ISSN 2186-7437

NII Shonan Meeting Report

No. 2016-10

Recent Advances in Randomized Numerical Linear Algebra

Ravindran Kannan
Michael Mahoney
David P. Woodruff

July 25–28, 2016

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Introduction and overview

Randomized Numerical Linear Algebra (RandNLA) is an interdisciplinary research area that exploits randomization as a computational resource to develop improved algorithms for large-scale linear algebra problems. From a foundational perspective, RandNLA has its roots in theoretical computer science (TCS), with deep connections to mathematics (convex analysis, probability theory, metric embedding theory) and applied mathematics (scientific computing, signal processing, numerical linear algebra). From an applied perspective, RandNLA is a vital new tool for machine learning, statistics, and data analysis. Well-engineered implementations have already outperformed highly-optimized software libraries for ubiquitous problems such as least squares (LS), with good scalability in parallel and distributed environments. RandNLA promises a sound algorithmic and statistical foundation for modern large-scale data analysis.

The great interdisciplinary strength of RandNLA is also one of the main challenges to its future progress. Researchers from diverse areas don't speak a common language, they don't publish in the same publication venues, they often can't evaluate the significance of important contributions of researchers from other areas. In this proposed workshop, we have assembled an interdisciplinary group of researchers who have made fundamental contributions to different aspects of RandNLA.

Our main goal is to advance mathematical aspects of RandNLA and to solve fundamental algorithmic and statistical challenges in making this theory useful in large-scale machine learning and data analysis applications. In particular, by bringing together these researchers in this workshop, we would like to develop a multi-faceted common theoretical foundation for state-of-the-art RandNLA methodologies and a detailed understanding of the complementary algorithmic and statistical issues that arise in their application to modern large-scale data analysis.

Background

Matrices are ubiquitous in computer science, statistics, and applied mathematics. An m x n matrix can encode information about m objects (each described by n features), or the behavior of a discretized differential operator on a finite element mesh; an n x n positive-definite matrix can encode the correlations between all pairs of n objects, or the edge-connectivity between all pairs of nodes in a social network; and so on. Motivated largely by technological developments that generate extremely large scientific and Internet data sets, recent years have witnessed exciting developments in the theory and practice of matrix algorithms. Particularly remarkable is the use of randomization typically assumed to be a property of the input data due to, e.g., noise in the data generation mechanism as an algorithmic or computational resource for the development of improved algorithms for fundamental matrix problems such as matrix multiplication, LS approximation, low-rank matrix approximation, Laplacian-based solvers, etc.

Each of the three organizers has written a separate NOW Publishers

monograph on different aspects of randomized numerical linear algebra, including: "Spectral Algorithms" by Kannan and Vempala in 2009, "Randomized Algorithms for Matrices and Data" by Mahoney in 2011, and "Sketching as a Tool for Numerical Linear Algebra" by Woodruff in 2014.

These monographs explain how sampling, sketching, and other randomized algorithmic techniques can be used for speeding up classical solutions to numerical linear algebra problems. Since it is an inherently interdisciplinary area, RandNLA can be approached from several different, yet complementary, perspectives.

Pure and applied mathematics. RandNLA has depended on progress in convex analysis and probability theory, making heavy use of, e.g., matrix measure concentration inequalities such as the matrix-Chernoff and matrix-Bernstein bounds. Dimensionality reduction ideas (such as fast random projections) have also been critical in RandNLA theory and practice and were used in NLA applications to construct highly efficient preconditioners for LS problems and to approximately solve low-rank approximation problems. Not surprisingly, there has been two-way traffic between RandNLA and the aforementioned areas, with progress in one area motivating and inspiring new results in the other.

Computer science. RandNLA initially grew out of work in TCS that analyzed the performance of algorithms that sample columns, rows, or elements from a matrix according to data-dependent non-uniform sampling probabilities. In parallel, but largely independently, ground-breaking progress was made on the theory and practice of solvers for systems of linear equations involving Laplacian matrices. Finally, sparse subspace-preserving embeddings have made RandNLA algorithms extremely efficient for massive sparse matrices.

