samsi
NSF·Duke·NCSU·UNC

Deep Learning Program Opening Workshop
August 12-16, 2019

SPEAKER TITLES/ABSTRACTS

Amitabh Basu
Johns Hopkins University

"Admissibility of Solution Estimators in Stochastic Optimization"

We consider stochastic optimization problems arising in deep learning and other areas of statistical and machine learning from a statistical decision theory perspective. In particular, we investigate the admissibility (in the sense of decision theory) of the sample average solution estimator. We show that this estimator can be inadmissible in very simple settings, a phenomenon that is derived from the classical James-Stein estimator. However, for many problems of interest, the sample average estimator is indeed admissible. We will end with several open questions in this research direction.

Mikhail Belkin
Ohio State University

"Fit without Fear: from classical statistics to modern machine learning"

"A model with zero training error is overfit to the training data and will typically generalize poorly" goes statistical textbook wisdom. Yet, in modern practice, over-parametrized deep networks with near perfect fit on training data still show excellent test performance. As I will discuss in my talk, this apparent contradiction is key to understanding modern machine learning.
While classical methods rely on the bias-variance trade-off where the complexity of a predictor is balanced with the training error, "modern" models are best described by interpolation, where a predictor is chosen among functions that fit the training data exactly, according to a certain inductive bias. Furthermore, classical and modern models can be unified within a single "double descent" risk curve, which extends the usual U-shaped bias-variance trade-off curve beyond the point of interpolation. This understanding of model performance delineates the limits of classical analyses and opens new lines of inquiry into computational,
statistical, and mathematical properties of models. A number of implications for model selection with respect to generalization and optimization will be discussed.

Anindya Bhadra
Purdue University

"Horseshoe Regularization for Machine Learning in Complex and Deep Models"

Since the advent of the horseshoe priors for regularization, global-local shrinkage methods have proved to be a fertile ground for the development of Bayesian theory and methodology in machine learning. They have achieved remarkable success in computation, and enjoy strong theoretical support. Much of the existing literature has focused on the linear Gaussian case. The purpose of the current talk is to demonstrate that the horseshoe priors are useful more broadly, by reviewing both methodological and computational developments in complex models that are more relevant to machine learning applications. Specifically, we focus on methodological challenges in horseshoe regularization in nonlinear and non-Gaussian models; multivariate models; and deep neural networks. We also outline the recent computational developments in horseshoe shrinkage for complex models along with a list of available software implementations that allows one to venture out beyond the comfort zone of the canonical linear regression problems.

Wyatt Bridgman & Sorin Mitran
University of North Carolina-Chapel Hill

"Information Geometric and Topological Approaches to Deep Learning"

Deep learning algorithms determine a high-dimensional data approximation through nonlinear optimization of a multiply-composite function (neural net layers). This talk presents a different viewpoint based upon interpretation of available data as a sample of some unknown statistical distribution. Successive data samples provide different approximations of the underlying distribution, interpreted as a point in the space of probability density functions (PDFs). The main object of interest becomes the support of the PDF points. The intrinsic dimension of this support is determined by persistent homology techniques from computational topology. A manifold of the resulting dimensionality is subsequently constructed and geodesic transport on this manifold in the Fisher metric is used to determine reduced models of lower dimensionality. Applications of the procedure to the problem of "learning" the behavior of complex physical systems are presented including protein dynamics and cellular motility.

Lawrence Carin
Duke University

"On Adversarial Learning"

Adversarial learning is examined from a statistical perspective, with connections made to traditional learning methods. Initial adversarial techniques have been developed under the assumption that data samples are available for learning to synthesize realistic samples. We extend this concept to learning to sample from an unnormalized distribution, based on adversarial methods. In addition to presenting the methods, several practical examples of the method are presented.

Shih-Kang Chao
University of Missouri

"Training DNN with Dynamic SMD"

Stochastic gradient descent (SGD) is a popular algorithm that can handle extremely large data sets due to its low computational cost at each iteration and low memory requirement. However, a major drawback of SGD is that it does not adapt well to the underlying structure of the solution, such as sparsity. Many variations of SGD have been developed based on the concept of stochastic mirror descent (SMD). In this paper, we develop diffusion approximation to the "dynamic" SMD with constant step size, using the local Bregman divergence. The dynamic SMD allows the regularizer of SMD to vary with time. The diffusion approximation results shed light on how to fine-tune an l1-norm based SMD algorithm which zeros inactive coefficients without suffering from bias. Numerical analysis with sparse principal component analysis and neural networks will be shown.

