly labeling the test points. We have introduced a new complexity measure of function classes called Permutational Rademacher Complexity, argued that it is better suited for analysis of transductive learning than other popular measures of complexity, and derived a general risk bound for transductive algorithms based on it [342].

Active learning exploits structure and information in unlabeled data to save label supervision. The goal is to ask for only a small subset of the data to be labeled while maintaining good performance. We provided novel statistical guarantees for a large class of practical cluster-based active learning procedures [340]. We further showed that an algorithmic technique from margin-based active learning yields a computationally efficient learner for linear separators under conditions that had earlier been shown to allow for fast statistical learning rates (Massart noise condition) [339].

We also provided a first formal analysis of the use of active learning for domain adaptation, that is, to use a learner’s active querying ability to adapt to changes in the data generation. We proposed a new active learning strategy and proved that it consistently learns in situations where no passive learning algorithm is consistent, while automatically adapting its amount of label queries to what is needed due to the underlying data shift [341].

More information: https://ei.is.tuebingen.mpg.de/project/learning-theory
Random geometric graphs

Random geometric graphs are built by first sampling a set of points from some underlying distribution, and then connecting each point to its k nearest neighbors. In this project we investigated the behavior of distance functions on random geometric graphs when the sample size n goes to infinity (and the connectivity parameter k scales appropriately).

It is well known that in graphs where the edges are suitably weighted according to their Euclidean lengths, the shortest path distance converges to the underlying Euclidean distance. However, it turned out this is not the case for unweighted kNN graphs \[461\]. In this case, the shortest path distance converges to a distance function that is weighted by the underlying density and takes wide detours to avoid high density regions (see left figure above). In machine learning applications, this behavior of the shortest path distance can be highly misleading. As an example, consider the Isomap algorithm and the data set shown in the middle figure. If we build an unweighted kNN graph based on this data and apply Isomap to recover the point configuration, we get the figure on the right. Obviously, it is grossly distorted and cannot serve as a faithful representation of the original data.

The commute distance (aka resistance distance) between vertex u and v is defined as the expected time it takes the natural random walk starting in vertex u to travel to vertex v and back. It is widely used in machine learning because it supposedly satisfies the following, highly desirable property: Vertices in the same cluster of the graph have a small commute distance, whereas vertices in different clusters of the graph have a large commute distance to each other. We studied the behavior of the commute distance as the number of vertices in the graph tends to infinity \[599\], proving that the commute distance between two points converges to a trivial quantity that only takes into account the degree of the two vertices. Hence, all information about cluster structure gets lost when the graph is large enough.

To alleviate this shortcoming, we proposed the family of \(p\)-resistances \[497\]. For \(p = 1\) it reduces to the shortest path distance, for \(p = 2\) it coincides with the resistance distance, and for \(p \to \infty\) it is related to the minimal s-t-cut in the graph. The family shows an interesting phase transition: there exist two critical thresholds \(p^*\) and \(p^{**}\) such that if \(p < p^*\), then the \(p\)-resistance depends on meaningful global properties of the graph, whereas if \(p > p^{**}\), it only depends on trivial local quantities and does not convey any useful information. In particular, the \(p\)-resistance for \(p = p^*\) nicely reveals the cluster structure.

More information: https://ei.is.tuebingen.mpg.de/project/distance-functions-on-random-geometric-graphs
Kernel methods

A Hilbert space embedding of distributions (KME)—which generalizes the feature map of individual points to probability measures—has emerged as a powerful machinery for probabilistic modeling, machine learning, and causal discovery. The idea behind this framework is to map distributions into a reproducing kernel Hilbert space (RKHS) endowed with a kernel $k$. It enables us to apply RKHS methods to probability measures and has given rise to a great deal of research and novel applications of kernel methods.

Given an i.i.d. sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, the most natural estimate of the embedding $\mu_p = \mathbb{E}_p[k(X, \cdot)]$ is an empirical average $\hat{\mu}_p = (1/n) \sum_{i=1}^n k(x_i, \cdot)$. In [35, 370], we showed that this estimator is not optimal in a certain sense. Inspired by James-Stein estimator, we proposed the so-called kernel mean shrinkage estimators (KMSEs) which improves upon the standard estimator. A suitable explanation for the improvement is a bias-variance tradeoff: the shrinkage estimator reduces variance substantially at the expense of a small bias. In addition, we presented a class of estimators called spectral shrinkage estimators (SSEs) which incorporates the RKHS structure via the eigenspectrum of the empirical covariance operator. Our empirical studies suggest that the proposed estimators are very useful for “large $p$, small $n$” situations (e.g. medical data, gene expression analysis, and text documents).

A natural application of KME is in testing for similarities between samples from distributions. We refer to the distance between two distribution embeddings as the maximum mean discrepancy (MMD). We have formulated a two-sample test [142] (of whether two distributions are the same), and showed that the independence test (of whether two random variables observed together are statistically independent) is a special case. A further application of the MMD as independence criterion is in feature selection, where we maximize dependence between features and labels [143]. We have further developed alternative independence tests based on space partitioning approaches and classical divergence measures (such as the $\ell_1$ distance and KL-divergence) [268]. Lastly, we also constructed the test for non-i.i.d. data such as time-series in [441].

Given that the MMD depends on the particular kernel that is chosen, we proposed two kernel selection strategies [494], the earlier one relying on a classification interpretation of the MMD, and the later one explicitly minimizing the probability of Type II error of the associated two-sample test (that is, the probability of wrongly accepting that two unlike distributions are the same, given samples from each).

We have also used the KME to develop a variant of an SVM which operates on distributions rather than points [478], permitting modeling of input uncertainties. One can prove a generalized representer theorem for this case, and in the special case of Gaussian input uncertainties and Gaussian kernel SVMs, it leads to a multiresolution SVM, akin to an RBF network with variable widths, which is still trained by solving a quadratic optimization problem. In [356], we applied this framework to perform bivariate causal inference between $X$ and $Y$ as a classification problem on joint distribution $P(X, Y)$. Another interesting application is in domain adaptation [407, 676]. This idea has also been extended to develop a variant of One-class SVM that operates on distributions, leading to applications in group anomaly detection [415].

A recent application uses kernel means in visualization. When using a power-of-cosine kernel...
for distributions on the projective sphere, the kernel mean can be represented as a symmetric tensor. In the context of diffusion MRI, this permits an efficient visual and quantitative analysis of the uncertainty in nerve fiber estimates, which can inform the choice of MR acquisition schemes and mathematical models [110, 388].

A natural question to consider is whether the MMD constitutes a metric on distributions, and is zero if and only if the distributions are the same. When this holds, the RKHS is said to be characteristic. We have determined necessary and sufficient conditions on translation invariant kernels for injectivity, for distributions on compact and non-compact subsets of $\mathbb{R}^d$ [253]: specifically, the Fourier transform of the kernel should be supported on all of $\mathbb{R}^d$. Gaussian, Laplace, and B-spline kernels satisfy this requirement. The MMD is a member of a larger class of metrics on distributions, known as the integral probability metrics (IPMs). In [16, 4], we provide estimates of IPMs on $\mathbb{R}^d$ which are taken over function classes that are not RKHSs, namely the Wasserstein distance (functions in the unit Lipschitz semi-norm ball) and the Dudley metric (functions in the unit bounded Lipschitz norm ball), and establish strong consistency of our estimators. Comparing the MMD and these two distances, the MMD converges fastest, and at a rate independent of the dimensionality $d$ of the random variables – by contrast, rates for the classical Wasserstein and Dudley metrics worsen when $d$ grows.

Embeddings of distributions can be generalized to yield embeddings of conditional distributions. The first application is to Bayesian inference on graphical models. We have developed two approaches: in the first [556, 602], the messages are conditional density functions, subject to smoothness constraints; these were orders of magnitude faster than competing nonparametric BP approaches, yet more accurate, on problems including depth reconstruction from 2-D images and robot orientation recovery. In the second approach [558], conditional distributions $P(Y|X = x)$ are represented directly as embeddings in the RKHS, allowing greater generality (for instance, one can define distributions over structured objects such as strings or graphs, for which probability densities may not exist). We showed the conditional mean embedding to be a solution to a vector valued regression problem [492], which allows us to formulate sparse estimates. The second application is to reinforcement learning. In [491], we estimate the optimal value function for a Markov decision process using conditional distribution embeddings, and the associated policy. This work was generalized to partially observable Markov decision processes in [462], where the kernel Bayes’ rule was used to integrate over distributions of the hidden states.

