Brummer & Partners MathDataLab mini course

RANDOMIZED METHODS IN LINEAR ALGEBRA AND THEIR APPLICATIONS IN DATA SCIENCE

Course content: The lectures will describe a set of recently developed randomized algorithms for accelerating matrix computations and the analysis of large data sets. A recurring theme will be the use of randomized embeddings that reduce the effective dimensionality of data sets while in certain respects preserving their geometric properties. We will describe how to use the methods in practice, and how their performance can be analyzed mathematically.

Lecturer: Per-Gunnar Martinsson, University of Texas at Austin.

Format: Three lectures delivered via Zoom, at 15:15 – 16:00 (CET), on November 17, 18, 19.

Target audience: The lectures are aimed at graduate students at the master and doctoral levels, as well as to researchers and practitioners interested in using these techniques. A solid foundation in linear algebra will be assumed (singular value and QR factorization, methods for least squares problems, etc.), as well as a basic understanding of probability theory (expectation, variation, independent linear variables, the normal distribution, etc.).

Web resources: Slides, links to surveys, etc, at: `https://users.oden.utexas.edu/~pgm/2020_kth_course/`

Tentative timeline:

LECTURE 1:

Random embeddings: A random embedding is a map that maps data in a high dimensional space to to a "sketch" that lives in a lower dimensional space. It should with high probability approximately preserve key properties of the data such as pairwise distances. We will discuss connections to Johnson-Lindenstrauss theory, to so call "fast J-L transforms", and discuss some applications such as nearest neighbor search.

The low rank approximation problem: A review of classical tools for constructing low rank approximations to matrices — the singular value decomposition (SVD), column pivoted QR, Krylov methods, etc.

Randomized methods for low rank approximation: We describe the "randomized SVD (RSVD)" algorithm that exploits random embeddings to compute a low rank approximation to a matrix. High practical speed is attained due to the fact that it requires less *communication* than traditional methods. Stochastic error analysis. Power iteration to improve accuracy.

LECTURE 2:

Streaming ("single-view") algorithms: Remarkably, randomization allows the construction of a low rank approximation to a matrix under the very strict constraint that each entry of the matrix can be viewed only once. This enables processing of datasets that are so huge that they cannot be stored.

Matrix approximation by sampling: It is sometimes possible to build an approximation to a matrix by looking at only a small subset of its entries. Such algorithms have sublinear complexity, and enable computations that would otherwise be impossible. However, such methods are less robust than methods based on random embeddings.

The CUR and interpolative decompositions: These matrix decompositions use a subset of the rows to span the row-space and/or a subset of the columns to span the column space. This leads to a reduction in storage requirements, and a decomposition that preserves properties of the original matrix such as sparsity or non-negativity. Applications to data interpretation.

LECTURE 3:

Randomized linear solvers: A very active area of current research is how to use randomization to efficiently solve a linear system $Ax = b$. Remarkable success has been achieved for special classes of problems such as solving overdetermined linear regression problems in a least square sense. Blendenpik. Randomized Kaczmarz method. Randomized solvers for graph Laplacians.

Analysis of the randomized SVD: If time permits, we will close the lectures by reviewing how results from random matrix theory can be applied to analyze the performance of randomized methods for low rank approximation.