Contents

Abstracts

Algorithms and Foundations for Data Science

(30 May–10 Jun 2022)

1 Sepehr Assadi

Rutgers University, USA
Brooks' theorem in graph streams

Abstract

Every graph with maximum degree Δ can be colored with $\Delta+1$ colors using a simple greedy algorithm. Remarkably, recent work has shown that one can find such a coloring even in the semi-streaming model: there exists a randomized algorithm that with high probability finds a $(\Delta+1)$-coloring of the input graph in only O(n polylog n) space assuming a single pass over the edges of the graph in any arbitrary order.

But, in reality, one almost never needs $\Delta+1$ colors to properly color a graph. Indeed, the celebrated Brooks' theorem states that every (connected) graph beside cliques and odd cycles can be colored with Δ colors. Can we find a Δ-coloring in the semi-streaming model as well?

In this talk, I will describe a semi-streaming algorithm to find a Δ-coloring and prove a "Streaming Brooks' Theorem".

Joint work with Pankaj Kumar and Parth Mittal.

2 Ainesh Bakshi

Carnegie Mellon University, USA
Low-rank approximation with $(1/\epsilon)^(1/3)$ matrix-vector products

Abstract

In this talk, we consider iterative methods based on Krylov subspaces for low-rank approximation under any Schatten-p norm. Here, given access to a matrix A through matrix-vector products, an accuracy parameter ϵ, and a target rank k, the goal is to find a rank-k matrix Z with orthonormal columns such that $\|A(I - ZZ^\top)\|_{S_p} \leq (1+\epsilon) \min_{U^\top U = I_k} \|A(I - UU^\top)\|_{S_p}$, where $\|M\|_{S_p}$ denotes the ℓ_p norm of the the singular values of M. For the special cases of $p = 2$ (Frobenius norm) and $p = \infty$ (Spectral norm), Musco and Musco (NeurIPS 2015) obtained an algorithm based on Krylov methods that uses $\tilde{O}(k/\sqrt{\epsilon})$ matrix-vector products, improving on the naïve $\tilde{O}(k/\epsilon)$ dependence obtainable by the power method.

Our main result is an algorithm that uses only $\tilde{O}(kp^{1/6}/\epsilon^{1/3})$ matrix-vector products, and works for *all*, not necessarily constant, $p \geq 1$. For $p = 2$ our bound improves the previous $\tilde{O}(k/\epsilon^{1/2})$ bound to $\tilde{O}(k/\epsilon^{1/3})$. Since the Schatten-$p$ and Schatten-∞ norms of any matrix are the same up to a $1+\epsilon$ factor when $p \geq (\log d)/\epsilon$, our bound recovers the result of Musco and Musco for $p = \infty$. Further, we prove a matrix-vector query lower bound of $\Omega(1/\epsilon^{1/3})$ for *any* fixed constant $p \geq 1$, showing that surprisingly $\tilde{\Theta}(1/\epsilon^{1/3})$ is the optimal complexity for constant k.

To obtain our results, we introduce several new techniques, including optimizing over multiple Krylov subspaces simultaneously, and pinching inequalities for partitioned operators. Our lower bound for $p \in [1, 2]$ uses the Araki-Lieb-Thirring trace inequality, whereas for $p > 2$, we appeal to a norm-compression inequality for aligned partitioned operators.

Based on joint work with Ken Clarkson and David Woodruff.

3 Petros Drineas

Purdue University, USA
Randomized linear algebra for interior point methods

Abstract

Linear programming is a central problem in computer science and applied mathematics with numerous applications across a wide range of domains, including machine learning and data science. Interior point methods (IPMs) are a common approach to solving linear programs with strong theoretical guarantees and solid empirical performance. The time complexity of these

methods is dominated by the cost of solving a linear system of equations at each iteration. In common applications of linear programming, particularly in data science and scientific computing, the size of this linear system can become prohibitively large, requiring the use of iterative solvers which provide an approximate solution to the linear system. Approximately solving the linear system at each iteration of an IPM invalidates common analyses of IPMs and the theoretical guarantees they provide. In this talk we will discuss how randomized linear algebra can be used to design and analyze theoretically and practically efficient IPMs when using approximate linear solvers.

