Probabilistic Signal Recovery and Random Matrices

Roman Vershynin
UNIVERSITY OF MICHIGAN

12/08/2016
Final Report

DISTRIBUTION A: Distribution approved for public release.
**Title and Subtitle:** Probabilistic Signal Recovery and Random Matrices

**Abstract:**
Our research program spanned several areas of mathematics and data science. In the area of high-dimensional inference, we showed that classical methods for linear regression (such as Lasso) are applicable for non-linear data. This surprising finding has already found several applications in the analysis of genetic, fMRI and proteomic data, compressed sensing, coding and quantization. In the area of network analysis, we showed how to detect communities in sparse networks by using semidefinite programming and regularized spectral clustering. In high dimensional convex geometry, we studied the complexity of convex sets. In numerical linear algebra, we analyzed the fastest known randomized approximation algorithm for computing the permanents of matrices with non-negative entries. In computational graph theory, we studied a randomized algorithm for estimating the number of perfect matchings in general graphs. In random matrix theory, we established delocalization of eigenvectors for a wide class of random matrices, proved a sharp invertibility result for sparse random matrices, showed how to improve the norm of a general random matrix by removing a small submatrix, and developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets. This has applications for dimension reduction, regression and compressed sensing.
<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
<th>17. LIMITATION OF ABSTRACT</th>
<th>18. NUMBER OF PAGES</th>
<th>19a. NAME OF RESPONSIBLE PERSON</th>
<th>19b. TELEPHONE NUMBER (Include area code)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclassified</td>
<td>Unclassified</td>
<td>Unclassified</td>
<td>UU</td>
<td></td>
<td>RIECKEN, RICHARD</td>
<td>703-941-1100</td>
</tr>
</tbody>
</table>

DISTRIBUTION A: Distribution approved for public release.
Our research program supported by this grant spanned several areas of mathematics and data science. It resulted in significant discoveries in high-dimensional inference and high-dimensional probability and lead to a variety of applications in statistics, biomedical data analysis, quantization, dimension reduction, and networks science.

1. HIGH-DIMENSIONAL INFERENCE AND GEOMETRY

Our main and surprising discovery was how that many classical methods that were designed for structured linear regression provably work even for non-linear data \cite{9, 13, 21}. The non-linearity can be very general: discontinuous, not one-to-one, and even unknown. In spite of this, we showed that methods for linear regression, such as Lasso, stay unharmed even in presence of such nonlinearities. This dramatically extends the range of statistical models for which data analysis can be done rigorously. Our findings have found a variety of applications for quantization and compressed sensing \cite{20}, as well as in the analysis of biomedical data \cite{3, 4}.

As an example, our results apply for binary, 0/1 measurements, which arise e.g. in classification problems and quantization. For such measurements, we also expanded the range of probability distributions the non-linear recovery results apply for. Our original analysis for non-linear data \cite{11, 12} was done under the premise of gaussian measurements. In the new work \cite{1}, we showed extended it to general nonlinear sub-gaussian measurements.

In the area of discrete and computational geometry, we analyzed how many random hyperplanes are needed to cut a given set $K$ in $\mathbb{R}^n$ into much smaller pieces \cite{10}. It turned out that the optimal number of cutting hyperplanes is proportional to the single geometric parameter of the set $K$, namely the effective dimension $d(K)$. This complexity parameter is also known to govern the efficacy of algorithms in high dimensional inference and compressed sensing. In particular, the optimal number of measurements in our work on non-linear data happened to be proportional to the same parameter – the effective dimension of the feasible set $K$, see \cite{21}. Through this link, our work on cutting hyperplanes has implications in quantization, coding, dimension reduction, and compressed sensing.

