Speakers

1 **Emmanuel Abbé**

2 **Afonso Bandeira**

3 **Yuheng Bu**

4 **Arnab Bhattacharyya**

5 **Clement Canonne**

6 **Wei-Ning Chen**

7 **Alex Dimakis**

8 **Amin Gohari**

9 **Aditya Gopalan**

10 **Deniz Gunduz**

11 **Reinhard Heckel**

12 **Oliver Johnson**

13 **P. N. Karthik**

14 **Hyeji Kim**

15 **Prashanth L.A.**

16 **Kangwook Lee**

17 **Yi Li**

18 **Yan Hao Ling**

19 **Marco Mondelli**

20 **Mehul Motani**

21 Frederique Oggier

22 Sewoong Oh

23 Ayfer Ozgur

24 Dimitris Papailiopoulos

25 Galen Reeves

26 Miguel Rodrigues

27 Cynthia Rush

28 Cong Shen

29 Yan Shuo Tan

30 Ali Tajer

31 Himanshu Tyagi

32 Antonios Varvitsiotis

33 Yao Xie

Abstracts

Information Theory and Data Science Workshop

(16–27 Jan 2023)

1 Emmanuel Abbé

EPFL, Switzerland
The leap complexity in neural network learning

Abstract

We propose a characterization of the sample complexity of learning with regular networks and stochastic gradient descent for isotropic data. The characterization brings to light a new complexity measure of target functions, the leap, which measures how "hierarchical" the data is. Joint work E. Boix (MIT) and T. Misiakiewicz (Stanford).

2 Afonso Bandeira

ETH Zurich, Switzerland
Statistical-to-Computational Gaps: The Low-Degree method and Free-Energy Barriers

Abstract

When faced with a data analysis, learning, or statistical inference problem, the amount and quality of data available fundamentally determines whether such tasks can be performed with certain levels of accuracy. With the growing size of datasets however, it is crucial not only that the underlying statistical task is possible, but also that is doable by means of efficient algorithms. In this talk we will discuss methods aiming to establish limits of when statistical

tasks are possible with computationally efficient methods or when there is a fundamental Statistical-to-Computational gap in which an inference task is statistically possible but inherently computationally hard. We will describe recent rigorous correspondences between two different approaches to argue computational hardness of high dimensional inference, the low degree method and methods based on free energy potentials. Based on joint work with Ahmed El Alaoui, Samuel B. Hopkins, Tselil Schramm, Alexander S. Wein, and Ilias Zadik.

3 Yuheng Bu

University of Florida, USA

From sensitivity-constrained information bottleneck to fair selective prediction

Abstract

This talk will discuss how to apply the information-theoretic approach to ensure fairness in machine learning. We first consider a generalized information bottleneck (IB) problem by introducing the notion of a sensitive attribute, where the goal is to construct representations of observations that are maximally informative about a target variable while also satisfying constraints with respect to a variable corresponding to the sensitive attribute. Such a formulation arises in a growing number of applications, e.g., fair machine learning and domain generalization.

In particular, we study the problem of fair selective prediction, where a reject option is allowed to improve the prediction performance at the cost of reducing coverage, i.e., predicting fewer samples. However, such a selective prediction method can lead to performance disparities between different sensitive groups. To address this issue, we show that the sufficiency criterion, which corresponds to a certain IB constraint, ensures that a smaller coverage will increase performance in all groups. We further constructed multiple upper bounds on the conditional mutual information, which can be used as regularizers during the training process of neural networks. The effectiveness of the proposed methods has been demonstrated over multiple real datasets with different modalities.

4 Arnab Bhattacharyya

National University of Singapore, Singapore
Near-optimal learning of tree-structured distributions

