Stat260/CS294: Randomized Algorithms for Matrices and Data

Instructor: Michael Mahoney (mmahoney@cs.stanford.edu)

Teaching Assistant: Yuchen Zhang (yuczhang@eecs.berkeley.edu)

Major research project: An important component of the class will be a major research project. The goal will be to drill down in much more detail on some topic related to what was covered in the lectures.

Important dates: Please email a ps or pdf of the following reports to the TA by 5PM on the date specified—*do not be late.*

- **Wed, Oct 23, 2013:** A brief statement stating (1) who you will be working with and (2) the project you plan to address. (One or two sentences as text is fine—we want to make sure people are working on a range of projects and will get back very soon if there is any problem.)
- **Wed, Oct 30, 2013:** Initial proposal, consisting of not more than a one page summary of the proposed project. (This will be mostly to make sure you are on the right track—we will get back to you within a few days if there are any issues, and we can also discuss any concerns you might have.)
- **Wed, Nov 20, 2013:** Mid-term report, consisting of approximately three to four pages with a brief summary of relevant literature, summary of proposed directions, and any questions or problems encountered.
- **Wed, Dec 18, 2013:** Final report, consisting of an eight to ten page report prepared in a format appropriate for publication.

Writing well (say, so that another person like the instructor or TA or a fellow student can read your paper and understand your ideas) is very hard, even after you have figured out the main ideas in the papers you will be summarizing or have the research results you want to present mostly completed. Thus, it would be wise to start the writing early and make a few iterations, perhaps running it by a fellow student, rather than waiting until the last minute.

More details: I suggest working in teams of two, in which case you should make clear who did what as a footnote in the final report. You may work individually or in a group of three, with an associated adjustment in expectations, but you should coordinate with the instructor first. The project will have two equally-important components:

- **Paper-reading component:** The goal will be to synthesize, summarize, and provide a detailed evaluation of some number of papers, most like roughly 2 to 4 papers, on some topic related to what we discussed in class. The final report should include a high-quality critique of these papers, much like but perhaps more detailed than a good review paper. It should place the papers in a broader context, and it should include a discussion of methods/results of the papers, of the strengths and weaknesses of the particular papers, as well as of their relationships with other related work.

- **Research component:** The goal is for you to perform new research, extending in a novel direction the papers you have read and reported upon. Ideally, you will obtain some interesting original theoretical or empirical results related to the topics we discussed in class and in doing so make substantial progress toward a nice conference paper. Since research sometimes does not succeed, success will certainly not be required to obtain a good grade on the final project. The final report should provide a detailed description of your methods and what novel theoretical or empirical results were obtained, your interpretation of the work, including why it succeeded or failed, what it reveals about the problem or the data or the techniques you used, and what its broader implications are.

The final report should be written in the form of a conference paper submission. Thus, it should include an introduction, a description of previous related work, a description of novel theoretical or empirical results that have been obtained, and a conclusion summarizing the results and further directions to follow.

The level of exposition of your report should be for one of your classmates, i.e., someone who has a good understanding of the area and of the lectures but who has not gone into detail on the particular topic you chose to address in detail. An example of a good final report (by Nikola Milosavljev in Tim Roughgarden's CS364A class in Fall 2004) may be found here:

- `http://theory.stanford.edu/(TILDE)tim/f06/nikola.pdf`

Clearly, your report will differ, depending among other things whether you are doing a more theoretical project or a more applied project, etc. Your report should, however, be at a similar level of depth, detail, and clarity.

Additional resources: There is slowly starting to be a body of publicly-available code for various randomized matrix algorithms. Here are several possibilities. While not necessary, it might help to start with one of these, rather than writing your own code from scratch.

