January 19, 2016

RandNLA, Pythons, and the CUR for Your Data Problems

Reporting from G2S3 2015 in Delphi

By Efstratios Gallopoulos, Petros Drineas, Ilse Ipsen, Michael W. Mahoney

G2S3 participants and organizers pose in front of the European Cultural Centre of Delphi.

A few dozen graduate students and Ph.D. candidates, selected from a pool of over 140 highly-qualified applicants from prestigious universities around the world, attended the 2015 Gene Golub SIAM Summer School (G2S3 2015) in Delphi, Greece, last summer. Co-organizers Ilse Ipsen (North Carolina State University), Petros Drineas (Rensselear Polytechnic Institute), Michael Mahoney (University of California, Berkeley), and Stratis Gallopoulos (University of Patras) served as instructors along with Ken Clarkson (IBM Research-Almaden).

The theme of this year's school, Randomized Numerical Linear Algebra (RandNLA), is an interdisciplinary research area that exploits randomness as an algorithmic resource for the development of improved matrix algorithms for ubiquitous problems in large-scale data analysis. It utilizes ideas from theoretical computer science, numerical linear algebra, high-performance computing, and machine learning and statistics to develop, analyze, implement, and apply novel matrix algorithms. These algorithms can then facilitate the manipulation and analysis of so-called big data in numerous areas. Many popular machine learning and data analysis computations can be formulated as problems in linear algebra, but the questions of interest in machine learning and data analysis applications are very different from those historically considered in numerical linear algebra.

For instance, NP-hard problems have made their way into numerical linear algebra, a significant paradigm shift for numerical analysts who have traditionally formulated their problems to be solvable in polynomial time. As an example, most formulations of the Column Subset Selection problem are intractable, as are most formulations of the

so-called Non-negative Matrix Factorization (NMF) problem.

Attendees listening attentively to the lectures. Forefront: (left to right) graduate students Hyunghoon Cho (MIT) and Michael Hynes (Waterloo).

Alternatively, matrix factorizations—if properly instructed—can be used to discover latent information in the data, thus providing qualitative insight and interpretability, which is often of interest in scientific data applications. In this case, CUR is a natural factorization that provides a low-rank approximation of the underlying matrix using a product of selected Columns C and Rows R with a possible middle factor U (a tri-factorization). When the problem is massive in size and a CUR decomposition is sought, randomization becomes essential. More generally, random sampling and projection methods allow one to design provably accurate algorithms for problems containing the following: matrices so large that they require novel models of data access; matrices that cannot be stored at all, or can only be stored in slow-memory devices or only accessed via oracle calls; and/or problems that are computationally expensive or NP-hard. Randomized CUR and other low-rank decompositions lead to tractable solutions for terabyte-sized data, and have been used in genetics, astronomy, climate science, and mass spectrometry imaging.

RandNLA algorithms have also led to the best worst-case bounds for problems such as least-squares approximations and variants of the low-rank matrix approximation problem. In recent years they have begun to penetrate numerical linear algebra in many ways, leading to new research and software such as Blendenpik and IBM's Skylark.

Outdoor problem solving, with a view of the gardens. Eugenia Kontopoulou (Patras) at the blackboard.

RandNLA has grown rapidly over the last 15 years. The 4th edition of Golub and van Loan's Matrix Computations contains a special section on randomized low rank approximations and the CUR, a strong indication that RandNLA is now reaching the mainstream. The massive scale of today's problems in applications ranging from scientific simulations to data analytics requires the development of novel, disruptive methods, and randomization via RandNLA has proven effective in the design and analysis of matrix algorithms. The area has achieved a level of maturity allowing basic methods to be taught to a broad range of graduate students; thus, the timing for G2S3 2015 was ideal.

G2S3 activities took place at the European Cultural Centre of Delphi (ECCD), which also housed participants. Fittingly, the ECCD was built to bring people together in the spirit of the ancient Delphic tradition of cultural exchange.

