The Alan Turing Institute

Statistics and computation

Abstracts

Implicit regularization for general norms and errors
Lorenzo Rosasco, Massachusetts Institute of Technology

Implicit regularization refers to the property of optimization methods to bias the search of solutions towards those with some small norm and ensure stability of the estimation process. While this idea is classic when considering Euclidean norms and quadratic error, much less is known for more general choices. In this talk we will discuss several results in this direction with an emphasis on accelerated optimization techniques.

Can learning theory resist deep learning?
Francis Bach, INRIA

Machine learning algorithms are ubiquitous in most scientific, industrial and personal domains, with many successful applications. As a scientific field, machine learning has always been characterized by the constant exchanges between theory and practice, with a stream of algorithms that exhibit both good empirical performance on real-world problems and some form of theoretical guarantees. Many of the recent and well publicized applications come from deep learning, where these exchanges are harder to make, in part because the objective functions used to train neural networks are not convex. In this talk, I will present recent results on the global convergence of gradient descent for some specific non-convex optimization problems, illustrating these difficulties and the associated pitfalls (joint work with Lénaïc Chizat and Edouard Oyallon)

A function space view of overparameterized neural networks
Rebecca Willet, University of Chicago

Contrary to classical bias/variance tradeoffs, deep learning practitioners have observed that vastly overparameterized neural networks with the capacity to fit virtually any labels nevertheless generalize well when trained on real data. One possible explanation of this phenomenon is that complexity control is being achieved by implicitly or explicitly controlling the magnitude of the weights of the network. This raises the question: What functions are well-approximated by neural networks whose weights are bounded in norm? In this talk, I will give some partial answers to this question. In particular, I will give a precise characterization of the space of functions realizable as a two-layer (i.e., one hidden layer) neural network with ReLU activations having an unbounded number of units, but where the Euclidean norm of the weights in the network remains bounded. Surprisingly, this characterization is naturally posed in terms of the Radon transform as used in computational imaging, and I will show how tools from Radon transform analysis yield novel insights about learning with two and three-layer ReLU networks. This is joint work with Greg Ongie, Daniel Soudry, and Nati Srebro.

The Alan Turing Institute

Benign overfitting
Peter Bartlett, University of California, Berkley

Classical theory that guides the design of nonparametric prediction methods like deep neural networks involves a tradeoff between the fit to the training data and the complexity of the prediction rule. Deep learning seems to operate outside the regime where these results are informative, since deep networks can perform well even with a perfect fit to noisy training data. We investigate this phenomenon of 'benign overfitting' in the simplest setting, that of linear prediction. We give a characterization of linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of effective rank of the data covariance. It shows that overparameterization is essential: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size. It also shows an important role for finite-dimensional data: benign overfitting occurs for a much narrower range of properties of the data distribution when the data lies in an infinite dimensional space versus when it lies in a finite dimensional space whose dimension grows faster than the sample size. We discuss implications for deep networks and for robustness to adversarial examples. Joint work with Phil Long, Gábor Lugosi, and Alex Tsigler.

Fast and optimal low-rank tensor regression via importance
Garvesh Raskutti, University of Wisconsin-Madison

In this talk, I present a procedure for low-rank tensor regression using importance sketching, namely Importance Sketching Low-rank Estimation for Tensors (ISLET). The central idea behind ISLET is carefully designed sketches based on both the responses and low-dimensional structure of the parameter of interest. Importantly, the proposed method is sharply minimax optimal in terms of the mean-squared error under low-rank Tucker assumptions and under randomized Gaussian ensemble design and provides significant computational advantages over existing methods. I demonstrate through numerical studies the computational advantages of our method. In particular our approach can perform reliable estimation for tensors up to dimension p = O(10^8) and is several orders of magnitude faster than baseline methods.

Big data is low rank
Madeleine Udell, Cornell University

Matrices of low rank are pervasive in big data, appearing in recommender systems, movie preferences, topic models, medical records, and genomics. While there is a vast literature on how to exploit low rank structure in these datasets, there is less attention on explaining why low rank structure appears in the first place. In this talk, we explain the abundance of low rank matrices in big data by proving that certain latent variable models associated to piecewise analytic functions are of log-rank. Any large matrix from such a latent variable model can be approximated, up to a small error, by a low rank matrix. Armed with this theorem, we show how to use a low rank modeling framework to exploit low rank structure even for datasets that are not numeric, with applications in the social sciences, medicine, and automated machine learning.