- Blendenpik, for least-squares (Avron, Maymounkov, and Toledo):
  `http://www.mathworks.com/matlabcentral/fileexchange/25241-blendenpik`
- LSRN, for least-squares (Meng, Saunders, and Mahoney):
  `http://www.stanford.edu/group/SOL/software/lsrn.html`
- Low-rank approximation via projections (Tygert):
  `http://cims.nyu.edu/(TILDE)tygert/software.html`
- Nystrom approximation of SPSD matrices (Gittens and Mahoney):
  `http://users.cms.caltech.edu/(TILDE)gittens/nystrombestiary`
- Least absolute deviations and quantile regression (Yang, Meng, and Mahoney):
  `http://www.stanford.edu/(TILDE)jiyan/quantreg_wp/quantreg.html`
- Sampling columns from orthonormal matrices (Ipsen and Wentworth):
  `http://www4.ncsu.edu/(TILDE)ipsen/kappa_SQ_v3.zip`

In addition, depending on what you want to work on, you can do something like implement one of the algorithms we have discussed on your own, or implement it in a framework such as with MPI or Hadoop or Spark from the AmpLab.

Potential project topics: Here are a few suggested topics. (Of course, you are free to suggest another.) Note that some of these topics might be more easily-addressable than others, depending on your background and interests. In addition to the references on the class web page, which you should use as a resource to get started finding relevant papers, a few additional pointers for some of the suggested topics are given. These should just get you started—it would also be good to look at other references.

- Random divide-and-conquer matrix factorization methods: [15]
- CUR decompositions and landmark selection for general matrices: [16]
- Nystrom approximation landmark selection and for SPSD kernels: [4, 3, 2, 10]
- Scientific computing, pde, etc. applications of randomized matrix algorithms: [19, 6]
- Statistical versus algorithmic approaches to randomized matrix algorithms: [14]
- Matrix multiplication and information retrieval applications: [9]
- Better bounds with rank, stable rank, etc.: [13, 1, 11]
- Constructing kernels with randomized methods: [18]
- Leverage scores and improved regression diagnostics for small and large data: [5, 8]
- Randomized matrix algorithms in streaming models: [17, 7, 8]
- Preconditioning and random projections and lasso: [12]
- Approximate matrix multiplication for spectral and trace norms with $\{\pm 1\}$ random variables.
- Approximate matrix multiplication with deterministic column selection and barrier functions.
- Approximate matrix multiplication in input-sparsity time with Matrix Bernstein bounds.
- Approximate matrix multiplication with Trace norm errors.

References

[1] D. Achlioptas, Z. Karnin, and E. Liberty. Near-optimal distributions for data matrix sampling. Manuscript. 2013.

[2] F. Bach. Sharp analysis of low-rank kernel matrix approximations. Technical report. Preprint: arXiv:1208.2015 (2012).

[3] M.-A. Belabbas and P. J. Wolfe. On landmark selection and sampling in high-dimensional data analysis. *Philosophical Transactions of the Royal Society, Series A*, 367:4295–4312, 2009.

[4] M.-A. Belabbas and P. J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proc. Natl. Acad. Sci. USA*, 106:369–374, 2009.

[5] S. Chatterjee and A.S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

[6] J. Chiu and L. Demanet. Sublinear randomized algorithms for skeleton decompositions. Technical report. Preprint: arXiv:1110.4193 (2011).

[7] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 205–214, 2009.

[8] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[9] S. Eriksson-Bique, M. Solbrig, M. Stefanelli, S. Warkentin, R. Abbey, and I.C.F. Ipsen. Importance sampling for a Monte Carlo matrix multiplication algorithm, with application to information retrieval. *SIAM Journal on Scientific Computing*, 33(4):1689–1706, 2011.

[10] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. Technical report. Preprint: arXiv:1303.1849 (2013).

[11] J. T. Holodnak and I. C. F. Ipsen. Randomized matrix multiplication: Exact computation and probabilistic bounds. Technical report. Preprint: arXiv:1310.1502 (2013).

[12] J. Jia and K. Rohe. Preconditioning to comply with the irrepresentable condition. Technical report. Preprint: arXiv:1208.5584 (2012).

[13] E. Liberty. Simple and deterministic matrix sketching. Technical report. Preprint: arXiv:1206.0594 (2012).

[14] P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. Technical report. Preprint: arXiv:1306.5362 (2013).

[15] L. Mackey, A. Talwalkar, and M. I. Jordan. Divide-and-conquer matrix factorization. Technical report. Preprint: arXiv:1107.0789 (2011).

[16] M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

[17] S. Muthukrishnan. *Data Streams: Algorithms and Applications.* Foundations and Trends in Theoretical Computer Science. Now Publishers Inc, Boston, 2005.

[18] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Annual Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 2008.

[19] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.