CS 270 Combinatorial Algorithms and Data Structures, Spring 2015

Lecture 2: Tail inequalities

We will see basic tail inequalities including Markov, Chebyshev and Chernoff, and then see some applications including a simple version of graph sparsification.

Concentration inequalities

On a high level, tail inequalities or concentration inequalies give upper bounds on the probability that the value of a random variable is far from its expected value, and these allow us to show that randomized algorithms behave like what we expect with high probability.

We will see the basic and most useful ones today. The simplest one is the Markov's inequality.

Markov's inequality Let X be a non-negative discrete random variable.

Then $\Pr(X \geq a) \leq E[X]/a$ for any $a > 0$.

Proof $E[X] = \sum_i i \cdot \Pr(X=i) \geq \sum_{i \geq a} i \cdot \Pr(X=i) \geq \sum_{i \geq a} a \cdot \Pr(X=i) = a \Pr(X \geq a)$. □

Quicksort: It is known that the expected runtime of randomized quicksort is $2n \ln n$.

Then Markov's inequality tells us that runtime is at least $2cn \ln n$ with probability $\leq \frac{1}{c}$.

Coin flipping: If we flip n fair coins, the expected number of heads is $\frac{n}{2}$, and Markov's inequality tells us that the probability that there are $\geq \frac{3n}{4}$ heads is at most $\frac{2}{3}$.

Remark: Markov's inequality is most useful when we have no information beyond the expected value (or when such information is difficult to obtain, e.g. when the random variable is complicated to analyze).

In the above examples, we could prove much sharper results using Chernoff bounds.

Note that Markov's inequality could be tight. Do you have an example?

Moments and variance To give better bounds, one needs to use more information about the random variable, and a commonly used quantity is the variance of the random variable, which measures the typical difference of a random variable to its expected value.

The k-th moment of a random variable X is defined as $E[X^k]$, e.g. second moment is $E[X^2]$

The k-th moment of a random variable X is defined as $E[X^k]$, e.g. second moment is $E[X^2]$

The variance of X is defined as $Var[X] = E[(X-E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - E[X]^2$.

The standard derivation of X is defined as $\sigma[X] = \sqrt{Var[X]}$.

(Ideally, we like to compute $E[|X-E[X]|]$ but $Var[X]$ and $\sigma[X]$ are often much easier to handle.)

The covariance of two random variables X, Y is defined as $Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$.

We say X, Y are positively correlated if $Cov(X,Y) > 0$, negatively correlated if $Cov(X,Y) < 0$.

The following are some simple facts whose proofs are left as exercises.

- $Var[X+Y] = Var[X] + Var[Y] + 2cov(X,Y)$.
- If X and Y are independent, then $Var[X+Y] = Var[X] + Var[Y]$.

We would like to distinguish distributions that are concentrated around its expected value and those that are not. One possible test is to compute $E[X^2]$ and see how far it is from $E[X]^2$.

Chebyshev's inequality provides such a bound.

<u>Chebyshev's inequality</u> For any $a > 0$, $Pr(|X-E[X]| \geq a) \leq Var[X]/a^2$.

<u>Proof</u> $Pr(|X-E[X]| \geq a) = Pr((X-E(X))^2 \geq a^2) \leq E[(X-E(X))^2]/a^2 = Var[X]/a^2$, where the inequality follows from Markov's inequality as $(X-E[X])^2$ is non-negative. $\square$

<u>Coin flipping</u> Let X be the number of heads in n independent fair coin flips.

Again we try to bound $Pr(X \geq 3n/4)$, but this time we use Chebyshev's inequality.

For this, we need to compute $Var[X]$.

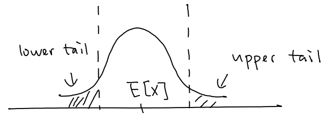
By independence, $Var[X] = \sum_{i=1}^{n} Var[X_i]$, where $X_i = \begin{cases} 1 & \text{if } i\text{-th coin flip is head} \\ 0 & \text{otherwise} \end{cases}$

So, $Var[X_i] = \frac{1}{2}(1-\frac{1}{2})^2 + \frac{1}{2}(0-\frac{1}{2})^2 = \frac{1}{4}$. (In general, if head with prob p, then $Var[X_i] = p(1-p)$)

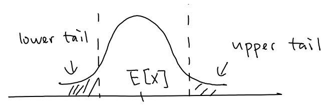
Hence, by Chebyshev, $Pr(X \geq 3n/4) \leq Pr(|X-E[X]| \geq \frac{n}{4}) \leq Var[X]/(\frac{n}{4})^2 = \frac{4}{n}$.

