Stat260/CS294: Randomized Algorithms for Matrices and Data

Lecture 3 - 09/11/2013

Lecture 3: Scalar and Matrix Concentration

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.

3 Scalar and Matrix Concentration

Here, we will give an aside on probabilities, and in particular on various ways to establish what is know as concentration. Given some information about a distribution, e.g., its mean or its variance or information about higher moments, there are various ways to establish bounds on the tails of sums of random variables from that distribution. That is, there are various ways to establish that estimates are close to their expected value, with high probability. Today and next time, we will cover several of these methods, both in the case where the random variables are scalar or real-valued and when the random variables are matrix-valued. The former can be used to bound that latter, e.g., by bounding every element of the random matrix individually, but the latter often provide tighter bounds in those cases.

Perhaps the simplest method to establish concentration goes by the name Markov's inequality. Although tight, given only information about the mean of a nonnegative random variable, it generally gives bounds that are too weak for what we will want. Thus, we will consider more sophisticated method that go by the name Chernoff (and related) bounds, that provide much stronger bounds, and that amount to applying Markov's inequality on the moment generating function of the random variable of interest. Since the Frobenius norm of a matrix is the sum of all the elements of the matrix, this method—or actually an extension of this basic idea that applies to non-independent random variables and that goes by the name Hoeffding-Azuma—applied to real-valued random variables will use to bound the Frobenius norm of the error in our approximate matrix multiplication result. To get bounds for the spectral norm (tighter than provided by bounding the spectral norm by the Frobenius norm), we will have to do something a little more sophisticated. Basically, we will have to consider the analogues of these results for matrix-valued random variables. This will provide tighter bounds, but at the expense of heavier machinery.

There is no particular reading for today, but if the material was too foreign, then take a look at the following reference.

- The first few chapters of "Probability and Computing," by Mitzenmacher and Upfal.

3.1 Scalar concentration: Markov's inequality

We will start with the following well-known result, known as *Markov's inequality*, which provides a bound on the probability that a nonnegative random variable deviates (on the high side) from it's expectation. It's proof is standard, but we include it for comparison with the operator-valued Markov inequality below.

Lemma 1 (Markov's Inequality) *Let X be a real-valued random variable such that $X \geq 0$. Then, $\forall a \geq 0$, $\mathbf{Pr}\left[X \geq a\right] \leq \frac{\mathbf{E}[X]}{a}$.*

Proof: For $a > 0$, let

$$\mathcal{X} = \begin{cases} 1 & \text{if } X \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

be the indicator variable for the event that $X \geq 0$; and note that, since $X \geq 0$, it follows that $\mathcal{X} \leq \frac{X}{a}$. Since $\mathcal{X}$ is a 0-1-valued random variable, it follows that $\mathbf{E}\left[\mathcal{X}\right] = \mathbf{Pr}\left[\mathcal{X} = 1\right] = \mathbf{Pr}\left[X \geq a\right]$, from which it follows that $\mathbf{Pr}\left[X \geq a\right] = \mathbf{E}\left[\mathcal{X}\right] \leq \mathbf{E}\left[\frac{X}{a}\right] = \frac{\mathbf{E}[X]}{a}$.

◇

3.2 Applying this to approximate matrix multiplication

As an aside, consider how this result can be applied back to our bound from last class on approximating the product of two matrices. We can use it to provide a result that holds with fixed probability, given out result in expectation. To see this, recall that last time we showed that

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

We could remove the expectation from this quantity, but for a cleaner comparison with results we will derive below, let's use Jensen's Inequality (if $\phi(\cdot)$ is a convex function, then $\phi(\mathbf{E}\left[X\right]) \leq \mathbf{E}\left[\phi(X)\right]$) to get

$$\mathbf{E}\left[\,\|AB - CR\|_F\right] \leq \frac{1}{\sqrt{\beta c}}\,\|A\|_F\,\|B\|_F \tag{1}$$

