Abstracts - Workshop on Algorithmic Challenges in Machine Learning

Wednesday January 7

Title: Hard Online Learning Problems
Speaker: Ofer Dekel, Microsoft Research

Abstract: The difficulty of an online learning problem is typically measured by its minimax regret. If the minimax regret grows sublinearly with the number of online rounds (denoted by T), we say that the problem is learnable. Until recently, we recognized only two classes of online learning problems: problems whose minimax regret grows at a slow rate of O(\sqrt(T)), and unlearnable problems with linear minimax regret. This talk is about a fascinating new class of problems whose minimax regret grows at a rate of \theta(T^{2/3}). These problems are still learnable, because their minimax regret is sublinear in the number of rounds, but they are strictly harder than standard online learning problems. I will present two problems from this class: the multi-armed bandit with switching costs and the family of online learning problems defined by a feedback graph.

Joint work with: Nicolo Cesa-Bianchi, Jian Ding, Tomer Koren, and Yuval Peres

Title: Efficient minimax strategies for online prediction
Speaker: Peter Bartlett, UC Berkeley and QUT

Abstract: Consider a prediction game in which, in each round, a strategy makes a decision, then observes an outcome and pays a loss. The aim is to minimize the regret, which is the amount by which the total loss incurred exceeds the total loss of the best decision in hindsight. We study the case where decisions and outcomes lie in a convex subset of a Hilbert space, and loss is squared distance. When the set is the simplex, this is the `Brier game', studied for the calibration of sequential probability forecasts; when it is the Euclidean ball, the game is related to sequential Gaussian density estimation. We show that the value of the game depends only on the radius of the smallest ball that contains the convex subset, and that the minimax optimal strategy is a simple shrinkage strategy that can be efficiently computed, given the center of the smallest ball.

Joint work with Wouter Koolen and Alan Malek

Title: Towards Optimal Algorithms for Prediction with Expert Advice
Speaker: Yuval Peres, Microsoft Research

Abstract: We study the classical problem of prediction with expert advice in the adversarial setting with a geometric stopping time. Cover (1965) gave the optimal algorithm for the case of 2 experts. In this paper, we design the optimal algorithm, adversary and regret for the case of 3 experts. Further, we show that the optimal algorithm for 2 and 3 experts is a probability matching algorithm (analogous to Thompson sampling) against a particular randomized adversary. Remarkably, it turns out that this algorithm is not only optimal against this adversary, but also minimax optimal against all possible adversaries. We establish a constant factor separation between the regrets achieved by the optimal algorithm and the widely used multiplicative weights algorithm. Along the way, we improve the regret lower bounds for the multiplicative weights algorithm for an arbitrary number of experts and show that this is tight for 2 experts. A novel aspect of our analysis is that upper and lower bounds are proved simultaneously, analogous to the primal-dual method. The analysis of the optimal adversary relies on delicate random walk estimates. We further use this connection to develop an improved regret bound for the case of 4 experts, and provide a general framework for designing the optimal algorithm for an arbitrary number of experts.

Joint work with Nick Gravin and Balu Sivan

Title: Volumetric Spanners: an Efficient Exploration Basis for Learning
Speaker: Zohar Karnin, Yahoo Labs

Abstract: Numerous learning problems require a mechanism for action space exploration at their core. Exploration broadly refers to the process of making choices in light of uncertainty. Such is the case of a designer of medical experiments choosing patients to examine, a router choosing a path in a network without knowing the congestion ahead of time, and a search engine displaying ads of a search result without knowing the user preference.

Numerous learning problems that contain exploration, such as experiment design, multi-arm bandits, online routing, search result aggregation and many more, have been studied extensively in isolation. In this paper we consider a generic and efficiently computable method for action space exploration based on convex geometry.

We define a novel geometric notion of an exploration mechanism with low variance called volumetric spanners, and give efficient algorithms to construct such spanners. We describe applications of this mechanism to the problem of optimal experiment design and the general framework for decision making under uncertainty of bandit linear optimization. For the latter we give efficient and near-optimal regret algorithm over general convex sets. Previously such results were known only for specific convex sets, or under special conditions such as the existence of an efficient self-concordant barrier for the underlying set.

Title: Beyond Disagreement-based Agnostic Active Learning
Speaker: Kamalika Chaudhuri, UCSD

Abstract: We study agnostic active learning, where the goal is to learn a classifier in a pre-specified hypothesis class interactively with as few label queries as possible, while making no assumptions on the true function generating the labels. The main algorithms for this problem are disagreement-based active learning, which has a high label requirement, and margin-based active learning, which needs fewer labels, but only applies to fairly restricted problems. Thus a challenge is to find an algorithm which achieves better label complexity, is consistent in an agnostic setting, and applies to general classification problems.