4 Dan Feldman

University of Haifa, Israel
Coresets for decision trees of signals

Abstract

A k-decision tree t (or k-tree) is a recursive partition of a matrix (2D-signal) into $k \geq 1$ block matrices (axis-parallel rectangles, leaves) where each rectangle is assigned a real label. Its regression or classification loss to a given matrix D of N entries (labels) is the sum of squared differences over every label in D and its assigned label by t. Given an error parameter $\varepsilon \in (0, 1)$, a (k, ε) -coreset C of D is a small summarization that provably approximates this loss to *every* such tree, up to a multiplicative factor of $1 \pm \epsilon$. We provide the first algorithm that outputs such a small (k, ε)-coreset for *every* such matrix D. This is by forging a link between decision trees from machine learning – to partition trees in computational geometry.

Based on a NeurIPS spotlight paper with Jubran, Newman, and Shayda.

5 Anna Gilbert

Yale University, USA
TBA

Abstract

6 Mohammad Taghi Hajiaghayi

University of Maryland, USA

Massively parallel algorithms for maximal matching and edit distance

Abstract

In this talk we will discuss the recent algorithmic progress made on the Massively Parallel Computations (MPC) model. The MPC model provides a clean theoretical abstraction of modern parallel computation frameworks such as MapReduce, Hadoop, Spark, etc., which have been extremely successful in processing large-scale data-sets.

Our main focus in the talk will be on the maximal matching problem. We give an outline of the analysis of an extremely simple algorithm, improving exponentially over previous maximal matching results. The analysis is based on a novel method of proving concentration bounds for algorithms satisfying a certain "locality" property, which we believe may find applications beyond the MPC model. We will also survey some other recent results in the area. Particularly, we will overview an algorithm for edit distance and longest common subsequence with almost tight bounds.

7 Rajesh Jayaram

Google Research NYC, USA

New streaming algorithms for high-dimensional EMD and MST

Abstract

In this talk, we discuss new algorithms and techniques for high-dimensional geometric streaming problems. Specifically, we study the problem of approximating the cost of a Minimum Spanning Tree (MST) of an n-point set $X \subset \{1, 2, \ldots, \Delta\}^d$, and computing the Earth Mover Distance (EMD) between two multi-sets $A, B \subset \{1, 2, \ldots, \Delta\}^d$ of size n, where points can be added and removed in an arbitrary order stream. We give the first streaming algorithms for these problems which achieve an approximation factor of $\tilde{O}(\log n)$ using $\mathrm{poly}(\log(n, d, \Delta))$-space. For MST, our algorithm uses only a single pass, and for EMD the algorithm uses two-passes, but can be compressed to a single pass with a small additive error. Previously, the best known sublinear-space streaming algorithms for either prob-

lem achieved an approximation of $O(\min\{\log n, \log(\Delta d)\} \log n)$) [Andoni-Indyk-Krauthgamer '08, Backurs-Dong-Indyk-Razenshteyn-Wagner '20].

Our algorithms are based on an improved analysis of a recursive space partitioning method known generically as the Quadtree. Specifically, we show that the Quadtree achieves an $\widetilde{O}(\log n)$ approximation for both EMD and MST, improving on the $O(\min\{\log n, \log(\Delta d)\} \log n)$ approximation of [Andoni-Indyk-Krauthgamer '08, Backurs-Dong-Indyk-Razenshteyn-Wagner '20].

Based on a joint work with Xi Chen, Amit Levi, and Erik Waingarten

8 Shaofeng Jiang

Peking University, China

Streaming facility location in high dimension

Abstract

In *Euclidean Uniform Facility Location*, the input is a set of clients in $\mathbb{R}^d$ and the goal is to place facilities to serve them, so as to minimize the total cost of opening facilities plus connecting the clients. We study the classical setting of dynamic geometric streams, where the clients are presented as a sequence of insertions and deletions of points in the grid $[\Delta]^d$, and we focus on the *high-dimensional regime*, where the algorithm's space complexity must be polynomial (and certainly not exponential) in $d \cdot \log \Delta$.