<u>Remark:</u> Chebyshev's inequality is most useful when we only have the second moment or when the second moment is easy to compute and is enough, e.g. second moment method, data streaming, etc.

<u>Sum of independent variables</u>



Sum of independent variables



The general question is to bound $\Pr(X > (1+\varepsilon)E[X])$ (upper tail) and $\Pr(X < (1-\varepsilon)E[X])$ (lower tail).

We consider the situation when X is the sum of many independent random variables, which is commonly seen in the analysis of randomized algorithms.

The law of large number asserts that the sum of n independent identically distributed variables is approximately $n\mu$, where μ is a typical mean.

The central limit theorem says that $\frac{X - n\mu}{\sqrt{n\sigma^2}} \to N(0,1)$. the deviations from $n\mu$ are typically of the order $\sqrt{n}$.

Chernoff bounds give us quantitative estimates of the probabilities that X is far from $E[X]$ for any (large enough) value of n.

Consider a simple setting where there are n coin flips, each is head with probability p.

The expected number of heads is np.

To bound the upper tail, in principle we just need to compute $\Pr(X \geq k) = \sum_{i \geq k} \binom{n}{i} p^i (1-p)^{n-i}$, and show that it is very small when k is much larger than np (say $k \geq (1+\varepsilon)np$), but this sum is not easy to work with and this method is not easy to be generalized.

Instead, we extend the approach of using Markov's inequality. The Markov's inequality is often too weak, but recall in the proof of Chebyshev's inequality we can strengthen it if we know the second moment of X.

To extend this, one can use the fourth moment or any $2k$-th moment to get (why even?)

$$\Pr(|X - E[X]| > a) = \Pr((X - E[X])^{2k} > a^{2k}) \leq E[(X - E[X])^{2k}] / a^{2k}$$

The idea in proving the Chernoff bounds is to consider:

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq E[e^{tX}] / e^{ta} \quad \text{for any } t > 0.$$

There are at least two reasons that we consider e^{tX}:

- Let $M_X(t) = E[e^{tX}] = E[\sum_{i \geq 0} \frac{t^i}{i!} X^i] = \sum_{i \geq 0} \frac{t^i}{i!} E[X^i]$ If we have $M_X(t)$, to compute $E[X^i]$, we can just compute $M_X^{(k)}(0)$, where $M_X^{(k)}(0)$ is the k-th derivative of $M_X(t)$ evaluated at $t=0$.

can just compute $M_X^{(k)}(0)$, where $M_X^{(k)}(0)$ is the k-th derivative of $M_X(t)$ evaluated at $t=0$.

So, $M_X(t)$ contains all the moments information, and is called the moment generating function.

It gives a strong bound when applying Markov's inequality, as the denominator is exponentially large.

- If $X = X_1 + X_2$ and X_1, X_2 are independent, then $E[e^{tX}] = E[e^{tX_1} e^{tX_2}] = E[e^{tX_1}]\, E[e^{tX_2}]$.

  So, this function is easy to compute when X is the sum of independent random variables.

Chernoff Bounds

Roughly speaking, Chernoff-type bounds are the bounds obtained by $\Pr(X \geq a) \leq E[e^{tX}]/e^{ta}$.

Let us consider a useful case when X is the sum of independent heterogenous coin flips.

Heterogenous coin flips:

Let $X_1, \ldots, X_n$ be independent random variables with $X_i = 1$ with probability p_i and $X_i = 0$ otherwise.

Let $X = \sum_{i=1}^n X_i$. Let $\mu = E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$ be the expected value.