In this talk, we provide such an algorithm. Our solution is based on two novel contributions -- a reduction from consistent active learning to confidence-rated prediction with guaranteed error, and a novel confidence-rated predictor.

This talk is based on joint work with Chicheng Zhang.

Thursday January 8

Title: Beyond Locality Sensitive Hashing
Speaker: Piotr Indyk, MIT

Abstract: Locality Sensitive Hashing (LSH) is a framework for designing data structures for the approximate Nearest Neighbor Search problem in high-dimensional spaces. It relies on the existence of efficiently computable random mappings (LSH functions) with the property that the probability of collision between two points is related to the distance between them. The framework is applicable to a collection of distances and similarity functions, including the Euclidean distance. For the latter metric, it is known that the "basic" application of the LSH function yields a 2-approximate algorithm with a query time of roughly dn^(1/4), for a set of n points in a d-dimensional space. It is also known that, for the "basic" LSH algorithm, this bound is tight.

In this talk I will give present new data structures that offer significant improvements over the aforementioned bounds. In the first part, I will describe the first algorithm that reduces the exponent in the query time below the aforementioned bound, for a large enough approximation factor (this is joint work with Alex Andoni, Huy Nguyen, and Ilya Razenshteyn, appearing in SODA'14). The improvement is achieved by performing *data-dependent* hashing, which enables overcoming the lower bounds. If time allows I will also give an overview of the very recent result due to Andoni and Razenshteyn that offers a 2-approximation guarantee with a dn^(1/7) query time. The latter result matches a natural "data dependent hashing" barrier.

Title: Property testing for machine learning problems
Speaker: Avrim Blum, Carnegie Mellon University

Abstract: One motivation for property testing is the idea that testing could serve as a cheap estimation step before learning. E.g., before embarking on some large medical study, perhaps with just a few samples we can first determine if we are measuring the right features. However, property testing algorithms generally require the ability to query for labels of arbitrary fictitious examples, making them unusable for most machine learning applications. Instead, we would like testing algorithms that only ask examples to be labeled from among those that actually appear in a polynomial-size unlabeled sample (i.e., as in active-learning rather than membership query learning).

In this work, we consider property testing under these constraints and show it can still yield significant benefits for interesting concept classes for learning, including unions of intervals and linear separators, as well as for various assumptions used in semi-supervised learning. For example, we show testing unions of d intervals can be done with O(1) label requests, independent of the VC-dimension d (which means there will exist a consistent union of intervals even when the tester outputs "no"). We also give new results for passive testing (considered by Kearns and Ron), where the algorithm must pay for labels on every example drawn from the underlying distribution. In the case of testing linear separators in R^n, we show that both active and passive testing can be done with O(sqrt{n}) queries, compared to the Omega(n) needed for learning. We also give a general combination result for active testing, as well as develop notions of the “testing dimension" of a given property that characterize the intrinsic number of label requests needed to test that property. This work brings up a number of interesting open problems about testing in this framework as well.

This is joint work with Nina Balcan, Eric Blais, and Liu Yang

Title: Towards Instance Optimal Testing and Learning
Speaker: Greg Valiant, Stanford University

Abstract: Much of the work on estimating and testing properties of distributions has focused on determining worst-case bounds on the sample complexities of these problems. Here we initiate a study of how to understand the difficulty of specific instances, with an eye towards developing a theory of ``instance optimal'' testing, estimation, and learning. We begin by revisiting one of the most basic problems in distributional property testing: verifying the identity of a distribution. Given the description of a distribution, p, over a discrete support, how many samples (independent draws) must one obtain from an unknown distribution, q, to distinguish, with high probability, the case that p=q from the case that the total variation distance (L1 distance) is at least eps? We resolve this question, up to constant factors, on an instance by instance basis: there exist universal constants c, c' and a function f(p,eps) on distributions and error parameters, such that our tester distinguishes the two cases using f(p) samples with success probability 2/3 , but no tester can distinguish the case that p=q from the case that the total variation distance is at least c*eps when given c' f(p,eps) samples. The talk will conclude with a discussion of several directions for future work on instance-optimal testing and learning.

Title: Foundations For Learning in the Age of Big Data: Interactive and Distributed Machine Learning
Speaker: Nina Balcan, Carnegie Mellon University

Abstract: With the variety of applications of machine learning across science, engineering, and computing in the age of Big Data, re-examining the underlying foundations of the field has become imperative. In this talk, I will describe new models and algorithms for important modern paradigms, specifically, interactive learning and distributed learning.

For active learning, where the algorithm itself can ask for labels of carefully chosen examples from a large pool of unannotated data with the goal of minimizing human labeling effort, I will present results giving computationally efficient, optimal label complexity algorithms. I will also discuss unexpected implications of these results for classic supervised learning paradigms.