We present a new algorithmic framework, based on importance sampling from the stream, for $O(1)$-approximation of the optimal cost using only $\mathrm{poly}(d \cdot \log \Delta)$ space. This framework is easy to implement in two passes, one for sampling points and the other for estimating their contribution. Over random-order streams, we can extend this to a one-pass algorithm by using the two halves of the stream separately. Our main result, for arbitrary-order streams, computes $O(d^{1.5})$-approximation in one pass by using the new framework but combining the two passes differently. This improves upon previous algorithms that either need space exponential in d or only guarantee $O(d \cdot \log^2 \Delta)$-approximation, and therefore our algorithms for high-dimensional streams are the first to avoid the $O(\log \Delta)$-factor in approximation that is inherent to the widely-used quadtree decomposition. Our improvement is achieved by introducing a geometric hashing scheme in

streaming that maps points in $\mathbb{R}^d$ into buckets of bounded diameter, with the key property that every point set of small-enough diameter is hashed into at most $\mathrm{poly}(d)$ distinct buckets.

Finally, we complement our results by showing 1.085-approximation algorithm requires space exponential in $\mathrm{poly}(d \cdot \log \Delta)$, even for insertion-only streams.

9 Michael Kapralov

EPFL, Switzerland

Factorial lower bounds for (almost) random order streams

Abstract

We introduce and study the StreamingCycles problem, a random order streaming version of the Boolean Hidden Hypermatching problem that has been instrumental in streaming lower bounds over the past decade. In this problem the edges of a graph G, comprising n/ℓ disjoint length-ℓ cycles on n vertices, are partitioned randomly among n players. Every edge is annotated with an independent uniformly random bit, and the players' task is to output, for some cycle in G, the sum (modulo 2) of the bits on its edges, after one round of sequential communication.

Our main result is an $\ell^{\Omega(\ell)}$ lower bound on the communication complexity of StreamingCycles, which is tight up to constant factors in the exponent. Applications of our lower bound for StreamingCycles include an essentially tight lower bound for component collection in (almost) random order graph streams, making progress towards a conjecture of Peng and Sohler [SODA'18] and the first exponential space lower bounds for random walk generation.

10 Rasmus Kyng

ETH Zurich, Switzerland

Scalar and matrix Chernoff bounds from ℓ_∞-independence

Abstract

We present new scalar and matrix Chernoff-style concentration bounds for a broad class of probability distributions over the binary hypercube $\{0, 1\}^n$.

Motivated by recent tools developed for the study of mixing times of Markov chains on discrete distributions, we say that a distribution is ℓ_∞-independent when the infinity norm of its influence matrix $\mathcal{I}$ is bounded by a constant. We show that any distribution which is ℓ_∞-independent satisfies a matrix Chernoff bound that matches the matrix Chernoff bound for independent random variables due to Tropp. Our matrix Chernoff bound is a broad generalization and strengthening of the matrix Chernoff bound of Kyng and Song (FOCS'18). Using our bound, we can conclude as a corollary that a union of $O(\log |V|)$ random spanning trees gives a spectral graph sparsifier of a graph with $|V|$ vertices with high probability, matching results for independent edge sampling, and matching lower bounds from Kyng and Song.

11 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, ϵ, δ, 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 $m = \Omega(\epsilon^{O(\delta)}d^2)$, improving on the previous $\Omega(\epsilon^2 d^2)$ lower bound due to Nelson and Nguyen (ICALP 2014).