Statistics and machine learning. RandNLA has received much attention in machine learning and statistics, despite the substantial "impedance mismatch" in problem formulation, parameterization, and objectives of interest. An obvious reason is that data sets are often modeled as matrices or graphs, and thus low-rank approximation, Laplacian-based problems, etc. arise naturally. In many cases, running time is a concern. See, e.g., for an example of applying RandNLA methods to CUR and Nystrom-based low-rank approximations in large-scale environments. A more subtle reason is that the randomization in RandNLA often "denoises" the input, and thus one observes implicit statistical regularization.

RandNLA in the field. Much of the interest in RandNLA has come from research areas that employ matrix algorithms as "black boxes." RandNLA has had important successes in such application domains for small- and medium-scale data, ranging from population genetics and biology to astronomy and mass spectroscopy imaging. RandNLA has thus far been studied from each of these very different perspectives largely in isolation. An important aspect of this workshop will be to address RandNLA in a more unified manner.

TALKS

[July 24 (Sun)]

Check in: 15:00-
Welcome Banquet: 19:00-21:00

[July 25 (Mon)]

Group Photo Shooting**:** 30min during lunch break

[July 25 (Mon)-28 (Thu)]*only AM for 28th

Coffee Break: 10:30/ 15:30
(Coffee and Snacks are served)

Talk Schedule:

Day 1:

Shonan video presentation 9:00-9:10

Ravi Kannan 9:10-10:10

Title: Some History and a story (of AA^T)

Abstract: Length-squared sampling was devised in the 90's to make matrix algorithms more efficient by sub-sampling the input matrix; it led to algorithms for Low-rank approximation, CUR approximation etc. Functional Analysts – Rudelson and Vershynin proved bounds on spectral norm error of length-squared sampling. Their original motivation was to pin down the number of samples to estimate the varience-covarience matrix of a probability density, which can be thought of as computing AA^T for an infinite matrix A. Their work was a crucial part of the Graph Sparsification results of Spielman and Srivatsava; they use what I call ``pre-conditioned length-squared sampling'', aka ``leverage scores'' (which will be discussed elsewhere in the workshop). An alternative to leverage scores is ``Volume Sampling'' which is a determinental process; recent advances prove that linear time suffices for this process.

Shengyu Zhang: 10:10-10:30

Title: Fast quantum algorithms for Least Squares Regression and Statistical Leverage Scores

Abstract: We present a quantum algorithm to generate a quantum state proportional to x^*, the unique solution to the least squares regression problem $\min \|Ax-b\|_2^2$. The algorithm takes $O(\log n)$ time for sparse and well-conditioned A. We will also give quantum algorithms for two regularized versions, i.e. ridge regression and δ-truncated SVD, with similar costs and solution approximation. Finally, we will show a quantum algorithm to approximately compute (or sample from the distribution of) statistical leverage score s_i in time $O(\log n)$ (for any fixed i) when A is sparse and well-conditioned.

Michael Kapralov: 10:50-11:25

Title: How to Fake Multiply by a Gaussian Matrix

Abstract: Have you ever wanted to multiply an $n \times d$ matrix X, with $n \gg d$, on the left by an $m \times n$ matrix $\tilde G$ of i.i.d. Gaussian random variables, but could not afford to do it because it was too slow? In this work we propose a new randomized $m \times n$ matrix T, for which one can compute $T \cdot X$ in only $O(\nnz(X)) + \tilde O(m^{1.5} \cdot d^{3})$ time, for which the total variation distance between the distributions $T \cdot X$ and $\tilde G \cdot X$ is as small as desired, i.e., less than any positive constant. Here $\nnz(X)$ denotes the number of non-zero entries of X. Assuming $\nnz(X) \gg m^{1.5} \cdot d^{3}$, this is a significant savings over the na\"ive $O(\nnz(X) m)$ time to compute $\tilde G \cdot X$. Moreover, since the total variation distance is small, we can provably use $T \cdot X$ in place of $\tilde G \cdot X$ in any application and have the same guarantees as if we were using $\tilde G \cdot X$, up to a small positive constant in error probability. We apply this transform to nonnegative matrix factorization (NMF) and support vector machines (SVM).

This is joint work with David Woodruff and Vamsi Potluru.

