The New York Academy of Sciences

14th Annual Machine Learning Symposium

March 13, 2020 | www.nyas.org/ml2020

POSTER ABSTRACTS

1. Leader Stochastic Gradient Descent for Distributed Training of Deep Learning Models
Yunfei Teng, PhD, Wenbo Gao, Francois Chalus, Anna Choromanska, Donald Goldfarb, Adrian Weller; NYU, Columbia University, University of Cambridge

2. Imitation-Projected Programmatic Reinforcement Learning
Abhinav Verma, Hoang M. Le, PhD[1], Yisong Yue, PhD[2], Swarat Chaudhuri, PhD[3]; Rice University, Microsoft Research AI[1], Caltech[2], University of Texas, Austin[3]

3. Exploring Structural Inductive Biases in Emergent Communication
Agnieszka Slowik, **PhD Student**, Abhinav Gupta, MSc[1], William L. Hamilton, PhD[2], Mateja Jamnik, PhD[3], Sean B. Holden, PhD[3], Christopher Pal, PhD[4]; University of Cambridge, 124Mila, McGill University[2], University of Cambridge[3], Ecole Polytechnique de Montreal[4]

4. Generative Models for Solving Nonlinear Partial Differential Equations
Ameya Joshi, Chinmay Hegde, Soumik Sarkar; New York University, Iowa State University

5. Neural Clustering Processes
Ari Pakman, Yueqi Wang, Catalin Mitelut, JinHyung Lee, Liam Paninski; Columbia University

6. Second-Order Information in Non-Convex Stochastic Optimization: Power and Limitations
Ayush Sekhari, Yossi Arjevani, Yair Carmon, Dylan Foster, Karthik Sridharan; Cornell University, NYU, Stanford University, MIT, Cornell University

7. Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control
Biswadip Dey, PhD, Yaofeng D. Zhong, MS[1], Amit Chakraborty, PhD[2]; Siemens Corporation, Corporate Technology, Princeton University, Princeton, New Jersey, United States[1]; Siemens Corporation, Corporate Technology, Princeton, New Jersey, United States[2]

8. Minimax Regret of Forecasting under Logarithmic Loss with Arbitrary Experts
Blair Bilodeau, PhD (in progress), Dylan J. Foster, PhD[1], Daniel M. Roy, PhD[2]; University of Toronto, Massachusetts Institute of Technology[1], University of Toronto & Vector Institute[2]

9. Corrupted Multidimensional Binary Search: Learning in the Presence of Irrational Agents
Chara Podimata, MSc, Akshay Krishnamurthy, PhD[1], Thodoris Lykouris, PhD[1]; Harvard, Microsoft Research NYC[1]

10. Robust Multi-agent Counterfactual Prediction
Christian Kroer, PhD, Alex Peysakhovich, Adam Lerer; Columbia University, Facebook Research

11. Optimal and efficient contextual bandits with regression oracles
Dylan Foster, PhD, Alexander Rakhlin; MIT

12. Bridging the Gap in Online Projection-free Optimization
Edgar Minasyan, PhD (in progress), Elad Hazan; Princeton University; Google AI Princeton

13. Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training
Giannis Karamanolakis, PhD Student, Daniel Hsu, Luis Gravano; Columbia University, Professor, Professor

14. On the role of data in PAC-Bayes bounds
Gintare Roy, PhD, Kyle Hsu, Waseem Gharbieh, Daniel M. Roy; Element AI, Vector Institute, University of Toronto

15. Fair Predictors under Distribution Shift
Harvineet Singh, MTech, Currently a PhD Student, Rina Singh, PhD, Vishwali Mhasawade, MSc and Rumi Chunara, PhD; New York University

16. Adaptivity in Adaptive Submodularity
Hossein Esfandiari, PhD, Vahab Mirrokni, Amin Karbasi; Google Research, Yale University

17. Zero-Shot AutoML
Iddo Drori, Professor, Lu Liu, Qiang Ma, Brandon Kates, Madeleine Udell; Columbia University and Cornell University

18. A/B test on bipartite graphs
Jean Pouget-Abadie, PhD, Kevin Aydin, Warren Schudy, Kay Brodersen, Vahab Mirrokni; Google

19. Generalization via rerandomization with applications to interpolating predictors
Jeffrey Negrea, PhD (in progress), Daniel M. Roy, PhD[1,2], Gintare Karolina Dziugaite, PhD[3]; University of Toronto, Vector Institute, University of Toronto[1], Vector Institute[2], Element AI[3]

20. Communication-Efficient Ensemble Methods for Federated Learning
Jenny Hamer, BS, Ananda Theertha Suresh, PhD, Mehryar Mohri, PhD; Google Research, New York

21. Stochastically Dominant Distributional Reinforcement Learning
John Martin, Michal Lyskawinski, Xiaohu Li, Brendan Englot; Stevens Institute of Technology

22. The Lottery Ticket Hypothesis: On Sparse, Trainable Neural Networks
Jonathan Frankle, PhD Student, Gintare Karolina Dziugaite, Daniel M. Roy, Alex Renda, Michael Carbin; MIT, Element AI, University of Toronto, MIT, MIT

23. Learning the Optimal Step Size for Gradient Descent on Convex Quadratics
Kiran Vodrahalli, Alexandr Andoni, Daniel Hsu, Tim Roughgarden; Columbia University

24. Feature Cross Search: Algorithms and Hardness
Lin Chen, MSc, Hossein Esfandiari, PhD[1], Thomas Fu, PhD[2], Vahab S. Mirrokni, PhD[3], Qian Yu; Yale University, Google

25. Reasoning About Generalization via Conditional Mutual Information
Lydia Zakynthinou, PhD, Thomas Steinke; Northeastern University, IBM Research – Almaden

26. LdSM: Logarithm-depth Streaming Multi-label Decision Trees
Maryam Majzoubi, PhD Student, Anna Choromanska; NYU Tandon

27. Langevin Dynamics as Nonparametric Variational Inference
Matthew Hoffman, Ph.D., Yian Ma; Google

28. Distillation under label noise
Michal Lukasik, PhD, Srinadh Bhojanapalli, Aditya Menon; Google Research

29. Deep clustering with measure propagation
Minhua Chen, PhD, Badrinath Jayakumar, Padmasundari Gopalakrishnan, Qiming Huang, Michael Johnston, Patrick Haffner; Interactions LLC

30. InvNet: Encoding Geometrical and Statistical Constraints in Deep Generative Models
Minsu Cho, Ameya Joshi, PhD, Chinmay Hegde, Prof; New York University

31. Boosting for Dynamical Systems
Nataly Brukhim, PhD, Naman Agarwal, PhD, Elad Hazan, Professor, Zhou Lu, PhD; Princeton University, Princeton University, Google Brain

32. Graph Attention Networks with Contextually Embedded Edge Features Predict Disease State from Single-Cell Data
Neal Ravindra, PhD[1,2], Arijit Sehanobish, PhD [1,2], David van Dijk, PhD [1,2]
Cardiovascular Research Center, Yale School of Medicine, New Haven, Connecticut, United States[1], Department of Computer Science, Yale University, New Haven, Connecticut, United States[2]

33. High Quality Real-Time Structured Debate Generation
Niles Christensen, Graduate Student, Eric Bolton, Alex Calderwood, Iddo Drori; Carnegie Mellon University, Columbia University, Cornell University, The Wall Street Journal

34. A mathematical theory of cooperative communication
Patrick Shafto, Pei Wang, Junqi Wang, Pushpi Paranamana; Rutgers University – Newark

35. No-Regret and Incentive Compatible Online Learning
Rupert Freeman, PhD, David M. Pennock, PhD1, Chara Podimata2, Jennifer Wortman Vaughan, PhD3; Microsoft Research, Rutgers University[1], Harvard University[2], Microsoft Research[3]

36. Multi-Task Learning (MTL) for Cross Corpus Speech Emotion Recognition (SER)
Shivali Goel, Masters in Computer Science, Shivali Goelâ€ , MSc, Homayoon Beigiâ€ ,â€¡, PhD
Columbia University (Department of Computer Science), New York, New York, United States
â€¡Recognition Technologies, Inc., South Salem, New York, United States

37. Learning Customer Journey Representation Through Wasserstein Sequence-to-Sequence Auto-Encoders
Shuai Zhao, Wen-Ling Hsu; George Ma; Guy Jacobson; Tan Xu; New Jersey Institute of Tech, AT&T Research Labs

38. Information Condensing Active Learning
Siddhartha Jain, PhD, Ge Liu, David Gifford; CSAIL, Massachusetts Institute of Technology

39. Are Transformers universal approximators of sequence-to-sequence functions?
Srinadh Bhojanapalli, PhD[2], Chulhee Yun, PhD[1], Ankit Singh Rawat, PhD[2], Sashank Reddi, PhD[2], Sanjiv Kumar, PhD[2]; MIT[1], Google Research[2]

40. Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs
Tankut Can, PhD, Kamesh Krishnamurthy, PhD[1], David J. Schwab, PhD[2]
Initiative for the Theoretical Sciences, The Graduate Center, CUNY, Joseph Henry Laboratories of Physics and PNI, Princeton University[1]; Initiative for the Theoretical Sciences, CUNY Graduate Center and Facebook AI Research[2]