This is joint work with Mingmou Liu

12 Michael W. Mahoney

University of California, Berkeley, USA
Random matrix theory and modern machine learning

Abstract

Random matrix theory (RMT) has a long history and has proven useful in many areas, most recently modern machine learning (ML). Some of the most interesting and promising uses of RMT in ML go beyond a direct application of popular Wigner and Marcenko-Pastur ideas to linear models in a high-dimensional regime, and instead must account for the peculiar properties of modern ML systems. These properties include: that the data dimension is not vanishingly small compared to the number of data points; that there are "layers" with intermediate features that can be large, comparable to, or smaller in size than the number of data points; that there are strongly non-linear models; and that, compared to parameter vectors or matrices, we are in general more interested in (scalar) functionals of those things (e.g., regression error and classification accuracy). Here, we provide an overview of recent developments in RMT that go beyond traditional ideas and that are well-suited for the analysis of modern ML models. Central to our approach is understanding different behaviors of linear models in the high-dimensional regime compared to their low-dimensional counterparts (where infinite norm and operator norm of matrices must not be considered "equivalent"), and how these ideas extend to non-linear models.

13 Andrew Mcgregor

University of Massachusetts Amherst, USA
A guide to estimating entropy for the forgetful and impatient

Abstract

An infinite stream consists of iid samples from an unknown and arbitrary discrete distribution. You want to additively estimate the entropy of this distribution. But you are impatient and forgetful: how long do you need to observe the stream given that you only have a constant amount of memory? The talk presents a new upper bound improving upon previous work by Acharya et al. (NeurIPS 2019). This is joint work with Maryam Aliakbarpour, Jelani Nelson, and Erik Waingarten.

14 Cameron Musco

University of Massachusetts Amherst, USA
Sublinear time eigenvalue approximation via random sampling

Abstract

We study the problem of approximating the eigenvalues of a symmetric $ntimesn$ matrix A, whose entries are all bounded in magnitude by 1. We show that simply uniformly subsampling an $s \times s$ principal submatrix of A for $s = poly(\log n, 1/\epsilon)$ suffices to estimate all eigenvalues up to additive error $\epsilon \cdot n$. This result can be viewed as a concentration bound on the full eigenspectrum of a random submatrix, extending known bounds on just the top eigenvalue. When A is sparse, and its rows can be efficiently sampled with probabilities proportional to their sparsity, we present an improved error bound of $\epsilon \cdot \sqrt{nnz(A)}$. Even for the strictly easier problem of testing the existence of large negative eigenvalues, our result is the first that takes advantage of sparsity in A. Surprisingly, in this setting, simply randomly sampling a principal submatrix does not suffice – we also must carefully zero out sampled entries that lie at the intersection of low probability rows and columns. We conclude by discussing several exciting open questions regarding sublinear time approximation algorithms for bounded entry matrices, and other restricted matrix classes.

15 Christopher Musco

New York University, USA
Linear and sublinear time spectral density estimation

Abstract

I will discuss new work on analyzing practically popular algorithms, including the kernel polynomial method (KPM) and moment matching method, for approximating the spectral density (eigenvalue distribution) of an n x n symmetric matrix A. We will see that natural variants of these algorithms achieve strong worst-case approximation guarantees: they can approximate any spectral density to epsilon accuracy in the Wasserstein-1 distance with roughly O(1/epsilon) matrix-vector multiplications with A. Moreover, we will show that the methods are robust to *in accuracy* in these matrix-vector

multiplications, which allows them to be combined with any approximation multiplication algorithm. As an application, we develop a randomized sublinear time algorithm for approximating the spectral density of a normalized graph adjacency or Laplacian matrices. The talk will cover the main tools used in our work, which include random importance sampling methods and stability results for computing orthogonal polynomials via three-term recurrence relations.

16 Jelani Nelson

University of California, Berkeley, USA
Optimal bounds for approximate counting

Abstract

Counting up to N deterministically of course takes Theta(log N) bits. In the first ever streaming algorithm, Morris in 1978 gave a randomized algorithm that improved upon this bound exponentially. The best known analysis of his algorithm shows that it gives a 1+eps approximation to N with probability at least 1-delta using O(loglog N + log(1/eps) + log(1/delta)) bits with high probability (the space usage is itself a random variable). We show that a very slight (but necessary) tweak of his algorithm actually achieves the better bound O(loglog N + log(1/eps) + loglog(1/delta)) bits, and we also show a new matching lower bound, establishing optimality. Our upper and lower bounds nearly match, up to small explicit constant factors.

