Stat260/CS294: Randomized Algorithms for Matrices and Data

Homework 1

Instructor: Michael Mahoney *Due: October 9, 2013*

Problem 1

(a) For $A \in \mathbb{R}^{m \times n}$, show that the matrix $A^T A$ is positive semi-definite; and that $A^T A$ is positive definite if and only if the columns of A are linearly independent.

(b) What is the SVD of a symmetric matrix? Of an orthoognal matrix? Of the identity matrix?

(c) For $p = 1, 2, \infty$, verify that the functions $||\cdot||_p$ are norms. Then, for a vector $x \in \mathbb{R}^n$, show that

$$||x||_\infty \leq ||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2 \leq n||x||_\infty,$$

and, for each inequality, provide an example demonstrating that the inequality can be tight.

(d) For vectors $x, y \in \mathbb{R}^n$, show that $|x^T y| \leq ||x||_2 ||y||_2$, with equality if and only if x and y are linearly dependent. More generally, show that $|x^T y| \leq ||x||_1 ||y||_\infty$. Note that this implies that $||x||_2^2 \leq ||x||_1 ||x||_\infty$; and that these are special cases of Hölder's inequality.

(e) For $A \in \mathbb{R}^{m \times n}$, show that $\text{Trace}(A^T A) = \sum_{ij} A_{ij}^2$, and show that $\sqrt{\sum_{ij} A_{ij}^2}$ is a norm on $m \times n$ matrices. This is the Frobenius norm, denoted $||\cdot||_F$. Show that, in addition to satisfying the defining properties of a norm, the Frobenius norm is a submultiplicative norm, in that

$$||AB||_F \leq ||A||_F ||B||_F,$$

whenever the dimensions are such that the product AB is defined.

(f) Recall the definition of the spectral norm of an $m \times n$ matrix A: $||A||_2 = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(X)$, where $\lambda_{max}(A^T A)$ is the largest eigenvalue of $A^T A$ and σ_{max} is the largest singular value of A. Show that the Frobenius norm and the spectral norm are unitarily invariant: if U and V are unitary (orthogonal in the real case) matrices, then $||U^T A V||_\xi = ||A||_\xi$, for $\xi = 2, F$.

(g) The relationship between the SVD of A and the spectral decomposition of $A^T A$ that was discussed in class allows one to obtain results on the SVD of general matrices from results on symmetric matrices. Here is another way. For $A \in \mathbb{R}^{m \times n}$, where assume $m \geq n$, let the SVD of A be

$$U^T A V = \text{diag}(\Sigma, 0),$$

where $\Sigma \in \mathbb{R}^{n \times n}$ and 0 is the all-zeros matrix of appropriate dimension. Then, show that the matrix

$$C = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$$

has eigenvalues $\pm\sigma_1, \ldots, \pm\sigma_n$, corresponding to the eigenvalues

$$\begin{pmatrix} U^{(i)} \\ \pm V^{(i)} \end{pmatrix},$$

for $i = 1, \ldots, n$, where $U^{(i)}$ is the i-th column of U and $V^{(i)}$ is the i-th column of V; and, in addition, C has $m - n$ zero eigenvalues whose eigenvectors are

$$\begin{pmatrix} U^{(i)} \\ 0 \end{pmatrix},$$

for $i = n + 1, \ldots, m$.

Problem 2

Suppose that we can obtain samples X_1, X_2, ... of a random variable X and that we want to use these samples to estimate $E[X]$. Using n samples, we can use $\left(\sum_{i=1}^{n} X_i\right)/n$ for our estimate of $E[X]$. We want the estimate to be within $\epsilon E[X]$ from the true value of $E[X]$ with probability at least $1-\delta$. We may not be able to use Chernoff's bound directly to bound how good our estimate is if X is not a 0-1 random variable, and we do not know its moment generating function. Here, we develop an alternative approach that requires having a bound on the variance of X. Let $r = \sqrt{Var[X]}/E[X]$.

(a) Show using Chebychev's inequality that $O(r^2/\epsilon^2\delta)$ samples are sufficient to solve the problem.

(b) Suppose that we need only a weak estimate that is within $\epsilon E[X]$ of $E[X]$ with probability at least $3/4$. Show that $O(r^2/\epsilon^2)$ samples are enough for this weak estimate.

