Stat260/CS294: Randomized Algorithms for Matrices and Data

Lecture 4 - 09/16/2013

Lecture 4: Concentration and Matrix Multiplication, Cont.

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

4 Concentration and Matrix Multiplication, Cont.

Today, we will continue with our discussion of scalar and matrix concentration, with a discussion of the matrix analogues of Markov's, Chebychev's, and Chernoff's Inequalities. Then, we will return to bounding the error for our approximating matrix multiplication algorithm. We will start with using Hoeffding-Azuma bounds from last class to get improved Frobenius norm bounds, and then (next time) we will describe how to use the matrix concentration results to get spectral norm bounds for approximate multiplication.

Here is the reading for today.

- Appendix of: Recht, "A Simpler Approach to Matrix Completion"
- Oliveira, "Sums of random Hermitian matrices and an inequality by Rudelson"
- Drineas, Kannan, and Mahoney, "Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication"

4.1 Matrix Concentration

We will now discuss several results having to do with concentration of matrix-valued random variables. We start with a matrix version of the Markov inequality.

Lemma 1 (Matrix Markov Inequality) *Let X be a random PSD matrix, and let A be a fixed PD matrix. Then, $\forall A$, $\mathbf{Pr}\left[X \npreceq A\right] \le \mathbf{Tr}\left(\mathbf{E}\left[X\right] A^{-1}\right)$.*

Proof: Consider the random variable $A^{-1/2}XA^{-1/2}$. Observe that, if $X \npreceq A$, then $A^{-1/2}XA^{-1/2} \npreceq I$. In this case,

$$1 < \left\| A^{-1/2}XA^{-1/2} \right\|_2 .$$

Let $\mathcal{X}_{X\npreceq A}$ be the characteristic/indicator function of the event $X \npreceq A$. Then, the claim is that

$$\mathcal{X}_{X\npreceq A} \le \mathbf{Tr}\left(A^{-1/2}XA^{-1/2}\right) .$$

To prove the claim, observe that the RHS ≥ 0. If the LHS $= 0$, then we are done. Otherwise, if the LHS $= 1$, then $1 < \left\| A^{-1/2}XA^{-1/2} \right\|_2 \leq \mathbf{Tr}\left(A^{-1/2}XA^{-1/2}\right)$. So,

$$\mathbf{Pr}\left[X \npreceq A\right] = \mathbf{E}\left[\mathcal{X}_{X \npreceq A}\right] \leq \mathbf{E}\left[\mathbf{Tr}\left(A^{-1/2}XA^{-1/2}\right)\right] = \mathbf{E}\left[\mathbf{Tr}\left(XA^{-1}\right)\right] = \mathbf{Tr}\left(\mathbf{E}\left[X\right]A^{-1}\right),$$

where the second equality follows from the cyclic properties of the trace, and where the last follows since the trace is linear.

$\diamond$

Although we will not use the matrix version of the Chebychev inequality in what follows, we include it for completeness and for comparison with the scalar version.

Lemma 2 (Matrix Chebychev Inequality) *Let X be a random PSD matrix, and let A be a fixed PD matrix. Then, $\forall A$, $\mathbf{Pr}\left[\left|X - \mathbf{E}\left[X\right]\right| \npreceq A\right] \leq \mathbf{Tr}\left(\mathbf{Var}\left[X\right]A^{-2}\right)$. XXX. CLARIFY WHAT THAT NORM THING IS ON LHS.*

Proof: First note that $(X - \mathbf{E}\left[X\right])^2 \preceq A^2$ implies that $\left|X - \mathbf{E}\left[X\right]\right| \preceq A$. The reason for this is that $\sqrt{\cdot}$ is operator monotone. (I.e., while it is obvious for numbers, it is true but non-obvious for matrices.) So,

$$\begin{aligned}
\mathbf{Pr}\left[\left|X - \mathbf{E}\left[X\right]\right| \npreceq A\right] &\leq \mathbf{Pr}\left[(X - \mathbf{E}\left[X\right])^2 \npreceq A^2\right] \\
&\leq \mathbf{Tr}\left(\mathbf{E}\left[(X - \mathbf{E}\left[X\right])^2\right]A^{-2}\right) \\
&= \mathbf{Tr}\left(\mathbf{Var}\left[X\right]A^{-2}\right).
\end{aligned}$$

$\diamond$

Next, what we really want to do is get a matrix analogue of the Chernoff bound. Here is one form of it; we will give more of a history below.