41. Postdoctoral researcher: Corruption robust exploration in episodic reinforcement learningThodoris Lykouris, PhD, Max Simchowitz, MsC[2], Aleksandrs Slivkins, PhD[1], Wen Sun, PhD[1]**;** Microsoft Research NYC, Microsoft Research NYC[1], UC Berkeley[2]

42. Rigorous Neural Network Training with the Duality Structure Gradient Descent Algorithm
Thomas Flynn, PhD; Brookhaven National Laboratory

43. Boosting with Online Convex Optimization
Xinyi Chen, Nataly Brukhim, Elad Hazan, Shay Moran; Princeton University, Google AI, Princeton

44. Low Transverse Momentum Upsilon Reconstruction Using Large Hadron Collider (LHC) Data
Ye Won Byun, Eleanor Eng, Jason Whang; Brown University, The European Organization for Nuclear Research (CERN), Whang

45. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks
Ziwei Ji, Matus Telgarsky, PhD; University of Illinois Urbana-Champaign

46. Overinterpretation Reveals Image Classification Model Pathologies
Brandon Carter, MEng, Siddhartha Jain, PhD[1], Jonas Mueller, PhD[1], David Gifford, PhD[1]
MIT CSAIL, Cambridge, Massachusetts, United States[1]

47. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples

Angie Boggust, BS, Brandon Carter, MEng[1], Arvind Satyanarayan, PhD[1]; MIT CSAIL, Cambridge, Massachusetts, United States[1]

48. MoFlow: An Invertible Flow Model for Generating Molecular Graphs

Chengxi Zang, PhD, Fei Wang, PhD[1]; Weill Cornell Medicine[1]

49. Neural Dynamics on Complex Networks

Chengxi Zang, PhD, Fei Wang, PhD[1]; Weill Cornell Medicine[1]

50. Disagreement-Regularized Imitation Learning

Mikael Henaff, PhD, Kiante Brantley, Wen Sun; Microsoft Research, University of Maryland

51. Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning

Mikael Henaff, PhD, Dipendra Misra, Akshay Krishamurthy, John Langford**;** Microsoft Research, Microsoft Research

52. Turbulence Forecasting via Neural ODEs

Peetak Mitra, PhD (in progress), Gavin Portwood, Mateus Dias Ribeiro, Michael Chertkov, Anima Anandkumar, Richard Baraniuk, David Schmidt**;** University of Massachusetts Amherst, Los Alamos National Laboratory, German Research Center for Artificial Intelligence, NVIDIA, Rice University, University of Massachusetts Amherst

53. Machine Learning for modeling Internal Combustion Engines

Peetak Mitra, PhD (in progress), Majid Haghshenas, Mateus Dias Ribeiro, David Schmidt University of Massachusetts Amherst, University of Massachusetts Amherst, German Research Center for Artificial Intelligence

1. Leader Stochastic Gradient Descent for Distributed Training of Deep Learning Models

Yunfei Teng, PhD, Wenbo Gao, Francois Chalus, Anna Choromanska, Donald Goldfarb, Adrian Weller
NYU, Columbia University, University of Cambridge

We consider distributed optimization under communication constraints for training deep learning models. Our method differs from the state-of-art parameter-averaging scheme EASGD in a number of ways: (i) our objective formulation does not change the location of stationary points compared to the original optimization problem; (ii) we avoid convergence decelerations caused by pulling local workers descending to different local minima to each other (i.e. to the average of their parameters); (iii) our update by design breaks the curse of symmetry (the phenomenon of being trapped in poorly generalizing sub-optimal solutions in symmetric non-convex landscapes); and (iv) our approach is more communication efficient since it broadcasts only parameters of the leader rather than all workers. We provide theoretical analysis of the batch version of the proposed algorithm, which we call Leader Gradient Descent (LGD), and its stochastic variant (LSGD). Finally, we implement an asynchronous version of our algorithm and extend it to the multi-leader setting, where we form groups of workers, each represented by its own local leader (the best performer in a group), and update each worker with a corrective direction comprised of two attractive forces: one to the local, and one to the global leader (the best performer among all workers).

2. Imitation-Projected Programmatic Reinforcement Learning

Abhinav Verma, Hoang M. Le, PhD[1], Yisong Yue, PhD[2], Swarat Chaudhuri, PhD[3]
Rice University, Microsoft Research AI[1], Caltech[2], University of Texas, Austin[3].

We study the problem of programmatic reinforcement learning, in which policies are represented as short programs in a symbolic language. Programmatic policies can be more interpretable, generalizable, and amenable to scalable formal verification than neural policies. Such policies can therefore be used in RL applications, where the opaqueness and unreliability of neural policies might be a critical impediment. However, designing rigorous learning approaches for programmatic policies remains a challenge. Our approach to this challenge - a meta-algorithm called PROPEL - is based on three insights. First, we view our learning task as optimization in policy space, modulo the constraint that the desired policy has a programmatic representation, and solve this optimization problem using a form of mirror descent that takes a gradient step into the unconstrained policy space and then projects back onto the constrained space. Second, we view the unconstrained policy space as mixing neural and programmatic representations, which enables employing state-of-the-art deep policy gradient approaches. Third, we cast the projection step as program synthesis via imitation learning, and exploit contemporary combinatorial methods for this task. We present theoretical convergence results for PROPEL and empirically evaluate the approach in three continuous control domains. The experiments show that PROPEL significantly outperform state-of-the-art approaches for learning programmatic policies.

3. Exploring Structural Inductive Biases in Emergent Communication

Agnieszka Slowik, **PhD Student**, Abhinav Gupta, MSc[1], William L. Hamilton, PhD[2], Mateja Jamnik, PhD[3], Sean B. Holden, PhD[3], Christopher Pal, PhD[4]; University of Cambridge, 124Mila, McGill University[2], University of Cambridge[3], Ecole Polytechnique de Montreal[4]

Human language and thought are characterized by the ability to systematically generate a potentially infinite number of complex structures (e.g., sentences) from a finite set of familiar components (e.g., words). Recent works in emergent communication have discussed the propensity of artificial agents to develop a systematically compositional language through playing co-operative referential games. The degree of structure in the input data was found to affect the compositionality of the emerged communication protocols. Thus, we explore various structural priors in multi-agent communication and propose a novel graph referential game. We compare the effect of structural inductive bias (bag-of-words, sequences and graphs) on the emergence of compositional understanding of the input concepts measured by topographic similarity and generalization to unseen combinations of familiar properties. We empirically show that graph neural networks induce a better compositional language prior and a stronger generalization to out-of-domain data. We further perform ablation studies that show the robustness of the emerged protocol in graph referential games.

4. Generative Models for Solving Nonlinear Partial Differential Equations

Ameya Joshi, Chinmay Hegde, Soumik Sarkar
New York University, Iowa State University

Partial differential equations (PDEs) describe a wide variety of physical systems. While there exist several numerical methods to solve PDEs, they are often computationally expensive, and solutions to varying boundary conditions and forcing functions need to be derived from scratch. We present a conditional generative modeling-based approach to solve families of PDEs parameterized by a distribution of boundary conditions and coefficients. We validate our approach by solving a family of nonlinear PDEs: the Burgers' equation with a single trained model. We also compare with other neural network-based solvers as well as standard numerical solvers and demonstrate comparable accuracy while being computationally more efficient.

5. Neural Clustering Processes

Ari Pakman, Yueqi Wang, Catalin Mitelut, JinHyung Lee, Liam Paninski
Columbia University

Probabilistic clustering models (or equivalently, mixture models) are basic building block in countless statistical models and involve latent random variables over discrete spaces. For these models, posterior inference methods can be inaccurate and/or very slow. In this work we introduce deep network architectures trained with labeled samples from any generative model of clustered datasets. At test time, the networks generate approximate posterior samples of cluster labels for any new dataset of arbitrary size. We develop two complementary approaches to this meta-learning task, requiring either O(N) or O(K) network forward passes per dataset, where N is the dataset size and K the number of clusters. Unlike previous approaches, our methods sample the discrete labels of all the data points from a well-defined posterior without limiting the number of mixture components, thus allowing to model nonparametric Bayesian posteriors. Moreover, the methods are easily parallelized with a GPU. As a scientific application, we present a novel approach to neural spike sorting for high-density multielectrode arrays.

6. Second-Order Information in Non-Convex Stochastic Optimization: Power and Limitations

Ayush Sekhari, Yossi Arjevani, Yair Carmon, Dylan Foster, Karthik Sridharan
Cornell University, NYU, Stanford University, MIT, Cornell University

We give new upper and lower bounds that characterize the complexity of finding \eps-stationary points of smooth non-convex functions using stochastic second-order methods and beyond. In the model where algorithms query an underlying function with Lipschitz Hessian through queries to unbiased stochastic gradient and Hessian oracles with bounded variance, we give the first method that obtains complexity \eps^{-3}; previously, such guarantees were only known to be possible under additional assumptions that facilitate the use of variance reduction. We prove a matching lower bound showing that \eps^{-3} queries is optimal in this model and that, perhaps more surprisingly, this rate is optimal for stochastic p^{th}-order methods for any p\geq2, even when the underlying function has Lipschitz p^{th} derivative. This ``elbow'' phenomenon, where higher order methods offer no improvement, stands in sharp contrast to the deterministic setting, where p^{th} order methods are known to attain improved rates for every p. Building on these results, we provide a number of extensions:

1. We show that stochastic first-order methods, under the same assumptions, can only obtain complexity \eps^{-3.5}; this establishes the optimality of recently proposed perturbed variants of stochastic gradient descent.

