6 Matrix Concentration Bounds

Concentration bounds are inequalities that bound probabilities of deviations by a random variable from some value, often its mean. Informally, they show the probability that a random variable deviates from its expectation is small. A basic example of random variables being concentrated around the mean is stated by "law of large numbers" that says under mild conditions, sum of *independent random variables* is close to their expectation with a high probability. Concentration bounds, which are also known as "tail bounds", are a major tool in analyzing average behaviour of algorithms, e.g. estimating the failure probability or establishing high probability bounds on running time and space usage.

In this lecture, we give an introduction to some of concentration bounds on random variables and random matrices.

6.1 Basic Tail Bounds

If X is a random variable and $a \in \mathbb{R}$ is a real value, then $\textsf{\textbf{Pr}}[X \geq a]$ and $\textsf{\textbf{Pr}}[X \leq a]$ are called *upper tail* and *lower tail* of the distribution on X, respectively. Concentration bounds might bound one of the tails or both. We start with the most basic yet fundamental tail bound, called as Markov's Inequality.

Theorem 6.1.1 (Markov's Inequality). *Let X be a non-negative random variable. Then for all $a > 0$*

$$\textbf{\textit{Pr}}(X \geq a) \leq \frac{\textbf{\textit{E}}[X]}{a}$$

Proof. Define an indicator random variable $I_a = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$. Note in both cases $X \geq aI_a$, therefore

$$\begin{aligned} \textbf{E}[X] &\geq a\ \textbf{E}[I_a] \\ &= a\ \textbf{Pr}(X \geq a) \end{aligned}$$

And that completes the proof. □

Markov's inequality is the best possible tail bound one can get if all he knows is non-negativity of random variable and its expectation, yet it is often too weak to yeild useful results. Tail bounds can be improved if more information about X, e.g. its variance, is available. Below, we state a significantly stronger bound that employs extra statistics.

Theorem 6.1.2 (Chebyshev's Inequality). *For any random variable X, and any $a > 0$*

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

Proof. Let's apply Markov's inequality on random variable $X = (Y - \textbf{E}[Y])^2$ and scalar $a = b^2$,

$$\textbf{Pr}\left((Y - \textbf{E}[Y])^2 \geq b^2\right) \leq \frac{\textbf{E}\left[(Y - \textbf{E}[Y])^2\right]}{b^2}$$

Note that $\textbf{Pr}\left((Y - \textbf{E}[Y])^2 \geq b^2\right) = \textbf{Pr}\left(|Y - \textbf{E}[Y]| \geq b\right)$, and $\textbf{E}\left[(Y - \textbf{E}[Y])^2\right] = \textbf{Var}(Y)$ is the definition of variance. Therefore

$$\textbf{Pr}\left(|Y - \textbf{E}[Y]| \geq b\right) \leq \frac{\textbf{Var}(Y)}{b^2}$$

□

The use of Chebyshev's inequality indicates that with probability larger than $1 - 1/a^2$, a random variable X will fall within a times the standard deviation around $\mathbf{E}[X]$. For example, 75% of the times a random value falls in the interval $[\mathbf{E}[X] - 2\mathbf{Var}(X), \mathbf{E}[X] + 2\mathbf{Var}(X)]$.

Both Markov's and Chebyshev's inequalities provide polynomially decaying bounds in amount of deviation (i.e. a in the formula). More interesting are concentration bounds in which deviation probabilities decay *exponentially* in the distance from mean.

Theorem 6.1.3 (Chernoff-Hoeffding Inequality). *Consider a set of independent random variables $\{X_1, X_2, \cdots, X_r\}$. If we know each random variable is bounded as $a_i \leq X_i \leq b_i$ with $\Delta_i = b_i - a_i$, then for $M = \sum_{i=1}^r X_i$ and any parameter $\varepsilon \in (0, 1/2)$*

$$\mathbf{Pr}(|M - \mathbf{E}[M]| > \varepsilon) \leq 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^r \Delta_i^2}\right)$$

Although Chernoff-Hoeffding inequality provides a stronger exponentially decreasing bound, it requires random variables to be independent a condition that neither the Markov's nor the Chebyshev's inequalities require. Before we move on to concentration bounds for matrices, we state a concrete version of Chernoff bound that has an analogous in matrices. This version bounds the tail distribution of a sum of independent $0, 1$ random variables that are not necessarily distributed identically.

Theorem 6.1.4 (Chernoff's Inequality). *Let $\{X_1, X_2, \cdots, X_r\}$ be a sequence of independent $0, 1$ random variables. Define $M = \sum_{i=1}^r X_i$ and $\mu = \mathbf{E}[M]$. Then for any $\varepsilon > 0$:*

$$\mathbf{Pr}(M \geq (1+\varepsilon)\mu) \leq \left(\frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^\mu$$

6.2 Matrix Concentration Bounds

In applications, it is common that a random matrix can be expressed as a sum of independent random matrices[1]. For example, the covariance of $X \in \mathbb{R}^{n \times d}$ can be written as $X^T X = \sum_{i=1}^n x_i^T x_i$ where x_i denotes i-th row of X. In this section, we state two common bounds on random matrices[1].

6.2.1 Matrix Chernoff Bound

Chernoff's Inequality has an analogous in matrix setting; the $0, 1$ random variables translate to positive-semidefinite random matrices which are uniformly bounded on their eigenvalues.

Theorem 6.2.1 (Matrix Chernoff bound). *Let $\{X_1, X_2, \cdots, X_r\}$ be a finite sequence of independent $d \times d$ dimensional random matrices such that each X_i is positive semi-definite ($X_i \succeq 0$) and is upper bounded on eigenvalues, i.e. $\lambda_{max}(X_i) \leq R$ for a positive constant R. Define $\mu_{min} = \lambda_{min}\left(\sum_{i=1}^r \mathbf{E}[X_i]\right)$ and $\mu_{max} = \lambda_{max}\left(\sum_{i=1}^r \mathbf{E}[X_i]\right)$, then for any $\varepsilon \in [0, 1]$*

$$\mathbf{Pr}\left(\lambda_{min}\left(\sum_{i=1}^r X_i\right) \leq (1-\varepsilon)\mu_{min}\right) \leq d. \left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{min}/R}$$

$$\mathbf{E}\left[\lambda_{min}\left(\sum_{i=1}^r X_i\right)\right] \geq 0.6\ \mu_{min} - R \log d$$

And for any $\varepsilon \geq 0$

$$\mathbf{Pr}\left(\lambda_{max}\left(\sum_{i=1}^r X_i\right) \leq (1+\varepsilon)\mu_{max}\right) \leq d. \left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{max}/R}$$

$$\boldsymbol{E}\left[\lambda_{max}\left(\sum_{i=1}^{r} X_i\right)\right] \leq 1.8\;\mu_{max} + R\log d$$

Here, we state an example to demonstrate the application of Matrix Chernoff bound.

Example [Random Submatrix]. [1] Consider a fix matrix $C = [c_1, c_2, \cdots, c_n] \in \mathbb{R}^{d\times n}$. Let's construct a submatrix $Z \in \mathbb{R}^{d\times n} \subseteq C$ by randomly picking columns of C with probability s/n. If column c_i is selected, we set i-th column of Z to c_i (i.e. $z_i = c_i$), otherwise $z_i = 0$. We are interested in finding the expected largest singular value of Z and expected smallest singular values of Z.

Solution: Let's define a random variable δ_i for each column c_i. If c_i is selected, we set $\delta_i = 1$ and otherwise we set it to zero. Therefore δ_i is a $0, 1$ random variable as following

$$\delta_i = \begin{cases} 1 & \text{with probability } s/n \\ 0 & \text{otherwise} \end{cases}$$

Putting them in a diagonal matrix $\Delta = diag(\delta_1, \delta_2, \cdots, \delta_n) \in \mathbb{R}^{n\times n}$ allows us to write Z as $Z = C\Delta$. In order to be able to use the Matrix Chernoff bound, we need to convert singular values to eigenvalues. Note that singular values of Z are related to eigenvalues of ZZ^T as $\sigma_i(Z)^2 = \lambda_i(ZZ^T)$; therefore if we consider singular values and eigen values indexed in descending order, we can bound largest singular value of Z as

$$\sigma_1(Z)^2 = \lambda_1(ZZ^T)$$

and smallest singular value as

$$\sigma_d(Z)^2 = \lambda_d(ZZ^T)$$

Next, we write ZZ^T in terms of sum of random matrices that satisfy conditions of theorem 6.2.1:

$$ZZ^T = (C\Delta)(C\Delta)^T = C\Delta^2C^T = C\Delta C^T = \sum_{i=1}^{n} \delta_i c_i c_i^T$$

If we denote each summand as X_i,

$$ZZ^T = \sum_{i=1}^{n} \delta_i c_i c_i^T = \sum_{i=1}^{n} X_i \qquad \mathbf{E}[ZZ^T] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \frac{s}{n}\sum_{i=1}^{n} c_i c_i^T = \frac{s}{n} CC^T$$

Note that X_is are independent, positive semi-definite, and bounded as $\lambda_{max}(X_i) \leq \|c_i\|^2$. So the maximum eigenvalue of all summands is bounded as $R = \max_{1\leq i\leq n} \|c_i\|^2$. The mean parameters satisfy

$$\mu_{max} = \lambda_n(\mathbf{E}[ZZ^T]) = \frac{s}{n}\sigma_1(C)^2 \qquad \mu_{min} = \lambda_d(\mathbf{E}[ZZ^T]) = \frac{s}{n}\sigma_d(C)^2$$

Applying Matrix Chernoff inequality we obtain

$$\mathbf{E}\left[\sigma_1(Z)^2\right] = \mathbf{E}\left[\lambda_d(ZZ^T)\right] \leq 1.8\;(\frac{s}{n})\;\sigma_1(C)^2 + \max_{1\leq i\leq n} \|c_i\|^2 \log d$$

and

$$\mathbf{E}\left[\sigma_d(Z)^2\right] = \mathbf{E}\left[\lambda_d(ZZ^T)\right] \geq 0.6\;(\frac{s}{n})\;\sigma_d(C)^2 - \max_{1\leq i\leq n} \|c_i\|^2 \log d$$

As this bound shows random matrix Z gets a share of the spectrum of C in proportion to the number of columns it picks. In other words, matrix C has n columns which random matrix Z includes about s of them in expectation, and the bound is showing each singular value $\sigma_i(Z)^2$ of Z inherits an s/n share of $\sigma_i(C)^2$ for all $1 \leq i \leq d$.

6.2.2 Matrix Bernstein Inequality

Bernstein Inequality drops the condition of positive semi-definitness and concerns a sum of zero-mean bounded random matrices.

Theorem 6.2.2 (**Matrix Bernstein Inequality**). *Let $\{X_1, X_2, \cdots, X_r\}$ be a set of independent $d_1 \times d_2$ dimensional random matrices with $\boldsymbol{E}[X_i] = 0$ and $\|X_i\|_2 \leq R$ for a positive constant R and for all X_is. Define variance parameter as $\gamma = \max\left\{\left\|\sum_{i=1}^r \boldsymbol{E}(X_i X_i^T)\right\|, \left\|\sum_{i=1}^r \boldsymbol{E}(X_i^T X_i)\right\|\right\}$, then for any $t \geq 0$*

$$\boldsymbol{Pr}\left[\left\|\sum_{i=1}^r X_i\right\| \geq t\right] \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{\gamma + Rt/3}\right)$$

$$\boldsymbol{E}\left[\left\|\sum_{i=1}^r X_i\right\|\right] \leq \sqrt{2\gamma\log(d_1 + d_2)} + \frac{R}{3}\log(d_1 + d_2)$$

Example [Randomize Matrix Multiplication]. [1] Consider matrix $B \in \mathbb{R}^{d_1 \times n}$ with unit norm columns (i.e. $\|b_i\| = 1, \forall 1 \leq i \leq n$) and matrix $C \in \mathbb{R}^{n \times d_2}$ with unit norm rows (i.e. $\|c^i\| = 1, \forall 1 \leq i \leq n$). We are interested in computing an approximation to the product $BC \in \mathbb{R}^{d_1 \times d_2}$ using a random sample of columns of B and rows of C.

solution: First note that if we construct a random variable $S = (n\ b_j\ c^j) \in \mathbb{R}^{d_1 \times d_2}$ where $j \in [1, n]$ is sampled uniformly at random, then S will be an unbiased estimator for the product BC. This is true because:

$$\mathbf{E}[S] = \mathbf{E}\left[n\ b_j\ c^j\right] = n\sum_{k=1}^n (b_k\ c^k\ \mathbf{Pr}(k = j)) = n\ \sum_{k=1}^n (b_k\ c^k\ (\frac{1}{n})) = \sum_{k=1}^n (b_k\ c^k) = BC$$

Although S is an unbiased estimator for BC, it has a large variance. In order to reduce the variance we compute m independent copy of S and average them as $Z = \frac{1}{m}\sum_{i=1}^m S_i$.Now Z is a better approximation for BC. The error we are interested to compute is

$$err = \mathbf{E}\left[\|Z - BC\|_2\right] = \mathbf{E}\left[\left\|\sum_{i=1}^m (\frac{1}{m}S_i - \frac{1}{m}BC)\right\|_2\right]$$

Therefore each summand is $X_i = \frac{1}{m}(S_i - BC)$. Note that summands are independent and $\mathbf{E}[X_i] = 0$. Before we bound spectral norm of summands, we use *symmetrization* technique to say

$$\mathbf{E}\left[\|Z - BC\|\right] \leq \frac{2}{m}\mathbf{E}\left[\left\|\sum_{i=1}^m \varepsilon_i S_i\right\|\right]$$

Where ε_i are independenet Rademacher random variables, independent of S_is. Therefore

$$\|X_i\|_2 \leq \|\varepsilon_i S_i\| \leq \|S_i\| = n\|b_j\|\|c^i\| = n$$

Therefore the upper bound R on spectral norm of any summand is n. Now we compute the variance in two steps:

$$\mathbf{E}[S_i S_i^T] = \sum_{j=1}^n n^2 (b_j c^j)(b_j c^j)^T \mathbf{Pr}(j = i) = \sum_{j=1}^n n^2 \|c^j\|^2 b_j b_j^T = n^2 BB^T$$

And

$$\mathbf{E}[S_i^T S_i] = \sum_{j=1}^n n^2 (b_j c^j)^T (b_j c^j)\mathbf{Pr}(j = i) = \sum_{j=1}^n n^2 \|b_j\|^2 {c^j}^T c^j = n^2 C^T C$$

Therefore $\gamma = \max\{\|mn\ BB^T\|, \|mn\ C^TC\|\} = mn \max\{\|B\|^2, \|C\|^2\}$. Applying the Bernstein Bound we obtain

$$\begin{aligned}
\mathbf{E}\left[\|Z - BC\|\right] &\leq \frac{2}{m}\mathbf{E}\left[\left\|\sum_{i=1}^{m} \varepsilon_i S_i\right\|\right] \\
&\leq \frac{2}{m}\left(\sqrt{2mn\log(d_1+d_2)}\max\{\|B\|_2, \|C\|_2\} + \frac{n}{3}\log(d_1+d_2)\right) \\
&= \sqrt{\frac{8n}{m}\log(d_1+d_2)}\max\{\|B\|_2, \|C\|_2\} + \frac{2n}{3m}\log(d_1+d_2)
\end{aligned}$$

In order to make sense out of this inequality, let's take a look at numeric rank of B and C. As we know, numeric rank of a matrix is defined as $nrank(A) = \|A\|_F^2/\|A\|_2^2$. Since B and C has unit columns and unit rows, respectively then $\|B\|_F^2 = \|C\|_F^2 = n$. Then the geometric mean of numeric ranks will be

$$\mu = \sqrt{nrank(B)\ nrank(C)} = \sqrt{\frac{n^2}{\|B\|_2^2\|C\|_2^2}} = \frac{n}{\|B\|_2\|C\|_2}$$

Converting the error bound to a relative scale we obtain

$$\begin{aligned}
\frac{\mathbf{E}\left[\|Z - BC\|\right]}{\|B\|_2\|C\|_2} &\leq \sqrt{\frac{8n}{m}\log(d_1+d_2)}\ \frac{\max\{\|B\|_2, \|C\|_2\}}{\|B\|_2\|C\|_2} + \frac{2}{3}\ \frac{n\log(d_1+d_2)}{m\|B\|_2\|C\|_2} \\
&\leq \sqrt{\frac{8n}{m}\log(d_1+d_2)}\ \frac{1}{\min\{\|B\|_2, \|C\|_2\}} + \frac{2}{3}\ \frac{\mu\log(d_1+d_2)}{m} \\
&= \sqrt{\frac{n}{\|B\|_2\|C\|_2}\ \frac{8}{m}\ \log(d_1+d_2)\ \frac{\|B\|_2\|C\|_2}{\min^2\{\|B\|_2, \|C\|_2\}}} + \frac{2}{3}\ \frac{\mu\log(d_1+d_2)}{m} \\
&\leq \sqrt{\frac{8\mu n}{m}\log(d_1+d_2)} + \frac{2}{3}\ \frac{\mu\log(d_1+d_2)}{m}
\end{aligned}$$

Note last inequality follows since $\min^2\{\|B\|_2, \|C\|_2\} = \min\{\|B\|_2^2, \|C\|_2^2\}$ and $1 \leq \|B\|_2, \|C\|_2 \leq n$.

Setting the relative error to be less than ε, the number of samples (m) will be $m = \Omega(\varepsilon^{-2}\mu n\log(d_1+d_2))$. This shows the *number of samples* we need to approximate product of two matrices is proportional to the *geometric mean of their numeric rank*. And that makes sense because if two matrices of high rank, it means there are many distinct directions in their spectrum, therefore we need to pick many samples to get a good approximation to their product.

Bibliography

[1] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.