(c) Show that, by taking the median of $O(\log(1/\delta))$ such weak estimates, we can obtain an estimate that is within $\epsilon E[X]$ of $E[X]$ with probability at least $1-\delta$. Conclude that only $O(r^2 \log(1/\delta)/\epsilon^2)$ samples is sufficient.

Next, assume that the X_i are independent 0-1 random variables, and suppose we want to estimate $X = \sum_{i=1}^{n} X_i$. Let $\mu = E[X]$, and let μ_L and μ_H be such that $\mu_L \leq \mu \leq \mu_H$.

(d) Show that, for and $\delta > 0$,

$$Pr\left[X \geq (1+\delta)\mu_H\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu_H}.$$

(d) Show that, for and $\delta \in (0,1)$,

$$Pr\left[X \leq (1-\delta)\mu_L\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu_L}.$$

Problem 3

In the class, we considered random projections with entries drawn from the standard normal distribution $N(0,1)$. Here, we consider a random projection matrix with entries that are sampled from a discrete distribution that is symmetric about the origin, and we consider how the random projection matrix can be sparsified if the input satisfies particular assumptions.

(a) Let the random $k \times d$ projection matrix R have entries drawn i.i.d. from $\{-1, +1\}$. In particular, $\{R_{ij}\}$ are independent random variables and

$$R_{ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

Given a vector u, we define the projection $f(u) = \frac{1}{\sqrt{k}} Ru$. Show that the Johnson-Lindenstrauss bound holds: if $k = \Omega(\log(n)/\epsilon^2)$, then for every set P of n points in $\mathbb{R}^d$, with probability at least $1 - 1/n$, for all $u, v \in P$

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2.$$

(b) If the input vectors are well-spread, then the input matrix can be sparsified. To define "well-spread" more concretely, we assume that the set of points P only contains unit vectors, and for any $u \in P$ it satisfies $\|u\|_\infty \leq 1/\sqrt{n}$. Let us set

$$q = \frac{c_1}{n} \log(n/\epsilon)$$

for a sufficiently large constant c_1, and let us assume that $q \leq 1$. Let $\{R_{ij}\}$ be independent random variables such that

$$R_{ij} = \begin{cases} +q^{-1/2} & \text{with probability } q/2, \\ -q^{-1/2} & \text{with probability } q/2, \\ 0 & \text{with probability } 1-q. \end{cases}$$

Show that the Johnson-Lindenstrauss bound in question (a) holds.

In the class, we also considered approximate matrix multiplication bounds when the sketching matrices were random sampling matrices constructed with judiciously-chosen sampling probabilities. Here, we consider the use of dense $\{-1, +1\}$ random projection matrices as the sketching matrices in the approximate matrix multiplication algorithm. More concretely, consider a $k \times n$ random matrix R where $\{R_{ij}\}$ are independent random variables such that

$$R_{ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

Given matrices $A \in \mathbb{R}^{m \times n}$ and define $C = \frac{1}{\sqrt{k}} AR^T$. We approximate the product AA^T by the product CC^T.

(c) Show that CC^T is an unbiased estimate with bounded variance:

$$\mathbb{E}[CC^T] = AA^T \quad \text{and} \quad \mathbb{E}[\|AA^T - CC^T\|_F^2] \leq \frac{1}{k}\|A\|_F^4.$$

(d) Let $\delta \in (0,1)$ and $\eta = 1 + \sqrt{8\log(1/\delta)}$. Show that with probability $1-\delta$, the approximation error has bounded Frobenius norm $\|AA^T - CC^T\|_F^2 \leq \frac{\eta^2}{k}\|A\|_F^4$.

(e) Now we turn to the spectral norm bound. Assume that $\|A\|_2 \leq 1$ and $\sum_{i=1}^n \|A^{(i)}\|_2 \leq M$ where $A^{(i)}$ is the i-th column of A. Let $\epsilon \in (0,1)$, show that $\|AA^T - CC^T\|_2 \leq \epsilon$ with probability at least $1 - (2k)^2 \exp\left(-\frac{k\epsilon^2}{16M^2 + 8M^2\epsilon}\right)$. Set the parameters of the algorithm so that this failure probability is less than δ. Is this bound tight, or can it be improved?