Then $E[e^{tX}] = E[e^{tX_1} e^{tX_2} \cdots e^{tX_n}] = \prod_{i=1}^n E[e^{tX_i}]$ by independence

$$= \prod_{i=1}^n \left(p_i e^{t\cdot 1} + (1-p_i) e^{t\cdot 0}\right) = \prod_{i=1}^n \left(1 + p_i(e^t - 1)\right) \leq \prod_{i=1}^n e^{p_i(e^t-1)} = e^{\mu(e^t-1)}.$$

We put in some specific parameters to get some useful bounds.

Theorem In the heterogenous coin flipping setting, we have:

① for $\delta > 0$, $\Pr(X \geq (1+\delta)\mu) < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$.

② for $0 < \delta < 1$, $\Pr(X \geq (1+\delta)\mu) < e^{-\delta^2\mu/3}$.

③ for $R \geq 6\mu$, $\Pr(X \geq R) \leq 2^{-R}$.

proof ① $\Pr(X \geq (1+\delta)\mu) \leq E[e^{tX}]/e^{t(1+\delta)\mu} \leq e^{\mu(e^t-1)}/e^{t(1+\delta)\mu}$

By elementary calculus, we find out that this term is minimized when $t = \ln(1+\delta)$, and

this implies that $\Pr(X \geq (1+\delta)\mu) \leq e^{\mu\delta}/(1+\delta)^{(1+\delta)\mu}$, proving ①.

② When $0 < \delta < 1$, it holds that $e^\delta/(1+\delta)^{1+\delta} \leq e^{-\delta^2/3}$.

This can be verified by taking log of both sides and letting $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3}$,

and show that $f'(\delta) \leq 0$ in the interval $[0,1]$, and thus $f(\delta) \leq 0$ in this interval

since $f(0) = 0$, and this implies the claim. (See MU Theorem 4.4 for details.)

since $f(0)=0$, and this implies the claim. (See MU Theorem 4.4 for details.)

③ Let $R=(1+\delta)\mu$. When $R \geq 6\mu$, we have $\delta \geq 5$.

Hence, $\Pr(X \geq (1+\delta)\mu) \leq \left(e^{\delta}/(1+\delta)^{1+\delta}\right)^{\mu} \leq \left(e/(1+\delta)\right)^{(1+\delta)\mu} \leq (e/6)^R \leq 2^{-R}$. □

Similar bounds hold for the lower tail; very similar proof (by setting $t<0$). (See MU Thm 4.5)

Theorem. In the heterogenous coin flipping setting, we have for $0<\delta<1$

① $\Pr(X \leq (1-\delta)\mu) \leq \left(e^{-\delta}/(1-\delta)^{1-\delta}\right)^{\mu}$

② $\Pr(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$.

Corollary In the heterogenous coin flipping setting, $\Pr(|X-\mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}$ for $0<\delta<1$.

Hoeffding extension The same bounds hold when each X_i is a random variable taking values in $[0,1]$ with mean p_i. This is because the function e^{tx} is convex, and thus it always lies below the straight line joining the endpoints $(0,1)$ and $(1,e^t)$. This line has the equation $y=\alpha x+\beta$ for $\alpha=e^t-1$ and $\beta=1$. Therefore, $E[e^{tX_i}] \leq E[\alpha X_i+\beta] = p_i(\alpha+\beta)+(1-p_i)\beta = 1+p_i(e^t-1)$, and the same calculations as above follow.

Remarks:

- The same method holds for other random variables, e.g. Poisson r.v., Gaussian r.v., etc.
- It is often an easier way to compute the moments by computing the moment generating functions.
- Chernoff bounds also hold for negatively correlated variables, because $E[e^{t(X+Y)}] \leq E[e^{tX}]E[e^{tY}]$ and then the same proof works, and this observation is very useful in some applications.

Basic Examples:

① Coin Flips: Consider n independent fair coin flips, so $\mu = n/2$.

$\Pr(|\#\text{heads} - \mu| \geq \delta\mu) \leq 2e^{-\delta^2\mu/3} = 2e^{-\frac{\delta^2 n}{6}}$.

So, by setting $\delta = \sqrt{\frac{60}{n}}$, this probability is at most $2e^{-10}$.

