CS 761 : Randomized Algorithms , Spring 2023 , Waterloo

Lecture 3 : Concentration Inequalities

Concentration inequalities are important tools in analyzing randomized algorithms.

We will start with the simple Markov's inequality. and then the Chebychev's inequality and the second moment method, and finally the Chernoff-Hoeffding inequality.

Concentration Inequalities

On a high level, tail inequalities or concentration inequalities 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 as what we expect with high probability (almost deterministic)

These are fundamental tools in analyzing randomized algorithms that we will use throughout the course.

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

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 $\leq \frac{1}{c}$ for $c \geq 1$.

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. the random variable is complicated to analyze). In the above examples, we can prove much sharper results using Chernoff bounds that we will see soon.

Questions: ① Is Markov's inequality tight? Can you give an example?

② Does it hold for general random variables (not just non-negative)?

③ Can it be modified to upper bound $\Pr(X \leq a)$ (e.g. $\Pr(X \leq E[X]/2)$) ?

Moments and Variance To give better bounds, we need 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 <u>k-th moment</u> of a random variable X is defined as $E[X^k]$, e.g. second moment is $E[X^2]$

The <u>variance</u> 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 <u>standard derivation</u> of X is defined as $\sigma[X] = \sqrt{Var[X]}$.

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

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

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

- $Var[X,Y] = Var[X] + Var[Y] + 2\,cov(X,Y)$.
- 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.

Second Moment Method

The first moment method is to compute the expected value of a random variable, and conclude that there is an outcome with value at least $E[X]$ or at most $E[X]$.

For a non-negative random variable X, we can use the Markov inequality to prove that $Pr(X \geq 1) \leq E[X]$, and thus conclude that $X=0$ with high probability if $E[X] \ll 1$ for an integral X.

Often we also want to prove that $\Pr(X \geq 1)$ is large. It is not enough to just show that $E[X]$ is large, as it could be the case that X is very large for a small fraction of the outputs, while $X=0$ for a large fraction of the outputs.

To exclude this case, we need some kind of concentration inequalities (e.g. variance of X is small), and the second moment method provides one way to establish this.

<u>Theorem</u> If X is an integral-valued random variable, then $\Pr(X=0) \leq \frac{\mathrm{Var}[X]}{(E[X])^2}$

<u>Proof</u> By Chebyshev's inequality, $\Pr(X=0) \leq \Pr(|X - E[X]| \geq E[X]) \leq \mathrm{Var}[X]/(E[X])^2$. □

<u>Corollary</u> If $\mathrm{Var}[X] = o((E[X])^2)$ or $E[X^2] = (1+o(1))(E[X])^2$, then $X > 0$ with high probability.

Threshold Behaviour in Random Graphs

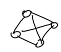
Let $G_{n,p}$ be a graph with n vertices where each edge appears with probability p.

A property has a <u>threshold behavior</u> if there is a function f such that:

- when $\lim_{n\to\infty} \frac{g(n)}{f(n)} = 0$, then almost surely $G(n, g(n))$ does not have this property.
- when $\lim_{n\to\infty} \frac{h(n)}{f(n)} = \infty$, then almost surely $G(n, h(n))$ has this property.

A property has a <u>sharp threshold behavior</u> if there is a function f such that for any $\varepsilon>0$,

- $G(n, (1-\varepsilon)f(n))$ almost surely does not have the property.
- $G(n, (1+\varepsilon)f(n))$ almost surely has the property.

For example, consider the property of having a clique of size 4, i.e. .

Let X be the number of 4-cliques in $G_{n,p}$. Then $E[X] = \binom{n}{4} p^6$.

When $p = o(n^{-\frac{2}{3}})$, then $E[X] \to 0$, and we can conclude that $\Pr(X=0) \to 1$ by the first moment method.

On the other hand, when $p = \omega(n^{-\frac{2}{3}})$, then $E[X] \to \infty$, and we can use the second moment method to conclude that $\Pr(X \geq 1) \to 1$, by showing that $\mathrm{Var}[X] = o((E[X])^2)$. See [MU] Chapter 6.5.1.

<u>Theorem</u> The property of having a clique of size 4 has a threshold function $f(n) = n^{-2/3}$.

Extending this result, one can prove that $G(n, \frac{1}{2})$ contains a clique of size $\sim 2\log_2 n$ with high probability. See Section 4.5 of Alon-Spencer.

Let's see a property with a sharp threshold.

Consider the property that a random graph has diameter less than or equal to two, i.e. for every pair of vertices there is a path of length at most two connecting them.

Theorem. The property that $G(n,p)$ has diameter at most two has a sharp threshold at $p = \sqrt{\frac{2\ln n}{n}}$.

We omit the proof this time. See 2018 L11 for a proof.

Random Graphs and Random Structures

Threshold phenomenons are common in random graphs and it is a subject of intense study.

An important result is the emergence of "giant component": When we sample a random graph with edge probability $p = \frac{1}{n}$, the largest components are of size $O(n^{\frac{2}{3}})$ and they are almost surely trees.