Theorem 1 (Matrix Chernoff Bound) *Let $X_1, \ldots, X_n$ be independent symmetric random matrices in $\mathbb{R}^{d \times d}$. Then, $\forall$ invertible $d \times d$ matrices T,*

$$\mathbf{Pr}\left[\sum_{k=1}^{n} X_k \npreceq nA\right] \leq d \prod_{k=1}^{n} \left\|\mathbf{E}\left[\exp\left(TX_kT^* - TAT^*\right)\right]\right\|_2,$$

where T^* denotes the transpose of the (real-valued) matrix T.

Proof: First, by the usual properties of the semi-definite ordering, we have that

$$\begin{aligned}
\mathbf{Pr}\left[\sum_{k=1}^{n} X_k \npreceq nA\right] &= \mathbf{Pr}\left[\sum_{k=1}^{n} (X_k - A) \npreceq 0\right] \\
&= \mathbf{Pr}\left[\sum_{k=1}^{n} T(X_k - A)T^* \npreceq 0\right] \\
&= \mathbf{Pr}\left[\exp\left(\sum_{k=1}^{n} T(X_k - A)T^*\right) \npreceq I_d\right].
\end{aligned}$$

By combining this with the Matrix Markov Inequality, and since the trace is linear, it follows that

$$\begin{aligned} \mathbf{Pr}\left[\sum_{k=1}^{n} X_k \not\preceq nA\right] &\leq \mathbf{Tr}\left(\mathbf{E}\left[\exp\left(\sum_{k=1}^{n} T(X_k - A)T^*\right)\right]\right) \\ &\leq \mathbf{E}\left[\mathbf{Tr}\left(\exp\left(\sum_{k=1}^{n} T(X_k - A)T^*\right)\right)\right]. \end{aligned}$$

Next, observe that we can peel apart the various terms as follows

$$\begin{aligned} \mathbf{Pr}\left[\sum_{k=1}^{n} X_k \not\preceq nA\right] &\leq \mathbf{E}\left[\mathbf{Tr}\left(\exp\left(\sum_{k=1}^{n-1} T(X_k - A)T^*\right)\exp\left(T(X_n - A)T^*\right)\right)\right] \\ &= \mathbf{E}\left[\mathbf{Tr}\left(\exp\left(\sum_{k=1}^{n-1} T(X_k - A)T^*\right)\mathbf{E}\left[\exp\left(T(X_n - A)T^*\right)\right]_n\right)\right]_{1,\cdots,n-1} \\ &\leq \left\|\mathbf{E}\left[\exp\left(T(X_n - A)T^*\right)\right]\right\|_2 \mathbf{E}\left[\mathbf{Tr}\left(\exp\left(\sum_{k=1}^{n-1} T(X_k - A)T^*\right)\right)\right]_{1,\cdots,n-1}, \end{aligned}$$

where the first line follows from the Golden-Thompson inequality; the second line follows from the independence of the X_k; and the third line follows by strong submultiplicitivity, i.e., since $\mathbf{Tr}\left(AB\right) \leq \mathbf{Tr}\left(A\right) \left\|B\right\|_2$, if A and B are SPSD. XXX. NEED TO FIX THAT NOTATION WITH EXPECTATION. By iterating this process it follows that

$$\begin{aligned} \mathbf{Pr}\left[\sum_{k=1}^{n} X_k \not\preceq nA\right] &\leq \prod_{k=2}^{n} \left\|\mathbf{E}\left[\exp\left(T(X_k - A)T^*\right)\right]\right\|_2 \mathbf{E}\left[\mathbf{Tr}\left(\exp\left(T(X_1 - A)T^*\right)\right)\right] \\ &\leq d\prod_{k=1}^{n} \left\|\mathbf{E}\left[\exp\left(T(X_k - A)T^*\right)\right]\right\|_2, \end{aligned}$$

where the last line follows since if A is PD, then $\mathbf{Tr}\left(A\right) = \sum_{i=1}^{d} \lambda_i(A) \leq d\lambda_{max}(A)$, where $\lambda_i(A)$ is the i^{th} eigenvalue of A and where $\lambda_{max}(A)$ is the largest eigenvalue of A. XXX. CAN I DO THOSE LAST STEPS IN ONE STEP.