From this it follows that if the number of samples $c \geq \beta/\epsilon^2$, then $\mathbf{E}\left[\,\|AB - CR\|_F\right] \leq \epsilon\,\|A\|_F\,\|B\|_F$. To "remove the expectation" from Eqn. (1) with Markov's inequality, let's set the failure probability to be δ as follows:

$$\delta = \mathbf{Pr}\left[\,\|AB - CR\|_F > \frac{\alpha}{\sqrt{\beta c}}\,\|A\|_F\,\|B\|_F\right],$$

where α is a parameter that will determine the how c needs to be chosen to get a fixed error probability. Thus,

$$\delta \leq \frac{\mathbf{E}\left[\,\|AB - CR\|_F\right]}{\frac{\alpha}{\sqrt{\beta c}}\,\|A\|_F\,\|B\|_F} \leq \frac{1}{\alpha},$$

where the first inequality follows by Markov's inequality, and where the second inequality follows from our result from Eqn. (1). Thus, $\alpha = \frac{1}{\delta}$; and so it follows that with probability $\geq 1 - \delta$

$$\begin{aligned} \|AB - CR\|_F &\leq \frac{1}{\sqrt{\delta^2 \beta c}}\,\|A\|_F\,\|B\|_F \\ &\leq \epsilon\,\|A\|_F\,\|B\|_F\,, \end{aligned}$$

where the second inequality holds if $c \geq \frac{\beta}{\delta^2 \epsilon^2}$.

That dependency on δ might not be a problem if one wants to choose $\delta = 0.1$. (Actually it might still be, since it implies that one needs to choose c to be a factor of 100 larger, but in cases where we just need some event to hold with constant probability, it is typically fine.) But, it certainly is a problem if one wants, say, $\delta = 10^{-6}$. In theses cases, we would like the dependence of c on δ to be $O(\log(1/\delta))$, rather than $\text{poly}(1/\delta)$. For this we need heavier machinery.

(Note that a popular and standard way to boost the success probability is to oversample by a factor of $O(\log(1/\delta))$, if concentration is supported, or to repeat the trials $O(\log(1/\delta))$ times and keep the best result, if there is an easy way to check which of the trials is the best. These methods won't work if concentration isn't supported and/or if there is not an easy way to check which of the trials is the best.)

3.3 More scalar concentration

First, recall the definition of the variance of a random variable.

Definition 1 $\mathbf{Var}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2$ *and StdDev*$(X) = \sigma(X) = \sqrt{\mathbf{Var}[X]}$.

Let's say we have information on the variance of X, e.g., we might have a bound on $\mathbf{Var}[X]$. In this case, we can get stronger result using Chebychev's inequality.

Lemma 2 (Chebychev's Inequality) $\forall A > 0$, $\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{Var}[X]}{a^2}$

Proof: First, observe that

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] = \mathbf{Pr}\left[(X - \mathbf{E}[X])^2 \geq a^2\right].$$

Then, since $(X - \mathbf{E}[X])^2$ is a nonnegative random variable, we can apply Markov's inequality to get that

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{a^2} \leq \frac{\mathbf{Var}[X]}{a^2}.$$

◇

Of course, there are other variants of Chebychev's Inequality that are parameterized slightly differently. For example, that $\forall t > 0$,

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq t\sigma(X)] \leq \frac{1}{t^2},$$

or that $\forall t > 0$,

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq t\mathbf{E}[X]] \leq \frac{\mathbf{Var}[X]}{t^2(\mathbf{E}[X])^2}.$$

Unfortunately, however, this is usually still not good enough for what we want. But, if we have bounds on higher moments of the random variable, then we can use "moment generating function methods" to get much stronger results that are qualitatively more like what one would get for the tail behavior of Gaussian random variables. In particular, Chernoff-style bounds are very powerful, providing exponentially-decreasing bounds on the tails of the distribution. Chernoff-style bounds do this basically by applying Markov's inequality on the moment generating function of a random variable.

Definition 2 *The* moment generating function *of a random variable X is $M_X(t) = \mathbf{E}\left[e^{tX}\right]$.*

We will mostly be interested in the existence and properties of this function in the neighborhood of $t = 0$. The basic idea of Chernoff-style bounds is to apply Markov's inequality to e^{tX} for a well-chosen value of t.

Lemma 3 (Vanilla Chernoff) *$\forall t > 0$: $\mathbf{Pr}\left[X \geq a\right] = \mathbf{Pr}\left[e^{tX} \geq e^{ta}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$. In particular, $\mathbf{Pr}\left[X \geq a\right] \leq \min_{t>0} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$.*

The same holds for the other direction.