The school was organized around talks on the following themes in RandNLA: sampling, numerical aspects, statistical and optimization aspects, random projections, and linear algebra and MATLAB tools. Many of the participants commented favorably on the opportunity to socialize and network with their peers and the instructors, an integral component of all G2S3s. Students participated in an open problems session and also presented their ongoing or planned research (sometimes outdoors, profiting from the art-filled yards and the breathtaking view of the valley at Delphi). Eugenia Kontopoulou demonstrated the RandNLA GUI for the TMG MATLAB toolbox built at the University of Patras.

In addition to the educational activities, participants also enjoyed a guided tour of ancient Delphi, visited the Delphi Museum and its Charioteer statue, and took a day-long bus ride along the Corinthian Bay to the Rio-Antirrio Bridge and the beautiful town of Lepanto.

G2S3 participants posing at Delphi's Temple of Apollo.

Mathworks offered all participants access to the latest MATLAB tools. Admittedly, several participants were spotted using Python, ipython, and other frameworks more popular in machine learning and data analysis. This was despite the fact that according to one legend, Apollo had to slay a serpent dragon guarding the cult center (in order to found his own temple), which rotted as its blood dried up under the sun. In classical Greek, to rot is "pytho" ($\Pi Y\Theta\Omega$), so the dragon became known as "python." The sunrays apparently turned the rotting dragon into Python!

Holding the RandNLA G2S3 at Delphi was a coincidence of sorts. Gene Golub was heavily involved in MMDS 2006, while several well-known contributors to RandNLA are Greek.

RandNLA techniques were recently used to negate a longstanding conjecture of the distinguished archeologist Sir Arthur Evans (who discovered the ancient palaces of the Minoans in Crete and defined the Linear A and Linear B scripts) regarding the ancestry of the early Cretans. According to a Homeric hymn, Apollo picked Delphi as a place "to make a glorious temple and an oracle for men." Delphi—the "navel of the world"—became a meeting point of pilgrims visiting from afar in search of oracular answers that would help cure their problems. Little did they know that some 2000 years later, Delphi would be the place where students of mathematics, computer science, and statistics would come to learn about the appropriate CUR for their data problems!

This year's G2S3 would not have been possible without SIAM. In spite of the maelstrom caused by the capital controls imposed on bank transactions in Greece the weekend following G2S3, the program was a success. Thanks are also due to the U.S. National Science Foundation for a significant grant supporting the travels of U.S.-based participants, ECCD staff for their hospitality and the University of Patras and its Computer Engineering and Informatics Department (CEID) for additional funding, efficient organization, and administration of finances.

To learn more:

Avron, H., Maymounkov, P., & Toledo, S. (2010). Blendenpik: Supercharging LAPACK's least squares solver. *SIAM J. Sci. Comput.*, 32(3), 1217-1236.

Dongarra, J., Kurzak, J., Luszczek, P., Moore, T., & Tomov, S. (19 Oct. 2015). On Numerical Algorithms and Libraries at Exascale. *HPC Wire.*

Golub, G.H. & Van Loan, C.F. (2013). *Matrix Computations* (4th ed.). Baltimore, MD: Johns Hopkins University Press.

Hughey, J.R., Paschou, P., Drineas, P., Mastropaolo, D., Lotakis, D.M., Navas, P.A., Michalodimitrakis, M.,

Stamatoyannopoulos, J.A., & Stamatoyannopoulos, G. (2013). A European population in Minoan Bronze Age Crete. *Nature Commun*., *4*, 1861.

Mahoney, M.W. (2011). Randomized Algorithms for Matrices and Data. *Found. Trends Mach. Learning*, 3, 123-224.

Randomized Numerical Linear Algebra for Large Scale Data Analysis. (2014). IBM Research. Retrieved from http://researcher.watson.ibm.com/researcher/view_group_pubs.php?grp=5131.

Scott, M. (2015). *Delphi: A History of the Center of the Ancient World*. Princeton, NJ: Princeton University Press.

Efstratios Gallopoulos is a professor and director of HPCLab in the Computer Engineering & Informatics Dept. at the University of Patras. Petros Drineas is an associate professor in the Computer Science Dept. at Rensselaer Polytechnic Institute. Ilse Ipsen is a professor of mathematics at North Carolina State University and associate director of the Statistical and Applied Mathematical Sciences Institute (SAMSI). Michael W. Mahoney is an associate professor in the Dept. of Statistics at the University of California, Berkeley.