Problem 4

Here, we will consider the empirical performance of random sampling and random projection algorithms for approximating the product of two matrices. You may use Matlab, or C, or R, or whatever software package you prefer to do your implementations, but be sure to describe what you used in sufficient detail that someone else could reproduce your results.

Let A be an $n \times d$ matrix, with $n \gg d$, and consider approximating the product A^TA as well as the product $U_A^TU_A$, where U_A is the $n \times d$ matrix consisting of the left singular vectors (or, if you prefer, the Q matrix from a QR decomposition) of A. Generate the matrices A from one of three different classes of distributions introduced below; and generate the matrices U_A by first generating A in the following manner and then performing the SVD or a QR decomposition of A.

- Generate a matrix A from multivariate normal $N(1_d, \Sigma)$, where the (i,j)th element of $\Sigma_{ij} = 2 \times 0.5^{|i-j|}$. (Refer to as GA data.)
- Generate a matrix A from multivariate t-distribution with 3 degree of freedom and covariance matrix Σ as before. (Refer to as T_3 data.)
- Generate a matrix A from multivariate t-distribution with 1 degree of freedom and covariance matrix Σ as before. (Refer to as T_1 data.)

To start, consider matrices of size $n \times d$ equal to 500×50. (So, you should have six matrices, one matrix A generated in each of the above ways, and one matrix U_A consisting of the left singular vectors of a matrix A generated in each of the above ways.)

(a) For each matrix, approximate the product ($A^T A$ or $U^T U$) with the random sampling algorithm we discussed in class, i.e., by sampling with respect to a probability distribution that depends on the norm-squared of the rows of the input matrix. Plot how uniform or nonuniform is that probability distribution. Plot the performance of the spectral and Frobenius norm error as a function of the number of samples.

(b) For each matrix, approximate the product ($A^T A$ or $U^T U$) with the random sampling algorithm we discussed in class, except that the uniform distribution, rather than the norm-squared distribution, should be used to construct the random sample. Plot the performance of the spectral and Frobenius norm error as a function of the number of samples. For which matrices are the results similar and for which are they different than when the norm-squared distribution is used.

(c) For each matrix A, approximate the product $A^T A$ with the random sampling algorithm we discussed in class, except that instead of the uniform or norm-squared distribution, you sample rows of A with respect to the norm-squared of the corresponding U matrix. Plot the performance of the spectral and Frobenius norm error as a function of the number of samples. For which matrices are the results similar and for which are they different than when the norm-squared distribution is used.

(d) For each matrix, approximate the product ($A^T A$ or $U^T U$) with the following variant of the random sampling algorithm we discussed in class. Instead of letting C be a sampled-and-rescaled version of a small number of rows of A, and approximating $A^T A$ by $C^T C$, let C be the result of performing a random projection, where the random projection matrix consists of scaled Gaussian entries or scaled $\{\pm 1\}$ entries (and still approximate $A^T A$ by this form of $C^T C$). Plot the performance of the spectral and Frobenius norm error as a function of the number of dimensions projected onto, and describe how the performance compares with the random sampling algorithm. Describe any differences you see between the use of a projection matrix consisting of Gaussian random variables and one consisting of $\{\pm 1\}$ random variables.

(e) Modify the previous projection-based variant to include "sparse" projection matrices by zeroing out some fraction of the entries of the projection matrix, as in the previous problem. (Be sure to rescale all the entries that you keep in the proper way.) Plot the performance of the spectral and Frobenius norm error as a function of the number of dimensions projected onto, and describe how the performance behaves for each of the above three data sets, as a function of the sparsity. For which data sets can you sparsity aggressively without loosing reconstruction quality, and for which do you have problems if you sparsity too aggressively. Describe the problems you encounter, and why you encounter them.

Next, describe how sensitive are your results to the particular random matrices that you constructed. In particular, describe quantitatively how much variability is there in your results if, for each class of matrices, you generate 20 matrices (rather tha 1 matrix) according to the above procedure; and describe quantitatively how your results change if you generate matrices of size 5000×50 or 500×5.