AFRL-RI-RS-TR-2021-050

ROBUST, EFFICIENT, AND LOCAL MACHINE LEARNING PRIMITIVES

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

MARCH 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-050 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MICHAEL J. MANNO
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief,
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MARCH 2021	FINAL TECHNICAL REPORT	MAR 2017 – MAR 2021

4. TITLE AND SUBTITLE

ROBUST, EFFICIENT, AND LOCAL MACHINE LEARNING PRIMITIVES

5a. CONTRACT NUMBER
FA8750-17-2-0122

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)

Michael Mahoney

5d. PROJECT NUMBER
D3ME

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

International Computer Science Institute
2150 Shattuck Avenue Suite 1100
Berkeley CA 94704

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIED
525 Brooks Road
Rome NY 13441-4505

DARPA/I2O
675 North Randolph Street
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2021-050

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this project, we develop, implement, and apply a suite of theoretically principled algorithmic and statistical primitives that are easy for the non-expert to use and that map cleanly to the intuition and understanding that domain experts have about their data and the processes generating their data. Most of our efforts focus on machine learning (ML) and data analysis (DA) primitives for analyzing data that are modeled by matrices or graphs, with an emphasis on primitives that (when combined appropriately) give complementary algorithmic and statistical advantage. Our main focus is on TA1, for which we develop a library of primitives, but we are also interested in TA2 questions having to do with how these primitives interact.

15. SUBJECT TERMS

Randomized matrix algorithms; local spectral graph algorithms; robust models; non-convex optimization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**MICHAEL J. MANNO**
U	U	U	UU	17	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

FINAL PROJECT REPORT:
Robust, Efficient, And Local Machine Learning Primitives

Michael W. Mahoney (PI)

International Computer Science Institute

Abstract. In this project, we develop, implement, and apply a suite of theoretically principled algorithmic and statistical primitives that are easy for the non-expert to use and that map cleanly to the intuition and understanding that domain experts have about their data and the processes generating their data. Most of our efforts focus on machine learning (ML) and data analysis (DA) primitives for analyzing data that are modeled by matrices or graphs, with an emphasis on primitives that (when combined appropriately) give complementary algorithmic and statistical advantage. Our main focus is on TA1, for which we develop a library of primitives, but we are also interested in TA2 questions having to do with how these primitives interact when composed into ML/DA pipelines used by domain scientists. To accomplish this goal, our emphasis is three-fold: theory for principled algorithmic and statistical primitives; implementations on single-machine and parallel/distributed environments; and applications in domains designed to stress-test our primitives. Our primitives are loosely organized around matrix models, robust statistical methods, graph models, and optimization methods for nonlinear models; our implementations consider how our primitives perform in isolation as well as when composed with other primitives in common domain science pipelines; applications include climate science, neuroscience, social network analysis, and image and video analysis, where we have prior experience, and other applications from D3M.

Contents

LIST OF FIGURES

1 SUMMARY

We develop a set of algorithmic/statistical primitives and integrate these primitives into a software stack that can be used on a single machine or in a parallel/distributed environment, where they can be easily used by domain scientists, such as our scientific collaborators and the TA2 and TA3 teams. Since matrices and graphs are ubiquitous data models— indeed, their use often operationally defines the model of interest more strongly than the often intractable statistical model purportedly of interest— several of our primitives focus on matrix and graph methods. Moreover, since statistical and optimization questions are often intertwined in domain applications in complicated ways, we also focus explicitly on statistical and optimization primitives.

2 INTRODUCTION

Here is a brief introduction to our main results.

- Our work on matrix models focuses on improved kernel algorithms using the Nyström and Random Feature Map methods and interpretable models that go beyond the usual global low-rank approximation to hierarchical and local methods.
- Our work on robust statistical methods develops scalable resampling methods to explore the predictivity-interpretability trade-off and sketching methods for GLMs and other nonlinear objectives.
- Our work on graph models develops local graph algorithms for unsupervised learning objectives, for identifying small, medium, and large-scale structure, for pre-processing and semi-supervised learning, and for graph visualization.
- Finally our work on optimization focuses on improved second-order and first-order methods and applications of sketching in tensor factorization.