Lemma 4 (Vanilla Chernoff, other direction) *$\forall t < 0$: $\mathbf{Pr}\left[X \leq a\right] = \mathbf{Pr}\left[e^{tX} \leq e^{ta}\right] \leq \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$. In particular, $\mathbf{Pr}\left[X \leq a\right] \leq \min_{t<0} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$.*

There are a lot of variants of this basic result, depending on what is known about the given distribution, how tight a bound one can provide for $\mathbf{E}\left[e^{tX}\right]$, etc. Here are two versions.

Theorem 1 (Hoeffding) *Let $\{X_i\}_{i=1}^n$ be r.v. such that $X_i \in [a_i, b_i]$, for all i, and let $X = \sum_{i=1}^n X_i$. Then*

$$\mathbf{Pr}\left[|X - \mathbf{E}\left[x\right]| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Theorem 2 (Bernstein) *Let $\{X_i\}_{i=1}^n$ be r.v. such that $\mathbf{E}\left[X\right]$, $\mathbf{E}\left[X^2\right] = \sigma^2$, $|X| \leq M$, X_i are i.i.d. copies. Then, for all $t > 0$,*

$$\mathbf{Pr}\left[|X - \mathbf{E}\left[x\right]| \geq t\right] \leq \exp\left(\frac{-t^2}{2n\sigma^2 + \frac{4}{3}tM}\right)$$

So, Chernoff-type results provide bounds for large deviations using higher moments for sums of independent random variables.

When applied to approximate matrix multiplication, we will be interested in providing both spectral and Frobenius norm bounds for the error. In order to provide spectral norm bounds for randomized matrix multiplication, we will need to use matrix versions of these Chernoff-style results; but to prove the Frobenius norm bounds we will need an extension that deals with random variables that are not quite independent. These latter extensions are known as Hoeffding-Azuma bounds, and they provide Chernoff-like bounds, with no independence assumptions, but assuming some sort of bounded difference form holds. We will do this in a standard way, by construction, and construct a martingale difference sequence with differences that are bounded in absolute value. To describe how we will do this, we start with the definition of a martingale.

Definition 3 *A sequence of random variables $Z_0, Z_1, \ldots$ is a* martingale with respect to a sequence $X_0, X_1, \ldots$ *if,* $\forall n \geq 0$,

- $Z_n = Z_n(X_0, X_1, \ldots, X_n)$*, i.e., it is a function of the* X_i
- $\mathbf{E}\left[|Z_n|\right] < \infty$
- $\mathbf{E}\left[Z_{n+1}|X_0 \cdots X_n\right] = Z_n$

A sequence is a martingale *if it is a martingale with respect to itself.*

The canonical example of a martingale is a gambler who plays a sequence of fair games. In this case, let X_i be the amount that the gambler wins in the i^{th} game; and let Z_i be the total winnings of the gambler after the i^{th} game. Since the game is fair, $\mathbf{E}\left[X_i\right] = 0$; and also

$$\mathbf{E}\left[Z_{i+1}|X_1 \cdots X_i\right] = Z_i + \mathbf{E}\left[X_{i+1}\right] = Z_i.$$

So, in this case, Z_i is a martingale with respect to X_i.

Martingales are powerful and ubiquitous in applications of probability, and in particular in the area of randomized algorithms, largely since they can be formed by nearly "any" random variable. We will use the usual approach, that in particular is common in the theory of algorithms. It involves constructing something called the Doob martingale.

Definition 4 *Let $X_0, X_1, \ldots, X_n$ be a sequence of random variables, and let Y be a random variable with $\mathbf{E}\left[|Y|\right] < \infty$. (Generally, Y is such that $Y = Y(X_1, \ldots, X_n)$. Then, consider*

$$Z_i = \mathbf{E}\left[Y|X_0 \cdots X_i\right] \quad i = 0, 1, \ldots, n.$$

This is a Doob Martingale.

The first thing to note is that this is a martingale.