2. We give an extension of our algorithm that finds an (\veps,\gamma)-second order stationary point using \eps^{-3} + \gamma^{-5} stochastic second-order queries, and show that this is optimal.

7. Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control

Biswadip Dey, PhD, Yaofeng D. Zhong, MS[1], Amit Chakraborty, PhD[2]
Siemens Corporation, Corporate Technology, Princeton University, Princeton, New Jersey, United States[1]; Siemens Corporation, Corporate Technology, Princeton, New Jersey, United States[2]

In recent years, deep learning has become very accurate and widely used in many application domains. To learn underlying patterns from data and enable generalization beyond the training dataset, it incorporates appropriate inductive bias which captures the ground reality for promoting one hypothesis over another. Inductive bias can be introduced as the prior in a Bayesian model, or via the choice of computation graph in a neural network. In a variety of problems, especially in robotics, laws of physics are primarily responsible for shaping the outcome, and hence generalization performance in these problems can be improved by exploiting underlying physics in the design of computation graphs. Here, by leveraging a generalization of the Hamiltonian dynamics, we propose Symplectic ODE-Net, a deep learning framework which can infer the dynamics of a system in a physically-consistent manner. As we also uncover relevant physical aspects of the system (such as mass and potential energy) from observed state and control trajectories, the learning approach is transparent by design. In addition, we propose a parametrization to enforce this Hamiltonian formalism even when the generalized coordinate data is embedded in a high-dimensional space or we can only access velocity data instead of generalized momentum. Finally, Symplectic ODE-Net, by offering transparent, physically-consistent models for physical systems, opens up new possibilities for synthesizing model-based control strategies.

8. Minimax Regret of Forecasting under Logarithmic Loss with Arbitrary Experts

Blair Bilodeau, PhD (in progress), Dylan J. Foster, PhD[1], Daniel M. Roy, PhD[2]
University of Toronto, Massachusetts Institute of Technology[1], University of Toronto & Vector Institute[2]

We study the classical problem of forecasting under logarithmic loss while competing against an arbitrary class of experts. Previous efforts to characterize minimax rates for this setting have succeeded only for certain expert classes -- primarily those which are low-dimensional and can be covered using the supremum norm -- and tight bounds for high-dimensional settings have remained elusive. We present a novel approach that exploits self-concordance of the log loss, and in particular introduce new variants of complexity measures including covering numbers and fat-shattering dimension that are tailored to the log loss. Using this strategy, we recover the best known rates for general expert classes and improve the best results for certain classes.

9. Corrupted Multidimensional Binary Search: Learning in the Presence of Irrational Agents

Chara Podimata, MSc, Akshay Krishnamurthy, PhD[1], Thodoris Lykouris, PhD[1]
Harvard, Microsoft Research NYC[1]

Standard game-theoretic formulations for settings like contextual pricing and security games assume that agents act in accordance with a behavioral model. In practice however, some agents may not prescribe to the dominant behavioral model or may react in ways that are arbitrarily inconsistent. Existing algorithms heavily depend on the model being (approximately) correct for all agents and have poor performance in the presence of even a few such unboundedly irrational agents. How do we design learning algorithms that are robust to the presence of arbitrarily irrational agents? We address this question for a number of canonical game-theoretic applications by designing a robust algorithm for the fundamental problem of multidimensional binary search. The performance of our algorithm degrades gracefully with the number of corrupted rounds, which correspond to irrational agents and need not be known in advance. As binary search is the key primitive in algorithms for contextual pricing, Stackelberg Security Games, and other game-theoretic applications, we immediately obtain algorithms for these settings. Our techniques draw inspiration from learning theory, game theory, high-dimensional geometry, and convex analysis, and may be of independent algorithmic interest.

10. Robust Multi-agent Counterfactual Prediction

Christian Kroer, PhD, Alex Peysakhovich, Adam Lerer
Columbia University, Facebook Research

We consider the problem of using logged data for counterfactual predictions in multi-agent systems. This task is difficult because in many cases we observe actions individuals take but not their private information. In addition, agents are strategic, so when the rules change, they will also change their actions. Existing methods (e.g. structural estimation, inverse reinforcement learning) make counterfactual predictions by constructing a model of the game, adding the assumption that agents' behavior comes from optimizing given some goals, and then inverting observed actions to learn agent's underlying utility function (a.k.a. type). This approach imposes heavy assumptions such as rationality of the agents being observed, correctness of the analyst's model of the environment/parametric form of the agents' utility functions, and various other conditions to make point identification possible. We

propose a method for analyzing the sensitivity of counterfactual conclusions to violations of these assumptions. We refer to this method as robust multi-agent counterfactual prediction (RMAC). We apply our technique to investigating the robustness of counterfactual claims for classic environments in market design: auctions, school choice, and social choice. Importantly, we show RMAC can be used in regimes where point identification is impossible (e.g. those which have multiple equilibria or non-injective maps from type distributions to outcomes).

11. Optimal and efficient contextual bandits with regression oracles

Dylan Foster, PhD, Alexander Rakhlin
MIT

A major challenge in contextual bandits and reinforcement learning is to develop flexible, general-purpose algorithms that are computationally no harder than classical supervised learning tasks such as classification and regression. Algorithms based on regression have shown promising empirical success, but theoretical guarantees have remained elusive except in special cases. We provide the first universal and optimal reduction from contextual bandits to online regression. Under a standard realizability assumption, we show how to transform any algorithm for online regression with a given class of value functions into an algorithm for contextual bandits with the induced policy class. We characterize the minimax rates for contextual bandits with general function classes, and show that the resulting contextual bandit algorithm is minimax optimal whenever the base algorithm obtains the optimal rate for regression. Compared to previous results, our algorithm requires no distributional assumptions beyond realizability, and works even when contexts are chosen adversarially.

12. Bridging the Gap in Online Projection-free Optimization

Edgar Minasyan, PhD (in progress), Elad Hazan
Princeton University; Google AI Princeton

In recent years, the field of online optimization has received a considerable amount of attention from the machine learning community given its strong theoretical guarantees in a minimal-assumption framework that covers settings such as stochastic optimization, statistical learning, zero-sum games and many more. This work presents a new efficient method for online projection-free constrained optimization that bypasses the projection operation in favor of, often less computationally expensive, linear optimization. Our method achieves a previously unknown T^2/3 regret bound for smooth convex loss functions while recovering the state-of-the-art T^3/4 regret in the general convex case.

The optimality of computationally efficient projection-free methods, for both online and stochastic cases, still remains an open problem since the corresponding lower bound is that of T^1/2 regret for online learning in general. Our work bridges the gap between the online and stochastic projection-free frameworks matching the T^2/3 regret performance of the best stochastic method known so far for smooth convex optimization. The proposed algorithm is an extension of the well-known Follow-The-Perturbed-Leader method with a sampling technique alleviating the computational intractability. Our use of duality in the analysis of randomized online algorithms against adaptive adversaries is, to the best of our knowledge, a novel approach and may potentially be of further use.

13. Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training

Giannis Karamanolakis, PhD Student, Daniel Hsu, Luis Gravano
Columbia University, Professor, Professor

User-generated reviews can be decomposed into fine-grained segments, each evaluating a different aspect of the principal entity (e.g., price, quality, appearance). Automatically detecting these aspects can be useful for both users and downstream opinion mining applications. Current supervised approaches for learning aspect classifiers require many fine-grained aspect labels, which are labor-intensive to obtain. And, unfortunately, unsupervised topic models often fail to capture the aspects of interest. In this work, we consider weakly supervised approaches for training aspect classifiers that only require the user to provide a small set of seed words (i.e., weakly positive indicators) for the aspects of interest. First, we show that current weakly supervised approaches do not effectively leverage the predictive power of seed words for aspect detection. Next, we propose a student-teacher approach that effectively leverages seed words in a bag-of-words classifier (teacher); in turn, we use the teacher to train a second model (student) that is potentially more powerful (e.g., a neural network using word embeddings). Finally, we show that iterative co-training can be used to cope with noisy seed words, leading to both improved teacher and student models. Our proposed approach consistently outperforms previous weakly supervised approaches in six different domains of product reviews and six multilingual datasets of restaurant reviews. This work is published in EMNLP2019.

14. On the role of data in PAC-Bayes bounds

Gintare Roy, PhD, Kyle Hsu, Waseem Gharbieh, Daniel M. Roy
Element AI, Vector Institute, University of Toronto

The dominant term in most PAC-Bayes bounds is the Kullback-Leibler divergence between the posterior and prior. For so-called linear PAC-Bayes bounds, naively, the risk bound for a fixed posterior kernel is minimized in expectation by choosing the prior to be the expected posterior (i.e., the oracle prior). Attempts to approximate the (distribution-dependent) oracle prior abound. In this work, we demonstrate the surprising fact that the optimal bound (in expectation) is not necessarily obtained by the expected posterior (i.e., oracle) prior. Instead, in some cases, one should remove data from the computation of the empirical risk and use this data to compute a conditional expectation of the posterior (data-dependent oracle prior). While using data to learn a prior is a known heuristic, its essential role in optimal bounds is new. In fact, we show that using data can mean the difference between vacuous and nonvacuous bounds. We apply this new principle in the setting of nonconvex learning, simulating data-dependent oracle priors on MNIST and Fashion MNIST with and without held-out data, and demonstrating (state-of- the-art) nonvacuous bounds in both cases.