The Alan Turing Institute

Data-driven regularisation for solving inverse problems
Carola-Bibiane Schönlieb, The Alan Turing Institute and University of Cambridge

Abstract: In this talk we discuss the idea of data- driven regularisers for inverse imaging problems. We are in particular interested in the combination of model-based and purely data-driven image processing approaches. In this context we will make a journey from “shallow” learning for computing optimal parameters for variational regularisation models by bilevel optimization to the investigation of different approaches that use deep neural networks for solving inverse imaging problems. This talk is based on a 2019 Acta Numerica paper written together with Simon Arridge, Peter Maass and Ozan Öktem.

Statistical Physics and Learning
Florent Krzakala, Sorbonne Université and Ecole Normale Superieure

Learning from ranks, learning to rank
Jean-Philippe Vert, Google Brain and Mines ParisTech

Permutations and sorting operators are ubiquitous in data science, e.g., when one wants to analyze or predict preferences. As discrete combinatorial objects, permutations do not lend themselves easily to differential calculus, which underpins much of modern machine learning. In this talk I will present several approaches to embed permutations to a continuous space, on the one hand, and to relax the ranking operator to be differentiable, on the other hand, in order to integrate permutations, sorting and ranking operators in differentiable architecture for machine learning.

Approximate cross validation for large data and high dimensions
Tamara Broderick, Massachusetts Institute of Technology

The error or variability of statistical and machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data. The ubiquitous tools of cross-validation (CV) and the bootstrap are examples of this technique. These methods are powerful in large part due to their model agnosticism but can be slow to run on modern, large data sets due to the need to repeatedly re-fit the model. We use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster than repeated re-fitting by orders of magnitude. This linear approximation is sometimes known as the "infinitesimal jackknife" (IJ) in the statistics literature, where it has mostly been used as a theoretical tool to prove asymptotic results. We provide explicit finite-sample error bounds for the infinitesimal jackknife in terms of a small number of simple, verifiable assumptions. Without further modification, though, we note that the IJ deteriorates in accuracy in high dimensions and incurs a running time roughly cubic in dimension. We additionally show, then, how dimensionality reduction can be used to successfully run the IJ in high dimensions in the case of leave-one-out cross validation (LOOCV). Specifically, we consider L1 regularization for generalized linear models. We prove that, under mild conditions, the resulting LOOCV approximation exhibits computation time and accuracy that depend on the (small) recovered support size rather than the full dimension. Simulated and real-data experiments support our theory.

The Alan Turing Institute

From causal inference to autoencoders, memorization and gene regulation
Caroline Uhler, Massachusetts Institute of Technology

Recent progress in genomics makes it possible to perform perturbation experiments at a very large scale. This motivates the development of a causal inference framework that is based on observational and interventional data. We characterize the causal relationships that are identifiable and present the first provably consistent algorithm for learning a causal network from such data. I will then couple gene expression with the 3D genome organization. In particular, we will discuss approaches for integrating different data modalities such as sequencing or imaging via autoencoders. We end by a theoretical analysis of autoencoders linking overparameterization to memorization. In particular, we will show that overparameterized autoencoders trained using standard optimization methods implement associative memory and provide a mechanism for memorization and retrieval of real-valued data.

Does learning require memorization? A short tale about a long tail
Vitaly Feldman, Google Research, Brain Team

Learning algorithms based on deep neural networks are well-known to (nearly) perfectly fit the training set and fit well even the random labels. This tendency to memorize the labels of the training data is not explained by existing theoretical analyses. Memorization of the training data also presents significant privacy risks when the training data contains sensitive personal information. We provide a simple conceptual explanation and a theoretical model demonstrating that for natural data distributions memorization of labels is necessary for achieving close-to-optimal generalization error. We complement the theoretical results with experiments on several standard benchmarks showing that memorization is an essential part of deep learning. Based in part on an ongoing work with Chiyuan Zhang.