Lemma 5 *Z_i constructed in this way is a martingale w.r.t. X_i.*

Proof:

$$\begin{aligned} \mathbf{E}\left[Z_{i+1}|X_0 \cdots X_i\right] &= \mathbf{E}\left[\mathbf{E}\left[Y|X_0 \cdots X_{i+1}\right]|X_0 \cdots X_i\right] \\ &= \mathbf{E}\left[Y|X_0 \cdots X_i\right] \\ &= Z_i, \end{aligned}$$

where the second of those inequalities used that $\mathbf{E}\left[Y|X_0 \cdots X_{i+1}\right]$ is a random variable, and that in general $\mathbf{E}\left[v|w\right] = \mathbf{E}\left[\mathbf{E}\left[v|u,w\right]|w\right]$.

$\diamond$

In most applications, one starts the Doob martingale with $Z_0 = \mathbf{E}\left[Y\right]$, which corresponds to X_0 being a trivial random variable, independent of Y; and then assume that we want to predict the value of Y and that $Y = Y(X_1, \ldots, X_n)$, then the sequence $Z_0, \ldots, Z_n$ is a sequence of more and more refined estimates of the value of Y, with $Z_0 = \mathbf{E}[Y]$, going all the way to $Z_n = Y$, in which case we fully know the value of the random variable.

There are a lot of applications of this idea. One big application of Doob martingales in the theory of algorithms is for the analysis of algorithms via Chernoff-like tail inequalities, that can apply, even when the underlying random variable is not independent. The basic form goes by the name Azuma-Heoffding.

Theorem 3 (Azuma-Hoeffding) *Let $X_0, \ldots, X_n$ be a martingale such that $|X_k - X_{k-1}| \le c_k$. Then, $\forall t \ge 0, \lambda > 0$,*

$$\mathbf{Pr}\left[|X_t - X_0| \ge \lambda\right] \le 2\exp\left(\frac{-\lambda^2}{\sum_{k=1}^{t} c_k^2}\right).$$

As a corollary, if $|X_k - X_{k-1}| \le c$, then

$$\mathbf{Pr}\left[|X_t - X_0| \ge \lambda c\sqrt{t}\right] \le 2\exp\left(-\lambda^2/2\right).$$

3.4 An aside on SPSD matrices

Recall that, although Azuma-Hoeffding is what we will use to get concentration for the Frobenius norm error of approximate matrix multiplication, we will use matrix analogues of Chernoff-style bounds to get concentration for the spectral norm error of approximate matrix multiplication. Although these bounds will apply to general matrices—by a relatively straightforward trick—they are most easily formulated in terms of symmetric positive semi-definite (SPSD) matrices, and so we will review the properties of that class of matrices, starting with the definition.

Definition 5 *A matrix $A \in \mathbb{R}^{n \times n}$ is SPSD if*

- $x^T A x \ge 0 \quad \forall x \in \mathbb{R}^n$
- *$A = U \Sigma U^T$ is the eigendecomposition, with all $\sigma_i \ge 0$.*

And it is SPD (symmetric positive definite) if those inequalities are tight, i.e., they are strict with no equalities.

There are several things to note about this definition.

- This is a generalization of $\mathbb{R}$ to matrices; and some, but certainly not all, properties of real numbers extend to SPSD matrices.
- One can define a partial order over the "cone" of SPSD matrices: $A \succ B$ iff $A - B \succ 0$ iff $A - B \in SPD$ and $A \succeq B$ iff $A - B \succeq 0$ iff $A - B \in SPSD$. This is not a total order unless we are dealing with 1×1 matrices, i.e., real numbers.
- $A \succeq 0$ iff all the eigenvalues are nonnegative. This is a set of d nonlinear inequalities; but it can be viewed as ∞-ly many linear inequalities, since $A \succeq 0$ iff $\forall \rho$ that are PSD matrices of trace 1, also known as "density operators," $\mathbf{Tr}(\rho A) \ge 0$) iff $\forall \pi$ that are one-dimensional projectors, $\mathbf{Tr}(\pi A) \ge 0$.

Since SPSD matrices are a generalization of $\mathbb{R}$, one can generalize many real functions to them. In particular, given a function $f : \mathbb{R} \to \mathbb{R}$, one can:

- Define a map on diagonal matrices by applying f to each diagonal entry
- Extend f to self-adjoint/Hermitian/symmetric matrices via the eigenvalue decomposition $f(A) = Qf(\Lambda)Q^T$, where $A = Q\Lambda Q^T$.
- Then, the spectral mapping theorem says: each eigenvalue of $f(A)$ is equal to $f(\lambda)$, for some eigenvalue λ of A.