Another important application of conditional mean embeddings is in testing for conditional independence (CI). We proposed a Kernel-based Conditional Independence test (KCI-test) [519] which avoids the classical drawbacks of CI testing. Most importantly, we further derived its asymptotic distribution under the null hypothesis, and provided ways to estimate such a distribution. Our method is computationally appealing and is less sensitive to the dimensionality of $Z$ compared to other methods. This is the first time that the null distribution of the kernel-based statistic for CI testing has been derived. Recently, we proposed a new permutation-based CI test [390] that easily allows the incorporation of prior knowledge during the permutation step, has power competitive with state-of-the-art kernel CI tests, and accurately estimates the null distribution of the test statistic, even as the dimensionality of the conditioning variable grows.

Lastly, we have recently leveraged the KME in computing functionals of random variables $Z = f(X_1, X_2, \ldots, X_n)$ [43], which is ubiquitous in various applications such as probabilistic programming. Our approach allows us to obtain the distribution embedding of $Z$ directly from the embeddings of $X_1, X_2, \ldots, X_n$ without resorting to density estimation. It is in principle applicable to all functional operations and data types, thank to the generality of kernel methods. Based on the proposed framework, we showed how it can be applied to non-parametric structural equation models, with an application to causal inference. As an aside, we have also developed algorithms based on distribution embedding for identifying confounders [422], which is one of the most fundamental problems in causal inference.

More information: https://ei.is.tuebingen.mpg.de/project/kernel-distribution-embeddings
Optimization and large scale learning

Optimization lies at the heart of most machine learning algorithms. The key aspects of the applications in which these algorithms are applied include: high-dimensional, noisy, and uncertain data; huge volumes of batch or streaming data; intractable models, low accuracy, and reliance on distributed computation or stochastic approximations. The success of most machine learning algorithms depends on how the optimization techniques can adapt and exploit these facets. Our interests are broadly divided into two categories, convex and non-convex methods.

Convex optimization In the realm of methods for convex optimization, we have addressed research challenges under various different problem settings. For large-scale problems, where scalability is an important aspect, a summary overview of large-scale aspects of convex optimization appears in our work [678]. A theoretically optimal large-scale convex method for problems with linear constraints is presented in [366] which develops a new stochastic alternating direction method of multipliers (ADMM) method that combines Nesterov’s accelerated gradient methods with ADMM.

Non-convex Optimization In the domain of non-convex optimization for large-scale problems, our work [673] presents a simplified analysis of what, to our knowledge, is the first non-convex, non-smooth incremental proximal method. This work started in 2011; interestingly, in recent years, the interest in incremental methods has sky-rocketed, though the analysis is limited only to the convex case. Finally, we mention a new direction in nonconvex optimization offered by our recent work [443], which introduces "Geometric optimization" on the manifold of positive definite matrices. The underlying idea is to develop a theory of convexity along geodesics on the Positive Semi-Definite manifold. This work also identifies some basic calculus rules for detection and construction of geodesically convex functions on the Positive Definite manifold, and as an application presents new algorithms for solving maximum likelihood estimation for elliptically contoured distributions, which despite non-convexity remain tractable thanks to geodesic convexity.
We are interested in developing machine learning and large-scale data mining methods for the analysis, modeling and control of large real-world networks and processes that take place over them. We addressed mainly two problems: network inference and influence maximization.

**Network inference** Observing a diffusion process often reduces to noting when nodes (people, blogs, etc.) reproduce a piece of information, get infected by a virus, or buy a product. Epidemiologists can observe when people become ill but they cannot tell who infected them or how many exposures and how much time was necessary for the infection to take hold. In information propagation, we observe when a blog mentions a piece of information. However if, as is often the case, the bloggers do not link to their sources, we do not know where they acquired the information or how long it took them to post it. Finally, viral marketers can track when customers buy products or subscribe to services, but typically cannot observe who influenced customers’ decisions, how long they took to make up their minds, or when they passed recommendations on to other customers. We observe when and when but not how or why information (be it in the form of a virus, a meme, or a decision) propagates through a population of individuals. The mechanism underlying the process is hidden. However, the mechanism is of outstanding interest, since understanding diffusion is necessary for stopping infections, predicting meme propagation, or maximizing sales of a product.

In our work, we formulated a generative probabilistic model of diffusion that aims to describe realistically how diffusion occurs over time in a network. First, we developed two algorithms, NetInf [133, 566] and MultiTree [457], which use submodular optimization to infer the structure of a diffusion network from diffusion traces. We demonstrate the effectiveness of our algorithm by tracing information cascades in a set of 170 million blogs and news articles from 3.3 million sites over a one year period. We find that the diffusion network of news tends to have a core-periphery structure with a small set of core media sites that diffuse information to the rest of the Web.

However, both algorithms force the transmission rates between all nodes to be fixed and static – and not inferred. To overcome this limitation, we then developed the algorithm NetRate [69, 375, 418, 512], which allows transmission at different rates across different edges, possibly time-varying, so that we can infer temporally heterogeneous interactions within a network. NetRate infers both structure and the temporal dynamics of the underlying network with provable guarantees using convex optimization. We study the evolution of information pathways in the online media space and find that information pathways for general recurrent topics are more stable across time than for on-going news events. Clusters of news media sites and blogs often emerge and vanish in matter of days for on-going news events. Major events involving civil population, such as the Libyan’s Civil War, lead to an increased amount of information pathways among blogs as well as in the overall increase in the network centrality of blogs and social media sites.

**Influence maximization** In this problem, one aims to select the most influential source node set of a given size in a diffusion network. A diffusion process that starts in such an influential set of nodes is expected to reach the greatest number of nodes in the network. Although the problem depends dramatically on the underlying temporal dynamics of the network, this was largely unexplored. In our work, we developed an influence estimation, InfluMax [458], which accounts for the temporal dynamics underlying diffusion processes [512]. Later, we developed a very efficient influence estimation method, ContInEst [429], which can be used in combination with InfluMax to scale up influence maximization to networks with million of nodes.

More information: [https://ei.is.tuebingen.mpg.de/project/structure-and-dynamics-of-diffusion-networks](https://ei.is.tuebingen.mpg.de/project/structure-and-dynamics-of-diffusion-networks)
Causal Inference

Causal Discovery In causal discovery, we try to learn a causal structure from data. This structure may be a directed acyclic graph (DAG) underlying a functional causal model (FCM), and possibly also the functions in the FCM. Without further assumptions, this goal is impossible to achieve: given only an observational distribution \( P \), we can find an FCM generating \( P \) for any graph \( G \) s.t. \( P \) is Markovian w.r.t. \( G \), i.e., the underlying graph is not identifiable. We investigate assumptions that make the graph identifiable from the observational distribution, and develop algorithms building on those assumptions. We now provide some examples, and refer to the literature for further papers that we cannot discuss below \([353, 358, 401, 422, 518]\).

Example (1): Additive Noise Models. In additive noise models, the functional assignments used in the FCM are of the form \( Z = f(X, Y) + N \). The subclass of linear functions and additive Gaussian noise does not lead to identifiability. This, however, constitutes an exceptional setting. If one assumes either (i) non-Gaussian noise, (ii) non-linear functions in the FCM \([84, 624]\) or (iii) all noise variables to have the same variance \([200]\), one can show that additive noise models are identifiable. Methods that are based on additive noise models perform above chance level not only on artificial data but also on the set of cause-effect pairs that we have collected over the last years \([36]\). A similar result holds if all variables are integer-valued \([585]\) or if we interpret the additivity in \( Z/kZ \) \([200]\). The concept of additive noise has been extended to time series, too \([438]\), as well as to the identification of confounders \([625]\).

Example (2): Information Geometric Causal Inference (IGCI). While the above methods inherently rely on noisy causal relations, statistical asymmetries between cause and effect also appear for deterministic relations. We have considered the case where \( Y = f(X) \) and \( X = f^{-1}(Y) \), for some invertible function \( f \), the task being to tell which variable is the cause. The general assumption of independence of causal mechanisms \([234]\) implies that \( P(X) \) and \( P(Y|X) \) should be independent if \( X \) causes \( Y \). Choosing \( P(X) \) and \( f \) independently implies that \( P(Y) \) tends to have high probability density in regions where \( f^{-1} \) has large Jacobian. This observation can be made precise within an information theoretic framework \([134, 562]\). Applying a non-linear \( f \) to \( P(X) \) decreases entropy and increases the relative entropy distance to Gaussians, provided that a certain independence between \( f \) and \( P(X) \) is postulated which can be phrased as orthogonality in information space.

A second approach, for linear invertible relations between multi-dimensional variables, is related in spirit: if the covariance matrix of \( X \) and the structure matrix relating \( X \) and \( Y \) are chosen independently, directions with high covariance of \( Y \) tend to coincide with directions corresponding to small eigenvalues of \( A^{-1} \), which can be checked by a formula relating traces of covariance matrices with traces of the product of structure matrices with their transpose \([518, 571]\).