$\diamond$

We will use this Matrix Chernoff Bound to establish an inequality that we will use. Note that, as in the scalar case, one can get lots of variations, and we will use Bernstein version due to Recht.

Theorem 2 (Noncommutative Bernstein Inequality) *Let $X_1, \ldots, X_L$ be independent zero-mean random matrices of dimension $d_1 \times d_2$. Suppose that $\rho_k = \max\{\left\|\mathbf{E}\left[X_k X_k^*\right]\right\|_2, \left\|\mathbf{E}\left[X_k^* X_k\right]\right\|_2\}$ and that $\left\|X_k\right\|_2 \leq M$ a.s., for all k. Then, $\forall \tau > 0$,*

$$\mathbf{Pr}\left[\left\|\sum_{k=1}^{L} X_k\right\|_2 > \tau\right] \leq (d_1 + d_2)\exp\left(\frac{-\tau^2/2}{\sum_{k=1}^{L} \rho_k^2 + M\tau/2}\right)$$

Before the proof, here are a few notes on this result.

- If $d_1 = d_2 = 1$, then this is just the 2-sided version of the standard Bernstein Inequality.

- If X_i are diagonal, then this is just the standard Bernstein Inequality applied and then do a union bound on the diagonal of the matrix sum.
- If $\tau \leq \frac{1}{M} \sum_{k=1}^{L} \rho_k^2$, then $RHS \leq (d_1 + d_2) \exp\left(\frac{-3\tau^2/8}{\sum_{k=1}^{L} \rho_k^2} \right)$.

Proof: Let $Y_k = \begin{bmatrix} 0 & X_k \\ X_k^* & 0 \end{bmatrix}$. Then, the Y_k are symmetric random functions, and $\forall k$, we have that

$$\begin{aligned} \left\| \mathbf{E}\left[Y_k^2\right] \right\|_2 &= \left\| \mathbf{E}\left[\begin{bmatrix} X_k X_k^* & 0 \\ 0 & X_k^* X_k \end{bmatrix} \right] \right\|_2 \\ &= \max\{ \left\| \mathbf{E}\left[X_k X_k^*\right] \right\|_2 , \left\| \mathbf{E}\left[X_k^* X_k\right] \right\|_2 \} \\ &= \rho_k^2. \end{aligned}$$

In addition, $\sigma_{max}(\sum_{i=1}^{L} X_k) = \lambda_{max}(\sum_{k=1}^{L} Y_k)$. By the Operator Chernoff Theorem, it follows that

$$\begin{aligned} \mathbf{Pr}\left[\left\| \sum_{k=1}^{L} X_k \right\|_2 > Lt \right] &= \mathbf{Pr}\left[\sum_{k=1}^{L} Y_k \not\preceq LtI \right] \\ &\leq (d_1 + d_2) \exp\left(-Lt\lambda\right) \Pi_{k=1}^{L} \left\| \mathbf{E}\left[\exp\left(\lambda Y_k\right)\right] \right\|_2 , \end{aligned}$$

$\forall \lambda > 0$. Then, $\forall k$, let $Y_k = U_k \Lambda_k U_k^*$ be the eigenvalue decomposition. Then, $\forall s > 0$, we have that

$$-M^s Y_k^s = -U_k M^s \Lambda_k^2 U_k^* \preceq U_k \Lambda_k^{s+2} U_k^2 = Y_k^{s+2} \preceq U_k M^s \Lambda_k^2 U_k^* = M^s Y_k^2,$$

where M is such that $\left\| X \right\|_2 \leq M$, forall k, which implies that

$$\left\| \mathbf{E}\left[Y_k^{s+2}\right] \right\|_2 \leq M^s \left\| \mathbf{E}\left[Y_k^2\right] \right\|_2 . \tag{1}$$