Several of our approaches in these technical domains rely upon greatly expanding the applicability of *sampling/sketching*, which has focused almost exclusively on linear problems, to nonlinear applications: e.g., we propose using sketching to robustify and decrease the computational cost of the fitting of nonlinear models. In addition to novel TA1 primitives, we will provide TA2 primitives that optimally compose mature approaches, such as Randomized Numerical Linear Algebra (RandNLA) sketching, with our TA1 primitives. See Figure 1 for a pictorial description of our overall approach.

A common theme is the provision of methods that are statistically robust, algorithmically efficient, and localizable in that they capture fine-scale properties. Each of our methodological contributions will be accompanied by implementations that are usable by non-experts in applications that leverage their domain-specific insights.

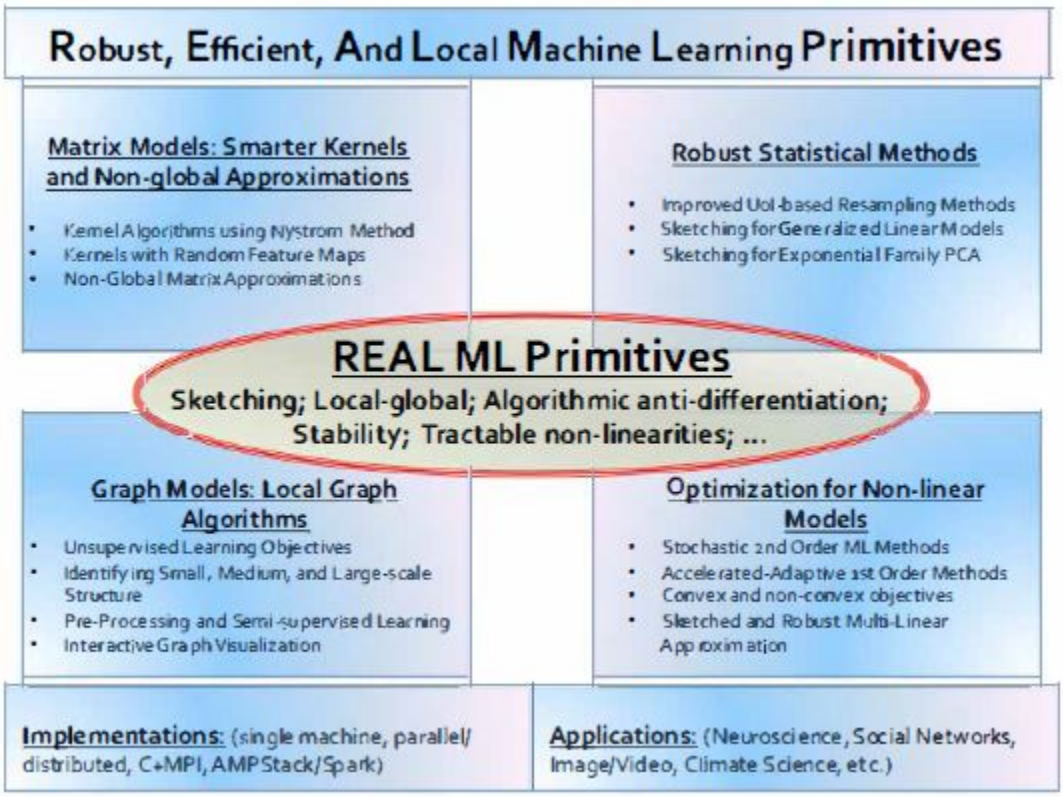

Figure 1: Introduction to overall approach: we will consider four classes of primitives, each grounded in implementations in RAM/parallel/distributed architectures, and each applied in real application domains.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

Matrix Models. Linear algebra (LA) has been used widely in ML/DA, but there has been little work on identifying LA primitives for matrix problems in ML/DA. One exception is *sampling/sketching* techniques, which have been used to great advantage in RandNLA. Recent work has shown that statistical considerations lead to improvements in practice and that robustified ridge leverage scores lead to improved low-rank approximation for both algorithmic and statistical objectives, but the statistical properties and implicit modeling implications of sketching/sampling are poorly understood.