Example (3): Invariant Prediction. In many situations, we are interested in the system’s behavior under a change of environment. Here,
causal models become important because they are often invariant under those changes [470]. Following the assumption of independence of causal mechanisms, localized changes of some noises or mechanisms of a causal model will often leave other conditionals invariant, and thus a causal prediction (which uses only direct causes of the target variable as predictors) may remain valid even if we intervene on predictor variables or change the experimental setting. We can exploit this for causal discovery: given data from different experimental settings, we use invariance as a criterion to estimate the set of causal predictors for a given target variable. This method also leads to valid confidence intervals for causal relations [20].

Example (4): Hidden Confounding in Time Series. Assume we are given a multivariate time series \( X_1, \ldots, X_L \) of measurements. In this project, our goal is to infer the causal structure underlying \( X_1, \ldots, X_L \), in spite of a potential unobserved confounder \( Z_t \in \mathbb{Z} \) (which other approaches such as Granger causality cannot handle). We assume a vector autoregressive causal model

\[
\begin{pmatrix} X_t \\ Z_t \end{pmatrix} := \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t.
\]

Restricting the model class to non-Gaussian independent noise \( (N_t)_{t \in \mathbb{Z}} \) makes \( B \) and \( C \) essentially identifiable [351]. We show that \( D = 0 \) is another sufficient restriction of the model class.

Example (5): Causal Strength. In real-world applications a measure of the strength of a causal influence is often required. This is a challenging question, even if the causal directed acyclic graph and the joint distribution are perfectly known. We have formulated a set of postulates that a measure of causal strength of an arrow (or of a set of arrows) in a causal network should satisfy [94]. We show that none of the measures in the literature satisfies all postulates, and describe examples where they therefore lead to poor results. This includes well-known approaches like Granger causality and transfer entropy for time series, and also measures for general graphs. The main problem is to quantify which part of the statistical dependences between two variables is due to a direct causal influence of one on the other, and which part is due to other causal paths. We propose an information-theoretic measure that satisfies all our postulates. It coincides with mutual information for a DAG with two nodes; for more complex DAGs it correctly removes the information propagated via alternative paths.

Causal Inference in Machine Learning We believe that causal knowledge is not only useful for predicting the effect of interventions, but that in some scenarios causal ideas can also improve the performance of classical machine learning methods. Again, we concentrate only on two examples and refer to some other papers [90, 470].

Example (6): Semi-supervised Learning. Our work [470] discusses several implications of the independence of cause and mechanism (as a special case of a more general independence principle) for standard machine learning. Let us assume that \( Y \) is predicted from \( X \). We have argued that semi-supervised learning (SSL) does not help if \( X \) is the cause of \( Y \) (“causal learning”), whereas it often helps if \( Y \) is the cause of \( X \) (“anticausal learning”). This is because additional observations of \( X \) only tell us more about \( P(X) \) – which is irrelevant in the case of causal prediction because the prediction requires information about the independent object \( P(Y|X) \). Our meta-study analyzing results reported in the SSL-literature supports this hypothesis: all cases where SSL helped where anticausal, confounded, or examples where the causal structure was unclear. To elaborate on the link between causal direction and performance of SSL, we studied the toy problem of interpolating a monotonically increasing function for the case where the relation between \( X \) and \( Y \) is deterministic [24]. In such a scenario \( P(X) \) can be shown to be beneficial for predicting \( Y \) from \( X \) whenever \( P(X|Y) \) and \( P(Y) \) satisfy a certain independence condition which coincides with the one postulated in our work on information-geometric causal inference [134].

Figure 1.4, finally, visualizes the idea of a recent method for correcting measurement errors called “half-sibling regression;” subtracting the conditional expectation of \( Y \) given \( X \) from \( Y \) can provide a better estimation of the quantity \( Q \) of interest than the noisy measurement \( Y \), which has been used in [349] to process astronomical light curves for the detection of exoplanets.

More information: [https://ei.is.tuebingen.mpg.de/project/causal-inference](https://ei.is.tuebingen.mpg.de/project/causal-inference)
Uncertainties have long been recognized as a key difficulty for control, deteriorating performance or even putting system safety at risk. This issue has been classically addressed by robust controller design, making use of a deterministic bound on the uncertainty and designing the controller for all possible uncertainty realizations. Tight uncertainty bounds are, however, difficult to obtain, limiting performance in practice, and robust techniques generally suffer from significant computational complexity.

The scope of this research is to develop new methods and tools for high performance and computationally efficient control of uncertain complex systems. We take a probabilistic approach, considering that the uncertainty affecting the system has a probabilistic nature. In contrast to common stochastic control methods, a key feature of the developed techniques is that they enable an online identification of the uncertainty distribution to maximize system performance, while providing (probabilistic) guarantees on the closed-loop system characteristics. The methods thereby leverage an interplay of probability, robustness, and adaption.

While uncertainties can enter controllers in various forms, we focus on model-based control techniques with model uncertainties due to high system complexity or due to the environment. The challenge is that learning happens in closed loop, i.e., 1) the resulting model is directly critical for both performance and safety at every time step and, 2) control and system identification have to be performed simultaneously.

**Learning-based model predictive control**

Model predictive control (MPC) makes use of predictions of the system and the environment in order to optimally choose a sequence of control inputs that provides safety in the form of stability and constraint satisfaction and minimizes a performance objective.

We investigate the use of learning-based identification methods for integration in an MPC controller, offering an ideal framework to incorporate probabilistic predictions based on online data. Due to their flexibility, Gaussian processes (GP) are used to model complex system behaviors and, most importantly, characterize the corresponding model uncertainty. The work [16] proposed a modeling and control approach using Gaussian processes to predict time-periodic errors that are then compensated by a predictive controller, which was demonstrated for high precision control of a telescope mount.

**Dual control** A key challenge of simultaneous control and system identification is that the control action influences not only the performance, but also the uncertainty in the dynamics. This is also called the exploration-exploitation trade-off. Dual control offers one way to address this trade-off by considering the control of the dynamical system augmented by the unknown parameters. While the exact dual controller is intractable, available approximations generally cannot maintain all core features: caution, exploration and the future value of information.

We have derived a tractable approximation of the dual control formulation that aims at maintaining these features, in particular the value of information, and have extended the technique to general nonlinear systems. The developed dual control framework has been applied in simulation to the problem of building control, providing simultaneous identification and control guided by electricity prices, weather, and occupancy.

More information: [https://ei.is.tuebingen.mpg.de/project/probabilistic-control](https://ei.is.tuebingen.mpg.de/project/probabilistic-control)
Probabilistic numerics

Artificial intelligent systems build models of their environment from observations, and choose actions that they predict will have beneficial effect on the environment’s state. The mathematical models used in this process call for computations that have no closed analytic solution. Learning machines thus rely on a whole toolbox of numerical methods: high-dimensional integration routines are used for marginalization and conditioning in probabilistic models. Fitting of parameters poses nonlinear (often non-convex) optimization problems. Predicting dynamic changes in the environment involves solving differential equations. In addition, there are special cases for each of these tasks in which the computation amounts to large-scale linear algebra (i.e. Gaussian conditioning, least-squares optimization, linear differential equations). Traditionally, machine learning researchers have served these needs by taking numerical methods “off the shelf” and treating them as black boxes.

Since the 1970s, researchers like Wahba, Diaconis, and O’Hagan repeatedly pointed out that, in fact, numerical methods can themselves be interpreted as statistical rules—more precisely, as acting machines, since they take decisions about which computations to perform: they estimate an unknown intractable quantity given known, tractable quantities. For example, an integration method estimates the value of an integral given evaluations of the integrand. This is an abstract observation, but Diaconis and O’Hagan separately made a precise connection between inference and computation in the case of integration: several classic quadrature rules, e.g. the trapezoid rule, can be interpreted as the maximum a posteriori (MAP) estimator arising from a family of Gaussian process priors on the integrand.

Over recent years, the research group on probabilistic numerics has been able to add more such bridges between computation and inference across the domains of numerical computation, by showing that various basic numerical methods are MAP estimates under equally basic probabilistic priors: quasi-Newton methods, such as the BFGS rule, arise as the mean of a Gaussian distribution over the elements of the inverse Hessian matrix of an optimization objective. This result can be extended to linear solvers, in particular the linear method of conjugate gradients (Gaussian regression on the elements of the inverse of a symmetric matrix). Regarding ordinary differential equations, some Runge-Kutta methods can be interpreted as autoregressive filters, returning a Gaussian process posterior over the solution of a differential equation.