Guang Cheng
Purdue University

"Modern Statistical Theory Inspired by Deep Learning"

Modern learning algorithms, such as deep learning, have gained great successes in real applications. However, some of their empirical behaviors may not be interpreted within the classical statistical learning framework. For example, deep learning algorithms achieve small testing error even when the training error is zero, i.e., over-fitting. Another phenomenon is observed in image recognition applications where a hardly noticeable change of data may lead to dramatic increase of mis-classification rates. Inspired by these observations, we attempt to illustrate new theoretical insights for data-interpolation and adversarial samples using the very simple nearest neighbor algorithms. In particular, we prove statistical optimality of interpolated nearest neighbor algorithms. More surprisingly, it is discovered that the classification performance, under a proper interpolation, is even better that the best kNN in terms of multiplicative constant. As for adversarial samples, we demonstrate that different adversarial mechanisms lead to different phase transition phenomena of mis-classification rate in terms of its upper bound. If time allowed, some adversarial robust and adaptive algorithms will be introduced.

Xiuyuan Cheng
Duke University

"Group-equivariant Representation by Jointly Decomposed Convolution"

Explicit encoding of group actions in data representation is desired for convolutional neural networks (CNNs) to successfully handle global deformations in input signals. In this talk, we introduce group-equivariant deep CNNs where the convolutional filters are jointly decomposed over steerable bases on the space and the group geometry simultaneously. This decomposition significantly reduces the model size and computational complexity while preserving network performance, and it also serves to regularize the convolutional filters by the truncation of bases expansion. The stability of the equivariant representation with respect to input variations is proved theoretically and also demonstrated on computer vision tasks where the datasets involve in-plane and out-of-plane object rotations. The work provides a general approach to achieve group equivariant features in deep CNNs with representation stability and computational efficiency.

Bianca Dumitrascu
SAMSI/Princeton

"Domain Adaptation Challenges in Genomics: a deep learning take on medical pathology"

Medical pathology images are visually evaluated by experts for disease diagnosis, but the connection between image features and the state of the cells in an image is typically unknown. To understand this relationship, we describe a multimodal modeling and inference framework that estimates shared latent structure of joint gene expression levels and medical image features. The method is built around probabilistic canonical correlation analysis (PCCA), which is jointly fit to image embeddings that are learned using convolutional neural networks and linear embeddings of paired gene expression data. We finally discuss a set of theoretical and empirical challenges in domain adaptation settings arising from genomics data.

(based on work in collab with Gregory Gundersen and Barbara E. Engelhardt)

Jianqing Fan
Princeton University

"Towards Deep Learning: Understanding Statistical Properties by Bridging Convex and Nonconvex Optimization"

This talk is on understanding statistical rates of convergence and asymptotic normality on noisy low-rank matrix completion. One of the most popular paradigms to tackle this problem is convex relaxation, which achieves remarkable efficacy in practice. However, the theoretical support of this approach is still far from optimal in the noisy setting, falling short of explaining the empirical success. When the rank of the unknown matrix is a constant, we demonstrate that the convex programming approach achieves near-optimal estimation errors --- in terms of the Euclidean loss, the entrywise loss, and the spectral norm loss --- for a wide range of noise levels. We further establish the asymptotic normality for entries of matrix and its associaaed local rank factors. Remarkably, we unveil their strong oracle and adaptive properties. All of these are enabled by bridging convex relaxation with the nonconvex Burer--Monteiro approach, a seemingly distinct algorithmic paradigm that is provably robust against noise. More specifically, we show that an approximate critical point of the nonconvex formulation serves as an extremely tight approximation of the convex solution, allowing us to transfer the desired statistical properties of the nonconvex approach to its convex counterpart.

(Joint work with Yuxin Chen, Cong Ma and Yuling Yan)

Guanghui (George) Lan
Georgia Institute of Technology

"Optimization and Learning with Nonconvex Functional Constraints"

Nonconvex optimization is becoming more and more important in machine learning. In spite of recent progresses, the development of provably efficient algorithms for optimization with nonconvex functional constraints remains open. Such problems have potential applications in risk-averse learning and adversarial learning among others. In this talk, we introduce a new proximal point type method for solving this important class of nonconvex problems by transforming them into a sequence of convex constrained subproblems. We show both the convergence and rate of convergence of our algorithm to a first-order KKT point under different types of constraint qualifications. In particular, we prove that our algorithm will converge to an ϵ-KKT point in $O(1/\epsilon)$ iterations. For practical use, we present inexact variants of this approach, in which approximate solutions of the subproblems are computed by either primal or primal-dual type algorithms, and establish their associated rate of convergence. To the best of our knowledge, this is the first time that proximal point type method is developed for nonlinear programing with nonconvex functional constraints, and all the complexity results seem to be new. Preliminary numerical results will also be presented.