For a fixed k, we have that

$$\begin{aligned} \left\| \mathbf{E}\left[e^{\lambda Y_k}\right] \right\|_2 &\leq \left\| I \right\|_2 + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} \left\| \mathbf{E}\left[Y_k^j\right] \right\|_2 \\ &\leq 1 + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} \left\| \mathbf{E}\left[Y_k^2\right] \right\|_2 M^{j-2} \\ &= 1 + \frac{\rho_k^2}{M^2} \sum_{j=2}^{\infty} \frac{\lambda^j}{j!} M^j \\ &= 1 + \frac{\rho_k^2}{M^2} \left(\exp(\lambda M) - 1 - \lambda M\right) \\ &\leq \exp\left(\frac{\rho_k^2}{M^2} \left(\exp(\lambda M) - 1 - \lambda M\right) \right) \end{aligned}$$

where the first inequality follows from the triangle inequality and since $\mathbf{E}\left[Y_k\right] = 0$; the second inequality follows from Eqn. (1); and the last inequality follows since $1 + x \leq e^x$. Thus,

$$\mathbf{Pr}\left[\left\| \sum_{k=1}^{L} X_k \right\|_2 > Lt \right] \leq (d_1 + d_2) \exp\left(-\lambda Lt + \frac{\sum_{k=1}^{L} \rho_k^2}{M^2} \left(\exp(\lambda M) - 1 - \lambda M\right) \right).$$

We can minimize this as a function of λ by choosing $\lambda = \frac{1}{M} \log\left(1 + \frac{tLM}{\sum_{k=1}^{L} \rho_k^2}\right)$, from which the result follows by tedious manipulations.

◇

4.2 Back to Frobenius norm matrix multiplication bounds

We will say that the sampling probabilities of the form

$$p_k = \frac{\left\|A^{(k)}\right\|_2 \left\|B_{(k)}\right\|_2}{\sum_{k'=1}^n \left\|A^{(k')}\right\|_2 \left\|B_{(k')}\right\|_2}$$

are the *optimal probabilities* since, as we saw before, they minimize $\mathbf{E}\left[\left\|AB-CR\right\|_F^2\right]$, which is one natural measure of the error caused by the random sampling process. In addition, we will also say that a set of sampling probabilities $\{p_i\}_{i=1}^n$ are *nearly optimal probabilities* if

$$p_k \geq \frac{\beta\left\|A^{(k)}\right\|_2 \left\|B_{(k)}\right\|_2}{\sum_{k'=1}^n \left\|A^{(k')}\right\|_2 \left\|B_{(k')}\right\|_2},$$

for some positive constant $\beta \leq 1$. Essentially, if we work with nearly optimal probabilities rather than the optimal probabilities, what this says is that we are working with probabilities that do not underestimate the optimal probability of choosing any column-row pair too much. The challenge with random sampling algorithms is ensuring that we find important samples, and so this is reasonable. In addition, as we will see below, if $\beta \neq 1$ then we suffer a small β-dependent loss in accuracy. That is, we will have to sample a little more, but if we do so then all of our bounds will work out. All of the results in which we will be interested will be robust if we work with nearly optimal probabilities, as opposed to exactly optimal probabilities, and we will gain a great deal of power and flexibility in doing so, so we will formulate the remainder of our results this semester in terms of nearly optimal probabilities (to such an extent that we will do so even when we don't make it explicit).

We now prove, for nearly optimal sampling probabilities, results analogous to those of Lemma **??**. In addition, we also prove that the corresponding results with the expectations removed hold with high probability. The proof of the latter will depend on the Hoeffding-Azuma inequality.