The picture emerging from these connections is a mathematically precise description of computation as the active collection of information. In this view, the analytic description of a numerical task provides a prior probability measure over possible solutions, which can be concentrated through conditioning on the result of tractable computations. Many concepts and philosophical problems from statistics carry over to computation quite naturally, with two notable differences: first, in numerical “inference” tasks, the validity of the prior can be analyzed to a higher formal degree than in inference from physical data sources, because the task is specified in a formal (programming) language. Secondly, since numerical routines are the bottom, “inner loop” layer of artificial intelligence, they must curtail computational complexity. This translates into a constraint on acceptable probabilistic models—most basic numerical methods make Gaussian assumptions.

In the machine learning context, the description of computation as the collection of informa-
tion has opened a number of research directions:

1. Once it is clear that a numerical method uses an implicit prior, it is natural to adapt this prior to reflect available knowledge about the integrand. This design of “customized numerics” was used in a collaboration with colleagues at Oxford to build an efficient active integration method that outperforms Monte Carlo integration methods in wall-clock time on problems of moderate dimensionality [396].

2. Many numerical problems are defined relative to a setup that is itself uncertain to begin with. Once numerical methods are defined as probabilistic inference, such uncertainty can often be captured quite naturally. In a collaboration with colleagues in Copenhagen, it was shown [354, 369, 399] how uncertainty arising from a medical imaging process can be propagated in an approximate inference fashion to more completely model uncertainty over neural pathways in the human brain.

3. Explicit representations of uncertainty can also be used to increase robustness of a computation itself. Addressing a pertinent issue in deep learning, we constructed a line search method [321]—a building block of nonlinear optimization methods—that is able to use gradient evaluations corrupted by noise. The resulting method automatically adapts step sizes for stochastic gradient descent.

4. More generally, it is possible to define probabilistic numerical methods: Algorithms that accept probability measures over a numerical problem as inputs, and return another probability measure over the solution of the problem, which reflects both the effect of the input uncertainty, and uncertainty arising from the finite precision of the internal computation itself. A position paper [37] motivates this class of algorithms, and suggests their use for the control of computational effort across composite chains of computations, such as those that make up intelligent machines. In collaboration with the Optimization group at the German Cancer Research Center we developed approximations to propagate physical uncertainties through the optimization pipeline for radiation treatment, to lower the risk of complications for patients [111, 437].

In a separate but related development, a community has also arisen around the formulation of global optimization as inference, and the formulation of sample-efficient optimization methods. These Bayesian Optimization methods can, for example, be used to structurally optimize and automate the design of machine learning models themselves. We contributed to this area with the development of the Entropy Search [135] algorithm that automatically performs experiments expected to provide maximal information about the location of a function’s extremum.

Probabilistic numerics is emerging as a new area at the intersection of mathematics, computer science and statistics. As co-founders, the research group on probabilistic numerics plays a central role in its development. The wider Intelligent Systems community, with their intractably large data streams and non-analytic model classes, are simultaneously contributors and beneficiaries: equipping computational routines with a meaningful notion of uncertainty stands to increase both the efficiency and reliability of intelligent systems at large.

More information: https://ei.is.tuebingen.mpg.de/project/probabilistic-numerics
Empirical Inference

1.2 Research Projects

Computational photography

Figure 1.6: In many real-world imaging applications, the common assumption of stationary blur does not hold. Examples that exhibit non-stationary blur include e.g. camera shake, optical aberrations, and atmospheric turbulence. We derived a mathematically sound and physically well-motivated model, which allows to express and efficiently compute spatially-varying blur. Our “Efficient Filter Flow” framework substantially broadens the application range of image deconvolution methods. In a number of challenging real-world applications we demonstrated both the validity and versatility of our approach.

Digital image restoration as a key area of signal and image processing aims at computationally enhancing the quality of images by undoing the adverse effects of image degradation such as noise and blur. It plays an important role in both scientific imaging and everyday photography. Using probabilistic generative models of the imaging process, our research aims to recover the most likely original image given a low-quality image, or image sequence. In the following we give a number of illustrative examples to highlight some of our work.

Spatially varying blurs: our work on blind deconvolution of astronomical image sequences can recover sharp images through atmospheric turbulence, but is limited to relatively small patches of the sky, since the image defect is modeled as a space-invariant blur. Images that cover larger areas require a convolutional model allowing for space-variant blur. In we proposed such a model based on a generalization of the short-time Fourier transform, called Efficient Filter Flow (EFF), which is illustrated in Fig. 1.6. Deconvolution based on EFF successfully recovers a sharp image from an image sequence distorted by air turbulence (see Fig. 1.7).

Super-resolution: we generalized our online method for the multi-frame blind deconvolution problem to account for the adverse effect of saturation and to enable super-resolution given a sequence of degraded images. By drawing a connection between Fraunhofer diffraction and kernel mean maps we are able to show that under certain imaging conditions imaging beyond the physical diffraction limit is in principle possible.

Removing camera shake: photographs taken with long exposure times are affected by camera shake creating smoothly varying blur. Real hand-held camera movement involves both translation and rotation, which can be modeled with our EFF framework for space-variant blur. We were also able to recover a sharp image from a single distorted image, using sparsity-inducing image priors and an alternating update algorithm. The algorithm can be made more robust by restricting the EFF to blurs consistent with physical camera shake. This leads to higher image quality and computational advantages. To foster and simplify comparisons between different algorithms removing camera shake, we created a public benchmark dataset and a comparison of current methods.
Empirical Inference

1.2 Research Projects

Correcting lens aberration: even good lenses exhibit optical aberration when used with wide apertures. Similar to camera shake, this creates a certain type of blur that can be modeled with our EFF framework; but optical aberrations affects each color channel differently (chromatic aberration). We measured these effects with a robotic setup and corrected them using a non-blind deconvolution based on the EFF framework \[535\]. We were also able to implement a blind method rectifying optical aberrations in single images \[472\]. An example is shown in Fig. 1.9. The key was to constrain the EFF framework to rotationally symmetric blurs varying smoothly from the image center to the edges. We have recently been able to formulate this as non-parametric kernel regression, enabling faithful optical aberration estimation and correction \[320\], which might lead to new approaches in lens design.

Denoising: another classical image distortion is noise. Noise can exhibit different structure, e.g., additive white Gaussian, salt-and-pepper, JPEG-artifacts, and stripe noise, depending on the application.

In astronomical imaging, dim celestial objects require very long exposure times, which causes high sensor noise. It is common to subtract a dark frame from the image — an image taken with covered lens, containing only sensor noise. The difficulty with this is that the sensor noise is stochastic, and image information is not taken into account. We studied the distribution of sensor noise generated by a specific camera sensor and proposed a parameterized model \[538\]. Combined with a simple image model for astronomical images, this gives superior denoising.

Multi-scale denoising: noise usually has a more damaging effect on the higher spatial frequencies, so most algorithms focus on those. However, if the noise variance is large, also lower frequencies are distorted. We were able to significantly improve the performance of many methods for that setting by defining a multi-scale meta-procedure, leading to a DAGM 2011 prize \[539\].

Image restoration as a learning problem: since denoising can be also seen as a non-trivial mapping from noisy to clean images as sketched in Fig. 1.10, we were particularly interested whether a learning-based approach can be applied. Using very large data sets, we trained a multi-layer perceptron (MLP) that is able to denoise images better than all existing methods, leading to the new state-of-the-art denoising method \[450\]. Such a discriminative approach turned out to be effective also for other image restoration tasks such as inpainting \[402\], non-blind deconvolution \[424\], as well as blind image deconvolution \[26, 320\].

More information: https://ei.is.tuebingen.mpg.de/project/computational-imaging
Medical and neuroscientific imaging

Modern medical imaging is a blend of 20th century breakthroughs in physics and information processing. Although there is an ongoing advancement of technology, many of the problems remain unsolved. Using machine learning and image processing techniques for more efficient and intelligent use of acquired data seems a promising solution to some of the obstacles. For example, understanding the statistical properties of medical images allows making certain prior assumptions on unobserved or corrupted aspects of the data, which makes it possible to improve the image quality and aid the medical diagnostics.

PET-MR systems combine functional information from Positron Emission Tomography (PET) with structural information from Magnetic Resonance (MR) Imaging. They demand PET attenuation (or $\mu$) map to be determined from MR data, which constitutes an ill-posed problem. We have developed a Gaussian process prediction framework, trained on co-registered MR and $\mu$-maps, and evaluated it on brain and whole body data. A collaborative effort with the university hospital, Tübingen, the method has been patented and licensed to Siemens AG. Further research focused on evaluating the influence of positioning aids [191], improving the robustness of the method in presence of MR image artifacts [106] and the assessment of PET quantification accuracy in a pediatric patient collective [28].

