APPM 4720/5720, Spring 2016

Open Topics in Applied Mathematics: Fast Algorithms for Big Data

Gunnar Martinsson
phone: 492-2646, email: martinss@colorado.edu, office: ECOT 233

(1). **Meeting times.** MWF 10.00 - 10.50, ECCR 150.

(2). **Office hours.** Monday 4:00pm - 4:50pm and Wednesday 3:00pm - 4:50pm.

(3). **Website.** All course material (lecture notes, homeworks, project descriptions, tutorial codes), and an up-to-date timeline will be posted at the following URL:

`http://amath.colorado.edu/faculty/martinss/Teaching/APPM5720_2016s/`

(4). **Course objectives.** The course is motivated by two important developments over the last decade. The first is the enormous and growing demand for techniques to extract information from very large data sets. The second is a change in computer hardware from an environment where the cost of *floating point operations (flops)* was the primary constraint to a modern environment characterized by multicore and parallel computers where the cost of *communication* has become the primary bottlebeck.

The course will describe powerful algorithms that scale well both with respect to the size of the data set, and with the number of processors available. A theoretically ideal algorithm should satisfy some relation like:

$$\text{Time to complete computation} \sim \frac{\text{Size of the data set}}{\text{Number of processors available}}.$$

Many of the algorithms we will discuss come very close to attaining this theoretical lower bound.

The techniques described will be applicable to a broad range of problems in data analysis, computational statistics, machine learning, and scientific computing.

Mathematically, the course will focus on linear algebraic techniques for dimension reduction. Matrix factorizations and randomized algorithms will play a prominent part. A more precise list of topics covered can be found in Section (11).

(5). **Prerequisites.** To take this course, it is essential that you know linear algebra well. A course such as APPM3310 (Matrix Methods) or the equivalent is absolutely necessary, as we will frequently use matrix factorizations such as the SVD and the QR. Familiarity with basic probability is also important (Gaussian and Bernoulli random numbers; the concepts of expectations, standard deviations, etc). Basic numerical analysis (accuracy, stability, floating point arithmetic, etc) is assumed.

Knowledge of basic programming in Matlab is required, as is basic knowledge of analysis of algorithms such as estimating asymptotic complexity, etc.

Finally, some familiarity with Fourier methods is very helpful. For one part of the course, knowledge of basic electrostatics and the properties of the Laplace and Helmholtz equations will be assumed, but this material covers only one or two weeks, so this is not an essential pre-requisite.

(6). **Grading.** There will be no regular exams. Instead, your grade will be based on projects, homeworks, etc, as follows:

- 20% for scribing, see Section (8).
- 30% for regular homeworks, see Section (9).
- 10% for the reference homework, see Section (9).
- 40% for a final project, see Section (10).

We will use a Google Spreadsheet to coordinate the scribing and the reference homeworks. You will receive an email inviting you to edit this spreadsheet in the first week. Please choose two lectures and one homework among the empty slots and enter your name.

Important: If you have not received the email invitation to edit the Google spreadsheet by Thursday, January 14, then please contact the instructor via email.

You can keep track of your scores for each component of the course that has been graded via the course D2L page. Please allow for at least 7 days after the deadline for scores to show up.

(7). **Text.** The course is defined by the material covered in class, and there is no “official” text book. Course notes and links to papers discussed in class will be posted on the course website.

(8). **Lecture notes / scribing.** As a participant in the course, you will be required to sign up for two lectures for which you will serve as a *scribe.* During the lecture when you are a scribe, you will take careful notes, type them up after the class (using latex if at all possible), and email them to the instructor within 48h. They will then be posted to the course webpage.

A template for the scribe notes can be downloaded from the course webpage.

(9). **Homeworks.** There will be 6 homeworks, due at the end of weeks 3, 5, 7, 9, 11, and 13. Working in groups is allowed and encouraged, with the maximal group size being 3 for 4720 and 2 for 5720.

Each individual in the course (not each group!) will be required to sign up for one homework problem and be responsible for producing a “reference solution.” This should be a typed solution, and should include Matlab codes where appropriate. The instructor will review the submitted reference homework, and suggest edits/corrections where appropriate. Once the reference homework is complete, it will be posted to the course webpage as a solution. You are allowed to work in your homework group to prepare the reference homework, but only one student will get credit for the problem.

Each regular homework set will be worth 5% of the grade. In addition, your reference homework problem will be worth 10%.

(10). **Project.** Your grade in this course will to 40% be based on a final project. You are allowed (and encouraged!) to work in pairs on the project. Groups of three students could be allowed if the project chosen is particularly labor intensive, but this requires instructor permission.

In the last week of the course, each group is expected to deliver a brief (10 – 15 minutes) presentation of the project, and to hand in a final project report.

A number of suggested projects will be listed on the course webpage. You are also very welcome to think of projects on your own; if you want to go with this option, you need to discuss the chosen project with the instructor to get it approved. Please initiate this discussion no later than March 15, if possible.