Therefore, we conclude that $\Pr(|\#\text{heads} - \frac{n}{2}| \geq \sqrt{60n}) \leq 2e^{-10}$

So, with high probability, the number of heads is within $O(\sqrt{n})$ of the expected

—

value, and this $\sqrt{n}$ term is something to remember, as it comes up in different places.

And this is the right bound as there is a constant probability that $|\#\text{heads} - \frac{n}{2}| \geq \sqrt{n}$.

Recall that Markov's inequality implies that $\Pr(\#\text{heads} \geq \frac{3n}{4}) \leq \frac{2}{3}$. Chebyshev's inequality implies that $\Pr(\#\text{heads} \geq \frac{3n}{4}) \leq \frac{4}{n}$

Chernoff's bound implies that $\Pr(\#\text{heads} \geq \frac{3n}{4}) \leq e^{-(\frac{n}{2})(\frac{1}{2})^2/3} = e^{-n/24}$, exponentially small.

② Probability amplication:

Recall that the success probability of a randomized algorithm with one-sided error can be amplified easily: say the algorithm is always correct when it says NO and is correct with prob p when it says YES. To decrease the failure probability, we just repeat the algorithm k times or until it says NO, then the failure probability is at most $(1-p)^k$ when it says YES k times for a NO instance. For constant p, repeating $\log n$ times will decrease the failure probability to $O(1/n)$.

Suppose the randomized algorithm is two-sided error, say it has 60% of giving the correct answer, but it could make mistakes when it says YES or NO. To decrease the failure probability, we run the algorithm for k times and output the majority answer. Say the instance is a YES instance. The majority answer is wrong when the randomized algorithm outputs NO for more than $k/2$ times. But the expected number of answering NO is equal to $0.4k$ by our assumption. So, by Chernoff bound, the majority answer is wrong is

$$\Pr(\#\text{NO} > (1+\tfrac{1}{4})E[\#\text{NO}]) \leq e^{-\mu\delta^2/3} = e^{-0.4k(1/4)^2/3} = e^{-k/120}.$$

Therefore, by repeating $k = O(\log n)$ times, the failure probability is at most $O(1/n)$.

This is of the same order as in the case of one-sided error.

This $O(\log n)$ term is another quantity to remember, and it will also come up in different places.

Graph Sparsification

Given an edge weighted undirected graph $G=(V,E,w)$, for a subset of vertices $S \subseteq V$, let $\delta_G(S)$ be the set of edges with one endpoint in S and another endpoint in $V-S$, and

Let $w(\delta_G(S))$ be the total weight of the edges in $\delta_G(S)$.

We say H is a $(1\pm\varepsilon)$-cut-approximator of G if $(1-\varepsilon)w(\delta_G(S)) \le w(\delta_H(S)) \le (1+\varepsilon)w(\delta_G(S))$ for all $S \subseteq V$. Note that H is on the same vertex set but may have different edge weights.

We are interested in finding a $(1\pm\varepsilon)$-cut-approximator of G that is sparse (having few edges).

Assumption: We consider a simple setting in which G is unweighted and has min-cut value $\Omega(\log n)$.

Algorithm: In this simple setting, the algorithm is very simple. Set a sampling probability p.

For every edge $e \in E(G)$, put it in H with weight $\frac{1}{p}$ with probability p.

Theorem Set $p = 9\ln n/(\varepsilon^2 c)$ where c is the min-cut value of G.

Then H is a $(1\pm\varepsilon)$-cut-approximator of G with $O(p \cdot |E(G)|)$ edges with prob $\ge 1-\frac{4}{n}$.

proof Consider a subset $S \subseteq V$. Say $\delta_G(S)$ has k edges. Note that $k \gtrsim c$ by definition.

By linearity of expectation, $E[|\delta_H(S)|] = pk$ and thus $E[w(\delta_H(S))] = p \cdot k \cdot \frac{1}{p} = k = |\delta_G(S)|$.

Since each edge is an independent 0-1 random variable, by Cheroff, we have

$$\Pr\left[\, \big|\, |\delta_H(S)| - pk \,\big| > \varepsilon pk \right) \le 2e^{-pk\varepsilon^2/3} = 2e^{-\left(\frac{9\ln n}{\varepsilon^2 c}\right)\left(\frac{k\varepsilon^2}{3}\right)} = 2e^{-\frac{3k\ln n}{c}}$$

Since $k \ge c$ by definition, this implies that this probability is at most $2/n^3$.