Patient motion during long MR scans leads to severe non-local degradations and can render images unacceptable for medical diagnosis. We developed a retrospective (post-processing) approach called GradMC, which operates on the motion-corrupted images [98]. It automatically corrects for motion artifacts by searching for a motion-generating function whose inverse application optimizes a quality measure based on image statistics. Our GPU implementation processes a full 3D volume in a few minutes, which is acceptable in routine clinical use. In a follow-up work, we have come up with a fully retrospective non-rigid motion correction scheme that only needs raw data as an input [50].

Diffusion MRI (dMRI) allows us to investigate brain tissue microstructure non-invasively. Recently, supervised classification has been used to recognize patterns in this rich and complex data that allow us to automatically detect neurological diseases. We have contributed to this field by developing improved methods for feature extraction [345], by applying machine learning to specific diseases [56], and via a survey that summarizes the state of the art.

More information: https://ei.is.tuebingen.mpg.de/project/machine-learning-for-medical-and-neuroscientific-imaging
Astronomy offers some of the most exciting opportunities for machine learning applications in the natural sciences. Satellite missions as well as ground-based observations have generated datasets of unprecedented quality and size. New subfields of astronomy such as exoplanet science have emerged, impossible without large scale data analysis. However, by and large, data analysis in astronomy is still rather conservative, and there are thus large opportunities for data-driven modeling and probabilistic inference. Our work in this field benefits from methods developed within several projects in the department, and some of it is thus discussed elsewhere (see page 23 where an astronomical image deconvolution method [184] is discussed).

**Combining heterogeneous imaging** The Web contains millions of astronomical images that contain scientifically valuable information about the night sky. However, these images are difficult to use for scientific purposes because images rendered for human visual consumption are usually processed with non-linear tone mappings. We developed a system that builds a high dynamic-range and wide-angle image of the night sky by combining a large set of such images [361]. It computes a “consensus” image whose pixel brightness rank information tries to agree with the individual images. The complexity of the algorithm is linear in the number of images. It permits discovery of astronomical objects or features that are not visible in any of the input images taken individually, as shown by a faint tidal stellar stream (see Figure 1.12). More importantly, however, it permits scientific exploitation of a huge source of astronomical images that would not be available to astronomical research without our automatic system. The system was featured on the BBC Stargazing Live program.

**Half-sibling regression for modeling systematic errors**

The Kepler space observatory, launched in 2009, observes a tiny fraction of the Milky Way in search for exoplanets, monitoring the brightness of 150,000 stars. Some of them exhibit periodic decreases of brightness in their light curves due to partial occlusions caused by exoplanets. These measurements, however, are corrupted with systematic noise due to the telescope. However, since the stars can be assumed to be causally independent of each other (they are light years apart) as well as of the instrument noise, it turns out that we can denoise the signal of a single star by removing all information that can be explained by the measurements of the other stars. Under the assumption that the systematic noise acts in an additive manner, we provide strong theoretical guarantees regarding the quality of the reconstruction [349].

**Exoplanet search & characterization** We applied the half-sibling regression model to perform the first systematic search for transiting exoplanet signals in new data from NASA’s K2 Mission. Using this method, we discovered 36 new planet candidates—of which 21 were subsequently validated as bona fide exoplanets (Monnet et al., Astrophysical Journal, 809, 25, 2015)—and demonstrated that these data could be used for planet detection despite the large systematic errors [38]. We used principal component analysis to reduce the dimensionality of the predictors and extended the half-sibling methodology to jointly fit for the regression weights and the exoplanet properties. This increased the sensitivity of our search to small signals and allowed the propagation of uncertainties in the systems model to our estimates of the physical parameters.

More information: https://ei.is.tuebingen.mpg.de/project/astronomy
Robot skill learning

Creating autonomous robots that can learn to assist humans in daily life situations is a fascinating challenge for machine learning. We focus on the first step of creating robots that can learn to accomplish many different tasks triggered by an environmental context or a higher-level instruction and plan to obtain a general approach to motor skill learning. We focus on (1) domain-appropriate machine learning approaches that allow for better control, imitation of behavior and self-improvement, as well as (2) new robotics approaches to create more appropriate systems for high-speed skill learning.

Starting from theoretically sound robotic control structures for task representation and execution, we replace analytic modules with more flexible learned ones [468]. To this end, we tackle problems such as accurate but compliant execution in joint-space [583] or task-space [150], learning of elementary behaviors using combination of imitation and reinforcement learning [178, 238], hierarchical composition of behaviors, and parsing complex demonstrations into elementary behaviors.

Mimicking how children learn new motor tasks, we have used imitation to initialize to learn libraries of elementary primitives, and subsequently reinforcement learning to improve the performance. We have learned elementary tasks such as Ball-in-a-Cup or bouncing a ball [178, 238] and gradually moved to more complex ones. As the benchmark of complex behavior, we chose the task of returning table tennis balls over the net. We created a parser that segments movements of a human teacher into elementary movements [594, 596]. These then train the single elementary movements [178, 238]. Novel behaviors, modulated by the opponent’s incoming ball, are composed by mixing motor primitives [465, 581]. With the help of this method, the robot table tennis player learnt with this method successfully returns of 97% of the balls played against a ball gun. Current approaches are likely to achieve even better results [347] and use accurate prediction of the human opponent’s behavior before the opponent even touched the ball [459].

More information: https://ei.is.tuebingen.mpg.de/project/robot-skill-learning
Reinforcement learning ranks among the biggest challenges for machine learning. Just controlling a known dynamical system is hard on its own – interacting with an unknown system poses even harder decision problems, such as the infamous exploration-exploitation tradeoff. Most research in this area is still confined to theoretical analysis and simplistic experiments, but the promise of autonomous machines justifies the effort. Over the past years, members of the department contributed to reinforcement learning in theory and experiment.

**Non-Parametric Dynamic Programming** [530] showed that a non-parametric kernel density representation of system dynamics unifies several popular policy evaluation methods: their Galerkin method joins Least-Squares Temporal Difference learning, Kernelized Temporal Difference learning, and a type of discrete-state Dynamic Programming, as well as a novel method of improved performance.

**EM-like Reinforcement Learning** Policy search, a successful approach to reinforcement learning, directly maximizes the expected return of a policy – in contrast to value function approximation, which derives policies from a learnt value function. However, few of its variants scale to many dimensions, as they are based on gradient descent over many trials. To improve efficiency, [178] reduced the problem to reward-weighted imitation, treating rewards received after actions as improper probabilities indicating the actions’ success. Their idea resembles Expectation Maximization, giving good actions a higher probability to be re-used. This framework also unifies previous algorithms, and allows the derivation of novel ones, such as episodic reward-weighted regression and PoWER.

**Relative Entropy Policy Search** Policy improvements in policy search often invalidate previously collected information, causing premature convergence and implausible solutions. These problems may be addressed by constraining the information loss. Relative Entropy Policy Search (REPS) bounds the information loss while maximizing expected return [586]. REPS differs significantly from previous policy gradient approaches. It yields an exact update shown to work well on reinforcement learning benchmarks. REPS can be generalized hierarchically [455] using a gating network to choose among several option policies. This hierarchical REPS learns versatile solutions while increasing learning speed and the quality of the learnt policy.

**Bayesian reinforcement learning** Probability theory gives a uniquely coherent answer to the exploration-exploitation dilemma: From the Bayesian perspective, reinforcement learning is about including possible future observations in considerations about optimal behavior. Since probabilistic models can predict future data, this process can be rigorously formalized. It amounts...
Empirical Inference

1.2 Research Projects

to modeling knowledge as an additional dynamic variable to be controlled. In general, the combinatorial number of possible futures is intractable; however, [540] showed that the Gaussian process (GP) framework, in which predictions involve linear algebra calculations, allows approximating optimal exploration-exploitation with classic numerical methods for the solution of stochastic differential equations.

**Reinforcement learning with Gaussian Processes:** [279] used GPs for approximate dynamic programming in reinforcement learning, as probabilistic function approximators for the value function, and as models of the system dynamics. Using the predictive uncertainty for guidance, active learning methods could explore the state space efficiently. [555] proposed a particularly efficient use of GPs for optimal control over continuous states for non-bifurcating systems with low sampling rate. In their work, GPs capture information gained, as well as remaining uncertainty due to noise and lack of experience. The system’s behavior is predicted by propagating state and action distributions through time, tractability is achieved approximating distributions by moment matched Gaussians. This “virtual simulation” is used to optimize the control policy. Their algorithms learn from even limited interactions with the environment due to the power of using probabilistic forward models for such indirect experience rehearsal.