15. Fair Predictors under Distribution Shift

Harvineet Singh, MTech, Currently a PhD Student, Rina Singh, PhD, Vishwali Mhasawade, MSc and Rumi Chunara, PhD
New York University

Recent work on fair machine learning adds to a growing set of algorithmic safeguards required for deployment in high societal impact areas such as medicine and public health. A fundamental concern

with putting predictive models in practice is to guarantee stable performance under changes in data distribution. For example, ensuring that a model retains good accuracy when deployed in different geographies or healthcare settings than the ones seen during training. Extensive work in domain adaptation from machine learning literature addresses this concern, albeit with the notion of stability limited to that of predictive performance. Our current work provides conditions under which a stable model, both in terms of predictive performance and fairness, can be trained. Using the causal mechanism underlying the data, we select a subset of features for training models with fairness constraints such that risk with respect to an unseen target data distribution is minimized. Advantages of the approach are demonstrated on the task of diagnosing acute kidney injury in a real-world dataset under an instance of measurement policy shift and selection bias.
A preliminary write-up of the work is available at https://arxiv.org/abs/1911.00677

16. Adaptivity in Adaptive Submodularity

Hossein Esfandiari, PhD, Vahab Mirrokni, Amin Karbasi
Google Research, Yale University

Adaptive sequential decision making is one of the central challenges in machine learning and artificial intelligence. In such problems, the goal is to design an interactive policy that plans for an action to take, from a finite set of n actions, given some partial observations. It has been shown that in many applications such as active learning, robotics, sequential experimental design, and active detection, the utility function satisfies adaptive submodularity, a notion that generalizes the notion of diminishing returns to policies. In this paper, we revisit the power of adaptivity in maximizing an adaptive monotone submodular function. We propose an efficient batch policy that with O(log n Ã— log k) adaptive rounds of observations can achieve an almost tight (1 âˆ’ 1/e âˆ’ Îµ) approximation guarantee with respect to an optimal policy that carries out k actions in a fully sequential setting. To complement our results, we also show that it is impossible to achieve a constant factor approximation with o(log n) adaptive rounds. We also extend our result to the case of adaptive stochastic minimum cost coverage. We first prove the conjecture by Golovin and Krause that the greedy policy achieves the asymptotically tight logarithmic approximation guarantee without resorting to stronger notions of adaptivity. We then propose a batch policy that provides the same guarantee in polylogarithmic adaptive rounds through a similar information-parallelism scheme.

17. Zero-Shot AutoML

Iddo Drori, Professor, Lu Liu, Qiang Ma, Brandon Kates, Madeleine Udell
Columbia University and Cornell University

We present a new zero-shot approach to automated machine learning (AutoML) that rapidly predicts a good model for a supervised learning task and dataset. Generally, automated machine learning systems require tens or hundreds of model evaluations. As machine learning datasets grow in size and machine learning models become more complex and computationally intensive to train, this runtime can be prohibitive. Our zero-shot approach accelerates AutoML by orders of magnitude. Our method uses a transformer-based language embedding for representing the dataset and algorithm description, a pre-trained neural network for representing the data, and allows training our model on four existing state-of-the-art AutoML systems. A fully connected neural network is trained to fuse these representations together, capturing the non-linear interactions between them. We form a graph on the fused

representation for each dataset. A graph attention network is trained on existing datasets and learns to predict the best machine learning algorithm for new test datasets. The graph neural network generalizes to new datasets and new sets of datasets. Our approach leverages the progress in natural language processing and transfer learning to provide a significant boost to AutoML. Performance is competitive with AutoML systems AutoSklearn, AlphaD3M, OBOE, and TPOT, while reducing running time from minutes to seconds and prediction time from minutes to milliseconds.

18. A/B test on bipartite graphs

Jean Pouget-Abadie, PhD, Kevin Aydin, Warren Schudy, Kay Brodersen, Vahab Mirrokni
Google

Causal inference in randomized experiments typically assumes that the units of randomization and the units of analysis are one and the same. In some applications, however, these two roles are played by distinct entities linked by a bipartite graph. The key challenge in such bipartite settings is how to avoid interference bias, which would typically arise if we simply randomized the treatment at the level of analysis units. One effective way of minimizing interference bias in standard experiments is through cluster randomization, but this design has not been studied in the bipartite setting where conventional clustering schemes can lead to poorly powered experiments. This paper introduces a novel clustering objective and a corresponding algorithm that partitions a bipartite graph so as to maximize the statistical power of a bipartite experiment on that graph. Whereas previous work relied on balanced partitioning, our formulation suggests the use of a correlation clustering objective. We use a publicly-available graph of Amazon user-item reviews to validate our solution and illustrate how it substantially increases the statistical power in bipartite experiments. This paper was accepted at NeurIPS 2019.

19. Generalization via rerandomization with applications to interpolating predictors

Jeffrey Negrea, PhD (in progress), Daniel M. Roy, PhD[1,2], Gintare Karolina Dziugaite, PhD[3]
University of Toronto, Vector Institute, University of Toronto[1], Vector Institute[2], Element AI[3]

We discuss a framework in which to study the generalization error of a learned predictor in terms of that of a surrogate (potentially randomized) classifier that is coupled to the original learned classifier and designed to trade empirical risk for control of generalization error.

In the case where the predictor interpolates the data, it is interesting to consider theoretical surrogate classifiers that are partially derandomized or rerandomized, e.g., fit to the training data but with modified label noise or modified feature noise.

We show that replacing a learned classifier by its conditional distribution with respect to an arbitrary sigma-field is a viable method to rerandomize.

We give an example, inspired by the work of Nagarajan and Kolter (2019), where the learned classifier interpolates the training data with high probability, has small risk, and, yet, does not belong to a nonrandom class with a tight uniform bound on two-sided generalization error.

At the same time, we bound the risk of the learned classifier in terms of a surrogate that is constructed by conditioning and is shown to belong to a nonrandom class with uniformly small generalization error.

We also apply our technique to overparameterized linear regression, showing that a specific derandomization leads to the decomposition used by Bartlett, Lugosi and Tsigler (2019), while rerandomization by conditioning on a choice sigma-field yields useful generalization bounds based on uniform convergence.

20. Communication-Efficient Ensemble Methods for Federated Learning

Jenny Hamer, BS, Ananda Theertha Suresh, PhD, Mehryar Mohri, PhD
Google Research, New York

Federated learning is a framework for learning global models on a central server from distributed data in a network of compute devices, where the devices are typically numerous and not distributed identically. This setting allows a central model to be learned without sending raw data from the devices to the server, but suffers from (i) a communication-bandwidth bottleneck between server and device, and (ii) limitations in on-device computing power and memory storage, (iii) a violation of the I.I.D. data assumption in statistical learning theory. In this work, we propose federated ensemble methods which offer communication efficiency and accommodate the non-identically distributed nature of local data. We present guarantees for learning federated ensembles in the standard federated learning and agnostic federated learning settings.

Results:
1. For any loss given by a Bregman divergence, we show that ensemble methods are optimal for both the SFL and AFL losses for density estimation. This result may be of independent interest for researchers in multiple source domain adaptation.
2. We propose a communication-efficient Mirror Descent algorithm and prove its convergence guarantee for both the SFL and AFL ensemble approaches. This guarantee characterizes the communication-convergence trade-off.
3. We modify these algorithms for cross-entropy loss by using bias correction on the stochastic gradients and provide improved convergence guarantees.

21. Stochastically Dominant Distributional Reinforcement Learning

John Martin, Michal Lyskawinski, Xiaohu Li, Brendan Englot
Stevens Institute of Technology

We describe a new approach for mitigating risk in the Reinforcement Learning paradigm. Instead of reasoning about expected utility, we use second-order stochastic dominance (SSD) to directly compare the dispersion of random returns induced by different actions. We frame the RL optimization within the space of probability measures to accommodate the SSD relation, treating the distributional Bellman equation as a potential energy functional. We map the distributional RL problem as a Wasserstein gradient flow, for which we prove optimality and convergence of our proposed algorithm. Our algorithm is a discrete-measure approximation method which we call the Dominant Particle Agent (DPA). In the full paper, we demonstrate how safety and performance are better balanced with DPA using SSD action selection than with other risk metrics.

The full paper can be found on arXiv, at https://arxiv.org/abs/1905.07318.

