Big Data 2023

Thursday, August 31, 2023

9:00 am	Breakfast
9:30 am	Introductions
9:45–10:45 am	**Albert-László Barabási (Northeastern, Harvard)** **Title:** From Network Medicine to the Foodome: The Dark Matter of Nutrition **Abstract:** A disease is rarely a consequence of an abnormality in a single gene but reflects perturbations to the complex intracellular network. Network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships between apparently distinct (patho) phenotypes. As an application, I will explore how we use network medicine to uncover the role individual food molecules in our health. Indeed, our current understanding of how diet affects our health is limited to the role of 150 key nutritional components systematically tracked by the USDA and other national databases in all foods. Yet, these nutritional components represent only a tiny fraction of the over 135,000 distinct, definable biochemicals present in our food. While many of these biochemicals have documented effects on health, they remain unquantified in any systematic fashion across different individual foods. Their invisibility to experimental, clinical, and epidemiological studies defines them as the 'Dark Matter of Nutrition.' I will speak about our efforts to develop a high-resolution library of this nutritional dark matter, and efforts to understand the role of these molecules on health, opening novel avenues by which to understand, avoid, and control disease.
10:45–11:00 am	Break
11:00 am –12:00 pm	**Rachel Cummings (Columbia)** **Title:** Differentially Private Algorithms for Statistical Estimation Problems **Abstract:** Differential privacy (DP) is widely regarded as a gold standard for privacy-preserving computation over users' data. It is a parameterized notion of database privacy that gives a rigorous worst-case bound on the information that can be learned about any one individual from the result of a data analysis task. Algorithmically it is achieved by injecting carefully calibrated randomness into the analysis to balance privacy protections with accuracy of the results. In this talk, we will survey recent developments in the development of DP

	algorithms for three important statistical problems, namely online learning with bandit feedback, causal interference, and learning from imbalanced data. For the first problem, we will show that Thompson sampling — a standard bandit algorithm developed in the 1930s — already satisfies DP due to the inherent randomness of the algorithm. For the second problem of causal inference and counterfactual estimation, we develop the first DP algorithms for synthetic control, which has been used non-privately for this task for decades. Finally, for the problem of imbalanced learning, where one class is severely underrepresented in the training data, we show that combining existing techniques such as minority oversampling perform very poorly when applied as pre-processing before a DP learning algorithm; instead we propose novel approaches for privately generating synthetic minority points. Based on joint works with Marco Avella Medina, Vishal Misra, Yuliia Lut, Tingting Ou, Saeyoung Rho, and Ethan Turok.
12:00–1:30 pm	Lunch
1:30–2:30 pm	**Morgane Austern (Harvard)** **Title:** To split or not to split that is the question: From cross validation to debiased machine learning **Abstract:** Data splitting is a ubiquitous method in statistics with examples ranging from cross-validation to cross-fitting. However, despite its prevalence, theoretical guidance regarding its use is still lacking. In this talk, we will explore two examples and establish an asymptotic theory for it. In the first part of this talk, we study the cross-validation method, a ubiquitous method for risk estimation, and establish its asymptotic properties for a large class of models and with an arbitrary number of folds. Under stability conditions, we establish a central limit theorem and Berry-Esseen bounds for the cross-validated risk, which enable us to compute asymptotically accurate confidence intervals. Using our results, we study the statistical speed-up offered by cross-validation compared to a train-test split procedure. We reveal some surprising behavior of the cross-validated risk and establish the statistically optimal choice for the number of folds. In the second part of this talk, we study the role of cross-fitting in the generalized method of moments with moments that also depend on some auxiliary functions. Recent lines of work show how one can use generic machine learning estimators for these auxiliary problems, while maintaining asymptotic normality and root-n consistency of the target parameter of interest. The literature typically requires that these auxiliary problems are fitted on a separate sample or in a cross-fitting manner. We show that when these auxiliary estimation algorithms satisfy natural leave-one-out stability properties, then sample splitting is not required. This allows for sample reuse, which can be beneficial in moderately sized sample regimes.

2:30–2:45 pm	Break
2:45–3:45 pm	**Ankur Moitra (MIT)** **Title:** Learning from Dynamics **Abstract:** Linear dynamical systems are the canonical model for time series data. They have wide-ranging applications and there is a vast literature on learning their parameters from input-output sequences. Moreover they have received renewed interest because of their connections to recurrent neural networks. But there are wide gaps in our understanding. Existing works have only asymptotic guarantees or else make restrictive assumptions, e.g. that preclude having any long-range correlations. In this work, we give a new algorithm based on the method of moments that is computationally efficient and works under essentially minimal assumptions. Our work points to several missed connections, whereby tools from theoretical machine learning including tensor methods, can be used in non-stationary settings.
3:45–4:00 pm	Break
4:00–5:00 pm	**Mark Sellke (Harvard)** **Title:** Algorithmic Thresholds for Spherical Spin Glasses **Abstract:** High-dimensional optimization plays a crucial role in modern statistics and machine learning. I will present recent progress on non-convex optimization problems with random objectives, focusing on the spherical p-spin glass. This model is related to spiked tensor estimation and has been studied in probability and physics for decades. We will see that a natural class of “stable” optimization algorithms gets stuck at an algorithmic threshold related to geometric properties of the landscape. The algorithmic threshold value is efficiently attained via Langevin dynamics or by a second-order ascent method of Subag. Much of this picture extends to other models, such as random constraint satisfaction problems at high clause density.

