Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nyström method

Michał Dereziński (speaker), Rajiv Khanna and Michael W. Mahoney

A key motivation for our work is interpretable data summarization, which is a core challenge in reasoning about large datasets and machine learning models. Data summarization can be used, for example, to select a representative subset of gene variants from a genetics dataset, or a collection of most informative documents from a text database. When data are represented numerically, they are often described via matrices, in which case linear algebra suggests a natural (and in a certain sense optimal) way of performing data summarization: find the principal components corresponding to the largest directions of variance. These principal components work well for black-box models that are evaluated only in terms of prediction quality, but they are generally not interpretable in terms of the domain from which the data are drawn. They do not correspond to, say, a particular document or a gene variant, but rather a complex mixture of them. Neural networks, on which the machine learning community increasingly relies, simply exacerbate this problem. A long-standing challenge has been to find summaries of data which mimic the numerical properties of principal components and which are also interpretable.

The cost of interpretability, when formulated this way, has strong connections to the so-called Column Subset Selection Problem in Randomized Numerical Linear Algebra (RandNLA). It has been extensively studied in the literature, resulting in optimal worst-case guarantees, developed for the first time nearly fifteen years ago. These results are still relied upon as an important technical tool in recent works, including the best paper at last year's International Conference on Machine Learning, which applied them to Gaussian Process regression. The worst-case results suggest that the cost of interpretability should become progressively higher as the data summaries get larger. To verify this, we carefully constructed a worst-case example for this problem; and we showed that, for this example, the cost of interpretability does reach the worse-case bound for one size of the summary, but then it rapidly drops down for both larger and smaller sizes. As it turns out, in this and other worst-case examples, the high cost of interpretability is only a corner case.

To explain this, we go beyond worst-case analysis, and our results are able to accurately capture the non-linear behavior of the cost of interpretability. Specifically, we conclude that, except for certain pathological examples which we can characterize, the cost of interpretability is far smaller than suggested by prior work, and it is often negligible for real-world problems. To construct our interpretable summaries, we use a randomized subset selection technique based on Determinantal Point Processes (DPPs), which provide a probabilistic model of diversity that has emerged across many scientific domains. DPPs were first discovered in physics as a model of fermions in thermal equilibrium, and they have since been used in random matrix theory, graph theory and quantum mechanics. More recently, thanks to the emergence of efficient algorithms for DPP sampling, they have also become popular in RandNLA and Machine Learning.

Our analysis reveals that the previously observed worst-case examples are part of a larger phenomenon, which we call the multiple-descent curve in data summarization. This curve represents

a family of phase transitions, observed through a spike in the cost of interpretability, which occur when the data exhibit an underlying hierarchical structure. The multiple-descent curve can be observed not only in artificially constructed pathological examples but also in real-world problems. However, it can generally be avoided by tuning model parameters.

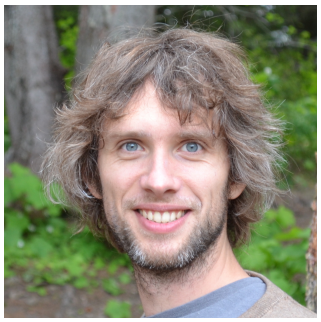
This phenomenon is very much reminiscent of the so-called double descent curve, which is exhibited by many machine learning models including deep neural networks. In the context of double descent, we distinguish two regimes of machine learning: the "classical" regime, where we have an abundance of training data relative to the number of model parameters we wish to learn; and the "modern" regime, which includes deep learning architectures, where there is far more parameters than training data. While each regime allows for constructing machine learning models that perform well on unseen data, the phase transition between the two regimes may lead to a spike in the error rate, resulting in poor performance.

Both the double descent curve in machine learning, and the new multiple-descent curve in data summarization, shown in our work, are related to fundamental phase transitions observed in the behavior of high-dimensional random matrices. Obtaining precise characterizations of these phenomena is of crucial importance to understanding modern machine learning as well as algorithmic-statistical tradeoffs that are central to the foundations of data science.

Media Contact: Michał Dereziński
Email: `mderezin@berkeley.edu`
Phone: 408-680-4652

Author Bios

Full name: **Michał Dereziński**
Title: Postdoctoral Researcher
Organization: University of California, Berkeley



Bio: Michał Dereziński is a Postdoctoral Researcher in the Department of Statistics at University of California, Berkeley. Previously, he was a research fellow at the Simons Institute for the Theory of Computing. He obtained his Ph.D. in Computer Science at University of California, Santa Cruz, advised by professor Manfred Warmuth. Michał's research is focused on developing scalable randomized algorithms with robust statistical guarantees for machine learning, data science and optimization. More information is available at: `https://users.soe.ucsc.edu/~mderezin/`.
Headshot: `https://users.soe.ucsc.edu/~mderezin/photo.jpg`

Full Name: **Rajiv Khanna**
Title: Postdoctoral Researcher
Organization: University of California, Berkeley

Bio: Rajiv Khanna is a Postdoctoral Researcher at the Foundations of Data Analysis Institute at University of California, Berkeley working with Michael Mahoney. He graduated with a PhD from UT Austin advised by Professors Joydeep Ghosh, and Alex Dimakis. His research interests include theoretical aspects of optimization and more recently generalization in neural networks.

Full name: **Michael W. Mahoney**
Title: Associate Professor
Organization: International Computer Science Institute and University of California at Berkeley
Bio: Michael W. Mahoney is at the University of California at Berkeley in the Department of Statistics and at the International Computer Science Institute (ICSI). He works on algorithmic and statistical aspects of modern large-scale data analysis. Much of his recent research has focused on large-scale machine learning, including randomized matrix algorithms and randomized numerical linear algebra, geometric network analysis tools for structure extraction in large informatics graphs, scalable implicit regularization methods, and applications in genetics, astronomy, medical imaging, social network analysis, and internet data analysis. He received him PhD from Yale University with a dissertation in computational statistical mechanics, and he has worked and taught at Yale University in the mathematics department, at Yahoo Research, and at Stanford University in the mathematics department. Among other things, he is on the national advisory committee of the Statistical and Applied Mathematical Sciences Institute (SAMSI), he was on the National Research Council's Committee on the Analysis of Massive Data, he co-organized the Simons Institute's fall 2013 and 2018 programs on the foundations of data science, and he runs the biennial MMDS Workshops on Algorithms for Modern Massive Data Sets. He is currently the Director of the NSF/TRIPODS-funded FODA (Foundations of Data Analysis) Institute at UC Berkeley. He holds several patents for work done at Yahoo Research and as Lead Data Scientist for Vieu Labs, Inc., a startup reimagining consumer video for billions of users. More information is available at: `https://www.stat.berkeley.edu/~mmahoney/`.
Headshot: `https://www.stat.berkeley.edu/~mmahoney/misc/mwm18-big.png`