Theorem 3 *Suppose $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $c \in \mathbb{Z}^+$ such that $1 \leq c \leq n$, and $\{p_i\}_{i=1}^n$ are such that $\sum_{i=1}^n p_i = 1$ and such that for some positive constant $\beta \leq 1$*

$$p_k \geq \frac{\beta\left\|A^{(k)}\right\|_2 \left\|B_{(k)}\right\|_2}{\sum_{k'=1}^n \left\|A^{(k')}\right\|_2 \left\|B_{(k')}\right\|_2}. \tag{2}$$

Construct C and R with the BasicMatrixMultiplication *algorithm, and let CR be an approximation to AB. Then,*

$$\mathbf{E}\left[\left\|AB-CR\right\|_F^2\right] \leq \frac{1}{\beta c}\left\|A\right\|_F^2 \left\|B\right\|_F^2. \tag{3}$$

Furthermore, let $\delta \in (0,1)$ and $\eta = 1+\sqrt{(8/\beta)\log(1/\delta)}$. Then, with probability at least $1-\delta$,

$$\left\|AB-CR\right\|_F^2 \leq \frac{\eta^2}{\beta c}\left\|A\right\|_F^2 \left\|B\right\|_F^2. \tag{4}$$

Proof: Following reasoning similar to that of Lemma **??**, and using the nearly-optimal sampling probabilities of Eqn. (2), we see that

$$\begin{aligned}
\mathbf{E}\left[\|AB-CR\|_F^2\right] &\le \frac{1}{c}\sum_{k=1}^{n}\frac{1}{p_k}\left\|A^{(k)}\right\|_2^2\left\|B_{(k)}\right\|_2^2 \\
&\le \frac{1}{\beta c}\left(\sum_{k=1}^{n}\left\|A^{(k)}\right\|_2\left\|B_{(k)}\right\|_2\right)^2 \\
&\le \frac{1}{\beta c}\|A\|_F^2\,\|B\|_F^2\,,
\end{aligned}$$

where the last inequality follows due to the Cauchy-Schwartz inequality. Next, we consider removing the expectation. To do so, define the event $\mathcal{E}_2$ to be

$$\|AB-CR\|_F \le \frac{\eta}{\sqrt{\beta c}}\|A\|_F\,\|B\|_F \tag{5}$$

and note that to prove the remainder of the theorem it suffices to prove that $\mathbf{Pr}\left[\mathcal{E}_2\right] \ge 1-\delta$. To that end, note that C and R and thus $CR = \sum_{t=1}^{c}\frac{1}{cp_{i_t}}A^{i_t}B_{i_t}$ are formed by randomly selecting c elements from $\{1,\ldots,n\}$, independently and with replacement. Let the sequence of elements chosen be $\{i_t\}_{t=1}^{c}$. Consider the function

$$F\left(i_1,\ldots,i_c\right) = \|AB-CR\|_F\,. \tag{6}$$

We will show that changing one i_t at a time does not change F too much; this will enable us to apply a martingale inequality. To this end, consider changing one of the i_t to i'_t while keeping the other i_t's the same. Then, construct the corresponding C' and R'. Note that C' differs from C in only a single column and that R' differs from R in only a single row. Thus,

$$\begin{aligned}
\left\|CR-C'R'\right\|_F &= \left\|\frac{A^{(i_t)}B_{(i_t)}}{cp_{i_t}}-\frac{A^{(i'_t)}B_{(i'_t)}}{cp_{i'_t}}\right\|_F && (7)\\
&\le \frac{1}{cp_{i_t}}\left\|A^{(i_t)}B_{(i_t)}\right\|_F+\frac{1}{cp_{i'_t}}\left\|A^{(i'_t)}B_{(i'_t)}\right\|_F && (8)\\
&= \frac{1}{cp_{i_t}}\left\|A^{(i_t)}\right\|_2\left\|B_{(i_t)}\right\|_2+\frac{1}{cp_{i'_t}}\left\|A^{(i'_t)}\right\|_2\left\|B_{(i'_t)}\right\|_2 && (9)\\
&\le \frac{2}{c}\max_{\alpha}\frac{\left\|A^{(\alpha)}\right\|_2\left\|B_{(\alpha)}\right\|_2}{p_\alpha}. && (10)
\end{aligned}$$