**Apprenticeship learning via inverse reinforcement learning** Unguided exploration can be hazardous for systems like robots. This issue is addressed by imitation learning from example actions provided by an expert, where the autonomous agent learns a policy generalizing the demonstrations to new states. This behavioral cloning may fail when the dynamics of expert and learner differ. Indeed, even simple repetition of the expert’s actions does not always yield the same results. An alternative is to infer the expert’s reward function from the expert’s behavior, then use it to learn in the new system. This avoids exhaustive exploration by searching for policies close to the expert’s. Previous work required a model of the expert’s dynamics, but [508] presented a model-free inverse reinforcement learning algorithm, using importance sampling to adapt expert examples to the learner’s dynamics. Tested on several benchmarks, the algorithm proved more efficient than the state of the art. Generalization in both forward and inverse reinforcement learning depends on the projection of states onto features to describe reward and value function. Features, especially visual ones, are often subject to noise, for example in robot grasping and manipulation tasks. To solve this problem, [463] combined control and structured output prediction over Markov Random Fields to represent the action distribution. Their method is robust to noise in a grasping task, and can also be used in other applications requiring control from vision.

**Data-dependent Analysis of Reinforcement Learning** Many analyses of reinforcement learning focus on worst-case scenarios, although reality is often not adversarial. [536] used PAC-Bayesian inequalities for martingales in a data-dependent analysis of the exploration-exploitation trade-off [138]. We studied stochastic multi-armed bandits with side information (also known as contextual bandits), a general framework where at each round of the game the agent is presented with side information (e.g., symptoms of a patient in a medical application) and has to find the best action (e.g., the best drug to prescribe given the symptoms). This model class is also used for personalized advertising on the internet. Our analysis includes the actual usage of side information by the algorithm, rather than the total amount of side information provided. This allows offering a lot of side information and letting the algorithm decide what is relevant, improving the run time of the algorithm exponentially over the state of the art.

More information: https://ei.is.tuebingen.mpg.de/project/reinforcement-learning
Brain-computer interfacing

Brain-computer interfaces (BCIs) translate neural recordings into signals that may be used for communication and/or the control of neuroprosthetic devices. Research in this domain poses interesting challenges to machine learning, because data is typically scarce, noisy, and non-stationary [186]. Furthermore, good decoding algorithms are contingent on domain knowledge that is not readily available and difficult to incorporate into traditional statistical methods. Accordingly, we employ machine-learning methods to study neural processes involved in BCI-control and use these insights to develop novel decoding algorithms and enhance experimental paradigms.

**Machine learning algorithms for brain-state decoding** A crucial aspect of our work is the development of algorithms for real-time brain-state decoding. We have developed a graphical model decoding framework for ERP-based visual speller systems [187]. Furthermore, we were the first to successfully apply the framework of multi-task learning to the domain of BCIs [560]. As the signal characteristics used by subjects to control a BCI share common aspects, the incorporation of data from previously recorded subjects substantially decreases calibration time and enhances overall decoding performance [13]. Besides working on methods for real-time decoding, we also develop tools to investigate the neural basis of disorders of cognition [14, 46]. This is essential to understand how the diseased brain differs from the one of healthy subjects, which has implications for the design of BCI systems for patient populations.

**Brain-computer interfaces for communication** Building upon our machine-learning methods, we have investigated the neural basis of the ability to operate a BCI in healthy subjects and in patient populations. We were able to show that the configuration of large-scale cortical networks, as represented in high-frequency gamma-oscillations of the brain’s electromagnetic field, influences a subject’s ability to communicate with a BCI [149, 174]. Building upon these insights, we have developed a novel class of BCIs for patients in late stages of amyotrophic lateral sclerosis [81, 346].

**Brain-computer interfaces for rehabilitation** While BCIs were initially conceived as communication devices for the severely disabled, we have argued that they can also be used for stroke rehabilitation [183]. By combining a BCI with a seven degrees-of-freedom robotic arm, serving an exoskeleton, we developed a brain-controlled rehabilitation robot supporting patients with chronic stroke in self-regulation of sensorimotor brain rhythms [194]. The concept of brain-controlled rehabilitation robotics can be extended to systems that monitor patients’ learning progress to adapt the rehabilitation exercise in real-time [65].

More information: [https://ei.is.tuebingen.mpg.de/project/brain-computer-interfaces](https://ei.is.tuebingen.mpg.de/project/brain-computer-interfaces)
Learning and inference for neuroscience

Mammalian brains efficiently combine perception, decision making, and motor commands in hundreds of milliseconds. Machine learning can help understand these fascinating distributed information processing capabilities.

A first application of machine learning algorithms to neuroscience is the extraction of information from complex brain signals, focusing on their dynamical aspects. This includes the design of non-parametric statistical dependency measures for time series using an implicit mapping in a Reproducing Kernel Hilbert Space to capture complex non-linear dependencies between brain rhythms [441], as well as dictionary learning techniques that automatically identify the transient dynamical patterns in ongoing Local Field Potentials (LFP) of a given brain structure that have an impact on the activity of the whole brain (see Figure 1.15) [160].

A second objective is to infer causal statements about the organization of underlying neural mechanisms. One key application is the estimation of the direction and strength of information flow across brain networks. We first studied the communication between multiple locations of the visual cortex using an information theoretic measure of Granger causality, Transfer Entropy, computed between LFP signals during visual stimulation [224]. Our results suggest the presence of waves propagating along the cortical tissue along the direction of maximal flow of information, routing visual information across the visual cortex [34]. We also study novel causality principles proposed by our department as an alternative to the Granger causality framework widely used in neuroscience. In particular, we developed a new causal inference method for time series based on the postulate of Independence of Cause and Mechanism. We provided theoretical guarantees for this approach, which outperformed Granger causality in experimental LFP signals [353].

Causal inference techniques can also be used to assess which aspect of brain activity affects the behavioral outcome of an experiment. Causal terminology is often introduced in the interpretation of neuroimaging data without considering the empirical support for such statements. We investigated which causal statements are warranted and which ones are not supported by empirical evidence [46]. Our work provides the first comprehensive set of causal interpretation rules for neuroimaging results.

Beyond providing powerful data analysis techniques, machine learning can help understand brain function from a theoretical perspective. We investigated a standard model of neuronal adaptation – spike-timing dependent plasticity – in this way and showed it can be viewed as a stochastic gradient descent of a neuromodulatory reward function implementing a form of empirical risk minimization [95, 473]. These results help understand optimality properties of biological learning.

More information: https://ei.is.tuebingen.mpg.de/project/machine-learning-and-neuroscience
This project investigates human perception, combining psychophysical experiments and computational modeling. Currently we have four main research foci:

First, we develop methods to gain more information from psychophysical data. We characterized serial dependencies in behavioral responses, and introduced methods to correct for them [63]. We wrote a software package to perform Bayesian inference for the psychometric function for non-stationary data. In addition, we introduced the use of spatial point processes to characterize eye-movement fixation patterns [41, 99]. Finally, we continue to improve machine learning methods to uncover the critical features observers use in complex perceptual tasks [118, 148]) (c.f. [289] and Figure 1.16).

Our second focus is the development of an 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 several data sets simultaneously, using a single set of parameters [108].

Third, the project seeks to understand the fundamental questions of characterizing the computational principles underlying lightness perception, treating lightness perception as one example of how to rigorously study the visual mechanisms translating ambiguous retinal input into perceptually and psychologically relevant categories [22, 23, 117].

Finally, we are moving towards understanding perceptual causality. Recently there has been considerable progress in understanding causal inference by viewing it as a machine learning problem. This is relevant to perception since animals cannot operate based on the assumption of independent and identically distributed (iid) data but need to employ suitable inference methods that work under changing distributions in order to produce robust perception. We study causal models of perception in a simple binary classification setting where the perceiver needs to distinguish cause and effect, or forward and backward, a task we recently studied in a computer vision context [363].

More information: https://ei.is.tuebingen.mpg.de/project/psychophysics-and-computational-models-of-behaviour
1.3 Equipment

Computing Infrastructure

The desktop computing environment of the Department of Empirical Inference is based on Intel PCs, and currently uses the centrally managed operating systems Ubuntu Linux, Microsoft Windows, and Mac OS X. Silent PCs with no moving parts are used to provide the best possible work environment.

For data storage the department currently offers 80TB of online file system storage on two file servers, used for shared data as well as personal home directories. Backups are done through a tape library system, as well as a disk based snapshotting system for the most important data. Off-site storage from RZ Garching is offered for long-term archival of scientific data.

For numerical experiments the group has access to a high performance computing cluster, shared by the departments of the MPI for Intelligent Systems. The cluster nodes and their supporting infrastructure are maintained by a central scientific facility. The Empirical Inference department is using a 45% share of the central cluster as of now. The cluster comprises 32 rack-mounted multi-processor nodes based on the x86_64 architecture, running Ubuntu Linux. The nodes are connected via 10 GBit/s fiber-optic connections. In total, the system consists of 1536 Intel and AMD 64-bit CPU cores with 18.5 terabytes of memory, as well as 74 modern nVidia GPUs and 100 terabytes of fast distributed storage space. The most powerful machines have 64 CPU cores and 1 terabyte of memory.