For distributed learning, I will discuss a model that for the first time addresses the core question of what are the fundamental communication requirements for achieving accurate learning. Broadly, we consider a framework where massive amounts of data is distributed among several locations, and our goal is to learn a low-error hypothesis with respect to the overall distribution of data using as little communication, and as few rounds of interaction, as possible. We provide broadly-applicable techniques for achieving communication-efficient learning both for supervised and unsupervised learning scenarios.

Title: The Reusable Holdout: Preserving Validity in Adaptive Data Analysis
Speaker: Moritz Hardt, IBM Research

Abstract: A great deal of effort has been devoted to reducing the risk of spurious scientific discoveries, from the use of holdout sets and sophisticated cross-validation techniques, to procedures for controlling the false discovery rate in multiple hypothesis testing. However, there is a fundamental disconnect between the theoretical results and the practice of science: the theory assumes a fixed collection of hypotheses to be tested, or learning algorithms to be applied, selected non-adaptively before the data are gathered, whereas science is by definition an adaptive process, in which data are shared and re-used, and hypotheses and new studies are generated on the basis of data exploration and previous outcomes.

Surprisingly, the challenges of adaptivity can be addressed using insights from differential privacy, a field of study supporting a definition of privacy tailored to private data analysis. As a corollary we show how to safely reuse a holdout set a great many times without undermining its validation power, even when hypotheses and computations are chosen adaptively. Armed with this technique, the analyst is free to explore the data ad libitum, generating and evaluating hypotheses, verifying results on the holdout, and backtracking as needed.

Joint work with Cynthia Dwork, Vitaly Feldman, Toni Pitassi, Omer Reingold and Aaron Roth

Friday January 9

Title: Follow the Leader with Dropout Perturbations
Speaker: Manfred Warmuth, University of California, Santa Cruz

Abstract: We consider online prediction with expert advice. Over the course of many trials, the goal of the learning algorithm is to achieve small additional loss (i.e.\ regret) compared to the loss of the best from a set of K experts. The two most popular algorithms are Hedge/Weighted Majority and Follow the Perturbed Leader (FPL). The latter algorithm first perturbs the losses of each expert (assumed to lie in [0,1]) by independent additive noise drawn from a fixed distribution, and then predicts with the expert of minimum perturbed loss (``the leader'') where ties are broken uniformly at random. To achieve the optimal worst-case regret as a function of the loss L^* of the best expert in hindsight, the two types of algorithms need to tune their learning rate or noise magnitude, respectively, as a function of L^*.

Instead of perturbing the losses of the experts in each trial with additive noise we use dropout, i.e. we randomly set each loss to 0 with probability half, and use Follow the Leader on the total dropout losses of the experts. For non-binary losses we need to make the losses binary and let dropout probability dependent on the losses.

All our proofs are rather elementary. We show that this simple, tuning-free version of the FPL algorithm achieves two feats: optimal worst-case O(\sqrt{L^* \ln K} + \ln K) regret as a function of L^* , and optimal O(\ln K) regret when the loss vectors are drawn i.i.d. from a fixed distribution and there is a gap between the expected loss of the best expert and all others.

A number of recent algorithms from the Hedge family (AdaHedge and FlipFlop) also achieve this, but they employ sophisticated tuning regimes. The dropout perturbation of the losses of the experts result in different noise distributions for each expert (because they depend on the expert's total loss) and curiously enough no additional tuning is needed: the choice of dropout probability only affects the constants.

Joint work with Tim Van Erven and Wojciech Kotlowski

Title: Simple, Efficient, and Neural Algorithms for Sparse Coding
Speaker: Ankur Moitra, MIT

Abstract: Sparse coding is a basic task in many fields including signal processing, neuroscience and machine learning where the goal is to learn a basis that enables a sparse representation of a given set of data, if one exists. Its standard formulation is as a non-convex optimization problem which is solved in practice by heuristics based on alternating minimization. There has been considerable recent work on designing algorithms for sparse coding with provable guarantees, but somewhat surprisingly these simple heuristics outperform them in practice. Here we give a general framework for understanding alternating minimization which we leverage to analyze existing heuristics and to design new ones also with provable guarantees.

We study this problem in a natural generative model, and obtain a variety of new algorithmic results: We give the first neurally plausible algorithm --- closely related to the original heuristic of Olshausen and Field --- that (provably) converges to a globally optimal sparse code. We also give the first algorithm for sparse coding that works almost up to the information theoretic limit for sparse recovery on incoherent dictionaries. All previous algorithms that approached or surpassed this limit run in time exponential in some natural parameter. Finally, our algorithms improve upon the sample complexity of existing approaches. We believe that our framework will have applications beyond sparse coding, and could be used to show that simple, iterative algorithms can be powerful in other contexts as well by suggesting new ways to analyze them.

This is based on joint work with Sanjeev Arora, Rong Ge and Tengyu Ma