Abstract

We provide finite sample guarantees for the classical Chow-Liu algorithm (IEEE Trans. Inform. Theory, 1968) to learn a tree-structured graphical model of a distribution. For a distribution P on Σ^n and a tree T on n nodes, we say T is an ε-approximate tree for P if there is a T-structured distribution Q such that $D(P \mid\mid Q)$ is at most ε more than the best possible tree-structured distribution for P. We show that if P itself is tree-structured, then the Chow-Liu algorithm with the plug-in estimator for mutual information with $\widetilde{O}(|\Sigma|^3 n\varepsilon^{-1})$ i.i.d. samples outputs an ε-approximate tree for P with constant probability. In contrast, for a general P (which may not be tree-structured), $\Omega(n^2\varepsilon^{-2})$ samples are necessary to find an ε-approximate tree. Our upper bound is based on a new conditional independence tester that addresses an open problem posed by Canonne, Diakonikolas, Kane, and Stewart (STOC, 2018): we prove that for three random variables X, Y, Z each over Σ, testing if $I(X; Y \mid Z)$ is 0 or $\geq \varepsilon$ is possible with $\widetilde{O}(|\Sigma|^3/\varepsilon)$ samples. Finally, we show that for a specific tree T, with $\widetilde{O}(|\Sigma|^2 n\varepsilon^{-1})$ samples from a distribution P over Σ^n, one can efficiently learn the closest T-structured distribution in KL divergence by applying the add-1 estimator at each node.

Joint work with Sutanu Gayen, Eric Price, Vincent Tan, and N.V. Vinodchandran

5 Clement Canonne

University of Sydney, Australia
Lower bounds for estimation under information constraints: general recipe, and new dishes

Abstract

In this talk, I will provide an overview of new techniques and recipes for distributed learning (estimation) under information constraints such as communication, local privacy, and memory constraints. Motivated by applications in machine learning and distributed computing, these questions lie at

the intersection of information theory, statistics, and theoretical computer science.

I will focus on the minimax setting, and cover a general and powerful approach to establishing information-theoretic lower bounds for estimation in interactive settings.

Based on joint work(s) with Jayadev Acharya, Ziteng Sun, and Himanshu Tyagi (abs/2010.06562).

6 Wei-Ning Chen

Stanford University, USA

Achieving joint privacy and communication efficiency in federated learning and analytics

Abstract

Two major challenges in federated learning (FL) and analytics (FA) are 1) preserving the privacy of the local samples; and 2) communicating them efficiently to a central server, while achieving high accuracy for the end-to-end task. In this talk, we will explore how to address these two challenges jointly without needing a trusted server, under local differential privacy (where the locally privatized message satisfies DP) and distributed differential privacy (where the securely aggregated local information satisfies DP).

In particular, we consider two canonical problems – mean and frequency estimation – that lie at the heart of FL and FA applications, respectively. We study the fundamental trade-offs between privacy, communication, and accuracy and show how to achieve them efficiently with tools including random sampling, random projection, sketching, secure aggregation, and distributed privacy mechanisms. Surprisingly, under both local and distributed DP, we show that the requirements for communication efficiency and privacy can be aligned. With more stringent privacy requirements, one can compress more aggressively and achieve higher communication efficiency while preserving roughly the same level of accuracy. Finally, we will conclude the talk with some recent results and open problems that go beyond the worst-case bounds, showing that by incorporating information about the structure of the problem, one can potentially improve the overly pessimistic global minimax bounds and obtain a better instance-dependent convergence.

7 Alex Dimakis

University of Texas at Austin, USA
Deep Generative models and Inverse problems

Abstract

Modern deep generative models like Score-based models and Diffusions are demonstrating excellent performance in representing high-dimensional distributions, especially for images. We will show how they can be used to solve inverse problems like denoising, filling missing data, and recovery from linear projections. We generalize compressed sensing theory beyond sparsity, extending Restricted Isometries to sets created by deep generative models. Our recent results include establishing theoretical results for Langevin sampling from full-dimensional generative models, generative models for MRI reconstruction, robustness out of distribution and fairness guarantees for inverse problems.

8 Amin Gohari

The Chinese University of Hong Kong, China
On the rate-distortion theory and the generalization error of learning algorithms

Abstract

Understanding generalization in modern machine learning settings has been one of the major challenges in statistical learning theory. Recent years have witnessed the development of various generalization bounds suggesting different complexity notions such as the mutual information between the data sample and the algorithm output, compressibility of the hypothesis space, and the fractal dimension of the hypothesis space. While these bounds have illuminated the problem at hand from different angles, their suggested complexity notions might appear seemingly unrelated. In this presentation, we explicitly relate the concepts of mutual information, compressibility, and fractal dimensions in a single mathematical framework through the lens of rate-distortion theory. This yields an operational connection between the generalization error in statistical learning theory and the rate-distortion theory from classical information theory. We also discuss other formal connections between generalization error and the rate-distortion theory using

f-divergences. Here, we introduce super-modular f-divergences and discuss some of their applications.