Robot Learning Lab: High-Speed Robot Arms

The Robot Learning Lab of the Empirical Inference Department focuses on finding task-appropriate machine learning methods for acquiring and refining motor skills. Current objectives include learning high-speed compliant control, learning simple skills (e.g., ball-in-a-cup or paddling balls), and larger composite tasks such as table tennis. These goals require a unique and maintenance-intensive set-up.

For fast as well as compliant control, we have a unique, custom-made high-voltage version of a Barrett WAM robot arm specifically designed for high-speed control while having seven degrees of freedom. Torque-level access to the robot, back drivability, and little backlash enables the use of this platform for learning control experiments and as a haptic input device during imitation learning or robot task assistance. The robot control makes use of a four camera high-speed (200 Hz) vision setup developed for that purpose. All software is based on a real-time Linux operating system and on the robot programming framework SL.

A tendon-driven pneumatic artificial muscle...
Empirical Inference

1.3 Equipment

A robot arm is being developed in order to study antagonistically actuated joints, allowing for lightweight segments, incorporating strong muscles, and using co-contraction for compliant control. We aim at approaching performance as observed in humans by designing such a system and enabling safer applications of learning control approaches whilst extending the variety of possible trajectories. We aim to use this new robot to show that learning can be particularly beneficial when classical methods fail.

Brain-Imaging Equipment

The QuickAmp 136 (Figure 1.19, left) is a high input-impedance amplifier capable of measuring and recording up to 128 channels of electroencephalographic (EEG) data, as well as eye (EOG), muscle (EMG), and other physiological signals. It is used for Brain-Computer Interface research, where brain signals with low noise are desirable, high temporal and spatial resolution, and where EOG and EMG signals are also required for artifact control. As of 2011, we also have a 128-channel active electrode system manufactured by BrainProducts GmbH, substantially reducing the preparation time of experimental studies and providing better signal-to-noise ratios outside of well-controlled laboratory environments (Figure 1.19, right). We have further acquired the Armeo Power, a robotic exoskeleton manufactured by Hocoma AG in Volketswil, Switzerland. This exoskeleton enables us to both support and perturb movements of the upper extremities during motor learning, which enables us to study the neural basis of motor learning in healthy subjects and patient populations with movement disorders.

Computational Imaging

In computational imaging it is only of limited value to work on simulated images. Instead it is important to develop deblurring and denoising algorithms that work on real data. To be able to take such photographs for varying setups and with controlled distortions, we use several SLR and CCD cameras (including cooled and uncooled Canon 5D Mk. II, Canon 5DS R, Sony A7R Mk. II, Finger Lakes Instruments ML-16803) with various lenses. For quality assessment and quantitative comparisons we use an image quality analysis system (iQ Analyzer by Image Engineering). To generate images and image sequences with controlled camera shake, we also use a Stewart platform (hexapod robot, see Figure 1.20) that allows a high accuracy for repeated movements.

In addition to usual photographic lenses we also make use of telescopes (including a Meade 12” Schmidt-Cassegrain) that allow us to work on real-world astronomical image sequences affected by turbulence. Furthermore we employ a setup with two telescopic mounts (ASA DDM60PRO, Vixen Sphinx) to study and develop reinforcement learning algorithms to improve their tracking precision.
1.4 Awards & Honors

2015

Shai Ben-David and **Ruth Urner** Best Paper Award at NIPS 2015 Workshop on "Transfer and Multitask Learning: Trends and New Perspectives".

**Jakob Zscheischler**: Köppen Award for his excellent doctoral dissertation (in climate and Earth system research) done at the MPI for Intelligent Systems and the MPI for Biogeochemistry in Jena.

**Vinay Jayaram**: Second prize in the "T3 Award" - Thesis in 3 - category PhD students. Title of the presentation: "Why can’t we control computers with our minds yet?"

**Leonardo Casarsa**: Second prize in the "T3 Award" - Thesis in 3 - category Master students. Title of the presentation: "Can computers treat sick brains?"

**Matej Balog**: Winner of Code Hunt programming contest at Microsoft Research PhD Summer School 2015

**Jakob Zscheischler**: Otto-Hahn-Medal for his PhD Thesis "A global analysis of extreme events and consequences for the terrestrial carbon cycle", performed at the MPI for Intelligent Systems, Tübingen and the MPI for Biogeochemistry, Jena.

**Katharina Muelling**: Award for Extraordinary Scientific Achievement for her Ph.D. thesis by the "Freunde der TU Darmstadt".

**Oliver Kroemer**: Runner Up for the 2015 Georges Giralt Award (Best European Robotics Ph.D. Thesis Award)

**Jan Peters**: ERC Starting Grant (2015-2020) for Research on Movements of Humanoid Robotics

**Philipp Hennig**: Emmy-Noether-Program "Probabilistic Numerics - Probabilistic Programming for Autonomous Systems"

2014

**Carl-Johann Simon-Gabriel**: Google European Doctorate Fellowship 2014-2016.


O. Kroemer (TU Darmstadt), H. van Hoof (TU Darmstadt), G. Neumann (TU Darmstadt), J. Peters: IROS Best Cognitive Robotics Paper Award Finalist

A. Abdolmaleki (Universidade de Aveiro), N. Lau (Universidade de Aveiro), G. Neumann (TU Darmstadt), J. Peters: 1st Place at the 3D Free Challenge of the RoboCup 2014.

**Sebastian Weichwald, Timm Meyer, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup**: best student paper award at the 4th International Workshop on Cognitive Information Processing (CIP 2014) for: "Decoding Index Finger Position From EEG Using Random Forests".

**Moritz Grosse-Wentrup and Daniel Braun**: Teaching Award Winter & Summer Term 2013/14 at Graduate School of Neural Information Processing, University of Tübingen

**Bernhard Schölkopf**: Royal Society Milner Award 2014
2013

**Manuel Gomez-Rodriguez**: outstanding paper award at NIPS 2013 for the paper: "Scalable Influence Estimation in Continuous-Time Diffusion Networks"

**Jan Peters**: Young Investigator Award of the International Neural Networks Society (INNS) in 2013.


**Jonas Peters**: ETH Silver Medal for his PhD thesis "Restricted Structural Equation Models for Causal Inference".

**Thomas Schultz**: MRM Distinguished Reviewer Award.

2012

**Ulrike von Luxburg**: Heisenberg Professorship


**Jan Peters**: Frontiers of Artificial Intelligence speaker at European Conference on Artificial Intelligence (ECAI) in 2012.

Christian Daniel, Gerhard Neumann, and **Jan Peters**: IROS CoTeSys Cognitive Robotics Best Paper Award, IROS 2012 Best Student Paper Award Finalist, and IROS 2012 Best Paper Award Finalist for "Learning Concurrent Motor Skills in Versatile Solution Spaces" at the International Conference on Intelligent Robot Systems (IROS)

**Stefan Harmeling**: Günter Petzow Prize of the Max-Planck-Institute for Intelligent Systems

**Bernhard Schölkopf**: Academy Prize 2012 of the Berlin-Brandenburg Academy of Sciences and Humanities

2011

**Moritz Grosse-Wentrup** and **Bernhard Schölkopf**: Annual BCI Research Award 2011 for the project "What are the neurophysiological causes of performance variations in brain-computer interfacing?"

**Thorsten Zander**: Willumeit award for his Ph.D. thesis entitled "Utilizing Brain-Computer Interfaces for Human-Machine Systems".

**Harold Christopher Burger**, and **Stefan Harmeling**: DAGM 2011 Prize for "Improving Denoising Algorithms via a Multi-Scale Meta-Procedure".

**Suvrit Sra**: Best Paper Runner up Award in Data Mining at The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2011 for "Fast Projections onto L1,q-Norm Balls for Grouped Feature Selection".


Gustavo Camps-Valls, Joris Mooij, and Bernhard Schölkopf: were selected for the 'IEEE Geoscience and Remote Sensing Society 2011 Letters’ Prize Paper Award for "Remote Sensing Feature Selection by Kernel Dependence estimation".

Bernhard Schölkopf: Annual Max Planck Research Award (shared with Sebastian Thrun, Stanford)

2010

Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause: Best Paper Award Honorable Mention at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD’2010) for the paper: "Inferring Networks of Diffusion and Influence."

Joris Mooij: Winner of two categories of the 2010 UAI Approximate Inference Challenge, 26th Conference on Uncertainty in Artificial Intelligence

Bernhard Schölkopf: Inclusion in the list of 'ISI Highly Cited Researchers’

Hirotaka Hachiya, Jan Peters, Mashashi Sugiyama: Young Researchers Award from IEEE Computational Intelligence Society of the Japan Chapter for their paper “Adaptive Importance Sampling with Automatic Model Selection in Reward Weighted Regression” presented at the Workshop of IEEE CIS Technical Committee on Neurocomputing (March 9-11, 2010).