Robust Statistical Methods. The main goal in many scientific and ML/DA applications is often either high *predictive* performance or high *interpretability*. In either case, it is desired that the underlying statistical procedure is *robust*, i.e., stable relative to "reasonable" perturbations to input data. Such statistical robustness is typically defined with respect to the particular goal in mind (e.g., interpretability or prediction). For example, in learning theory, the main objective is good generalization performance, and robustness is defined relative to prediction. On the other hand, in many scientific applications interpretability and correct model selection, i.e., estimation of discrete structures as in variable selection or clustering, is the primary objective, in which case robustness is defined relative to such model estimation. In general, there is a trade-off between predictivity and interpretability, and for applications which require both methods that robustly satisfy both are lacking. In this light, it is highly desirable to use sketching, bootstrapping, or sub-sampling to design new statistical methods that are robust both in terms of interpretability and predictivity, and are also easy-to-use for practitioners who are not data science experts.

Graph Models. *Local spectral graph methods*—algorithms that do not touch all of a large input graph, but come with provable guarantees analogous to the more common global spectral graph methods—have recently emerged. They have strong connections with heuristics practicing domain scientists actually perform, and come with algorithmic and statistical guarantees. They were originally of theoretical interest, where they were used for linear-time Laplacian solvers, but they are also powerful primitives for exploratory ML/DA on graphs. Recent work has shown that these local spectral methods can be improved by flow-based methods, that they have sparsity-inducing statistical properties, that they can be used for data preprocessing in ML/DA pipelines, that they map well to matrix-vector multiplication primitives on sparse graphs, and that they can be implemented on billion-node graphs.

Optimization Methods for NonLinear Models. Going beyond graph and matrix models, ML/DA applications involving general nonlinear models are typically cast as complex optimization problems. First-order methods (i.e., those that use only gradient information) have much lower per-iteration cost relative to their second-order (i.e., those algorithms that employ Hessian as well as gradient information) counterparts, and thus there is a plethora of *deterministic and stochastic first-order optimization algorithms* for ML/DA problems; and a significant amount of effort has been made to *accelerate* the convergence rate of these first-order methods. Indeed, first-order methods have taken center stage as the primary workhorse for solving ML/DA

problems. While low-precision solutions suffice for certain ML/DA problems, for many scientific problems (where one is more interested in the domain than the methods) one needs *medium to high precision solutions*. However, for even moderately ill-conditioned problems, e.g., those with highly correlated data and small regularization or when one is interested in medium-scale clusters, first-order methods are simply incapable of achieving such solutions within a reasonable amount of time. On the other hand, *second-order methods* exhibit greater *insensitivity* to problem conditioning.

4 RESULTS AND DISCUSSION

Our results will be organized loosely into the following four main topics: Matrix Models; Robust Statistical Methods; Graph Models; and Optimization Methods for NonLinear Models. In this section, we will summarize and discuss our main published results; and then we will summarize our main publicly-available code.

4.1 Matrix Models

Along this general direction, our main results include:

- A scalable method for k-means and related problems under sketching techniques: [23].
- A system for the analysis of large-scale data sets that calls MPI from within Spark: [11].
- A bootstrap method to estimate the error in randomized matrix algorithms: [15].
- A detailed empirical analysis of the empirical spectral distribution of neural network weight matrices, which provides the basis for an operational theory that uses heavy-tailed random matrix theory to describe implicit regularization in neural network learning: [17].
- A method to use random matrix ideas to predict trends in test accuracies for very large pre-trained neural network models: [18].
- A method to perform fluid flow reconstruction in an environment where there is limited sensors and labeled data: [5].
- Tutorial paper on "Statistical Mechanics Methods for Discovering Knowledge from Modern Production Quality Neural Networks": [19].
- An algorithm to analyze time series data using ideas from Koopman Autoencoder theory: [1].