This is a joint work with Digvijay Boob and Qi Deng.

Jason Klusowski
Rutgers University

"Complexity Bounds for Deep Learning Networks via the Probabilistic Method"

It has been experimentally observed in recent years that multi-layer neural networks have a surprising ability to generalize, even when trained with far more parameters than observations. Is there a theoretical basis for this? As a partial answer, we show that the mean squared generalization error for multi-layer Lipschitz networks is of order $[(L^3 V^2 \log(d)) / n]^{1/2}$, where L is the number of layers, V is a complexity constant that coincides with the 1-norm of the path weights, d is the number of inputs per layer, and n is the sample size. The key idea is a probabilistic reformulation of any multi-layer Lipschitz network which, in turn, motivates a proof strategy based purely on the probabilistic method instead of the usual route via Rademacher analysis.

This is joint work with Andrew R. Barron from Yale University.

Faming Liang
Purdue University

"An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Global Optimization in Deep Learning"

We propose an adaptively weighted stochastic gradient MCMC algorithm for Bayesian learning. The proposed algorithm possesses a self-adjusting mechanism for escaping local traps; it is essentially immune to local traps and can converge quickly to global optimal solutions. Theoretically, we establish the convergence of the proposed algorithm and provide an upper bound for its hitting time for a wide class of non-convex functions. The proposed algorithm has a much smaller order of hitting time than the stochastic gradient Langevin dynamics algorithm and simulated annealing stochastic gradient Langevin dynamics algorithm. We test the proposed algorithm on multiple benchmark data sets including CIFAR10 and CIFAR100. The numerical results indicate the superiority of the proposed algorithm over the existing state-of-the-art algorithms in training deep neural networks.

Ruiqi Liu
Indiana University

"Deep Instrumental Variables Estimator"

Estimation based on deep neural nets has seen advantages when the underlying function demonstrates certain complicated structures. In this talk, I will discuss how to use this truth to make statistical inferences and achieve statistical efficiency.

Poh-Ling Loh
University of Wisconsin

"Robust Information Bottleneck"

We derive bounds for a notion of adversarial risk, designed to characterize the robustness of linear and neural network classifiers to adversarial perturbations. Specifically, we introduce a new class of function transformations with the property that the risk of the transformed functions upper-bounds the adversarial risk of the original functions. This reduces the problem of deriving bounds on the adversarial risk to the problem of deriving risk bounds using standard learning-theoretic techniques. We then derive bounds on the Rademacher complexities of the transformed function classes, obtaining error rates on the same order as the generalization error of the original function classes. We also discuss extensions of our theory to multiclass classification and regression. Finally, we provide two algorithms for optimizing the adversarial risk bounds in the linear case, and discuss connections to regularization and distributional robustness.

Andrew Zammit Mangion
University of Wollongong

"Deep Compositional Spatial Models"

Nonstationary, anisotropic spatial processes are often used when modelling, analysing and predicting complex environmental phenomena. One such class of processes considers a stationary, isotropic process on a warped spatial domain. The warping function is generally difficult to fit and often results in `space-folding.' Here, we propose modelling a class of parsimonious warping functions constructed through a composition of multiple elemental functions in a deep-learning framework. We consider two cases; first, when these functions are known up to some weights that need to be estimated, and, second, when the weights in each layer are random. Inspired by recent methodological and technological advances in deep learning and deep Gaussian processes, we employ approximate Bayesian methods to make inference with these models using graphical processing units. Through simulation studies in one and two dimensions we show that the deep compositional spatial models are quick to fit, and are able to provide better predictions and uncertainty quantification than other deep stochastic models of similar complexity. We also show their remarkable capacity to model highly nonstationary, anisotropic spatial data using radiances from the MODIS instrument aboard the Aqua satellite.

Deanna Needell
University of California, Los Angeles

"Deep Models for Improved Topic Recovery"

We introduce a new method for detecting latent hierarchical structure in data based on non-negative matrix factorization. Datasets with hierarchical structure arise in fields as diverse as document classification, image processing, and bioinformatics. Our method recursively applies topic modeling in layers to discover overarching topics encompassing the lower-layer features. By computing the general form of the derivative of the function that defines the relationship between the layers, we derive a backwards propagation scheme, thus framing our method as a neural network. We test our method on a synthetic and real data; numerical results demonstrate the efficacy and promise of our method.