Carl Edward Rasmussen and Chris Williams: 2009 DeGroot Prize for textbooks or monographs concerned with fundamental issues of statistical inference, decision theory and/or statistical applications, and chosen based on their novelty, thoroughness, timeliness, and importance of their intellectual scope: Gaussian Processes for Machine Learning, MIT Press 2006.

Oliver Kroemer, Renaud Detry, Justus Piater, Jan Peters: Best Paper Award of the 7th International Conference on Informatics in Control, Automation and Robotics (ICINCO) for: "Grasping with Vision Descriptors and Motor Primitives".

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, Bernhard Schölkof: Best student paper award for: "Inferring Deterministic Causal Relations", 26th Conference on Uncertainty in Artificial Intelligence (UAI)

Jan Peters, Jun Morimoto, Nick Roy, Jun Nakanishi: Most Active Technical Committee Award of the IEEE Robotics and Automation Society at ICRA 2010

José M. Leiva, Suzanne M.M. Martens: winning team of the Machine Learning for Signal Processing (MLSP) 2010 Competition: Mind Reading

2009

Reshad Hosseini, Matthias Bethge: prize winning method for the highest performance in the main challenge of GREAT08 PASCAL competition (Gravitational Lensing and Accuracy Testing 2008)


Christian Walder, Martin Breidt, Heinrich Bülthoff, Bernhard Schölkopf, Cristobal Curio: Markerless 3D Face Tracking, Runner up prize of the conference of the German Association for Pattern Recognition (DAGM)

Michael Hirsch, Stefan Harmeling, Bernhard Schölkopf: top score in ‘low noise’ GREAT08 PASCAL competition (Gravitational Lensing and Accuracy Testing 2008)

Stefan Harmeling, Michael Hirsch, Suvrit Sra, Bernhard Schölkopf: Best Poster Award for "Online Blind Deconvolution with Super-Resolution & Saturation Correction", at the 1st International Conference on Cosmology and Statistics (CosmoStats09)

Matthias Hofmann, Florian Steinke, Verena Scheel, Guillaume Charpiat, Jason Farquhar, Philip Aschoff, Michael Brady, Bernhard Schölkopf and Bernd J. Pichler: MRI-Based Attenuation Correction for PET/MRI: A Novel Approach Combining Pattern Recognition and Atlas Registration, Best Paper Awards 2008 of The Journal of Nuclear Medicine
1.5 Director profile: Bernhard Schölkopf

Bernhard Schölkopf studied Physics, Mathematics and Philosophy in Tübingen and London. In 1994 he joined Bell Labs to work on a Ph.D. with Vladimir Vapnik. Following researcher positions at GMD, Microsoft Research, and a biotech startup, Schölkopf started his lab at the Max Planck Institute for Biological Cybernetics (Tübingen) in 2002. In 2011, he became a founding director of the Max Planck Institute for Intelligent Systems.

Bernhard Schölkopf has been program chair of NIPS and COLT and is currently co-editor-in-chief of the flagship journal in machine learning (JMLR). He has been elected to the boards of the NIPS foundation and of the International Machine Learning Society. With Alex Smola, he initiated the Machine Learning Summer Schools series in 2002, which has meanwhile been organized, by various teams, 30 times. Many of his past students and postdocs have gone into academia (around 30 tenured or tenure-track positions) as well as to R&D labs (around 25 tenured), and he is one of the most highly cited researchers in Computer Science worldwide.4

Appointments

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015 – present</td>
<td>Co-Director of the Max Planck ETH Center for Learning Systems</td>
</tr>
<tr>
<td>2015 – present</td>
<td>Member of the Max Planck Campus Triumvirat, Tübingen</td>
</tr>
<tr>
<td>2011 – 2013</td>
<td>Managing Director, MPI for Intelligent Systems, Stuttgart &amp; Tübingen</td>
</tr>
<tr>
<td>2011 – present</td>
<td>Director, Max Planck Institute for Intelligent Systems</td>
</tr>
<tr>
<td>2012 – present</td>
<td>Guest Professor, ETH Zürich, Computer Science Department</td>
</tr>
<tr>
<td>2010 – present</td>
<td>Honorarprofessor, Tübingen University, Department of Mathematics and Physics</td>
</tr>
<tr>
<td>2002 – present</td>
<td>Honorarprofessor, Technical University Berlin, Department of Computer Science</td>
</tr>
</tbody>
</table>

Awards & Honors (2009 – 2015)

<table>
<thead>
<tr>
<th>Year</th>
<th>Award</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Overseas Visiting Scholarship, St. John’s College, Cambridge, UK</td>
</tr>
<tr>
<td>2014</td>
<td>Royal Society Milner Award, London, UK</td>
</tr>
<tr>
<td>2013</td>
<td>XXVIIIth Courant Lectures, New York University</td>
</tr>
<tr>
<td>2012</td>
<td>Academy Prize, Berlin-Brandenburg Academy of Sciences and Humanities</td>
</tr>
</tbody>
</table>

4http://web.cs.ucla.edu/~palsberg/h-number.html
Current Memberships

SimTECH (Excellence Cluster for Simulation Technology, Stuttgart University); Center for Integrative Neuroscience (Excellence Cluster, Tübingen University); Bernstein Focus for Neurotechnology (Freiburg/Tübingen); Bernstein Center for Computational Neuroscience (BCCN, Tübingen); German Association for Pattern Recognition (DAGM); Deutsche Mathematiker Vereinigung (DMV); Association for Computing Machinery (ACM); Institute of Electrical and Electronics Engineers (IEEE, Senior Member); European Academy of Sciences and Arts.

Committees, Service, Board Memberships

Co-founder of DALI – Data, Learning, and Inference (2015 –); Core Committee Member, MPI for Biological Cybernetics (2015 –); The Future of AI – A New York University Symposium on Science, Technology, Reason and Ethics (2016); ACM Heidelberg Laureate Forum Committee (2014–), Section Panel for Mathematics in Science and Technology, International Congress of Mathematicians (ICM) 2014; General Chair of the International Conference on Artificial Intelligence and Statistics (AISTATS) 2012; Initial Training Network for Machine Learning for Personalized Medicine (MLPM); PASCAL/PASCAL2 EU Network of Excellence; Forum Scientiarum at the University of Tübingen; Snowbird Learning workshop; Machine Learning Summer Schools (MLSS); International Machine Learning Society (IMLS); Neural Information Processing Systems Foundation (NIPS).

Review Panels

Member of the pool of experts in the process to establish the Alan Turing Institute; Centre for Doctoral Training (CDT) in Data Science, Edinburgh, UK: Italian Institute of Technology (IIT); Computer Science Department, École Normale Supérieure, Paris; Neural Computation and Adaptive Perception Program of the Canadian Institute of Advanced Research (Chair); Gatsby Computational Neuroscience Unit, Quinquennial Review 2010; Machine Learning Program, NICTA (Sydney).

Editorial Board


Invited Talks 2009 – 2015 (Selection)

1.6 Publications

1.6.1 Books

2014


2013


2011


2010


1.6.2 Proceedings

2013


2012


2011


2010

1.6.3 Journal Articles

2016


2015


1 Empirical Inference
1.6 Publications
1.6.3 Journal Articles


2014


1.6 Publications

1.6.3 Journal Articles


2013


1.6 Publications

1.6.3 Journal Articles


1.6 Publications

1.6.3 Journal Articles


1 Empirical Inference
1.6 Publications
1.6.3 Journal Articles


2010


<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
<th>Title</th>
<th>Journal/Citation</th>
</tr>
</thead>
</table>


1.6.3 Journal Articles


2009


1.6.4 Conference Papers

2016


2015


1.6 Publications

1.6.4 Conference Papers


1.6 Publications

1.6.4 Conference Papers


1.6 Publications

1.6.4 Conference Papers


2014


1.6 Publications

1.6.4 Conference Papers


2013


1 Empirical Inference
1.6 Publications
1.6.4 Conference Papers


2012


1.6 Publications

1.6.4 Conference Papers


2011


1.6.4 Conference Papers


1.6.4 Conference Papers


1.6.4 Conference Papers


2009


1.6 Publications

1.6.4 Conference Papers


1.6 Publications

1.6.4 Conference Papers


1.6.5 Book Chapters

2015


2014


2013


2012


1 Empirical Inference

1.6 Publications

1.6.5 Book Chapters


2011


2010


2009


### 1.6.6 Theses

#### PhD Theses


1 Empirical Inference

1.6 Publications

1.6.6 Theses


Master Theses


**Diploma Theses**