Jelani Nelson: 11:30-12:00

Title: Optimal approximate matrix product in terms of stable rank

Abstract: We prove, using the subspace embedding guarantee in a black box way, that one can achieve the spectral norm guarantee for approximate matrix multiplication with a dimensionality-reducing map having m = O(r~ / eps^2) rows. Here r~ is the maximum stable rank, i.e. squared ratio of the Frobenius norm to the operator norm, of the two matrices being multiplied (it is always at most the rank). This resolves the main open problem of (Magen, Zouzias SODA 2011). Our main theorem, via connections shown in prior work, implies quantitative improvements for approximate least squares regression and low rank approximation, has been applied by (Cohen et al, STOC 2015) in the analysis of dimensionality reduction for k-means clustering, and can be used in the context of dimensionality reduction applied to nonparametric regression by an analysis of (Yang, Pilanci, Wanwright Ann. Stat. 2015). We also separately point out that the proof of the "BSS" deterministic row-sampling result of can be modified to show that for any matrices A,Bof stable rank at most r~, one can achieve the spectral norm guarantee for approximate matrix multiplication of A^T B by deterministically sampling O(r~ / eps^2) rows that can be found in polynomial time. The original result of BSS was for rank instead of stable rank.

Joint work with Michael B. Cohen and David P. Woodruff.

Cameron Musco: 2:00-2:45

Title: Fast Low-Rank Approximation and PCA: Beyond Sketching

Abstract: I will discuss recent work on randomized algorithms for low-rank approximation and principal component analysis (PCA). The talk will focus on efforts that move beyond the extremely fast, but relatively crude approximations offered by random sketching algorithms.

In particular, we will see how advances in Johnson-Lindenstrauss projection methods have provided tools for improving the analysis of classic iterative SVD algorithms, including the block power method and block Krylov methods. The key insight is to view the iterative algorithms as denoising procedures for coarse sketching methods.

I will discuss how this view can be used to analyze a simple block Krylov method, showing that the algorithm gives (1+epsilon) near optimal PCA and low-rank approximation in just O(1/sqrt{epsilon}) iterations. Despite their long history, this analysis is the first of a Krylov subspace method that does not depend on the matrix's spectral gaps.

I will also survey promising work on approximate PCA via stochastic optimization, and faster techniques for singular value decomposition targeted at specific downstream tasks, such as principal component regression.

Ken Clarkson: 2:45-3:30

Title: Regularized Regression and Low-Rank Approximation via Sketching

Christopher Musco: 4:00-5:00

Title: Ridge Leverage Score Sampling for Low-rank Matrix and Kernel Approximation

Abstract: I'll discuss recent work on a new algorithm for finding a near optimal low-rank approximation of a matrix A in O(nnz(A)) time. The method is based on a recursive sampling scheme for computing a representative subset of A's columns, which is then used to find a low-rank approximation. In particular, sampling is performed using recently introduced ridge leverage scores, which we adapt for low-rank problems.

This approach differs substantially from prior O(nnz(A)) time algorithms, which are all based on fast Johnson-Lindenstrauss random projections. It matches the guarantees of these methods while offering a number of advantages.

Not only are sampling algorithms faster for sparse and structured data, but they can also be applied in settings where random projections cannot. As an example, I will discuss how our techniques can be used to give the first subquadratic time algorithms for provably approximating kernel matrices without any coherence or regularity assumptions. For standard problems like kernel ridge regression, kernel PCA, and kernel k-means clustering, the runtime of these kernel approximation methods is only linear in the number of data points being processed.

David Woodruff: 5:00-6:00

Title: A Story of Principal Component Analysis in the Distributed Setting

Abstract: We consider an illustrative problem in distributed machine learning - computing a low rank approximation in the arbitrary partition model. In this model each of s servers holds an n x d matrix A^i, where each entry is an O(log nd)-bit word, and we let A = A^1 + ... + A^s. We would like each server to output the same k x d matrix V, so that V is an approximation to the top k principal components of A, in the sense that projecting onto V provides a (1+eps)-approximate low rank approximation. We give the first communication optimal protocol for this problem, namely a protocol with communication O(skd) + poly (sk/eps) words, improving upon several previous works, and show how to implement the protocol in input sparsity time up to low order terms. Importantly our results do not make any condition number assumptions, yet still achieve the desired bit complexity.

Based on work with Christos Boutsidis and Peilin Zhong.

