Randomized Numerical Linear Algebra for Scipy

Randomized matrix algorithms have shown their utility in large-scale machine learning and statistical data analysis applications. Scipy has support for Randomized Numerical Linear Algebra since version 1.0, although its support is really limited (yet).

The **clarkosn_woodruff_transfrorm (CTW)** method was introduced under de scipy. linalg package. Given a matrix of size **m x n** the CWT reduces the dimensionality of the vector space to an embedded space. We say that given a matrix A, CWT creates a sketch of A, called A', with dimension **m' x n** where m' < m. In some cases, sketches can be used to get faster ways to find high-precision solutions to the problem. In other cases, sketches can be used to summarize the data by identifying the most important rows or columns.

I would implement **Blendenpik**, a least-squares solver for dense highly overdetermined systems. Blendenpik is based on these techniques. It outperforms LAPACK by significant factors and also scales significantly better than any QR-based solver. Some time ago Blendenpik was implemented successfully in Matlab, but there is non-open-source implementation widely available.

I already contacted Haim Avron, the author of Belndenpik. He was really excited about getting Blendenpik in Scipy. He showed interest and could help mentoring me without any problems.

Related with the Randomized Numerical Linear Algebra support I have some future work in mind. On the one hand, CWT could take advantage of sparse matrix representation but it does not exploit it in the current implementation. On the other hand, there is another method for sketching matrices called Fast Johnson-Lindenstrauss Transform (FJLT), which could be implemented too. By doing so, Blendenpik could adapt and have support for sparse and dense matrices with minimal changes. However, this is out of the scope of this proposal, but it is good to keep in mind this roadmap for a near future.

I used to work at IBM Almaden Research Center in San Jose. During the last year, I worked on a xdata open source project called libSkylark. The library is suitable for general statistical data analysis and optimization applications, but it is heavily focused on distributed systems. LibSkylark is a high-quality implementation of Randomized numerical methods, but it is not a good fit outside HPC systems.

If you have any further questions or comments you would like to discuss in depth any aspect of this proposal, send me an email to jomsdev@gmail.com .

Thank you for your time,

Jordi Montes.