Joint work with Huacheng Yu.

17 Huy Le Nguyen

Northeastern University, USA
Private frequency estimation via projective geometry

Abstract

In this work, we propose a new algorithm ProjectiveGeometryResponse (PGR) for locally differentially private (LDP) frequency estimation. For a universe size of k and with n users, our epsilon-LDP algorithm has communication cost

log(k) bits in the private coin setting and epsilon*log(e)+O(1) in the public coin setting, and has computation cost O(n+k*exp(epsilon)*log(k)) for the server to approximately reconstruct the frequency histogram, while achieving the state-of-the-art privacy-utility tradeoff. In many parameter settings used in practice this is a significant improvement over the $O(n + k^2)$ computation cost that is achieved by the recent PI-RAPPOR algorithm (Feldman and Talwar; 2021). Our empirical evaluation shows a speedup of over 50x over PI-RAPPOR while using approximately 75x less memory for practically relevant parameter settings. In addition, the running time of our algorithm is within an order of magnitude of HadamardResponse (Acharya, Sun, and Zhang; 2019) and RecursiveHadamardResponse (Chen, Kairouz, and Ozgur; 2020) which have significantly worse reconstruction error. The error of our algorithm essentially matches that of the communication- and time-inefficient but utility-optimal SubsetSelection (SS) algorithm (Ye and Barg; 2017). Our new algorithm is based on using Projective Planes over a finite field to define a small collection of sets that are close to being pairwise independent and a dynamic programming algorithm for approximate histogram reconstruction on the server side. We also give an extension of PGR, which we call HybridProjectiveGeometryResponse, that allows trading off computation time with utility smoothly. This is based on joint work with Vitaly Feldman, Jelani Nelson, and Kunal Talwar.

18 Rasmus Pagh

University of Copenhagen, Denmark
Differentially private CountSketch

Abstract

CountSketch is a core algorithm for data streams. In particular, it yields an efficient estimation of the number of occurrences of a given item in a stream. There are known approaches to making CountSketch differentially private with respect to stream elements, namely to initialize the empty sketch with independent noise instead of the zero vector. This can be shown to be differentially private if the magnitude of the noise is chosen proportional to the number of repetitions in CountSketch. It would thus seem that there is a trade-off between the number of repetitions (determining the error probability of CountSketch) and the error incurred by the noise. However, using

Gaussian noise we are able to show that the final error of the median estimator can be bounded independently of the number of repetitions, essentially matching the noise of (non-private) CountSketch and the noise needed to make a single counter differential private.

Joint work with Mikkel Thorup

19 Ely Porat

Bar-Ilan University, Israel
TBA

Abstract

20 Eric Price

University of Texas at Austin, USA
Finite-sample maximum likelihood estimation of location

Abstract

We consider 1-dimensional location estimation, where we estimate a parameter λ from n samples $\lambda+\eta_i$, with each η_i drawn i.i.d. from a known distribution f. For fixed f the maximum-likelihood estimate (MLE) is well-known to be optimal *in the limit* as $n \to \infty$: it is asymptotically normal with variance matching the Cramér-Rao lower bound of $\frac{1}{nI}$, where I is the Fisher information of f. However, this bound does not hold for finite n, or when f varies with n. We show for arbitrary f and n that one can recover a similar theory based on the Fisher information of a *smoothed* version of f, where the smoothing radius decays with n.

21 Chris Schwiegelshohn

Aarhus University, Denmark
Recent developments on coresets for clustering

Abstract

We consider center-based clustring problem where we are given a set of points P in some metric space and the objective consists of finding a set of at most k centers minimizing the sum of z-th powers of distances of every point to its assigned center. Popular special cases include k-median (z=1) and k-means (z=2). Our task is to compute a small weighted sketch S of the points P such that the weighted clustering cost of S approximates the clustering cost of P up to a (1+eps) factor for *any* candidate set of at most k centers. Such sketches are known as coresets and are a cornerstone of big data algorithms for many problems beyond clustering. In this talk will survey some recent results in this line of work and highlight potential future directions.