Overall, the matrix methods we developed went well beyond the previous scalable randomized matrix algorithms to include domain structure, fine-scale properties of matrices that are not well-approximated by low-rank models, and the application in several domain areas.

4.2 Robust Statistical Methods

Along this general direction, our main results include:

- A bootstrapping method to get high-quality and sparse predictors: [2].
- A bootstrapping method to perform error estimation for randomized least-squares: [16].
- An algorithm for using second order analysis methods for robust large batch training of neural networks that generates more robust models: [28].

- A demonstration that second order optimization methods can perform well for neural network training when coupled with adversarial training to develop more robust models: [26].
- A method to perform a very simple "retrofit" strategy to improve the robustness of learned ML models: [6].

Overall, the robust statistical methods we developed use principled matrix, graph, and optimization techniques to develop and analyze robust models at scale.

4.3 Graph Models

Along this general direction, our main results include:

- A novel local form of diffusion that is particularly well-suited for noisy graphs: [22].
- A model for locality and structure aware graph embedding: [7].
- A method to construct an out-of-sample extension of graph adjacency spectral embedding: [14].
- A tutorial / introduction paper to local spectral graph methods: [9].
- A more general theoretical examination of limit theorems for out-of-sample extensions of the adjacency and Laplacian spectral embeddings: [13].
- A review / overview on local graph algorithms, "Flow-based Algorithms for Improving Clusters: A Unifying Framework, Software, and Performance" [10].

Overall, the local graph methods we developed provide principled methods and a strong code base which did not exist previously to perform local and locally-biased graph analytics at scale.

4.4 Optimization Methods for NonLinear Models

Along this general direction, our main results include:

- A scalable method that uses novel sampling techniques for non-convex optimization: [24].
- A detailed empirical evaluation of second-order methods for non-convex problems: [25].
- An algorithm for communication-avoiding methods for first order optimization to speed up the graph and optimization pipelines: [3].
- An implementation and evaluation of subsampled Newton methods on a GPU, illustrating their practical advantages: [12].
- An implementation of subsampled Newton methods in distributed environments: [8].
- An algorithm to perform Newton based optimization for problems for which smoothness and convexity fail to hold: [20].

- An algorithm using second order optimization methods including trust region methods to perform adversarial attacks on neural networks: [30].
- An algorithm to use approximate second order optimization methods to quantize models: [21].
- A paper to describe software to use second order optimization methods efficiently and at scale: [27].
- A method to use second order optimization for more improved quantization: [4].

Overall, the optimization methods we developed went well beyond prior work on first order methods to provide composable models for use by downstream domain scientists.

4.5 Additional results winding down the program

We also made improvements to Ristretto, LocalGraphClustering, PyHessian and other software. Both Ristretto and LocalGraphClustering provide high-performance implementations of the primitives, with an emphasis on a user-friendly interface for domain-scientists in various fields, e.g., including neuroscience, fluid flow dynamics and astronomy. We used these two software packages in Python as a starting point for their D3M primitive implementations. We have also implemented various primitives as part of the D3M software package. All of our submitted unsupervised primitives (and corresponding demo pipelines) have successfully passed the recent dry-run evaluations. An exception to this are the pipelines of the TensorMachinesBinaryClassification primitive, which require the sklearn imputer primitive. This was put on hold as the program wound down. We also refined the pipelines, coordinating with other members of the program.

4.6 Publicly-available code beyond pipeline development

We developed a large body of code. Beyond our work on pipeline development, here are examples of code that was developed and made available during this project.