22. The Lottery Ticket Hypothesis: On Sparse, Trainable Neural Networks

Jonathan Frankle, PhD Student, Gintare Karolina Dziugaite, Daniel M. Roy, Alex Renda, Michael Carbin
MIT, Element AI, University of Toronto, MIT, MIT

Neural network training is a computationally intensive procedure with substantial financial and environmental costs. In this work, we explore the extent to which it would have been possible to instead train much smaller subnetworks of conventional, overparameterized models. In standard image classification models on MNIST, CIFAR-10, and ImageNet, we retroactively find small (5-10x) subnetworks that emerge early (0-6%) in training and are capable of training in isolation to the same accuracy as the original network. These subnetworks have a combination of connectivity and weights that makes them particularly adept at learning: randomly reinitializing these subnetworks or pruning randomly substantially reduces accuracy. We generalize these results into the "lottery ticket hypothesis," which postulates the existence of these "winning lottery ticket" subnetworks in all standard dense models.

We have followed up on these observations several further empirical results. These sparse subnetworks only train to high accuracy when they find the same, linearly connected optimum despite the noise of stochastic gradient descent. In addition, our procedure for uncovering winning lottery tickets produces sparse networks that match or exceed the performance of state-of-the-art pruning methods. Overall, these findings highlight the opportunity to prune neural networks early in training and raise new questions about the role of overparameterization in neural network optimization.

23. Learning the Optimal Step Size for Gradient Descent on Convex Quadratics

Kiran Vodrahalli, Alexandr Andoni, Daniel Hsu, Tim Roughgarden
Columbia University

We analyze the optimal step size for L steps of gradient descent from a fixed initialization for convex quadratic minimization problems, and characterize the improvement in error compared to using the textbook step size corresponding to the inverse of the maximum eigenvalue of the Hessian of the convex quadratic. In particular, we bound both above and below the ratio of the errors attained by using the optimal step size and the textbook step size. Furthermore, we are able to extend the analysis to the setting of learning the optimal learning rate over a distribution of convex quadratic minimization problems: We show (1) the step size is efficient to learn both in sample complexity and computationally, (2) upper and lower bounds on the performance improvement due to learning which depend on the ratio between the maximum and second largest eigenvalues of the Hessian, and (3) that learning the optimal step size yields a gradient descent algorithm which achieves a distance to the optimal error exponentially smaller in L compared to the textbook step size, even as the ratio between the eigenvalues of the Hessian goes to 1 or infinity.

24. Feature Cross Search: Algorithms and Hardness

Lin Chen, MSc, Hossein Esfandiari, PhD[1], Thomas Fu, PhD[2], Vahab S. Mirrokni, PhD[3], Qian Yu
Yale University, Google

We study feature cross search as a fundamental primitive in feature engineering. The importance of feature cross search especially for the linear model has been known for a while. In this problem, we need to select a small subset of features from a larger set of features and combine them to form a new feature (called the crossed feature) by considering their Cartesian product. The goal is to find feature crosses to learn an accurate model. In particular, we study the problem of maximizing a normalized AUC of the linear model trained on the crossed feature column.

First, we show that it is impossible to provide an $n^{1/\log\log n}$-approximation algorithm unless the exponential time hypothesis fails. This result also rules out the possibility of solving this problem in polynomial time unless P=NP. On the positive side, by assuming some property for the probability distributions of the feature values, we show that there exists a polynomial-time algorithm that achieves (1-1/e)-approximation. This result is established by relating the AUC to the total variation of the commutator of two probability measures and showing that the total variation of the commutator is monotone submodular. To show this, we relate the submodularity to the positive semi-definiteness of a kernel matrix. Then, we use Bochner's theorem to prove the positive semi-definiteness by showing that its inverse Fourier transform is non-negative. Our techniques and structural results may be of independent interest.

25. Reasoning About Generalization via Conditional Mutual Information

Lydia Zakynthinou, PhD, Thomas Steinke
Northeastern University, IBM Research – Almaden

How can we ensure that a machine learning system produces an output that generalizes to the underlying distribution, rather than overfitting its training data? This is perhaps the fundamental question for the science of statistical machine learning. A vast array of methods have been proposed to answer this question. Most notably, the theory of uniform convergence shows that, if the output is sufficiently “simple,” then it cannot overfit too much. A more recent line of work has used distributional stability (in the form of differential privacy) to provide generalization guarantees that compose adaptively. Other methods for proving generalization include compression schemes and uniform stability. Unfortunately, these methods are largely disconnected from one another; it is, in general, not possible to compare or combine techniques.

In this paper, we provide a framework for studying the generalization properties of machine learning algorithms, which ties together existing approaches, using the unifying language of information theory. Specifically, we use Conditional Mutual Information (CMI) to quantify how well the input (i.e., the training data) can be recognized given the output (i.e., the trained model) of the algorithm. We show that bounds on CMI can be obtained from VC dimension, compression schemes, differential privacy, and other methods. We then show that bounded CMI implies various forms of generalization, applications of which include AUROC generalization guarantees.

26. LdSM: Logarithm-depth Streaming Multi-label Decision Trees

Maryam Majzoubi, PhD Student, Anna Choromanska
NYU Tandon

We consider multi-label classification where the goal is to annotate each data point with the most relevant subset of labels from an extremely large label set. Efficient annotation can be achieved with balanced tree predictors, i.e. trees with logarithmic-depth in the label complexity, whose leaves correspond to labels. Designing prediction mechanism with such trees for real data applications is non-trivial as it needs to accommodate sending examples to multiple leaves while at the same time sustain high prediction accuracy. In this paper we develop the LdSM algorithm for the construction and training of multi-label decision trees, where in every node of the tree we optimize a novel objective function that favors balanced splits, maintains high class purity of children nodes, and allows sending examples to multiple directions but with a penalty that prevents tree over-growth. Each node of the tree is trained once the previous node is completed leading to a streaming approach for training. We analyze the proposed objective theoretically and show that minimizing it leads to pure and balanced data splits. Furthermore, we show a boosting theorem that captures its connection to the multi-label classification error. Experimental results on benchmark data sets demonstrate that our approach achieves high prediction accuracy and low prediction time and position LdSM as a competitive tool among existing state-of-the-art approaches.

27. Langevin Dynamics as Nonparametric Variational Inference

Matthew Hoffman, Ph.D., Yian Ma
Google

Variational inference (VI) and Markov chain Monte Carlo (MCMC) are approximate posterior inference algorithms that are often said to have complementary strengths, with VI being fast but biased and MCMC being slower but asymptotically unbiased. In this paper, we analyze gradient-based MCMC and VI procedures and find theoretical and empirical evidence that these procedures are not as different as one might think. In particular, a close examination of the Fokker-Planck equation that governs the Langevin dynamics (LD) MCMC procedure reveals that LD implicitly follows a gradient flow that corresponds to a variational inference procedure based on optimizing a nonparametric normalizing flow. This result suggests that the transient bias of LD (due to too few warmup steps) may track that of VI (due to too few optimization steps), up to differences due to VI's parameterization and asymptotic bias. Empirically, we find that the transient biases of these algorithms (and momentum-accelerated versions) do evolve similarly. This suggests that practitioners with a limited time budget may get more accurate results by running an MCMC procedure (even if it's far from burned in) than a VI procedure, as long as the variance of the MCMC estimator can be dealt with (e.g., by running many parallel chains).

28. Distillation under label noise

Michal Lukasik, PhD, Srinadh Bhojanapalli, Aditya Menon
Google Research

Knowledge distillation is a simple yet effective means of improving the performance of one model (the ``student'') using probabilistic outputs from another (the ``teacher'').

A common intuition as to the success of distillation is that the teacher provides ``dark knowledge'' in the form of incorrect labels for a given instance; however, are there settings where this knowledge might be *harmful*?

In this paper, we study the behaviour of distillation under the pervasive setting of *label noise*. We show that for training samples that are even mildly noisy, naive distillation offers minimal benefits over simply training the student. Surprisingly, we also show that correcting for noise in training the teacher model can *harm* performance.

To resolve this, we propose a simple means of performing distillation while correcting for label noise, inspired by loss correction techniques in the label noise literature.

Experiments on a range of synthetic and real-world datasets confirm the efficacy of this approach.

29. Deep clustering with measure propagation

Minhua Chen, PhD, Badrinath Jayakumar, Padmasundari Gopalakrishnan, Qiming Huang, Michael Johnston, Patrick Haffner
Interactions LLC

Deep models have improved state-of-the-art for both supervised and unsupervised learning. For example, deep embedded clustering (DEC) has greatly improved the unsupervised clustering performance, by using stacked autoencoders for representation learning. However, one weakness of deep modeling is that the local neighborhood structure in the original space is not necessarily preserved in the latent space. To preserve local geometry, various methods have been proposed in the supervised and semi-supervised learning literature (e.g., spectral clustering and label propagation) using graph Laplacian regularization. In this paper, we combine the strength of deep representation learning with measure propagation (MP), a KL-divergence based graph regularization method originally used in the semi-supervised scenario. The main assumption of MP is that if two data points are close in the original space, they are likely to belong to the same class, measured by KL-divergence of class membership distribution. By taking the same assumption in the unsupervised learning scenario, we propose our Deep Embedded Clustering with Measure Propagation (DEC-MP) model. We evaluate DEC-MP on short text clustering tasks. On three public datasets, DEC-MP performs competitively with other state-of-the-art baselines. As an example, on the Stackoverflow dataset, DEC-MP achieved a clustering accuracy of 79%, which is about 4% higher than all competing baselines.