The point is that symmetric and SPSD matrices are *much* more structured objects than general matrices, and you get much nicer and cleaner results. We will see the same things for this in general versus symmetric/SPSD matrix perturbation results, where in the latter case we get much better results. This is familiar to NLA people, so others don't think general matrices are so nice. Also, a lot of data matrices are symmetric or SPSD, or we are interested in robustness, and so we consider singular vectors/values, rather than eigen vectors/values via $A \to A^TA \quad \text{and} \quad AA^T$.

We can define the exponential of s.a. matrix A by the spectral mapping theorem, with $f(\lambda) = e^\lambda$, or we can define it as

$$\exp(A) = I + \sum_{i=1}^{\infty} \frac{A^i}{i!} = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

By the spectral mapping theorem, we know that this matrix is PD. Note that $I + A \preceq e^A$. Also, note that this expansion of symmetric matrices is more generally-used in machine with kernel methods.

Here is a fact.

Fact 1 $\mathbf{Tr}\,(\exp(A))$ *is a convex function; and it is monotone with respect to the semidefinite order, i.e.,* $A \preceq B \quad \to \quad \mathbf{Tr}\,(\exp(A)) \leq \mathbf{Tr}\,(\exp(B))$*. Note that the first inequality is over SPSD matrices, while the second inequality is over numbers.*

We can define the logarithm as the functional inverse of the matrix exponentials: $\log(\exp(A)) \equiv A \quad \forall \quad \text{s.a./Hermitian/symmetric} \quad A$.

Definition 6 *A function f is* operator monotone *with respect to the semidefinite order if* $0 \preceq A \preceq B$ *implies that* $f(A) \preceq f(B)$*. (Note that both inequalities are inequalities over SPSD matrices.) A function f is* operator concave *with respect to the semidefinite order if* $cf(A) + (1-c)f(B) \preceq f(cA + (1-c)B)$*, for all PD A, B and for all $c \in [0, 1]$.*

These generalize properties for the analogous things for real numbers, but note that operator monotone and operator convex functions are not so common. But here are some examples we will encounter.

- $A \to \log(A)$ and also $A \to A^s$, for $s \in [0, 1]$, are operator monotone and operator con*cave.*
- $A \to \exp(A)$ and also $A \to A^s$, for $s > 2$, are neither operator monotone nor operator con*vex.*
- As an aside, $A \to A^s$, for $s \in [0, 1]$, is operator convex, but is not operator monotone.

In particular, $A \to A^{1/2} = \sqrt{A}$ *is* operator monotone, which is a fact we will use below.

We should note that, although we are working with symmetric matrices, many of the results extend easily to general matrices. In particular, given a rectangular matrix $A \in \mathbb{R}^{m \times n}$, we can define a matrix $B \in \mathbb{R}^{(m+n)\times(m+n)}$ as

$$B = \left[\begin{array}{cc} 0 & A \\ A^T & 0 \end{array} \right] .$$

Then,

$$B^2 = \left[\begin{array}{cc} AA^T & 0 \\ 0 & A^T A \end{array} \right] .$$

And the eigen vectors/values of B are related to the eigen vectors/values or A. For the eigenvectors, see the first homework. For the eigenvalues, we will need that $\lambda_{max}(B) = \|B\|_2 = \|A\|_2$.

Two other things to note.

- Expectation is a convex combination and the PSD cone is convex, so $X \preceq Y$ a.s. $\to \mathbf{E}\left[X\right] \preceq \mathbf{E}\left[Y\right]$.
- Every operator convex function admits a Jensen's inequality, and so since the matrix square is operator convex, we have that $\left(\mathbf{E}\left[X\right]\right)^2 \preceq \mathbf{E}\left[X^2\right]$.

Finally, for numbers $a, b \in \mathbb{R}$, we have that $e^{a+b} = e^a e^b$. The matrix exponential does *not* convert sums into products in an analogous way (unless the matrices commute). But there is something weaker that will still be good enough for some purposes.

Lemma 6 (Golden-Thompson Inequality) *For A, B that are SPSD matrices, we have that* $\mathbf{Tr}\left(\exp(A+B)\right) \leq \mathbf{Tr}\left(\exp(A)\exp(B)\right)$.