But when $p \geq (1+\varepsilon)n$, then there is a unique giant component of size $\Omega(n)$, while all other components are of size $O(\log n)$.

Other random structures also have this phenomenon of "phase transition".

For random 3-SAT formula where each clause has three random variables (or their negations).

It is conjectured and generally believed that when the clause-to-variable ratio is less than 4.2, then the formula is almost surely satisfiable, and when this ratio is greater than 4.2, then the formula is almost surely unsatisfiable.

The same is conjectured for k-SAT, and the conjecture is very recently settled by Ding, Sly, and Sun in the paper "Proof of the satisfiability conjecture for large k".

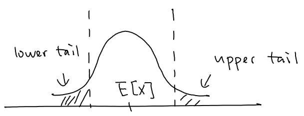
Algorithmic Issues

The second moment method can be used to show that $G_{n,1/2}$ has a clique of size $2\log_2 n$ almost surely.

While it is easy to find a clique of size $\log_2 n$ (a simple greedy algorithm would work), it is not known how to find a clique of size $(1+\varepsilon)\log_2 n$ in polynomial time, and in fact there is some evidence suggesting it may be computationally hard.

Similarly, there is no known polynomial time algorithms to determine whether a random 3-SAT formula with clause-to-variable ratio 4.2 is satisfiable or not. In fact, these are some hardest instances for 3-SAT that we know how to generate efficiently.

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 the mean of a random variable

The central limit theorem says that $\frac{X-n\mu}{\sqrt{n\sigma^2}} \to N(0,1)$. the deviation from $n\mu$ are typically of the order $\sqrt{n}\sigma$. where σ is the standard derivation of a random variable.

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 $t=0$.

  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 for Bounded Variables

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.

<u>Heterogenous coin flips</u>:

Let $X_1, \dots, X_n$ 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

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

using $1 + x \le e^x$

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

<u>Theorem</u> In the heterogenous coin flipping setting, we have:

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

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

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

<u>Proof</u> ① $\Pr(X \ge (1+\delta)\mu) \le E[e^{tX}] / e^{t(1+\delta)\mu} \le e^{\mu(e^t-1)} / e^{t(1+\delta)\mu}$

By elementary calculus, we find out that this term is minimized when $t = \ln(1+\delta)$, and this implies that $\Pr(X \ge (1+\delta)\mu) \le e^{\mu\delta} / (1+\delta)^{(1+\delta)\mu}$, proving ①.

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

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 $f'(\delta) \le 0$ in the interval $[0,1]$, and thus $f(\delta) \le 0$ in this interval since $f(0) = 0$, and this implies the claim. (see MU Theorem 4.4 for details.)

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

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

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

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

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

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

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

<u>Hoeffding extension</u> 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 $E[e^{tX_i}] \le E[\alpha X_i + \beta] = p_i(\alpha + \beta) + (1-p_i)\beta = 1 + p_i(e^t - 1)$ ...

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, $E[e^{tX_i}] \le E[\alpha X_i + \beta] = p_i(\alpha+\beta) + (1-p_i)\beta = 1 + p_i(e^t - 1)$. and then the same calculations as above follow.

Remarks:

- The same method holds for other random variables, e.g. Poisson random variables, Gaussian random variables, 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 $E[e^{t(X+Y)}] \le E[e^{tX}] \cdot E[e^{tY}]$ and then the same proof works. and this observation is very useful in some applications.

For example, it is known that two edges appear in a random spanning tree are negatively correlated, and thus Chernoff bounds apply to analyze random spanning trees even though the variables are dependent.

Basic Examples

① Coin Flips: Consider n independent fair coin flips, so the expected # of heads is $\mu = \frac{n}{2}$.

$$\Pr(\,|\#\text{heads} - \mu| \ge \delta\mu\,) \le 2e^{-\mu\delta^2/3} = 2e^{-n\delta^2/6}.$$

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

Therefore, we conclude that $\Pr(\,|\#\text{heads} - \frac{n}{2}| \ge \sqrt{15n}\,) \le 2e^{-10}$.

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 $|\#\text{heads} - \frac{n}{2}| \ge \sqrt{n}$.

Recall that Markov's inequality implies $\Pr(\#\text{heads} \ge \frac{3n}{4}) \le \frac{2}{3}$, and Chebyshev's inequality implies that $\Pr(\#\text{heads} \ge \frac{3n}{4}) \le \frac{4}{n}$

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

② Probability amplification

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

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

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.

③ Balls and bins: We throw n balls into n bins, where each ball goes to a uniformly random bin independently.

Exercise: Give an upper bound on the maximum load of a bin that holds with high probability.

References: Chapter 3 and 4 of Mitzenmacher and Upfal.

- The book "Probabilistic methods" by Alon and Spencer is a great resource.
- One important example of the first moment method is the work of Shannon who showed the existence of good error correcting codes. See [MU].