30. InvNet: Encoding Geometrical and Statistical Constraints in Deep Generative Models

Minsu Cho, Ameya Joshi, PhD, Chinmay Hegde, Prof
New York University

Generative Adversarial Networks (GANs), while widely successful in modeling complex data distributions including face images, content generation, image translation, style transfer, and others, GANs have not yet been sufficiently leveraged in scientific computing and design. Reasons for this include the lack of flexibility of GANs to represent discrete-valued image data, as well as the lack of control over physical properties of generated samples. We propose a new conditional generative modeling approach (InvNet) that efficiently enables modeling discrete-valued images, while allowing control over their parameterized geometric and statistical properties. We evaluate our approach on a synthetic and a real world problem. We first present a toy example which our model navigates manifolds of geometric shapes with desired sizes and locations. Also we apply our InvNet to a (challenging) material science problem of generating multi-orientation polycrystalline microstructures with statistical constraint control knobs to explore data spaces that are independent of the training set.

31. Boosting for Dynamical Systems

Nataly Brukhim, PhD, Naman Agarwal, PhD, Elad Hazan, Professor, Zhou Lu, PhD
Princeton University, Princeton University, Google Brain.

The application of boosting has transformed machine learning across a variety of applications, mostly in supervised learning: classification, regression, online learning, and many more.

The same exact motivation for boosting exists in dynamical systems: it is often easy to come up with a reasonable controller for a dynamical system, or a reasonable predictor in time series analysis. However, the theory and practice of boosting faces significant challenges in these settings by the existence of state. Taking control of dynamical systems as an example, a controller affects the state of the system, and it is not a-priori clear how to obtain a meaningful guarantee when shifting between different controllers.

In this paper we study a framework for theoretically sound boosting in the presence of state. We show how using techniques from learning with memory and online gradient boosting gives rise to provable guarantees for learning and control in stateful frameworks.

We conclude with experimental evaluation of our methods on a variety of control and learning tasks in dynamical systems. As weak learners, we use both provable controllers, such as recent improvements of the Linear Quadratic Regulator, as well as deep recurrent neural networks, and show how boosting improves the performance of both.

32. Graph Attention Networks with Contextually Embedded Edge Features Predict Disease State from Single-Cell Data

Neal Ravindra, PhD[1,2], Arijit Sehanobish, PhD [1,2], David van Dijk, PhD [1,2]
Cardiovascular Research Center, Yale School of Medicine, New Haven, Connecticut, United States[1],
Department of Computer Science, Yale University, New Haven, Connecticut, United States[2]

Graph Attention Networks (GAT) have proven to be effective for a wide range of tasks by learning from node features and graph structures. Edge features have recently been used to improve performance of graph attention networks. However, the sequence of edge features contains additional information which has not previously been used in graph neural networks. We develop a new architecture to incorporate fully-contextually embedded edge features using an attention bi-LSTM into a GAT. We apply this model to various node classification tasks. In particular, we achieve 92.3% accuracy in predicting multiple sclerosis (MS) at the single-cell level, outperforming other state-of-the-art methods such as a graph convolutional networks. Single-cell RNA sequencing (scRNA-seq) provides insight into healthy and diseased tissues but has not previously been used for disease prediction. We train our model on scRNA-seq data obtained from blood and cerebrospinal fluid from seven multiple sclerosis (MS) patients and six healthy adults. Further, we use our model to learn new graphs from the scRNA-seq data, providing insight into the cell types and genes that are important for this prediction. Our model also allows us to infer a new low-dimensional feature space for the cells that emphasizes the differences between disease states. To the best of our knowledge, this is the first effort to use deep learning to predict disease state from single-cell data.

33. High Quality Real-Time Structured Debate Generation

Niles Christensen, Graduate Student, Eric Bolton, Alex Calderwood, Iddo Drori
Carnegie Mellon University, Columbia University, Cornell University, The Wall Street Journal

Automatically generating debates is a challenging task that requires an understanding of arguments and how to negate or support them. Large-scale transformer-based generative language models have achieved high semantic coherence, consistent theme and tone over long durations, and basic adherence to grammar rules. Generating structured debates presents an additional requirement: that the model understands the argumentative flow of a debate. A model capable of debate must be able to interpret supporting and refuting claims alike, and understand the semantic context of topics of debate. To model the space of paths through a debate, we define a new data structure we call debate trees. We generate a corpus of tree-structured debates to develop a framework for synthesizing plausible debates, which is agnostic to the language model. Our results demonstrate the ability to generate debates in real-time on complex topics at a quality that is on par with humans, as evaluated by the style, content, and strategy metrics used for judging competitive human debates.

34. A mathematical theory of cooperative communication

Patrick Shafto, Pei Wang, Junqi Wang, Pushpi Paranamana
Rutgers University – Newark

Cooperative communication plays a central role in theories of human cognition, language, development, culture, and improving human-algorithm and human-robot interaction. Existing models of cooperative communication are algorithmic in nature and do not explain why cooperation may yield effective belief transmission or what limitations may arise due to differences in beliefs between agents. We present mathematical analysis proving that three classes of algorithmic models are approximations of Entropy regularized Optimal Transport. We derive a statistical interpretation of previous algorithms as approximate maximum likelihood belief transfer plans, and conditions under which cooperative communication is robust to deviations of beliefs between communicating agents. Our results show that cooperative communication provably enables effective, robust information transmission which is required to explain feats of human learning and improve human-machine interaction.

35. No-Regret and Incentive Compatible Online Learning

Rupert Freeman, PhD, David M. Pennock, PhD1, Chara Podimata2, Jennifer Wortman Vaughan, PhD3
Microsoft Research, Rutgers University[1], Harvard University[2], Microsoft Research[3]

We study the problem of prediction with expert advice in a setting in which experts act strategically to maximize their influence on the learning algorithm's prediction by potentially misreporting their beliefs about the binary event to be realized. Our goal is twofold. On the one hand, we want a learning algorithm to be no-regret when comparing against the best fixed expert in hindsight. On the other, we want to guarantee that each expert's best strategy, irrespective of the strategies of other experts, is to report their true belief about the realization of the event. Towards achieving this goal, we first show that any expert learning algorithm's update rule has an equivalent interpretation as a wagering mechanism, a type of multi-agent scoring rule. When experts are myopic (i.e., wishing to maximize their influence

while looking only one step into the future), we show that using a variant of a known wagering mechanism, one can achieve both incentive-compatibility and asymptotically optimal regret guarantees. When experts are not myopic, we identify an incentive-compatible algorithm with low regret in practice. Further, we extend our results to the bandit setting in which the learner is only able to observe the loss of a single expert at each round.

36. Multi-Task Learning (MTL) for Cross Corpus Speech Emotion Recognition (SER)

Shivali Goel, Masters in Computer Science, Shivali Goelâ€ , MSc, Homayoon Beigiâ€ ,â€¡, PhD
Columbia University (Department of Computer Science), New York, New York, United States
â€¡Recognition Technologies, Inc., South Salem, New York, United States

Majority of existing speech emotion recognition models are trained and evaluated on a single corpus and a single language setting. These systems do not perform as well in a cross-corpus and cross-language scenario. MTL for SER with gender, naturalness, and arousal as auxiliary tasks has shown to enhance the generalisation capabilities of emotion models. This work introduces language ID as another auxiliary task in the MTL framework to explore the role of spoken language on emotion recognition which has not been studied before. Due to the lack of adequately sized emotion corpora in many languages, researchers have previously tried training emotion recognition models on cross-corpus data, i.e. training with data in one or more languages and testing on another. This approach sounds valid only if we consider that expression of emotion is the same in all languages. To verify this hypothesis, we came up with a multi-task framework composed of stacked LSTM that jointly learns to predict emotion and the language in which the emotion is being expressed. We perform experiments on 5 acted speech datasets in 4 languages: Emodb (German), Emovo (Italian), Mandarin Affective Speech Corpus (Mandarin), and Savee and IEMOCAP (English). We also compare the SER performance of using training data from all languages and training a single classifier vs. using training data from all languages in a multi-task setting.

37. Learning Customer Journey Representation Through Wasserstein Sequence-to-Sequence Auto-Encoders

Shuai Zhao, Wen-Ling Hsu; George Ma; Guy Jacobson; Tan Xu
New Jersey Institute of Tech, AT&T Research Labs

A customer journey is the complete path of experiences that a customer contacts with a company. Each contact is a record including the customer's identity, contact channel, timestamp, and contact reasons. Studying the path of customer contacts can provide a better context to understand customers' behavior, increase customer loyalty and lead to business cost reduction. However, customer journey analysis is challenging due to the heterogeneity of user behavior. First, each customer journey can be regarded as a complex variant-length, multiple-attributed sequence. Second, there are usually a large number of categories in contact reasons. Traditional one-hot encoding on categorical variables may lead to a large feature space, which creates a high computation burden. To address those challenges and inspired by the sequence modeling methods utilized in language translation work, we proposed to learn efficient journey embedding through the sequence-to-sequence learning framework. Specifically, we propose a novel Wasserstein sequence-to-sequence autoencoders (WSSAE) method to convert each customer journey into same-length latent embeddings. The experimental results over a real dataset provided from a large telecom company demonstrate the effectiveness of WSSAE compared with the state-of-the-art

embedding algorithms. In addition, we further cluster the journey embedding learned by our model and investigate the intrinsic properties of customer contact sequences.