Day 2:

Matan Gavish 9:00-9:45

TBA

Haim Avron: 9:45-10:15

Faster Kernel Ridge Regression Using Sketching and Preconditioning

Random feature maps, such as random Fourier features, have recently emerged as a powerful technique for speeding up and scaling the training of kernel-based methods such as kernel ridge regression. However, random feature maps only provide crude approximations to the kernel function, so delivering state-of-the-art results requires the number of random features to be very large. Nevertheless, in some cases, even when the number of random features is driven to be as large as the training size, full recovery of the performance of the exact kernel method is not attained. In order to address this issue, we propose to use random feature maps to form preconditioners to be used in solving kernel ridge regression to high accuracy. We provide theoretical conditions on when this yields an effective preconditioner, and empirically evaluate our method and show it is highly effective for datasets of up to one million training examples.

Joint work with Ken Clarkson and David Woodruff

Richard Peng: 11:00-12:00

Title: L_p Row Sampling by Lewis Weights

Abstract:

We give a generalization of statistical leverage scores to non-linear settings: L_p-norm Lewis weights. When 0 < p < 2, sampling the rows of an n-by-d matrix A by these weights gives A' with about dlogd rescaled rows of A such that |Ax|_p is close to |A'x|_p for all vectors x.

These weights can be computed iteratively, leading to input-sparsity time algorithms. We also give an elementary proof for the convergence of sampling by L_1 Lewis weights that's arguably simpler than proofs of L_2 matrix concentration bounds.

Joint work with Michael Cohen

Prateek Jain: 1:30-2:30

Title: Near-Optimal Robust Matrix Completion via Non-convex Optimization

Abstract: Several important applications require completion a low-rank matrix in presence of gross outliers. Examples include robust PCA with missing entries, robust recommendation system, foreground background separation in sublinear time etc. Existing solutions for this problem do not scale for large data and do not have optimal sample complexity/outlier complexity bounds.

In this talk, we will present a simple non-convex optimization based method that is scalable, easy to implement and needs few parameters. Moreover, despite using a non-convex optimization approach, we show that our method achieves linear rate of convergence and solves the problem in nearly optimal time, using nearly optimal number of samples and outliers, thus significantly improving upon the existing state-of-the-art results. Moreover, our experiments show that the method is an order of magnitude faster than the existing methods and leads to a video foreground extraction technique that significantly outperforms existing standard techniques for the problem.

Mahdi Soltanolkotabi: 2:30-3:30

Title: Generic chaining meets (non)convex optimization

Abstract: (Non)convex iterative shrinkage schemes are the working-horse of signal processing and machine learning, yet our mathematical understanding of such algorithms is still in its infancy. In this talk I will discuss some recent results demonstrating that the global optimum of a variety of (non)convex problems can be found efficiently via local search heuristics. These results hold as long as the coefficients of the objective functions are sufficiently randomized e.g. are functions of Gaussian random variables. At the heart of this new analysis are powerful results for concentration of Gaussian stochastic processes. Surprisingly, these stochastic processes seem to exhibit universality behavior: it seems that essentially the same concentration results hold for many non-Gaussian and non-i.i.d. random processes. I will end by discussing our recent efforts towards settling this conjecture for a wide variety of multiplication friendly matrices (e.g. Fourier matrices) via generic chaining methods.

Srinadh Bhojanapalli: 4:00-5:00

A New Sampling Technique for Tensors

Abstract: In this work we propose new techniques to sample elements of arbitrary third-order tensors, with an objective of speeding up tensor algorithms that have recently gained popularity in machine learning. Our main contribution is a new way to select, in a biased random way, only $O(n^{1.5}/\epsilon^2)$ of the possible n^3 elements while still achieving each of the following goals:

(a) tensor sparsification: compute few elements of a tensor from samples to get a good spectral approximation for arbitrary orthogonal tensors, (b) tensor factorization: compute approximate factors of a low-rank tensor corrupted by noise and (c) tensor completion: recover an exactly low-rank tensor from a small number of elements via alternating least squares.

Our sampling can be used along with existing tensor-based algorithms to speed them up, removing the computational bottleneck in these methods.

Yannis Koutis: 5:00-6:00

Title: On fully dynamic graph sparsifiers

Abstract: We present fully dynamic algorithms for graph sparsification problems. The algorithms allow both edge insertions and edge deletions. For cut sparsification, the algorithm takes poly-logarithmic time for each update. For spectral sparsification, the algorithm takes poly-logarithmic time on average. We also discuss an application of these sparsification algorithms in a fully dynamic algorithm for maintaining an approximate minimum cut in an unweighted, undirected, bipartite graph.