Junier Oliva
University of North Carolina-Chapel Hill

"Improving Generative Models"

Unsupervised generative methods have undergone a recent renaissance, spurred on in large part by impressive photo-realistic results in image applications. These generative methods seek to yield models that understand data by learning how to generate samples through implicit and explicit likelihood optimization. However, despite the surge in interest, these models are limited in several key aspects. First, although methods with an explicit likelihood are, in principle, able to perform additional tasks like anomaly detection and imputation, biases in the learned likelihood render these models useless for such important tasks. For example, recent work has shown that modern methods lead to high out-of-distribution likelihoods for data that is unlike seen training instances. Secondly, most current generative methods are limited to fixed-length vector or sequential data, leaving a substantial gap for the analysis of exchangeable data like sets and graphs. I.e., modern generative models excel at modeling dependencies among features in a point, but are lacking in modeling dependencies among points in a collection. In this talk I discuss these shortcomings and suggest some possible avenues for improvement.

Veronika Rockova
University of Chicago

"Posterior Concentration for Sparse Deep Learning"

We introduce Spike-and-Slab Deep Learning (SS-DL), a fully Bayesian alternative to dropout for improving generalizability of deep ReLU networks. This new type of regularization enables provable recovery of smooth input-output maps with unknown levels of smoothness. Indeed, we show that the posterior distribution concentrates at the near minimax rate for ?-Ho?lder smooth maps, performing as well as if we knew the smoothness level ? ahead of time. Our result sheds light on architecture design for deep neural networks, namely the choice of depth, width and sparsity level. These network attributes typically depend on unknown smoothness in order to be optimal. We obviate this constraint with the fully Bayes construction. As an aside, we show that SS-DL does not overfit in the sense that the posterior concentrates on smaller networks with fewer (up to the optimal number of) nodes and links. Our results provide new theoretical justifications for deep ReLU networks from a Bayesian point of view. (joint work with Nicholas Polson)

Johannes Schmidt-Hieber
University of Twente

"Deep ReLU Networks Viewed as a Statistical Method"

We provide a review of the recent literature on statistical risk bounds for deep neural networks. We also discuss some theoretical results that compare the performance of deep ReLU networks to other methods such as wavelets and spline-type methods. The talk will moreover highlight some open problems and sketch possible new directions.

Deborshee Sen
Duke University

“Neural Network Density Estimation”

In conducting non-linear dimensionality reduction and feature learning, it is common to suppose that the data lie near a lower-dimensional manifold. However, there are very few model-based approaches for density estimation that can accommodate such structure. One such class includes latent variables in an unknown non-linear regression function; this includes Gaussian process latent variable models (GP-LVMs) and variational auto-encoders (VAEs) as special cases. VAEs are similar to GP-LVMs, but instead of using a GP to model the unknown regression function, one uses neural networks and additionally employs approximations to make the computation tractable. Current implementations of such frameworks lack adequate uncertainty quantification in estimating the unknown density and the lower-dimensional subspace, and can be unstable and lack reproducibility in practice. We attempt to solve this problem by designing Markov chain Monte Carlo (MCMC) sampling algorithms for fully Bayesian inferences in neural network models with latent variables. Most sampling algorithms tend to have very poor mixing â€œoff-the-shelfâ€• for non-linear regression with latent variables, but building on Hamiltonian Monte Carlo (HMC) approaches, we develop efficient algorithms that are shown to produce good performance in a variety of settings. This approach can be used for not only uncertainty quantification in density estimation, but also to conduct inferences on any functional of the density as well as prediction. We also discuss issues of identifiability.

Rui Song
North Carolina State University

“Statistical Inference for Online Decision Making via Stochastic Gradient Descent”

Online decision making aims to learn the optimal decision rule by making personalized decisions and updating the decision rule recursively. It has become easier than before with the help of big data, but new challenges also come along. Since the decision rule should be updated once per step, an offline update which uses all the historical data is inefficient in computation and storage. To this end, we propose a completely online algorithm that can make decisions and update the decision rule online via stochastic gradient descent.
It is not only efficient but also supports all kinds of parametric reward models. Focusing on the statistical inference of online decision making, we establish the asymptotic normality of the parameter estimator produced by our algorithm and the online inverse probability weighted value estimator we used to estimate the optimal value. Online plugin estimators for the variance of the parameter and value estimators are also provided and shown to be consistent, so that interval estimation and hypothesis test are possible using our method.
The proposed algorithm and theoretical results are tested by simulations and a real data application to news article recommendation.