38. Information Condensing Active Learning

Siddhartha Jain, PhD, Ge Liu, David Gifford
CSAIL, Massachusetts Institute of Technology

We introduce Information Condensing Active Learning (ICAL), a batch mode model agnostic Active Learning (AL) method targeted at Deep Bayesian Active Learning that focuses on acquiring labels for points which have as much information as possible about the still unacquired points. We show that employing the popular strategy of acquiring labels for points that maximize the mutual information with respect to the model parameters does not always minimize the uncertainty of the model's predictions averaged over the still unlabeled points post acquisition. This suboptimal uncertainty can negatively affect test accuracy. Motivated by this observation, we propose acquiring points a batch of points B such that the model's predictions on B have as high a statistical dependency as possible with the model's predictions on the entire unlabeled set U. Thus we want a batch B that condenses the most amount of information about the model's predictions on U. ICAL uses the Hilbert Schmidt Independence Criterion (HSIC) to measure the strength of the dependency between a candidate batch of points and the unlabeled set. We develop key optimizations that allow us to scale our method to large unlabeled sets. We show significant improvements in terms of model accuracy and negative log likelihood (NLL) on several image classification tasks compared to state of the art batch mode AL methods for deep learning.

39. Are Transformers universal approximators of sequence-to-sequence functions?

Srinadh Bhojanapalli, PhD[2], Chulhee Yun, PhD[1], Ankit Singh Rawat, PhD[2], Sashank Reddi, PhD[2], Sanjiv Kumar, PhD[2].
MIT[1], Google Research[2]

Despite the widespread adoption of Transformer models for NLP tasks, the expressive power of these models is not well-understood. In this paper, we establish that Transformer models are universal approximators of continuous permutation equivariant sequence-to-sequence functions with compact support, which is quite surprising given the amount of shared parameters in these models. Furthermore, using positional encodings, we circumvent the restriction of permutation equivariance, and show that Transformer models can universally approximate arbitrary continuous sequence-to-sequence functions on a compact domain. Interestingly, our proof techniques clearly highlight the different roles of the self-attention and the feed-forward layers in Transformers. In particular, we prove that fixed width self-attention layers can compute contextual mappings of the input sequences, playing a key role in the universal approximation property of Transformers. Based on this insight from our analysis, we consider other architectures that can compute contextual mappings and empirically evaluate them.

40. Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs

Tankut Can, PhD, Kamesh Krishnamurthy, PhD[1], David J. Schwab, PhD[2]

Initiative for the Theoretical Sciences, The Graduate Center, CUNY, Joseph Henry Laboratories of Physics and PNI, Princeton University[1]; Initiative for the Theoretical Sciences, CUNY Graduate Center and Facebook AI Research[2]

Recurrent neural networks (RNNs) are powerful dynamical models for data with complex temporal structure. However, training RNNs has traditionally proved challenging due to exploding or vanishing of gradients. RNN models such as LSTMs and GRUs significantly mitigate the issues associated with training RNNs by introducing various types of gating units into the architecture. While these gates empirically improve performance, how the addition of gates influences the dynamics and trainability of GRUs and LSTMs is not well understood. Here, we take the perspective of studying randomly-initialized LSTMs and GRUs as dynamical systems, and ask how the salient dynamical properties are shaped by the gates. We leverage tools from random matrix theory and mean-field theory to study the state-to-state Jacobians of GRUs and LSTMs. We show that the update gate in the GRU and the forget gate in the LSTM can lead to an accumulation of slow modes in the dynamics. Moreover, the GRU update gate can poise the system at a marginally stable point. The reset gate in the GRU and the output and input gates in the LSTM control the spectral radius of the Jacobian, and the GRU reset gate also modulates the complexity of the landscape of fixed-points. Furthermore, for the GRU we obtain a phase diagram describing the statistical properties of fixed-points. Finally, we compare training performance in different dynamical regimes, and use this to inform effective choices for network initialization.

41. Postdoctoral researcher: Corruption robust exploration in episodic reinforcement learning

Thodoris Lykouris, PhD, Max Simchowitz, MsC[2], Aleksandrs Slivkins, PhD[1], Wen Sun, PhD[1]
Microsoft Research NYC, Microsoft Research NYC[1], UC Berkeley[2]

We initiate the study of multi-stage episodic reinforcement learning under adversarial manipulations in both the rewards and the transition probabilities of the underlying system. Existing efficient algorithms heavily rely on the "optimism under uncertainty" principle which dictates their behavior and does not allow flexibility to perform corruption-robust exploration. We address this by departing from the optimistic behavior and creating a general framework that incorporates the principle of action-elimination. This principle has been essential for corruption-robust exploration in multi-armed bandits, a degenerate special case of episodic reinforcement learning. Despite constructing a lower bound for a straightforward implementation of action-elimination, we provide a clean and modular way to transfer it to episodic reinforcement learning. Our algorithm enjoys near-optimal guarantees in the absence of adversarial manipulations, has performance that degrades gracefully as the amount of corruption increases, and does not need to know this amount. Our results shed light on the broader question of robust exploration and suggest a way to address a rather daunting mismatch between optimistic algorithms and algorithms with higher flexibility. To demonstrate the applicability of our framework, we provide a second instantiation thereof, showing how it can provide efficient guarantees for the stochastic setting despite doing almost uniform exploration across plausibly optimal actions.

42. Rigorous Neural Network Training with the Duality Structure Gradient Descent Algorithm

Thomas Flynn, PhD
Brookhaven National Laboratory

The training of deep neural networks is typically carried out using some form of gradient descent, often with great success. However, it is unclear if existing analyses of gradient descent can explain the success

of the algorithm in the context of these complex models. We argue that existing performance guarantees for gradient descent rely on assumptions that are too strong to be applicable in the case of deep neural networks. To address this, we propose an algorithm, duality structure gradient descent (DSGD), that is amenable to a non-asymptotic performance analysis, under mild assumptions on the training set and network architecture. The algorithm can be viewed as a form of layer-wise coordinate descent, where at each iteration the algorithm chooses one layer of the network to update. The decision of what layer to update is done in a greedy fashion, based on a rigorous lower bound on the improvement of the objective function for each choice of layer. In the analysis, we bound the time required to reach approximate stationary points, in both the deterministic and stochastic settings. The convergence is measured in terms of a Finsler geometry that is derived from the network architecture and designed to confirm a Lipschitz-like property on the gradient of the training loss function. Numerical experiments suggest that the algorithm is a promising step towards methods for training neural networks that are both rigorous and efficient.

43. Boosting with Online Convex Optimization

Xinyi Chen, Nataly Brukhim, Elad Hazan, Shay Moran
Princeton University, Google AI, Princeton

Online Boosting is a theoretically sound methodology to improve the accuracy of online predictors. However, state-of-the-art online boosting algorithms apply only in the realizable setting, which assumes that a near-perfect predictor exists. In this work we consider online boosting in the more general, agnostic, setting. Our main result is an algorithm that is capable of improving the accuracy of weak online learners even when a near-perfect predictor does not exist. To the best of our knowledge, our algorithm is the first result in online agnostic boosting.

Our algorithm is based on online convex optimization, and we generalize this method to capture other known boosting methods. Specifically, we give a simple meta-algorithm that unifies realizable and agnostic boosting, in both the statistical and the online settings. These settings have been studied extensively, but approaches typically differ for each case. Our framework gives a single approach to apply boosting in the statistical realizable, statistical agnostic, online realizable, and online agnostic settings.

44. Low Transverse Momentum Upsilon Reconstruction Using Large Hadron Collider (LHC) Data

Ye Won Byun, Eleanor Eng, Jason Whang
Brown University, The European Organization for Nuclear Research (CERN), Whang

Work in particle physics often aims to detect new particles. An early discovery tool for new particles is a model that can accurately reconstruct the particle (after it is fired through a collider) and identify its mass value. A mass value that does not correspond to an existing particle signals potential for an unknown particle. We present a neural network which reconstructs low transverse momentum upsilon particles using Large Hadron Collider (LHC) data from The European Organization for Nuclear Research (CERN). Upsilon particles are used because they decay to tau particles (made up of three pions and a neutrino) with relative regularity. Properties of tau particles are well-known; hence, we can simulate them very accurately. Our model successfully predicts the invariant mass of the neutrino and three pions, and thus reconstructs the upsilon particle and accurately identifies its mass value. We demonstrate the effectiveness of our approach using three different upsilon particle masses: 5 GeV, 9.5 GeV, and 15 GeV. Our approach improves mass invariance and surpasses preliminary studies that can

only reconstruct upsilon particles for a single mass value. We utilize a deep neural network with a custom local coordinate system and a custom loss function which weighs features of the neutrino according to their importance. Though this model does not entirely prove the existence of new particles, it helps us know where to look, which is critical in the field of particle physics.

45. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks

Ziwei Ji, Matus Telgarsky, PhD
University of Illinois Urbana-Champaign

Recent theoretical work has guaranteed that overparameterized networks trained by gradient descent achieve arbitrarily low training error, and sometimes even low test error. The required width, however, is always polynomial in at least one of the sample size n, the (inverse) target error 1/epsilon, and the (inverse) failure probability 1/delta. This work shows that \tilde{O}(1/epsilon) iterations of gradient descent with \tilde{Î©}(1/epsilon^2) training examples on two-layer ReLU networks of any width exceeding polylog(n,1/epsilon,1/delta) suffice to achieve a test misclassification error of epsilon. The analysis further relies upon a margin property of the limiting kernel, which is guaranteed positive, and can distinguish between true labels and random labels.

46. Overinterpretation Reveals Image Classification Model Pathologies

Brandon Carter, MEng, Siddhartha Jain, PhD[1], Jonas Mueller, PhD[1], David Gifford, PhD[1]
MIT CSAIL, Cambridge, Massachusetts, United States[1]

Image classifiers are typically scored on their test set accuracy, but high accuracy can mask a subtle type of model failure. We find that high scoring convolutional neural networks (CNN) exhibit troubling pathologies that allow them to display high accuracy even in the absence of semantically salient features. When a model provides a high-confidence decision without salient supporting input features we say that the classifier has overinterpreted its input, finding too much class-evidence in patterns that appear nonsensical to humans. Here, we demonstrate that state of the art neural networks for CIFAR-10 and ImageNet suffer from overinterpretation, and find CIFAR-10 trained models make confident predictions even when 95% of an input image has been masked and humans are unable to discern salient features in the remaining pixel subset. Although these patterns portend potential model fragility in real-world deployment, they are in fact valid statistical patterns of the image classification benchmark that alone suffice to attain high test accuracy. We find that ensembling strategies can help mitigate model overinterpretation and classifiers which rely on more semantically meaningful features can improve accuracy over both the test set and out-of-distribution images from a different source than the training data.

47. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples

Angie Boggust, BS, Brandon Carter, MEng[1], Arvind Satyanarayan, PhD[1]
MIT CSAIL, Cambridge, Massachusetts, United States[1]

Embeddings -- mappings from high-dimensional discrete input to lower-dimensional continuous vector spaces -- have been widely adopted in machine learning, linguistics, and computational biology as they

often surface interesting and unexpected domain semantics. Through semi-structured interviews with embedding model researchers and practitioners, we find that current tools poorly support a central concern: comparing different embeddings when developing fairer, more robust models. In response, we present the Embedding Comparator, an interactive system that balances gaining an overview of the embedding spaces with making fine-grained comparisons of local neighborhoods. For a pair of models, we compute the similarity of the k-nearest neighbors of every embedded object, and visualize the results as Local Neighborhood Dominoes: small multiples that facilitate rapid comparisons. Using case studies, we illustrate the types of insights the Embedding Comparator reveals including how fine-tuning embeddings changes semantics, how language changes over time, and how training data differences affect two seemingly similar models.

48. MoFlow: An Invertible Flow Model for Generating Molecular Graphs

Chengxi Zang, PhD, Fei Wang, PhD[1]
Weill Cornell Medicine[1]

Generating molecular graphs with desired chemical property driven by deep generative models is a very promising way to accelerating existing long and costly drug discovery process.

Such generative models usually consist of learning latent representations and generation of molecular graphs. However, how to generate novel and chemically-valid molecular graphs from latent representations is very challenging because of combinatorial nature and chemical constraints of molecular graphs.

We propose MoFlow, which is a flow-based model, to learn invertible mapping between molecular graphs and their latent representations. Our MoFlow generates molecules by first generating bond skeleton through a Glow based model, then generating atoms given bonds by a novel graph conditional flow, and finally combing them into a molecule with postmortem validity correction.

Our MoFlow has merits of exact and tractable likelihood training, efficient one-pass embedding and generation, chemical validity guarantees, 100\% reconstruction of training data, and good generalization property. We validate our model through four tasks: molecular generation and reconstruction, visualizing continuous latent space, property optimization, and constrained property optimization. Our MoFlow achieves many new state-of-the-art performance, implying the potentials to explore large chemical space for drug discovery.

49. Neural Dynamics on Complex Networks

Chengxi Zang, PhD, Fei Wang, PhD[1]
Weill Cornell Medicine[1]

Learning dynamics on complex networks governed by differential equation systems is crucial for understanding, predicting, and controlling complex systems in science and engineering. However, this task is very challenging due to the intrinsic complexities in the structures of the high dimensional systems, their elusive continuous-time nonlinear dynamics, and their structural-dynamic dependencies.

To address these challenges, we propose a differential deep learning model to learn continuous-time dynamics on complex networks in a data-driven way. We model differential equation systems by graph

neural networks. Instead of mapping through a discrete number of hidden layers in the forward process, we solve the initial value problem by integrating the neural differential equation systems over time numerically. In the backward process, we learn the optimal parameters by back-propagating against the forward integration.

We validate our model by learning and predicting various real-world dynamics on different complex networks in both (continuous-time) network dynamics learning setting and (regularly-sampled) structured sequence learning setting, and then apply our model to graph semi-supervised classification tasks (a one-snapshot case). The promising experimental results demonstrate our model's capability of jointly capturing the structure, dynamics, and semantics of complex systems in a unified framework.

50. Disagreement-Regularized Imitation Learning

Mikael Henaff, PhD, Kiante Brantley, Wen Sun
Microsoft Research, University of Maryland

We present a simple and effective algorithm designed to address the covariate shift problem in imitation learning. It operates by training an ensemble of policies on the expert demonstration data, and using the variance of their predictions as a cost which is minimized with RL together with a supervised behavioral cloning cost. Unlike adversarial imitation methods, it uses a fixed reward function which is easy to optimize. We prove a regret bound for the algorithm in the tabular setting which is linear in the time horizon multiplied by a coefficient which we show to be low for certain problems in which behavioral cloning fails. We evaluate our algorithm empirically across multiple pixel-based Atari environments and continuous control tasks, and show that it matches or significantly outperforms behavioral cloning and generative adversarial imitation learning.

51. Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning

Mikael Henaff, PhD, Dipendra Misra, Akshay Krishamurthy, John Langford
Microsoft Research, Microsoft Research

We present an algorithm, HOMER, for exploration and reinforcement learning in rich observation environments that are summarizable by an unknown latent state space. The algorithm interleaves representation learning to identify a new notion of kinematic state abstraction with strategic exploration to reach new states using the learned abstraction. The algorithm provably explores the environment with sample complexity scaling polynomially in the number of latent states and the time horizon, and, crucially, with no dependence on the size of the observation space, which could be infinitely large. This exploration guarantee further enables sample-efficient global policy optimization for any reward function. On the computational side, we show that the algorithm can be implemented efficiently whenever certain supervised learning problems are tractable. Empirically, we evaluate HOMER on a challenging exploration problem, where we show that the algorithm is more sample efficient than standard reinforcement learning baselines.

52. Turbulence Forecasting via Neural ODEs

Peetak Mitra, PhD (in progress), Gavin Portwood, Mateus Dias Ribeiro, Michael Chertkov, Anima Anandkumar, Richard Baraniuk, David Schmidt

University of Massachusetts Amherst, Los Alamos National Laboratory, German Research Center for Artificial Intelligence, NVIDIA, Rice University, University of Massachusetts Amherst

Fluid turbulence is characterized by strong coupling across a broad range of scales. Furthermore, besides the usual local cascades, such coupling may extend to interactions that are non-local in scale-space. As such the computational demands associated with explicitly resolving the full set of scales and their interactions, as in the Direct Numerical Simulation (DNS) of the Navier-Stokes equations, in most problems of practical interest are so high that reduced modeling of scales and interactions is required before further progress can be made. While popular reduced models are typically based on phenomenological modeling of relevant turbulent processes, recent advances in machine learning techniques have energized efforts to further improve the accuracy of such reduced models. In contrast to such efforts that seek to improve an existing turbulence model, we propose a machine learning(ML) methodology that captures, de novo, underlying turbulence phenomenology without a pre-specified model form. To illustrate the approach, we consider transient modeling of the dissipation of turbulent kinetic energy, a fundamental turbulent process that is central to a wide range of turbulence models using a Neural ODE approach. After presenting details of the methodology, we show that this approach outperforms state-of-the-art approaches.

53. Machine Learning for modeling Internal Combustion Engines

Peetak Mitra, PhD (in progress), Majid Haghshenas, Mateus Dias Ribeiro, David Schmidt
University of Massachusetts Amherst, University of Massachusetts Amherst, German Research Center for Artificial Intelligence

In Internal Combustion Engines, turbulence plays a key role in the fuel/air mixing, improving overall efficiency and reducing emissions. Modeling these environments involve dealing with turbulence, multiphase flow, combustion, and moving boundaries. Because of the wide variations in length and time scales, the high fidelity models such as Large Eddy Simulations impose strict resolution requirements making the computations both expensive and doomed to omit critical information that cannot be resolved in a pragmatic design cycle. Here we propose using a data driven method for learning optimized approximations to these unresolved features trained on in-house generated high fidelity dataset. In this hybrid PDE-ML framework developed in OpenFOAM the large scale, resolvable features are to be obtained by solving the governing flow/energy equations (PDE) and the machine learning is to be only applied to the small, unresolved scales. A key aspect in developing this framework is that the machine learning model respects the rotational as well as Galilean invariance of the Reynolds stress models and use local quantities to construct the feature set for the data-driven model thereby improving model performance on a low resolution grid, and thus providing a pathway to coarse-graining methods