Another natural measure of complexity of a convex set $K$ is the number of faces of a polytope $P$ that approximates $K$ to within a constant precision. To encode a high-dimensional convex body in a form allowing computer processing, one has to construct an oracle, i.e., an algorithm that using coordinate of a point as an input, outputs whether the point is contained in...
the body, or not. Construction of an oracle for a general convex body is known to be computationally hard. A potentially possible way to bypass this obstacle was suggested by Barvinok. He proposed to approximate a given body by a projection of a section of a simplex in a higher dimension. This new body, being the feasible set of a linear programming problem, allows an efficient construction of the oracle. The complexity of this construction is determined by the dimension of the simplex. An approximation with a simplex of the dimension polynomial in the dimension of the original body would have allowed to circumvent the computational hardness of the oracle construction. In the previous work of the co-PI in collaboration with A. Litvak and N. Tomczak-Jaegermann showed that, in general, such approximation is impossible. This, however, left open a possibility of constructing such approximation for some important classes of convex bodies, primarily, for convex bodies with coordinate symmetries. Nevertheless, we showed in [15] that even such highly symmetric convex bodies require a simplex of exponential dimension to produce such approximation, making bypassing the hardness obstacle impossible.

2. Networks

In the area of network analysis, we developed and rigorously analyzed algorithmic methods for finding structure in sparse networks [6, 7, 5]. There had been an abundance of algorithmic methods for data mining in relatively dense networks, where an average vertex has degree $\gg \log n$, i.e. is connected to at least $\gg \log n$ other vertices or so. Most of these methods, including the most popular Principal Component Analysis (PCA), manifestly fail for sparser networks, in particular for those with constant average degrees.

Practitioners had suggested that the problem for sparse networks lies in the vertices of abnormally high degrees, and suggested that regularizing those vertices by pruning or lowering their weight could solve the problem. We confirmed this rigorously by showing a very general result: any regularization pre-processing which brings the degrees down to normal provably, leads to spectral concentration, and therefore makes spectral methods like PCA work [7].

In a related development [5], we proved for the first time that methods based on semidefinite programming also work for structure discovery in sparse networks. Our analysis is based on Grothendieck’s inequality. In all previous applications in theoretical computer science had only yielded approximation to within some fixed constant factor. We demonstrated a new method where Grothendieck’s inequality can be used to give an arbitrarily fine approximation.
For both methods, our theory is applicable for a far wider class of networks than the benchmark class of stochastic block models that is usually discussed in network science results.

3. PERMANENTS, HAFNIANS AND PERFECT MATCHINGS

In numerical linear algebra, we studied the fastest known randomized approximation algorithm for computing the permanents of matrices with non-negative entries, namely the Barvinok-Godsil-Gutman estimator. The permanent is an important computational characteristic which counts, for instance, the number of perfect matchings in a bipartite graph. Besides this, permanents arise naturally in the study of contingency tables, evaluation of the expected product of dependent normal random variables, etc. It is known that the evaluation of a permanent is a \#-P hard problem, so taking into account the limitations of the computing power, one can hope only to estimate it. Barvinok-Godsil-Gutman estimator probabilistic estimator is the fastest known means of estimating the permanent. In the worst-case scenario, it outputs the permanent with the multiplicative error which is exponential in the size of the matrix. Yet, it has been observed that, typically, the actual performance of this estimator is much better than the worst case. We discovered a sufficient condition on a deterministic graph or matrix guaranteeing a smaller error for estimating the permanent \[19\].

In a related development in computational graph theory \[16\], we analyzed a probabilistic algorithm for estimating the number of perfect matchings in general graphs. Unlike bipartite graphs, where the number of perfect matchings is represented by the permanent of the adjacency matrix, in a general case, it is evaluated by a much more complex quantity, namely the hafnian of the same matrix. Because of this additional complexity, almost all known methods of estimating the number of perfect matchings which were developed for bipartite graphs fail for the general ones. The only exception is the Barvinok estimator which is currently the unique polynomial time probabilistic estimator for the number of perfect matchings. This fact makes the error analysis for this estimator especially important. As for bipartite graphs, the worst case error of this estimator is exponential in the size of the graph. We showed that if the graph possesses certain expansion properties, then the error of the Barvinok estimator is much smaller than in the worst case.

4. RANDOM MATRIX THEORY

Several significant advances were made in random matrix theory and its applications. We established delocalization of eigenvectors for a wide class of random matrices \[17\] \[18\]. This means that with high probability, every eigenvector of a random matrix is delocalized in the sense that any subset of
its coordinates carries a non-negligible portion of its $L^2$ norm. Our results pertain to a wide class of random matrices, including matrices with independent entries, symmetric and skew-symmetric matrices, as well as some other naturally arising ensembles.

Next, we analyzed in [2] the condition number of sparse random matrices, a quantity important for controlling the running time and the precision of various numerical linear algebra algorithms. This is an important problem especially for sparse random matrix, which are among the basic tools in statistics, computer sciences and signal processing. As we increase sparsity, we found that the condition number stays nearly the same as for a dense matrix almost until the transition point where an entire zero row appears (at which point it obviously jumps to infinity).

Furthermore, we showed how to improve the behavior of a random matrix by modifying a small fraction of its entries [14]. We studied the conditions where the operator norm of a random matrix $A$ can be reduced to the optimal order by zeroing out a small submatrix of $A$. We found that this is possible if and only if the entries of $A$ have zero mean and finite variance. Moreover, we obtained an almost optimal dependence between the size of the removed submatrix and the resulting operator norm.

Finally, we developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets [8]. This new deviation inequality unified many existing results, such as Johnson-Lindenstrauss Lemma which plays a major role in dimension reduction, $M^*$ bound in high-dimensional convex geometry, and a non-asymptotic version of Bai-Yin limiting law in random matrix theory. On top of that, our deviation inequality led to several new applications, in particular for dimension reduction, model selection, structured regression and compressed sensing [8].

REFERENCES


1.

Report Type
Final Report

Primary Contact Email
Contact email if there is a problem with the report.
romanv@umich.edu

Primary Contact Phone Number
Contact phone number if there is a problem with the report
7349294092

Organization / Institution name
University of Michigan

Grant/Contract Title
The full title of the funded effort.
Probabilistic Signal Recovery and Random Matrices

Grant/Contract Number
AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".
FA9550-14-1-0009

Principal Investigator Name
The full name of the principal investigator on the grant or contract.
Roman Vershynin

Program Officer
The AFOSR Program Officer currently assigned to the award
Richard Riecken

Reporting Period Start Date
01/01/2014

Reporting Period End Date
12/31/2016

Abstract
Our research program spanned several areas of mathematics and data science. In the area of high-dimensional inference, we showed that classical methods for linear regression (such as Lasso) are applicable for non-linear data. This surprising finding has already found several applications in the analysis of genetic, fMRI and proteomic data, compressed sensing, coding and quantization. In the area of network analysis, we showed how to detect communities in sparse networks by using semidefinite programming and regularized spectral clustering. In high dimensional convex geometry, we studied the complexity of convex sets. In numerical linear algebra, we analyzed the fastest known randomized approximation algorithm for computing the permanents of matrices with non-negative entries. In computational graph theory, we studied a randomized algorithm for estimating the number of perfect matchings in general graphs. In random matrix theory, we established delocalization of eigenvectors for a wide class of random matrices, proved a sharp invertibility result for sparse random matrices, showed how to improve the norm of a general random matrix by removing a small submatrix, and developed a simple and general tool for bounding the deviation of random matrices on arbitrary geometric sets. This has applications for dimension reduction, regression and compressed sensing.
DISTRIBUTION A: Distribution approved for public release.


New discoveries, inventions, or patent disclosures:
Do you have any discoveries, inventions, or patent disclosures to report for this period?

No

Please describe and include any notable dates

Do you plan to pursue a claim for personal or organizational intellectual property?

Changes in research objectives (if any):

Change in AFOSR Program Officer, if any:

Dr. Tristan Nguyen to Dr. Richard Riecken

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

<table>
<thead>
<tr>
<th></th>
<th>Starting FY</th>
<th>FY+1</th>
<th>FY+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equipment/Facilities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supplies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

DISTRIBUTION A: Distribution approved for public release.
Appendix Documents

2. Thank You

E-mail user

Nov 10, 2016 09:28:01 Success: Email Sent to: romanv@umich.edu

DISTRIBUTION A: Distribution approved for public release.