- ADAHESSIAN: https://github.com/amirgholami/adahessian
- PyHessian: https://github.com/amirgholami/pyhessian
- LocalGraphClustering: https://github.com/kfoynt/LocalGraphClustering
- HessianFlow: https://github.com/amirgholami/HessianFlow
- Alchemist: https://github.com/alexgittens/alchemist
- Distributed Second-order Convex Optimization https://github.com/fang150/Newton_ ADMM
- GPU-accelerated Sub-sampled Newton's Method https://github.com/kylasa/NewtonCG/
- Second-Order Optimization for Non-Convex Machine Learning https://github.com/ git-xp/Non-Convex-Newton

5 CONCLUSIONS

We developed and implemented algorithmic/statistical primitives—based on Matrix Models, Robust Statistical Methods, Graph Models, and Optimization Methods for NonLinear Models—providing publicly-available software and integrating many of these primitives into a software stack that can be used on a single machine or in a parallel/distributed environment, where they can easily be used by domain scientists and performers in the program. A main focus of the work involved developing methods that were composable but that also came with strong algorithmic and statistical guarantees.

6 RECOMMENDATIONS

Based on our results, several directions should be recommended for future consideration. First, statistical aspects of randomized numerical linear algebra algorithms deserve further development. These algorithms come with strong worst-case theory and can be implemented at scale, but they have non-trivial statistical properties that have important consequences for downstream domain scientists who want to use them in an automated way. Second, local graph algorithms deserve further development. These algorithms provide a principled way for downstream domain scientists to explore and identify structure in up to tera-byte size-scale graphs (an, in principle, beyond, but larger ones tend not to be publicly-available), and our local graph partitioning package opens the door to further development. Third, stochastic second order optimization algorithms deserve further development. These algorithms provide an important point in trade-off space that is currently underserved by the machine learning and artificial intelligence communities, e.g., in particular for analytics on data in which one has deep domain knowledge.

7 REFERENCES

[1] O. Azencot, N. B. Erichson, V. Lin, and M. W. Mahoney. Forecasting sequential data using consistent Koopman autoencoders. Technical report, 2020. Preprint: arXiv:2003.02236.

[2] K. E. Bouchard, A. F. Bujan, F. Roosta-Khorasani, S. Ubaru, Prabhat, A. M. Snijders, J.-H. Mao, E. F. Chang, M. W. Mahoney, and S. Bhattacharyya. Union of Intersections (UoI) for Interpretable Data Driven Discovery and Prediction. Technical Report Preprint: arXiv:1705.07585, 2017.

[3] A. Devarakonda, K. Fountoulakis, J. Demmel, and M. W. Mahoney. Avoiding synchronization in firstorder methods for sparse convex optimization. Technical report, 2017. Preprint: arXiv:1712.06047.

[4] Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. Technical Report Preprint: arXiv:1911.03852, 2019.

[5] N. B. Erichson, L. Mathelin, Z. Yao, S. L. Brunton, M. W. Mahoney, and J. N. Kutz. Shallow learning for fluid flow reconstruction with limited sensors and limited data. Technical report, 2019. Preprint: arXiv:1902.07358.

[6] N. B. Erichson, Z. Yao, and M. W. Mahoney. JumpReLU: A retrofit defense strategy for adversarial attacks. Technical Report Preprint: arXiv:1904.03750, 2019.

[7] E. Faerman, F. Borutta, K. Fountoulakis, and M. W. Mahoney. LASAGNE: Locality and structure aware graph node embedding. Technical report, 2017. Preprint: arXiv:1710.06520.

[8] C.-H. Fang, S. B. Kylasa, F. Roosta, M. W. Mahoney, and A. Grama. Newton-ADMM: A distributed GPU-accelerated optimizer for multiclass classification problems. Technical Report Preprint: arXiv:1807.07132, 2018.

[9] K. Fountoulakis, D. F. Gleich, and M. W. Mahoney. A short introduction to local graph clustering methods and software. Technical Report Preprint: arXiv:1810.07324, 2018.