22 David P. Woodruff

Carnegie Mellon University, USA

Memory bounds for the experts problem

Abstract

Online learning with expert advice is a fundamental problem of sequential prediction. In this problem, the algorithm has access to a set of n "experts" who make predictions on each day. The goal on each day is to process these predictions, and make a prediction with the minimum cost. After making a prediction, the algorithm sees the actual outcome on that day, updates its state, and then moves on to the next day. An algorithm is judged by how well it does compared to the best expert in the set.

The classical algorithm for this problem is the multiplicative weights algorithm. Variations of this algorithm have been applied to and optimized for a broad range of problems, including boosting an ensemble of weak-learners in machine learning, and approximately solving linear and semi-definite programs. However, every application, to our knowledge, relies on storing weights for every expert, and uses Omega(n) memory. There is little work on understanding the memory required to solve the online learning with expert advice problem (or run standard sequential prediction algorithms, such as multiplicative weights) in natural streaming models, which is important when the number of experts, as well as the number of days on which the experts make predictions, is large.

We initiate the study of the learning with expert advice problem in the streaming setting, and show lower and upper bounds. Our lower bound for

i.i.d., random order, and adversarial order streams uses a novel masking technique and distributional detection game to show a smooth trade-off for regret versus memory. Our upper bounds in adversarial and random-order streams show ways to run standard sequential prediction algorithms in rounds on small "pools" of experts, thus reducing the necessary memory. For random-order streams, we show that our upper bound is tight up to low order terms.

Joint work with Vaidehi Srinivas, Ziyu (Neil) Xu, and Samson Zhou

23 Qin Zhang

Indiana University Bloomington, USA
Collaborative learning with limited communication

Abstract

In this talk, I will introduce my recent work on collaborative (reinforcement) learning (CL), in which multiple agents work together to learn an objective function. We are particularly interested in a scenario in which agent communication is very expensive. Our goal is to identify the tradeoffs between the speedup of the collaboration and the communication cost among the agents. We convey the following messages using a basic problem in bandit theory as a vehicle: (1) adaptive CL is more powerful than non-adaptive CL; (2) CL with non-IID data is harder than that with IID data; and (3) problems incomparable in the single-agent learning model can be separated in the CL model.

24 Samson Zhou

Carnegie Mellon University, USA
Near-linear sample complexity for L_p polynomial regression

Abstract

We study L_p polynomial regression. Given query access to a function $f : [-1, 1] \to$, the goal is to find a degree d polynomial $\widehat{q}$ such that, for a given parameter $\varepsilon > 0$,

$$\|\widehat{q} - f\|_p \leq (1 + \epsilon) \cdot \min_{q:\deg(q) \leq d} \|q - f\|_p. \quad (1)$$

Here $\cdot_p$ is the L_p norm, $g_p = (\int_{-1}^{1} |g(t)|^p dt)^{1/p}$. We show that querying f at points randomly drawn from the Chebyshev measure on $[-1, 1]$ is a near-optimal strategy for polynomial regression in any L_p norm. In particular, to output $\hat{q}$ satisfying 1, it suffices to sample $\frac{d,\text{poly}(p,\log d)}{\text{poly}\,\varepsilon}$ points from $[-1, 1]$ with probabilities proportional to this measure. While the polynomial regression problem is well understood for L_2 and L_∞, prior approaches for general p either only gave sample complexity linear in d for $1 \leq p \leq 2$, requiring a sub-optimal $\Omega(d^2)$ samples for general p, or only gave results for constant ε. We simultaneously overcome both of these limitations.

One of our main technical contributions is to provide explicit bounds on the L_p *Lewis weight function* of an infinite linear operator underlying the polynomial regression problem. Using tools from the orthogonal polynomial literature, we show that this function is closely related to the Chebyshev density. Our approach advances prior work, which studies explicit bounds on the L_2 leverage scores of infinite linear operators. A second contribution is to prove tighter bounds on L_p Lewis weight sampling for the polynomial operator than hold for general linear operators or matrices.

Joint work with Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff