Compact Information Representations

Martin Wells
CORNELL UNIVERSITY

08/02/2016
Final Report

DISTRIBUTION A: Distribution approved for public release.
Numerous modern applications in the context of network traffic, information retrieval, and databases are faced with very large, inherently high-dimensional, or naturally streaming datasets. This proposal aims at developing mathematically rigorous and general-purpose statistical methods based on stable random projections, to achieve compact information representations, for solving very large-scale engineering problems in data stream computations, real-time network monitoring & anomaly detections (e.g., DDoS attacks), machine learning, databases, and search. Fundamentally, compact data representations are highly beneficial because they could substantially reduce memory or disk storage, facilitate efficient data transmission over the networks, accomplish time-critical missions, improve experience in user-facing applications, reduce energy consumptions, etc. The proposed research topics largely fall into three categories: (i) Data stream algorithms for network anomaly detections; (ii) Probabilistic quantization for compact information storage, indexing, and search; and (iii) Effective sparse recovery from (quantized) stable random projections. The proposed research is highly interdisciplinary, across statistics, theoretical & applied computer science, and applied math. Within the scope of this proposal, the focus is preliminarily on the fundamental, theoretical research which lies in the mission of AFOSR.
<table>
<thead>
<tr>
<th>19b. TELEPHONE NUMBER (include area code)</th>
</tr>
</thead>
<tbody>
<tr>
<td>703-696-5999</td>
</tr>
</tbody>
</table>
Final Report for AFOSR-FA9550-13-1-0137

Compact Information Representations

Principal Investigator : Ping Li

Department of Statistics and Biostatistics
Department of Computer Science
Rutgers University, the State University of New Jersey
Piscataway, NJ 08854, USA
pingli@stat.rutgers.edu

May 2016
# Contents

1 Training for Ph.D. Students and Postdoc Researchers  

2 Papers  

3 Summary of Proposed Research: Compact Information Representations  

4 Research Results  
   4.1 Sign Cauchy Random Projections (Published in NIPS 2013)  
   4.2 Coding for Gaussian Random Projections (Published in ICML 2014)  
   4.3 b-Bit Marginal Regression (Published in NIPS 2015)  
   4.4 Binary and Multi-Bit Coding for Stable Random Projections  
   4.5 CoRE Kernels (Published in UAI 2014)  

5 Future Work
1 Training for Ph.D. Students and Postdoc Researchers

The following students and postdoc researchers were partially supported by this grant.

- Jun Hu, Ph.D. student in CS
- Jie Shen, Ph.D. student in CS
- Liang Wang, Ph.D. student in Statistics
- Ruijun Ma, Ph.D. student in Statistics
- Jing Wang (female), Postdoc researcher
- Anshumali Shrivastava, now Assistant Professor, Department of Computer Science, Rice University
- Martin Slawski, (to start in August 2016) Assistant Professor, Department of Statistics, George Mason University
- Xiao-tong Yuan, now Professor, Nanjing University of Information Science & Technology
- Guangcan Liu, now Professor, Nanjing University of Information Science & Technology
- Jian Wang, now Professor, Nanjing University of Information Science & Technology
- Tung-Lung Wu, now Assistant Professor, Dept. of Math and Stat, Mississippi State Univ

2 Papers

In this section, we list the papers which acknowledged the (partial) support from this grant.

* indicates a co-author is my Ph.D. student, graduate research assistant, or postdoc researcher.

1. Ping Li, *Binary and Multi-Bit Coding for Stable Random Projections*, to be submitted
2. Jun Hu* and Ping Li, *DMF: A Decomposed Ordinal Matrix Factorization Approach for Improving Rating Prediction*, to be submitted
3. Ping Li, Syama Sundar Rangapuram, Martin Slawski*, *Methods for Sparse and Low-Rank Recovery under Simplex Constraints*, to be submitted
4. Xiao-tong Yuan*, Ping Li, and Tong Zhang, *Semiparametric Pairwise Graphical Models for Learning with Nonlinear Sufficient Statistics*, to be submitted
5. Ping Li, Michael Mitzenmacher, Anshumali Shrivastava*, *2-Bit Random Projections, NonLinear Estimators, and Approximate Near Neighbor Search*, to be submitted
7. Tung-Lung Wu* and Ping Li, *Tests for High-Dimensional Covariance Matrices Using Random Matrix Projection*, to be submitted
11. Ping Li, *One Scan 1-Bit Compressed Sensing*, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2016
12. Jie Shen* and Ping Li, *Learning Structured Low-Rank Representation via Matrix Factorization*, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2016
15. Ping Li, *0-Bit Consistent Weighted Sampling*, Knowledge Discovery and Data Mining (KDD), 2015
17. Anshumali Shrivastava* and Ping Li, *Improved Asymmetric Locality Sensitive Hashing (ALSH) for Maximum Inner Product Search (MIPS)*, Uncertainty in Artificial Intelligence (UAI) 2015.
27. Ping Li, Michael Mitzenmacher, and Anshumali Shrivastava, Coding for Random Projections, International Conference on Machine Learning (ICML), 2014


30. Ping Li, CoRE Kernels, Uncertainty in Artificial Intelligence (UAI), 2014

31. Anshumali Shrivastava and Ping Li, Improved Densification of One Permutation Hashing, Uncertainty in Artificial Intelligence (UAI), 2014

32. Anshumali Shrivastava and Ping Li, In Defense of Minhash over Simhash, International Conference on Artificial Intelligence and Statistics (AISTATS), 2014

33. Anshumali Shrivastava and Ping Li, A New Space for Comparing Graphs, IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM), 2014

34. Xiao-tong Yuan and Ping Li, Sparse Additive Subspace Clustering, European Conference on Computer Vision (ECCV), 2014

35. Jinhua Ma and Ping Li, Semi-Supervised Ranking for Re-Identification with Few Labeled Image Pairs, Asian Conference on Computer Vision (ACCV), 2014

36. Jinhua Ma and Ping Li, Query Based Adaptive Re-Ranking for Person Re-Identification, Asian Conference on Computer Vision (ACCV), 2014

37. Zuofeng Shang and Ping Li, Bayesian ultrahigh-dimensional screening via MCMC, Journal of Statistical Planning and Inference (JSPI), 2014

38. Zuofeng Shang and Ping Li, High-Dimensional Bayesian Inference in Nonparametric Additive Models, Electronic Journal of Statistics (EJS), 2014


41. Ping Li and Cun-Hui Zhang, Exact Sparse Recovery with L0 Projections, in Knowledge Discovery and Data Mining (KDD), 2013

3 Summary of Proposed Research: Compact Information Representations

Numerous modern applications in the context of network traffic, information retrieval, and databases are faced with very large, inherently high-dimensional, or naturally streaming datasets. This proposal aims at developing mathematically rigorous and general-purpose statistical methods based on stable random projections, to achieve compact information representations, for solving very large-scale engineering problems in data stream computations, real-time network monitoring & anomaly detections (e.g., DDoS attacks), machine learning, databases, and search. Fundamentally, compact data representations are highly beneficial because they could substantially reduce memory or disk storage, facilitate efficient data transmission over the networks, accomplish time-critical
missions, improve experience in user-facing applications, reduce energy consumptions, etc. The proposed research topics largely fall into three categories: (i) Data steam algorithms for network anomaly detections; (ii) Probabilistic quantization for compact information storage, indexing and search; and (iii) Effective sparse recovery from (quantized) stable random projections. The proposed research is highly interdisciplinary, across statistics, theoretical & applied computer science, and applied math. Within the scope of this proposal, the focus is preliminarily on the fundamental, theoretical research which lies in the mission of AFOSR.

4 Research Results

The proposed research goals have been largely accomplished. More than 40 papers have been published (including two best paper awards) or will be soon submitted. Several students and post-doc researchers who were (partially) supported by this grant landed faculty positions to start their independent research careers.

Selected research accomplishments are summarized in this report.

4.1 Sign Cauchy Random Projections (Published in NIPS 2013)

The method of stable random projections is useful for efficiently approximating the $l_\alpha$ distance ($0 < \alpha \leq 2$) in high dimension and it is naturally suitable for data streams. In this work, we propose to use only the signs of the projected data and we analyze the probability of collision (i.e., when the two signs differ). Interestingly, when $\alpha = 1$ (i.e., Cauchy random projections), we show that the probability of collision can be accurately approximated as functions of the chi-square ($\chi^2$) similarity. In text and vision applications, the $\chi^2$ similarity is a popular measure when the features are generated from histograms (which are a typical example of data streams). Experiments confirm that the proposed method is promising for large-scale learning applications.

4.2 Coding for Gaussian Random Projections (Published in ICML 2014)

The method of random projections has become very popular for large-scale applications in statistical learning, information retrieval, bio-informatics and other applications. Using a well-designed coding scheme for the projected data, which determines the number of bits needed for each projected value and how to allocate these bits, can significantly improve the effectiveness of the algorithm, in storage cost as well as computational speed. In this paper, we study a number of simple coding schemes, focusing on the task of similarity estimation and on an application to training linear classifiers. We demonstrate that uniform quantization outperforms the standard existing influential method. Indeed, we argue that in many cases coding with just a small number of bits suffices. Furthermore, we also develop a non-uniform 2-bit coding scheme that generally performs well in practice, as confirmed by our experiments on training linear support vector machines (SVM). Our work also includes the analysis for the 1-bit scheme. Overall, our paper provides the guideline for choosing the coding schemes in practice.
4.3 b-Bit Marginal Regression (Published in NIPS 2015)

We consider the problem of sparse signal recovery from \( m \) linear measurements quantized to \( b \) bits. b-bit Marginal Regression is proposed as recovery algorithm. We study the question of choosing \( b \) in the setting of a given budget of bits \( B = m \cdot b \) and derive a single easy-to-compute expression characterizing the trade-off between \( m \) and \( b \). The choice \( b = 1 \) turns out to be optimal for estimating the unit vector corresponding to the signal for any level of additive Gaussian noise before quantization as well as for adversarial noise. For \( b \geq 2 \), we show that Lloyd-Max quantization constitutes an optimal quantization scheme and that the norm of the signal can be estimated consistently by maximum likelihood.

4.4 Binary and Multi-Bit Coding for Stable Random Projections

We develop efficient binary (i.e., 1-bit) and multi-bit coding schemes for estimating the scale parameter of \( \alpha \)-stable distributions. The work is motivated by the recent work on one scan 1-bit compressed sensing (sparse signal recovery) using \( \alpha \)-stable random projections, which requires estimating the scale parameter at bits-level. Our technique can be naturally applied to data stream computations for estimating the \( \alpha \)-th frequency moment. In fact, the method applies to the general scale family of distributions, not limited to \( \alpha \)-stable distributions. Due to the heavy-tailed nature of \( \alpha \)-stable distributions, using traditional estimators will potentially need many bits to store each measurement in order to ensure sufficient accuracy. Interestingly, our paper demonstrates that, using a simple closed-form estimator with merely 1-bit information does not result in a significant loss of accuracy if the parameter is chosen appropriately. For example, when \( \alpha = 0+, 1, \) and 2, the coefficients of the optimal estimation variances using full (i.e., infinite-bit) information are 1, 2, and 2, respectively. With the 1-bit scheme and appropriately chosen parameters, the corresponding variance coefficients are 1.544, \( \pi^2/4 \), and 3.066, respectively. Theoretical tail bounds are also provided. Using 2 or more bits per measurements reduces the estimation variance and importantly, stabilizes the estimate so that the variance is not sensitive to parameters. With look-up tables, the computational cost is minimal. Extensive simulations are conducted to verify the theoretical results. The estimation procedure is integrated into the sparse recovery with one scan 1-bit compressed sensing. One interesting observation is that the classical “Bartlett correction” (for MLE bias correction) appears particularly effective for our problem when the sample size (number of measurements) is small.

4.5 CoRE Kernels (Published in UAI 2014)

The term “CoRE kernel” stands for correlation-resemblance kernel. In many real-world applications (e.g., computer vision), the data are often high-dimensional, sparse, and non-binary. We propose two types of (nonlinear) CoRE kernels for non-binary sparse data and demonstrate the effectiveness of the new kernels through a classification experiment. CoRE kernels are simple with no tuning parameters. However, training nonlinear kernel SVM can be costly in time and memory and may not be always suitable for truly large-scale industrial applications (e.g., search). In
order to make the proposed CoRE kernels more practical, we develop basic probabilistic hashing (approximate) algorithms which transform nonlinear kernels into linear kernels.

5 Future Work

This is a great time for research in big data and machine learning, as many urgent practical problems can be efficiently solved by effect & compact data representations and simple & robust learning algorithms. Under the support of this AFOSR grant, a lot of excited research problems have been solved and many more arise. We will continue many research topics we have started and expect much more will be accomplished in the near future. We highly appreciate AFOSR for this generous support and we look forward to working with AFOSR again soon.