[10] K. Fountoulakis, M. Liu, D. F. Gleich, and M. W. Mahoney. Flow-based algorithms for improving clusters: A unifying framework, software, and performance. Technical Report Preprint: arXiv:2004.09608, 2020.

[11] A. Gittens, K. Rothauge, M. W. Mahoney, S. Wang, L. Gerhardt, Prabhat, J. Kottalam, M. Ringenburg, and K. Maschhoff. Alchemist: An Apache Spark <=> MPI interface. Technical Report Preprint: arXiv:1806.01270, 2018.

[12] S. B. Kylasa, F. Roosta-Khorasani, M. W. Mahoney, and A. Grama. GPU accelerated sub-sampled Newton's method. Technical Report Preprint: arXiv:1802.09113, 2018.

[13] K. Levin, F. Roosta, M. Tang, M. W. Mahoney, and C. E. Priebe. Limit theorems for out-ofsample extensions of the adjacency and Laplacian spectral embeddings. Technical Report Preprint: arXiv:1910.00423, 2019.

[14] K. Levin, F. Roosta-Khorasani, M. W. Mahoney, and C. E. Priebe. Out-of-sample extension of graph adjacency spectral embedding. Technical report, 2018. Preprint: arXiv:1802.06307.

[15] M. E. Lopes, S. Wang, and M. W. Mahoney. A bootstrap method for error estimation in randomized matrix multiplication. Technical report, 2017. Preprint: arXiv:1708.01945.

[16] M. E. Lopes, S. Wang, and M. W. Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. Technical Report Preprint: arXiv:1803.08021, 2018.

[17] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. Technical Report Preprint: arXiv:1810.01075, 2018.

[18] C. H. Martin and M. W. Mahoney. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. Technical Report Preprint: arXiv:1901.08278, 2019.

[19] C. H. Martin and M. W. Mahoney. Statistical mechanics methods for discovering knowledge from modern production quality neural networks. In *Proceedings of the 25th Annual ACM SIGKDD Conference*, pages 3239–3240, 2019.

[20] F. Roosta, Y. Liu, P. Xu, and M. W. Mahoney. Newton-MR: Newton's method without smoothness or convexity. Technical Report Preprint: arXiv:1810.00303, 2018.

[21] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer. Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. Technical Report Preprint: arXiv:1909.05840, 2019.

[22] D. Wang, K. Fountoulakis, M. Henzinger, M. W. Mahoney, and S. Rao. Capacity releasing diffusions for speed and locality. Technical report, 2017. Preprint: arXiv:1706.05826.

[23] S. Wang, A. Gittens, and M. W. Mahoney. Scalable kernel k-means clustering with Nystrom approximation: Relative-error bounds. Technical report, 2017. Preprint: arXiv:1706.02803.

[24] P. Xu, F. Roosta-Khorasani, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. Technical report, 2017. Preprint: arXiv:1708.07164.

[25] P. Xu, F. Roosta-Khorasani, and M. W. Mahoney. Second-order optimization for non-convex machine learning: An empirical study. Technical report, 2017. Preprint: arXiv:1708.07827.

[26] Z. Yao, A. Gholami, K. Keutzer, and M. W. Mahoney. Large batch size training of neural networks with adversarial training and second-order information. Technical Report Preprint: arXiv:1810.01021, 2018.

[27] Z. Yao, A. Gholami, K. Keutzer, and M. W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. Technical Report Preprint: arXiv:1912.07145, 2019.

[28] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. Technical report, 2018. Preprint: arXiv:1802.08241.

[29] Z. Yao, A. Gholami, S. Shen, K. Keutzer, and M. W. Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. Technical Report Preprint: arXiv:2006.00719, 2020.

[30] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. W. Mahoney. Trust region based adversarial attack on neural networks. Technical Report Preprint: arXiv:1812.06371, 2018.