Friday, September 1, 2023

9:00 am	Breakfast
9:30 am	Introductions
9:45–10:45 am	**Jacob Andreas (MIT)** **Title:** What Learning Algorithm is In-Context Learning? **Abstract:** Neural sequence models, especially transformers, exhibit a remarkable capacity for "in-context" learning. They can construct new predictors from sequences of labeled examples (x,f(x)) presented in the input without further parameter updates. I'll present recent findings suggesting that transformer-based in-context learners implement standard learning algorithms implicitly, by encoding smaller models in their activations, and updating these implicit models as new examples appear in the context, using in-context linear regression as a model problem. First, I'll show by construction that transformers can implement learning algorithms for linear models based on gradient descent and closed-form ridge regression. Second, I'll show that trained in-context learners closely match the predictors computed by gradient descent, ridge regression, and exact least-squares regression, transitioning between different predictors as transformer depth and dataset noise vary, and converging to Bayesian estimators for large widths and depths. Finally, we present preliminary evidence that in-context learners share algorithmic features with these predictors: learners' late layers non-linearly encode weight vectors and moment matrices. These results suggest that in-context learning is understandable in algorithmic terms, and that (at least in the linear case) learners may rediscover standard estimation algorithms. This work is joint with Ekin Akyürek at MIT, and Dale Schuurmans, Tengyu Ma and Denny Zhou at Stanford.
10:45–11:00 am	Break
11:00 am–12:00 pm	**Tommi Jaakkola (MIT)** **Title:** Generative modeling and physical processes **Abstract:** Rapidly advancing deep distributional modeling techniques offer a number of opportunities for complex generative tasks, from natural sciences such as molecules and materials to engineering. I will discuss generative approaches inspired from physical processes including diffusion models and more recent electrostatic models (Poisson flow), and how they relate to each other in terms of embedding dimension. From the point of view of applications, I will highlight our recent work on SE(3) invariant distributional modeling over backbone 3D structures with ability to generate designable monomers without relying on pre-trained protein structure prediction methods as well as state of the art image generation capabilities (Poisson flow). Time permitting, I will also discuss recent analysis of efficiency of sample generation in such models.
12:00–1:30 pm	Lunch

1:30–2:30 pm	**Marinka Zitnik (Harvard Medical School)** **Title:** Multimodal Learning on Graphs **Abstract:** Understanding biological and natural systems requires modeling data with underlying geometric relationships across scales and modalities such as biological sequences, chemical constraints, and graphs of 3D spatial or biological interactions. I will discuss unique challenges for learning from multimodal datasets that are due to varying inductive biases across modalities and the potential absence of explicit graphs in the input. I will describe a framework for structure-inducing pretraining that allows for a comprehensive study of how relational structure can be induced in pretrained language models. We use the framework to explore new graph pretraining objectives that impose relational structure in the induced latent spaces—i.e., pretraining objectives that explicitly impose structural constraints on the distance or geometry of pretrained models. Applications in genomic medicine and therapeutic science will be discussed. These include TxGNN, an AI model enabling zero-shot prediction of therapeutic use across over 17,000 diseases, and PINNACLE, a contextual graph AI model dynamically adjusting its outputs to contexts in which it operates. PINNACLE enhances 3D protein structure representations and predicts the effects of drugs at single-cell resolution.
2:30–2:45 pm	Break
2:45–3:45 pm	**Jianqing Fan (Princeton)** **Title:** UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation **Abstract:** Uncertainty quantification for prediction is an intriguing problem with significant applications in various fields, such as biomedical science, economic studies, and weather forecasts. Numerous methods are available for constructing prediction intervals, such as quantile regression and conformal predictions, among others. Nevertheless, model misspecification (especially in high-dimension) or sub-optimal constructions can frequently result in biased or unnecessarily wide prediction intervals. In this work, we propose a novel and widely applicable technique for aggregating multiple prediction intervals to minimize the average width of the prediction band along with coverage guarantee, called Universally Trainable Optimal Predictive Intervals Aggregation (UTOPIA). The method also allows us to directly construct predictive bands based on elementary basis functions. Our approach is based on linear or convex programming which is easy to implement. All of our proposed methodologies are supported by theoretical guarantees on the coverage probability and optimal average length, which are detailed in this paper. The effectiveness of our approach is convincingly demonstrated by applying it to synthetic data and two real datasets on finance and macroeconomics. (Joint work Jiawei Ge and Debarghya Mukherjee).

3:45–4:00 pm	Break
4:00–5:00 pm	**Melissa Dell (Harvard)** **Title:** Efficient OCR for Building a Diverse Digital History **Abstract:** Many users consult digital archives daily, but the information they can access is unrepresentative of the diversity of documentary history. The sequence-to-sequence architecture typically used for optical character recognition (OCR) – which jointly learns a vision and language model - is poorly extensible to low-resource document collections, as learning a language-vision model requires extensive labeled sequences and compute. This study models OCR as a character level image retrieval problem, using a contrastively trained vision encoder. Because the model only learns characters' visual features, it is more sample efficient and extensible than existing architectures, enabling accurate OCR in settings where existing solutions fail. Crucially, it opens new avenues for community engagement in making digital history more representative of documentary history.