Day 3:

Rachel Ward: 9:00-9:30

"How should we sample entries for low-rank matrix completion?"

Fred Roosta: 9:30-10:30

Title: Sub-sampled Newton Methods: Uniform and Non-Uniform Sampling

Abstract: Many data analysis applications require the solution of optimization problems involving a sum of large number of functions. We consider the problem of minimizing a sum of n functions over a convex constraint set. Algorithms that carefully sub-sample to reduce n can improve the computational efficiency, while maintaining the original convergence properties. For second order methods, we first consider a general class of problems and give quantitative convergence results for variants of Newtons methods where the Hessian or the gradient is uniformly sub-sampled. We then show that, given certain assumptions, we can extend our analysis and apply non-uniform sampling which results in modified algorithms exhibiting more robustness and better dependence on problem specific quantities, such as the condition number.

Michael Mahoney: 11:00-12:00

Title: Randomized Linear Algebra and Sub-sampled Newton Methods

Day 4:

Alex Gittens: 9:00-9:45

Title: Low-rank matrix factorizations at scale, using Apache Spark

Abstract: We explore the trade-offs of performing linear algebra in Apache Spark versus the traditional C+MPI approach by examining three widely-used matrix factorizations: NMF (for physical plausibility), PCA (for its ubiquity) , and CX (for model interpretability). We apply these methods to TB-scale problems in particle physics, climate modeling, and bio-imaging using algorithms that map nicely onto Spark's data-parallelism model. We perform scaling experiments on up to 1600 Cray XC40 nodes, describe the sources of slowdowns, and provide tuning guidance to obtain better performance.

Harsha Simhadri: 9:45-10:15

Title: The Nested Dataflow model (based on http://dx.doi.org/10.1145/2935764.2935797)
David Dinh, Harsha Vardhan Simhadri, Yuan Tang

Abstract: Implicitly parallel programming models provide high-level templates (e.g. Fork-Join, Map-Reduce) for programmers to specify algorithms. The templates used by the programmer define a directed acyclic graph (DAG) of function calls and dependencies between them, while the underlying runtime system maps the DAG to a parallel machine. The DAG thus specified often differs from the true dependence graph of the algorithm as the templates provided by the programming model introduce unnecessary "artificial" dependencies. These artificial dependencies affect the efficiency of the system by reducing the amount of parallelism and the locality of the algorithm exposed at runtime. We consider the case of classical numerical linear algebra in the nested-parallel programming model and quantify these inefficiencies. We extend the programming model to the *nested dataflow* programming model, which can accurately capture the dependence patterns that arise in numerical linear algebra, among other classes of algorithms. We present a scheduler to map these algorithms to shared memory parallel machines, and prove that it can make use of the extra parallelism exposed by the nested dataflow model.

Tamas Sarlos: 11-12

Title: Approximating non-linear computations in machine learning

Abstract: It is https://arxiv.org/abs/1605.09046 building on older http://research.google.com/pubs/pub41466.htmland Fast Locality-Sensitive

Hashing, KDD 2011.

Open Questions:

Please contact the participant for the precise question:

1. Jelani - Distributional sparse JL

2. Srinadh - Overcomplete tensor factorization

3. Ravi - noise model for which large subsets are good + other conditions => NMF. What are general deterministic conditions?

4. Richard - leverage score sampling for graphs reduces to solving well-conditioned linear systems. combinatorial algortihm interpreted algebraically - is this more general than graphs? Is it easier to sample than to solve?

5. Sivan - can sketching be useful for recovering groups of singular vectors in other parts of the spectrum? not only the top singular vectors

6. Chris - k-means clustering 1+eps-approximation, (log k)/eps^2 gives a constant-factor?

7. Cameron - adaptive sampling, can you do the adaptive sampling algorithm - sample one column at a time, and instead of getting k! or 2^k approximation, see if it can give poly(k)

8. David - weighted low rank approximation, better bounds

9. Richard - l_1 gradient descent with log(1/eps) convergence

10 Prateek - recover x* from b = A x* + c, with |c|_0 < n/2-eps

Google Document with Talk/Open Question Comments:

https://docs.google.com/document/d/1Mh1Gt8-D8j_LKRRCNkD45iI67SQ7eqaW60fMpurdxPk/edit?usp=sharing