9 Aditya Gopalan

Indian Institute of Science, India
Free Inference for Bandits with Rich Actions

Abstract

It is well-known that when playing a multi-armed bandit with finitely many arms, there is a tension between minimizing regret and performing statistical inference on the arms' rewards (e.g., best arm identification). Here, regret-optimal algorithms pull suboptimal arms only log(n) times in n rounds, which is not fast enough to estimate rewards with polynomially decaying error. We show that the situation is rather different for bandits with rich or continuous arm sets. Specifically, in linearly parameterized bandits with action sets that are smooth enough, we show that the minimum eigenvalue of the expected design matrix grows as Omega(sqrtn) whenever the cumulative regret of the algorithm is optimal at O(sqrtn). The argument involves a new application of the standard change-of-measure (data processing) inequality for bandits together with matrix eigenspace perturbation tools such Weyl's inequality and the Davis-Kahan sine-theta theorem. The result opens up the possibility of faster learning with standard regret-optimal algorithms in structured problems such as model selection and clustering in linear bandits, without any additional forced exploration. (Joint work with Debanshu Banerjee, Sayak Ray Chowdhury and Avishek Ghosh)

10 Deniz Gunduz

Imperial College London, UK
Semantic and Pragmatic Compression: New Problems and Information Theoretic Bounds

Abstract

Semantic communication problems are attracting increasing interest mainly driven by the advances in machine learning algorithms and applications. In

principle, they can be considered as generalized compression problems with distortion measures that go beyond single-letter additive measures considered in classical information theory. As an example, I will present the remote contextual multi-armed bandit problem, in which a decision-maker observes the context and the reward, and must communicate the actions to be taken by the agents over a rate-limited communication channel. I will present information theoretic bounds on the rate of communication required to achieve a sub-linear regret in this problem. I will then introduce a general formulation of semantic communications with constraints on the encoder and/or decoder, modelling a prescribed language, and identify achievable rate-distortion pairs. Next, I will introduce the concept of pragmatic compression, which takes 'time' into consideration, and highlight its differences from semantic compression. I will finally present a new sequential compression problem with decoding constraints as an example of pragmatic communications, which can model controlling a remote agent over a rate-limited link. I will show how this formulation generalizes existing lossless and lossy compression problems, and provide bounds on the optimal performance.

11 Reinhard Heckel

Technical University of Munich, Germany

The role of data and models for deep-learning based image reconstruction

Abstract

Deep-learning methods give state-of-the-art performance for a variety of imaging tasks, including accelerated magnetic resonance imaging. In this talk we discuss whether improved models and algorithms or training data are the most promising way forward. First, we ask whether increasing the model size and the training data improves performance in a similar fashion as it has in domains such as language modeling. We find that scaling beyond relatively few examples yields only marginal performance gains. Second, we discuss the robustness of deep learning based image reconstruction methods. Perhaps surprisingly, we find no evidence for neural networks being any less robust than classical reconstruction methods (such as l1 minimization). However, we find that both classical and deep learning based approaches perform significantly worse under distribution shifts, i.e., when trained (or tuned) and tested on slightly different data. Finally, we show that the out-

of-distribution performance can be improved through more diverse training data, or through an algorithmic intervention called test-time-training.

12 Oliver Johnson

University of Bristol, UK
Information-theoretic limit theorems old and new

Abstract

Dating back to the classic work of Barron, it is well known that the Central Limit Theorem holds in the sense of KL divergence. This gives a strong form of convergence, with optimal rates in a certain sense, and characterizing the Gaussian via the property of linear Fisher score can provide insights into why the CLT is true. I will describe a number of results of this flavour, both from my work and that of others, in settings that include Poisson convergence, exchangeability, extreme value theory and order statistics.

13 P. N. Karthik

National University of Singapore, Singapore
Almost Cost-Free Communication in Federated Best Arm Identification

Abstract

In this talk, I shall describe some of our recent results on the problem of finding the best arm in a federated learning setting with a central server and M clients. The setup is that each client is associated with a K-armed bandit with arms generating independent Gaussian rewards, and interested in finding the arm with the largest mean among its arms (the client's "local best arm"). Additionally, each client communicates with the server on a dedicated uplink that entails a cost of C units per usage. The server is interested in finding the "global best arm", defined as the arm with the largest average mean across all the clients. The goal is to identify the local best arms and the global best arm with minimal total cost, defined as the sum of the total number of arm selections at all the clients and the total communication cost, subject to an upper bound on the error probability (fixed-confidence regime).

We propose a novel algorithm called FEDELIM that is based on successive elimination and communicates only in exponential time steps, and obtain a

high probability instance-dependent upper bound on its total cost in terms of the error probability. We study separately the cases when C=0 and C¿0, and demonstrate that for any C¿0, for all sufficiently small error probabilities, the total cost of FEDELIM is at most 3 times that under an algorithm that communicates at every time step when C=0. That is, communication is almost cost-free for the problem of federated best arm identification. We evaluate the performance of our algorithm on two synthetic datasets and the real-world MovieLens dataset. We also compare our algorithm's performance with that of (a) an algorithm that communicates periodically every H time instants for some H¿0, and (b) an algorithm that communicates at super-exponential time steps, and observe empirically that our algorithm strikes a balanced trade-off between the total number of arm selections and the communication cost.

14 Hyeji Kim

University of Texas at Austin, USA
Advancing information theory and coding via deep learning

Abstract

The design of codes for communication and data compression is an important endeavor involving deep mathematical research and wide-ranging practical applications. This talk will discuss a family of codes obtained via deep learning, which notably outperforms state-of-the-art codes for challenging communication and data compression scenarios. For communications, we take an autoencoder-based approach to learn novel codes for multi-terminal communications (e.g., interference channels) and bi-directional communications. For data compression, we use generative models for realizing the result of Wyner-Ziv on the distributed compression of various practical data.

Based on joint work with Alliot Nagle, Anish Acharya, Jay Whang, Karl Chahine, Rajesh Mishra, Yihan Jiang, Alex Dimakis, Sriram Viswanath, and Syed Jafar.

15 Prashanth L.A.

Indian Institute of Technology Madras, India
A Wasserstein Distance Approach for Concentration of Empirical Risk Esti-

mates

Abstract

This talk presents a unified approach based on Wasserstein distance to derive concentration bounds for empirical estimates for two broad classes of risk measures defined in the paper referenced below. The classes of risk measures introduced include as special cases well known risk measures from the finance literature such as conditional value at risk (CVaR), optimized certainty equivalent risk, spectral risk measures, utility-based shortfall risk, cumulative prospect theory (CPT) value, rank dependent expected utility and distorted risk measures. Two estimation schemes are considered, one for each class of risk measures. One estimation scheme involves applying the risk measure to the empirical distribution function formed from a collection of i.i.d. samples of the random variable (r.v.), while the second scheme involves applying the same procedure to a truncated sample. The bounds provided apply to three popular classes of distributions, namely sub-Gaussian, sub-exponential and heavy-tailed distributions. The bounds are derived by first relating the estimation error to the Wasserstein distance between the true and empirical distributions, and then using recent concentration bounds for the latter. Previous concentration bounds are available only for specific risk measures such as CVaR and CPT-value. The bounds derived are shown to either match or improve upon previous bounds in cases where they are available. The usefulness of the bounds is illustrated through an algorithm and the corresponding regret bound for a stochastic bandit problem involving a general risk measure from each of the two classes introduced in the paper referenced below.

16 Kangwook Lee

University of Wisconsin-Madison, USA

Score-based Generative Modeling Secretly Minimizes the Wasserstein Distance

Abstract

Score-based generative models are shown to achieve remarkable empirical performances in various applications such as image generation and audio synthesis. However, a theoretical understanding of score-based diffusion models

is still incomplete. Recently, Song et al. showed that the training objective of score-based generative models is equivalent to minimizing the Kullback-Leibler divergence of the generated distribution from the data distribution. In this work, we show that score-based models also minimize the Wasserstein distance between them. Specifically, we prove that the Wasserstein distance is upper bounded by the square root of the objective function up to multiplicative constants and a fixed constant offset. Our proof is based on a novel application of the theory of optimal transport, which can be of independent interest to the society. Our numerical experiments support our findings. By analyzing our upper bounds, we provide a few techniques to obtain tighter upper bounds.

17 Yi Li

Nanyang Technological University, Singapore
Lower Bounds for Sparse Oblivious Subspace Embeddings

Abstract

An oblivious subspace embedding (OSE), characterized by parameters $m, n, d, \epsilon, \delta$, is a random matrix $\Pi \in \mathbb{R}^{m\times n}$ such that for any d-dimensional subspace $T \subseteq \mathbb{R}^n$, $\Pr_\Pi[\forall x \in T, (1-\epsilon)\|x\|_2 \leq \|\Pi x\|_2 \leq (1+\epsilon)\|x\|_2] \geq 1-\delta$. For ϵ and δ at most a small constant, we show that any OSE with one nonzero entry in each column must satisfy that $m = \Omega(d^2/(\epsilon^2\delta))$, establishing the optimality of the classical Count-Sketch matrix. When an OSE has $1/(9\epsilon)$ nonzero entries in each column, we show it must hold that $m = \Omega(d^2/\epsilon^{1-O(\delta)})$, which is the first lower bound with both d^2 and $1/\epsilon$ as multiplicative factors.

18 Yan Hao Ling

National University of Singapore, Singapore
Optimal Rates of Teaching and Learning Under Uncertainty

Abstract

In this talk, we consider a simple model of teaching and learning under uncertainy in which a teacher receives independent observations of a single bit corrupted by binary symmetric noise, and sequentially transmits to a student

through another binary symmetric channel based on the bits observed so far. After a given number n of transmissions, the student outputs an estimate of the unknown bit, and we are interested in the exponential decay rate of the error probability as n increases. We propose a novel block-structured teaching strategy in which the teacher encodes the number of 1s received in each block, and show that the resulting error exponent is the binary relative entropy $D\big(\frac{1}{2}\|\max(p,q)\big)$, where p and q are the noise parameters. This matches a trivial converse result based on the data processing inequality, and settles two conjectures of [Jog and Loh, 2020] and [Huleihel, Polyanskiy, and Shayevitz, 2019]. In addition, we show that the computation time required by the teacher and student is linear in n. We also study a more general setting in which the binary symmetric channels are replaced by general binary-input discrete memoryless channels. We provide an achievability bound and a converse bound, and show that the two coincide in certain cases, including (i) when the two channels are identical, and (ii) when the student-teacher channel is a binary symmetric channel. In addition, we give sufficient conditions under which our achievable learning rate is the best possible for block-structured protocols.

19 Marco Mondelli

Institute of Science and Technology Austria, Austria

Inference in High Dimensions for (Mixed) Generalized Linear Models: the Linear, the Spectral and the Approximate

Abstract

In a generalized linear model (GLM), the goal is to estimate a d-dimensional signal x from an n-dimensional observation of the form f(Ax, w), where A is a design matrix and w is a noise vector. Well-known examples of GLMs include linear regression, phase retrieval, 1-bit compressed sensing, and logistic regression. We focus on the high-dimensional setting in which both the number of measurements n and the signal dimension d diverge, with their ratio tending to a fixed constant. Linear and spectral methods are two popular solutions to obtain an initial estimate, which are also commonly used as a 'warm start' for other algorithms. In particular, the linear estimator is a data-dependent linear combination of the columns of the design matrix, and its analysis is quite simple; the spectral estimator is the principal eigenvector

of a data-dependent matrix, whose spectrum exhibits a phase transition. In this talk, I will start by discussing the emergence of this phase transition and provide precise asymptotics on the high-dimensional performance of the spectral method. Next, I will show how to optimally combine the linear and spectral estimators. Finally, I will add a 'twist' to the problem and consider the recovery of two signals from unlabeled data coming from a mixed GLM. Approximate message passing (AMP) algorithms (often used for high-dimensional inference tasks) will provide a powerful analytical tool to solve these problems.

20 Mehul Motani

National University of Singapore, Singapore
Studying Generalization in Deep Neural Networks

Abstract

What makes a learning algorithm have the ability to generalize, i.e., predict beyond the training data? Can we predict when a learning algorithm will generalize well? We believe that a clear answer to these questions is still elusive. In this talk, we will share our perspectives on understanding generalization in deep neural networks. This includes approaches based on a recently proposed novel complexity measure, called Kolmogorov Growth (KG), and a recently proposed information theoretic measure, called sliced mutual information (SMI). Our work is aimed towards combating the high complexity and dimensionality of neural networks, understanding how neural networks generalize, and improving generalization performance.

21 Frederique Oggier

Nanyang Technological University, Singapore
Entropy-based Centrality and Clustering

Abstract

Centrality is a concept used in the context of complex networks, to measure node importance. Graph clustering addresses the questions of grouping nodes that are 'most similar'. This talk will present different variations of centrality

based on Shannon's entropy, and some applications of these centralities for graph clustering.

22 Sewoong Oh

University of Washington, USA
The power of adaptivity in representation learning: From meta-learning to federated learning

Abstract

A central problem in machine learning is as follows: How should we train models using data generated from a collection of heterogeneous tasks/environments, if we know that these models will be deployed in a new and unseen environment? In the setting of few-shot learning, a prominent approach is to develop a modeling framework that is "primed" to adapt, such as Model Adaptive Meta Learning (MAML) and then fine tune the model for the deployment environment. We study this approach in the multi-task linear representation setting. We show that the reason behind generalizability of the models in new environments is that the dynamics of training induces the models to evolve toward the common data representation among the various tasks. The structure of the bi-level update at each iteration (an inner and outer update with MAML) holds the key — the diversity among client data distributions are exploited via inner/local updates. This is the first result that formally shows representation learning, and derives exponentially fast convergence to the ground-truth representation. I will conclude by making a connection between MAML and Federated Average (FedAvg) in the context of personalized federated learning, where the the local and global updates of FedAvg exhibits the same representation learning. This is based on joint work with Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Paper: https://arxiv.org/abs/2202.03483

23 Ayfer Ozgur

Stanford University, USA
Information Constrained Optimal Transport: From Relay to Generative Adversarial Networks

Abstract

In this talk, we will discuss the notion of information constrained Wasserstein distance. We will present an upper bound on this quantity that corresponds to a generalization and strengthening of a celebrated inequality by Talagrand and show how it leads to the solution of a problem on the relay channel posed by Cover in 1980's. We will then discuss how the information constraint can be used as a regularizer for training Wasserstein GANs. We will investigate the impact of the regularization on the optimal solution in a benchmark setting and show that it improves the quality of the generator learned from empirical data by removing the curse of dimensionality.

24 Dimitris Papailiopoulos

University of Wisconsin-Madison, USA

Looped Transformers are Universal Computers and Prompts are their Programs

Abstract

We demonstrate the potential of transformer (TF) networks to serve as universal computers by hard-coding them with specific weights and placing them inside a loop, where prompts serve as programs. We show how a constant number of TF layers can emulate basic compute blocks, including lexicographic operations, non-linear functions, function calls, a program counter, and a conditional branch command. By piecing together these blocks and placing them inside a loop, we demonstrate the programmability and versatility of transformer networks. Although the Turing Completeness of these models is known, in this work we provide explicit, deterministic constructions of TFs that emulate complex compute boxes such as a basic calculator, a basic linear algebra library, a learning algorithm, and a single-instruction, universal computer. We construct these models without any training, but rather by providing explicit weight matrices that emulate the desired functionality. We hope that our results provide new possibilities for the use of transformer networks as programmable compute boxes.

25 Galen Reeves

Duke University, USA

Inference from heterogeneous pairwise data

Abstract

High-dimensional inference problems involving heterogeneous pairwise observations arise in a variety of applications, including covariance estimation, clustering, and community detection. In this talk I will present a unified approach for the analysis of these problems that yields exact formulas for both the fundamental and algorithmic limits. The high-level idea is to model the observations using a linear Gaussian channel whose input is the tensor product of the latent variables. The limits of this general model are then described by a finite-dimensional variational formula, which provides a decoupling between the prior information about the latent variables (usually a product measure) and the specific structure of the observations. I will discuss some of the key ideas in the proof of the fundamental limits, which is based on the adaptive interpolation method. I will also provide examples of how this approach can be applied to problems involving covariate assisted clustering and spiked matrix models with heteroscedastic noise. Based on joint work with Josh Behne and Riccardo Rossetti.

26 Miguel Rodrigues

University College London, UK

Generalization Behaviour of Learning Algorithms: Recent Information-Theoretic Advances

Abstract

This talk overviews recent advances in the characterization of the generalization behaviour of learning algorithms using information-theoretic oriented tools. It covers results in the area of supervised learning, transfer learning, and meta-learning.

27 Cynthia Rush

Columbia University, USA

On the Robustness to Misspecification of α-Posteriors and Their Variational

Approximations

Abstract

Variational inference (VI) is a machine learning technique that approximates difficult-to-compute probability densities by using optimization. While VI has been used in numerous applications, it is particularly useful in Bayesian statistics where one wishes to perform statistical inference about unknown parameters through calculations on a posterior density. In this talk, I will review the core concepts of VI and introduce some new ideas about VI and robustness to model misspecification. In particular, we will study -posteriors, which distort standard posterior inference by downweighting the likelihood, and their variational approximations. We will see that such distortions, if tuned appropriately, can outperform standard posterior inference when there is potential parametric model misspecification. This is joint work with Marco Avella, Jose Montiel Olea, and Amilcar Velez.

28 Cong Shen

The University of Virginia, USA

The Role of Random Orthogonality in Federated Learning

Abstract

Federated learning (FL) focuses on many clients collaboratively training a machine learning model under the coordination of a central server while keeping the local data private at each client. It is an emerging distributed machine learning paradigm that has many attractive properties, bridging different disciplines such as machine learning, communications, networking, and security/privacy. In this talk, we highlight the benefits and potential of this interdisciplinary view by introducing random orthogonality, a concept that originates from physical-layer communications and signal processing but is under-explored in FL.

In the first example of random orthogonality, we show that by tightly coupling FL and two unique characteristics of massive MIMO – channel hardening and favorable propagation – we are able to achieve natural over-the-air model aggregation without requiring transmitter side channel state information (CSI) for the uplink phase of FL, while significantly reducing the channel

estimation overhead at the receiver. We extend this principle to the downlink communication phase and develop a simple but highly effective model broadcast method for FL. We also relax the massive MIMO assumption by proposing an enhanced random orthogonalization design for both uplink and downlink FL communications, that does not rely on channel hardening or favorable propagation.

In the second example, we ask how we can achieve random orthogonality for FL when there is no "naturally born" orthogonal signals. We borrow the ideas from CDMA and proactively introduce orthogonality to the uplink model aggregation using orthogonal sequences. We propose FLORAS, a privacy-preserving random orthogonality method for FL model aggregation. FLORAS enjoys most of the benefits of the first example (with massive MIMO), and we further prove that it achieves a variety of differential privacy (DP) guarantees. In particular, we theoretically show an interesting tradeoff between the FL model convergence rate and achieved DP guarantee levels, via simple adjustment of the FLORAS parameter configuration.

29 Yan Shuo Tan

National University of Singapore, Singapore
Understanding and overcoming the statistical limitations of decision trees

Abstract

Decision trees are important both as interpretable models, amenable to high-stakes decision-making, and as building blocks of ensemble methods such as random forests and gradient boosting. Their statistical properties, however, are not yet well understood. In particular, it is unclear why there is a prediction performance gap between them and powerful but uninterpretable machine learning methods including random forests, gradient boosting, and deep learning.

We partially explain this performance gap by proving sharp squared error generalization lower bounds for any decision tree fitted to a sparse additive generative model. By connecting decision tree estimation with rate-distortion theory, we establish a bound that is surprisingly much worse than the minimax rate for estimating such models, which is attainable via penalized kernel methods. This inefficiency is not due to any facet of the tree-growing procedure, but to the loss in power for detecting global structure when we average

responses solely over each leaf. Empirical results demonstrate that random forests also suffer a similar weakness.

Using this new insight, we propose Fast Interpretable Greedy-Tree Sums (FIGS), a method which grows a flexible number of trees simultaneously via a form of adaptive backfitting. We show theoretically and via simulations that FIGS can disentangle additive components in a model without knowing the block structure, thereby achieving better prediction performance than methods based on single trees. In particular, we give theoretical evidence that it can achieve the minimax rate of $O(d * n^{-2/3})$ for additive models with C^1 component functions. Furthermore, extensive experiments demonstrate that it also achieves impressive prediction performance on real world data sets while maintaining interpretability via a small number total splits.

30 Ali Tajer

Rensselaer Polytechnic Institute, USA
Causal Bandits

Abstract

In this talk, we provide an overview of the causal bandit problems. The purpose of causal bandit settings is to formalize theoretically principled frameworks for the experimental design when the experiments involve an array of parameters that causally affect one another. The key objective of causal bandits is to leverage causal relationships to design effective experiments judiciously. Designing causal bandit algorithms critically hinges on the extent of information available about the (i) causal structure and (ii) the interventional distributions. Based on the availability of information on each of these two dimensions, there are, broadly, four possible model combinations. The existing literature, for the most part, focuses on settings in which the interventional distributions are known (with or without knowing the causal structure). First, we provide an overview of the existing literature on the existing literature. Secondly, motivated by the fact that acquiring the interventional distributions is often infeasible, we address the following question: is it possible to achieve the optimal regret scaling rates without knowing the interventional distributions? We address this question affirmatively in the case of linear structural equation models when the causal structure is known.