(7) follows by construction and (9) follows since $\left\|xy^T\right\|_F = \|x\|_2\,\|y\|_2$ for $x\in\mathbb{R}^n$ and $y\in\mathbb{R}^n$. Thus, using the probabilities (2) and employing the Cauchy-Schwartz inequality we see that

$$\begin{aligned}
\left\|CR-C'R'\right\|_F &\le \frac{2}{\beta c}\sum_{k=1}^{n}\left\|A^{(k)}\right\|_2\left\|B_{(k)}\right\|_2 && (11)\\
&\le \frac{2}{\beta c}\|A\|_F\,\|B\|_F\,. && (12)
\end{aligned}$$

Therefore, using the triangle inequality we see that

$$\begin{aligned}
\|AB-CR\|_F &\le \left\|AB-C'R'\right\|_F+\left\|C'R'-CR\right\|_F \\
&\le \left\|AB-C'R'\right\|_F+\frac{2}{\beta c}\|A\|_F\,\|B\|_F\,. && (13)
\end{aligned}$$

By similar reasoning, we can derive

$$\left\|AB - C'R'\right\|_F \leq \left\|AB - CR\right\|_F + \frac{2}{\beta c} \left\|A\right\|_F \left\|B\right\|_F . \tag{14}$$

Define $\Delta = \frac{2}{\beta c} \left\|A\right\|_F \left\|B\right\|_F$; thus,

$$\left|F\left(i_1, \ldots, i_k, \ldots, i_c\right) - F\left(i_1, \ldots, i'_k, \ldots, i_c\right)\right| \leq \Delta. \tag{15}$$

Let $\gamma = \sqrt{2c\log(1/\delta)}\Delta$ and consider the associated Doob martingale. By the Hoeffding-Azuma inequality [**?**],

$$\mathbf{Pr}\left[\left\|AB - CR\right\|_F \geq \frac{1}{\sqrt{\beta c}} \left\|A\right\|_F \left\|B\right\|_F + \gamma\right] \leq \exp\left(-\gamma^2/2c\Delta^2\right) = \delta \tag{16}$$

and theorem follows.

◇

An immediate consequence of Theorem 3 is that by choosing enough column-row pairs, the error in the approximation of the matrix product can be made arbitrarily small. In particular, if $c \geq 1/\beta\epsilon^2$ then by using Jensen's inequality it follows that

$$\mathbf{E}\left[\,\left\|AB - CR\right\|_F\right] \leq \epsilon \left\|A\right\|_F \left\|B\right\|_F \tag{17}$$

and if, in addition, $c \geq \eta^2/\beta\epsilon^2$ then with probability at least $1 - \delta$

$$\left\|AB - CR\right\|_F \leq \epsilon \left\|A\right\|_F \left\|B\right\|_F . \tag{18}$$

In certain applications, we will be interested in an application of Theorem 3 to the case that $B = A^T$, i.e., one is interested in approximating $\left\|AA^T - CC^T\right\|_F^2$. In this case, sampling column-row pairs corresponds to sampling columns of A, and nearly optimal probabilities will be those such that $p_k \geq \frac{\beta\left\|A^{(k)}\right\|_2}{\left\|A\right\|_F}$ for some positive $\beta \leq 1$. By taking $B = A^T$ and applying Jensen's inequality, we have the following theorem as a corollary of Theorem 3.

Theorem 4 *Suppose $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{Z}^+$, $1 \leq c \leq n$, and $\{p_i\}_{i=1}^n$ are such that $\sum_{i=1}^n p_i = 1$ and such that $p_k \geq \frac{\beta\left\|A^{(k)}\right\|_2^2}{\left\|A\right\|_F^2}$ for some positive constant $\beta \leq 1$. Furthermore, let $\delta \in (0,1)$ and $\eta = 1 + \sqrt{(8/\beta)\log(1/\delta)}$. Construct C (and $R = C^T$) with the* BasicMatrixMultiplication *algorithm, and let CC^T be an approximation to AA^T. Then,*

$$\mathbf{E}\left[\,\left\|AA^T - CC^T\right\|_F\right] \leq \frac{1}{\sqrt{\beta c}} \left\|A\right\|_F^2 \tag{19}$$

and with probability at least $1 - \delta$,

$$\left\|AA^T - CC^T\right\|_F \leq \frac{\eta}{\sqrt{\beta c}} \left\|A\right\|_F^2 . \tag{20}$$