So, the probability that one cut is violated is pretty small but there are exponentially many cuts. The important observation here is that the probability that a large cut is violated is much smaller, and there are not too many small cuts.

Claim: The number of cuts with at most αc edges for $\alpha \ge 1$ is at most $n^{2\alpha}$.

proof: It follows from the same analysis in the random contraction algorithm that a cut with αc edges survive with probability at least $1/n^{2\alpha}$. (We've seen the argument for $\alpha = 1$.) $\square$

Now, with the claim, we can bound

$$
\begin{aligned}
&\Pr(\text{some cut } S \text{ is violated}) \\
&\le \sum_{S \subseteq V} \Pr(\text{cut } S \text{ is violated}) \qquad (\text{this is called the union bound}) \\
&\le \sum_{\alpha \ge 1} \sum_{S \subseteq V \,:\, |\delta_G(S)| \le \alpha c} \Pr(\text{cut } S \text{ is violated}) \\
&\le \sum_{\alpha \ge 1} n^{2\alpha} \Pr(\text{cut } S \text{ is violated} \mid |\delta_G(S)| \le \alpha c) \\
&\le \sum_{\alpha \ge 1} n^{2\alpha} 2e^{-3\alpha \ln n}
\end{aligned}
$$

$$= \sum_{\alpha \geq 1} 2n^{-\alpha} \leq 4/n.$$

It is easy to see that H has $O(p \cdot |E(G)|)$ edges with high probability, and the theorem follows. $\square$

Applications : We briefly sketch some applications of the theorem.

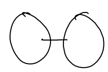
① Approximate min s-t cut : Given G, fix ε (say to 0.01), construct H with $\tilde{O}(\frac{m}{c})$ edges.

Solve min s-t cut in H. The same cut is a $(1+\varepsilon)$-approx solution in G.

② Approximate max s-t flow: Randomly color the edges by $\frac{1}{p}$ colors.

Each color subgraph has $O(p|E(G)|)$ edges, and the max s-t flow in each subgraph is at least $(1-\varepsilon)p \cdot OPT$ where OPT is the max s-t flow value in G, as each color subgraph has the same distribution as a random sampled subgraph of G.

Solve max s-t flow in each subgraph and combine the solutions to get a $(1-\varepsilon)$-approximate max s-t flow in G.

Improvements: Without the minimum cut assumption, then it is easy to see that a uniform sampling algorithm won't work, e.g. in , it is very likely that the cut edge is not picked.

Benczur and Karger designed a very clever non-uniform sampling algorithm, where the sampling probability for each edge is inversely proportional to the "connectivity" of the two endpoints.

They defined a notion called "strong connectivity" and proved that sampling inversely proportional to it will result in a $(1 \pm \varepsilon)$-cut-approximator with $O(n \log n)$ edges.

Furthermore, they showed that there is an almost linear-time algorithm to estimate the strong connectivity which is good enough for the purpose, leading to an $\tilde{O}(n^2)$-algorithm for approx min s-t cut.

The definition of "strong connectivity" is a bit unnatural, and they conjectured that it can be replaced by the more natural edge-connectivity (i.e. the max u-v flow value for edge uv).

This Conjecture is proven by Fung, Hariharan, Harvey and Panigrahi in 2011.

Spectral sparsification : Spielman and Tang defined a notion called "spectral sparsifier" that is stronger then that of a "cut sparsifier". Spielman and Srivastava proved that a spectral sparsifier with $O(n\log n)$ edges alway exists. The algorithm is also by random sampling, by assigning each edge a probability proportional to the effective resistance of the edge.

I plan to prove this result in the third part of the course, which has implications in solving linear equations. We won't discuss Benczur-Karger in more details.

References The content of this lecture is based on chapter 3 and 4 of the book "probability and computing" by Mitzenmacher and Upfal. This book is easy to read with detailed explanations and calculations and is good for self-study.

There is very nice mathematics in the theory of graph sparsifications. This is an excellent topic for course project. You can read the survey paper mentioned in the course project page for all the references that were discussed above.