We discuss the design and performance of algorithms for the frequentist and Bayesian settings.

31 Himanshu Tyagi

Indian Institute of Science, India
Lower Bounds for Discrete Distribution Testing Under Information Constraints

Abstract

Discrete distribution testing problems are composite hypothesis testing problems where we seek to determine if the observed samples are from a given distribution or not. A modern variant of is where the tests cannot access the complete samples, but can use only limited information about them. For instance, the tests can access samples perturbed with noise to maintain privacy or quantized to respect communication constraints. Over the past few years, we have derived a general recipe for establishing tight lower bounds for such problems. Our recipe involves information-theoretic methods to carefully quantify how much the distances between distributions shrink due to information constraints. In this talk, you will see applications of our recipe to obtain results for uniformity testing under communication and privacy constraints, using both non-interactive and interactive protocols. You will see that public-coin protocols are far superior to private-coin protocols and that there are constraints for which interactive protocols are far superior to even public-coin protocols. You will also realise that this is an on-going research thread with many open directions.

This work is a part of a long-term collaboration with Jayadev Acharya and Clément Canonne, with many other collaborators for specific results.

32 Antonios Varvitsiotis

Singapore University of Technology and Design, Singapore
Multiplicative Updates for Symmetric-cone Factorizations

Abstract

Given an m-by-n nonnegative matrix X with non-negative entries, the factorization problem over a cone K is to compute $\{a_1, \ldots, a_m\} \subseteq K$ and

$\{b_1, \ldots, b_n\} \subseteq K^*$ belonging to its dual so that $X_{ij} = \langle a_i, b_j \rangle$ for all $i \in [m], j \in [n]$. Cone factorizations are fundamental to mathematical optimization as they allow us to express convex bodies as feasible regions of linear conic programs. In this paper, we introduce and analyze the symmetric-cone multiplicative update (SCMU) algorithm for computing cone factorizations when K is symmetric; i.e., it is self-dual and homogeneous. Symmetric cones are of central interest in mathematical optimization as they provide a common language for studying linear optimization over the nonnegative orthant (linear programs), over the second-order cone (second order cone programs), and over the cone of positive semidefinite matrices (semidefinite programs). The SCMU algorithm is multiplicative in the sense that the iterates are updated by applying a meticulously chosen automorphism of the cone computed using a generalization of the geometric mean to symmetric cones. Using an extension of the Lieb's concavity theorem and the von Neumann's trace inequality to symmetric cones, we show that the squared loss objective is non-decreasing along the trajectories of the SCMU algorithm. When specialized to the nonnegative orthant, the SCMU algorithm corresponds to the seminal algorithm by Lee and Seung for computing Nonnegative Matrix Factorizations. Lastly, we use the SCMU algorithm to formulate conjectures on the (non) existence of hybrid lifts (i.e., extended formulations over products of positive semidefinite and second-order cones) for regular polygons. Joint work with Yong Sheng Soh.

33 Yao Xie

Georgia Institute of Technology, USA

Invertible Neural Networks for Graph Prediction

Abstract

Graph prediction problems prevail in data analysis and machine learning. The inverse prediction problem, namely to infer input data from given output labels, is of emerging interest in various applications. In this work, we develop *invertible graph neural network* (iGNN), a deep generative model to tackle the inverse prediction problem on graphs by casting it as a conditional generative task. The proposed model consists of an invertible sub-network that maps one-to-one from data to an intermediate encoded feature, which allows forward prediction by a linear classification sub-network as well as

efficient generation from output labels via a parametric mixture model. The invertibility of the encoding sub-network is ensured by a Wasserstein-2 regularization which allows free-form layers in the residual blocks. The model is scalable to large graphs by a factorized parametric mixture model of the encoded feature and is computationally scalable by using GNN layers. The existence of invertible flow mapping is backed by theories of optimal transport and diffusion process, and we prove the expressiveness of graph convolution layers to approximate the theoretical flows of graph data. The proposed iGNN model is experimentally examined on synthetic data, including the example on large graphs, and the empirical advantage is also demonstrated on real-application datasets of solar ramping event data and traffic flow anomaly detection. This is a joint work with Chen Xu at Georgia Tech, and Xiuyuan Cheng at Duke University.