Quoc Tran-Dinh
University of North Carolina-Chapel Hill

"ProxSARAH Algorithms for Stochastic Composite Nonconvex Optimization"

We propose a new stochastic first-order algorithmic framework to solve stochastic composite nonconvex optimization problems that covers both finite-sum and expectation settings. Our algorithms rely on the SARAH estimator and consist of two steps: a proximal gradient and an averaging step making them different from existing nonconvex proximal-type algorithms. The algorithms only require an average smoothness assumption of the nonconvex objective term and additional bounded variance assumption if applied to expectation problems. They work with both constant and adaptive step-sizes, while allowing single sample and mini-batches. In all these cases, we prove that our algorithms can achieve the best-known complexity bounds. One key step of our methods is new constant and adaptive step-sizes that help to achieve desired complexity bounds while improving practical performance. Our constant step-size is much larger than existing methods including proximal SVRG schemes in the single sample case. We also specify the algorithm to the non-composite case that covers existing state-of-the-arts in terms of complexity bounds.

Our update also allows one to trade-off between step-sizes and mini-batch sizes to improve performance. We test the proposed algorithms on two composite nonconvex problems and neural networks using several well-known datasets.

This is a joint work with Nhan Pham (UNC-Chapel Hill), Lam M. Nguyen (IBM Research), and Dzung Phan (IBM Research).

Rebecca Willett
University of Chicago

"Learning to Solve Inverse Problems in Imaging"

Many challenging image processing tasks can be described by an ill-posed linear inverse problem: deblurring, deconvolution, inpainting, compressed sensing, and superresolution all lie in this framework. Traditional inverse problem solvers minimize a cost function consisting of a data-fit term, which measures how well an image matches the observations, and a regularizer, which reflects prior knowledge and promotes images with desirable properties like smoothness. Recent advances in machine learning and image processing have illustrated that it is often possible to learn a regularizer from training data that can outperform more traditional regularizers. I will describe an end-to-end, data-driven method of solving inverse problems inspired by the Neumann series, called a Neumann network. Rather than unroll an iterative optimization algorithm, we truncate a Neumann series which directly solves the linear inverse problem with a data-driven nonlinear regularizer. The Neumann network architecture outperforms traditional inverse problem solution methods, model-free deep learning approaches, and state-of-the-art unrolled iterative methods on standard datasets. Finally, when the images belong to a union of subspaces and under appropriate assumptions on the forward model, we prove there exists a Neumann network configuration that well-approximates the optimal oracle estimator for the inverse problem and demonstrate empirically that the trained Neumann network has the form predicted by theory. This is joint work with Davis Gilton and Greg Ongie.

Yao Xie
Georgia Institute of Technology

"ReLu regression: Complexity and Approximation Algorithms"

ReLU regression problem is related to fundamental tasks in machine learning, such as training deep neural networks and performing variable selection in embedding using neural networks. We study this problem from the algorithmic complexity perspective and how to possibly solve the non-convex problem more efficiently and reliably. First, we show that ReLu regression is NP-hard in general. When the number of features is p, and the number of samples is n, there exists a polynomial algorithm that achieves the global optimal solution in O(n^p) running time. Second, we present an integer programming (IP) framework, which can produce dual bounds and feasible upper bounds. Moreover, we present a polynomial-time iterative n-Approximation Algorithm based on convex relaxation and statistical intuition, which performs well in practice, as demonstrated by numerical studies.

Authors: Guanyi Wang, Santanu Dey, and Yao Xie

Harrison Zhou
Yale University

"Statistical and Computational Guarantees of EM with Random Initialization"

This talk considers parameter estimation in the two-component symmetric Gaussian mixtures in d dimensions with n independent samples. We show that, even in the absence of any separation between components, with high probability, the EM algorithm converges to an estimate in at most $O(\sqrt{n} \log n)$ iterations, which is within $O((d/n)^{1/4} (\log n)^{3/4})$ in Euclidean distance to the true parameter, provided that $n=\Omega(d \log^2 d)$. This is within a logarithmic factor to the minimax optimal rate of $(d/n)^{1/4}$. The proof relies on establishing (a) a non-linear contraction behavior of the population EM mapping (b) concentration of the EM trajectory near the population version, to prove that random initialization works. This is in contrast to previous analysis in Daskalakis, Tzamos, and Zampetakis (2017) that requires sample splitting and restart the EM iteration after normalization, and Balakrishnan, Wainwright, and Yu (2017) that requires strong conditions on both the separation of the components and the quality of the initialization. Furthermore, we obtain the asymptotic efficient estimation when the signal is stronger than the minimax rate.