The expectation is that the project is based on material covered in the first two thirds of the course, and will be completed during the last third. You must pick a project and notify the instructor of what your project is by March 19.

(11). **Time line.** The plan is to cover the following topics:

Week:	*Material covered:*
1:	Low-rank approximation – the problem formulation and a brief survey of applications. The Singular Value Decomposition (SVD) and the Eckart-Young theorem on optimality.
2:	Power iterations, and Krylov methods. Gram-Schmidt and the QR factorization. How to cheaply get an approximate SVD from a QR factorization.
3:	Interpretation of data. The interpolative decomposition (ID) and the CUR decomposition.
4,5:	Randomized methods for computing low-rank approximations. Connection to Quick-Sort and Monte Carlo.
6,7:	Applications of low-rank approximation: Principal Component Analysis (PCA), Latent Semantic Indexing (LSI), eigenfaces, pagerank, potential evaluation.
8,9:	Linear regression problems. L^2 and L^1 minimization. Brief introduction to linear programming.
10:	Johnson-Lindenstrauss methods; random projections as a tool for dimensionality reduction.
11:	Nearest neighbor search.
12:	Clustering and the "k-means" problem.
13:	The Fast Fourier Transform.
14:	The Fast Multipole Method.
15:	Project presentations.

(12). **This is a new course!** This is the first time this course is run. As a consequence, the timeline given above is just a rough guide — we may find that some topics require more or less time as we go. The course webpage will be updated during the semester to show the actual pace of progress.

Moreover, all homeworks and projects are also new. This means that they might be easier or harder than intended. If something seems odd, then please notify the instructor asap — you might very well have found a mistake!

Disability: If you qualify for accommodations because of a disability, please submit to your professor a letter from Disability Services in a timely manner (for exam accommodations provide your letter at least one week prior to the exam) so that your needs can be addressed. Disability Services determines accommodations based on documented disabilities. Contact Disability Services at 303-492-8671 or by e-mail at `dsinfo@colorado.edu`. If you have a temporary medical condition or injury, see Temporary Injuries guidelines under the Quick Links at the Disability Services website and discuss your needs with your instructor. See:
http://www.colorado.edu/disabilityservices/students/temporary-medical-conditions

Religious holidays: Campus policy regarding religious observances requires that faculty make every effort to deal reasonably and fairly with all students who, because of religious obligations, have conflicts with scheduled exams, assignments or required attendance. In this class, if you would like to submit a request for an accommodation, please send a written request (email is ok) to the instructor within the first two weeks of class. For full details on the campus policy, see:
http://www.colorado.edu/policies/observance-religious-holidays-and-absences-classes-andor-exams

Classroom behavior: Students and faculty each have responsibility for maintaining an appropriate learning environment. Students who fail to adhere to such behavioral standards may be subject to discipline. Faculty have the professional responsibility to treat all students with understanding, dignity and respect, to guide classroom discussion and to set reasonable limits on the manner in which they and their students express opinions. Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender variance, and nationalities. Class rosters are provided to the instructor with the student's legal name. I will gladly honor your request to address you by an alternate name or gender pronoun. Please advise me of this preference early in the semester so that I may make appropriate changes to my records. See polices at:
http://www.colorado.edu/policies/student-classroom-and-course-related-behavior
http://www.colorado.edu/osc/sites/default/files/attached-files/studentconductcode_15-16.pdf

Sexual harassment: The University of Colorado Boulder (CU-Boulder) is committed to maintaining a positive learning, working, and living environment. CU-Boulder will not tolerate acts of sexual misconduct, discrimination, harassment or related retaliation against or by any employee or student. CUs Sexual Misconduct Policy prohibits sexual assault, sexual exploitation, sexual harassment, intimate partner abuse (dating or domestic violence), stalking or related retaliation. CU-Boulders Discrimination and Harassment Policy prohibits discrimination, harassment or related retaliation based on race, color, national origin, sex, pregnancy, age, disability, creed, religion, sexual orientation, gender identity, gender expression, veteran status, political affiliation or political philosophy. Individuals who believe they have been subject to misconduct under either policy should contact the Office of Institutional Equity and Compliance (OIEC) at 303-492-2127. Information about the OIEC, the above referenced policies, and the campus resources available to assist individuals regarding sexual misconduct, discrimination, harassment or related retaliation can be found at:
http://www.colorado.edu/institutionalequity/

Honor code: All students enrolled in a University of Colorado Boulder course are responsible for knowing and adhering to the academic integrity policy of the institution. Violations of the policy may include: plagiarism, cheating, fabrication, lying, bribery, threat, unauthorized access, clicker fraud, resubmission, and aiding academic dishonesty. All incidents of academic misconduct will be reported to the Honor Code Council (honor@colorado.edu; 303-735-2273). Students who are found responsible of violating the academic integrity policy will be subject to nonacademic sanctions from the Honor Code Council as well as academic sanctions from the faculty member. Additional information regarding the academic integrity policy can be found at:
http://honorcode.colorado.edu