Lecture 4: Concentration inequalities

Rajat Mittal

IIT Kanpur

We learned about random variables, their expectation and their variance in previous lectures.

- In many applications (given expectation and variance) we are interested in bounding the probability that the random variables is close to expectation. If variance is small, intuitively, the random variable should be close to the expectation.
- One interpretation of expectation was, if the random variable is repeated large number of times, then the average is close to the expected value with high probability.

Such results are called *concentration inequalities*, they allow us to lower bound probability that a random variable is close to a chosen quantity (in other words, random variable is concentrated around a quantity).

1 Concentration inequalities

A concentration inequality allows us to obtain bounds on the distribution of the random variable given aggregate properties (like expectation and variance) and possibly some restrictions on the behaviour of the random variable. An intuitive way to look at many different concentration inequalities is, stronger the conditions on the behaviour or aggregate properties, we can obtain stronger bounds on the distribution.

Markov's inequality will be covered first, a basic inequalities about the distribution using just its expected value. Using both, expectation and variance, we will derive Chebyshev inequality. Finally, we will look at repeating a random experiment (or a random variable). Two results, *law of large numbers* and *Chernoff bound* will be shown, formalizing the interpretation of expectation discussed in the previous lectures.

Another important result in this area is known as *central limit theorem (CLT)*. It states that, the normalized sum of independent copies of a random variable tends towards a normal distribution (with the same expectation as the original random variable) as the number of copies increase (irrespective of the original distribution of the random variable). The variance goes down as the number of copies increase and is related to the variance of the original random variable.

It is important to remember the essence of central limit theorem. Though, we will not cover its proof in this course.

1.1 Markov inequality

We start with *Markov's inequality*. It uses only the expectation of the random variable, making it very general but very weak. It almost follows from the definition of expectation.

Theorem 1 (Markov's inequality). *Given a positive random variable X and $a > 0$,*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Note 1. If the random variable is not positive then, $Pr(|X| \geq a) \leq \frac{E[|X|]}{a}$, by applying Markov's inequality to $|X|$.

Exercise 1. Before looking at the proof, why do you think the inequality should be true? Can you come up with a proof?

Proof. The main idea is, there will be two positive contributions to the expectation, one from values higher than a and other from values lower than a. If lot of weight (probability) is placed on values higher than a, already the contribution will be more than the expectation.

Exercise 2. Why do we need X to be positive?

Formally, the result will be proved by contradiction. Assume that the converse holds, $Pr(X \geq a) > \frac{E[X]}{a}$.

$$\begin{aligned} E[X] &= \sum_x P(X = x)x \\ &\geq \sum_{x<a} P(X = x).0 + \sum_{x \geq a} P(X = x).a \\ &= a \sum_{x \geq a} P(X = x) \\ &> E[x] \end{aligned} \tag{1}$$

Where the last inequality follows from assumption. So the assumption is false and hence Markov inequality is proved. □

Markov inequality does not use variance. How can we formalize the statement, if the variance is low, we should be close to expectation with high probability. The idea is to consider the random variable $|X - E[X]|$, this gives *Chebyshev's inequality.*

Theorem 2 (Chebyshev's inequality). *Let X be a random variable and $a > 0$ be a positive real number. Then,*

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

You will prove this inequality in the assignment.

1.2 Law of large numbers

We move to the first theorem about repetition of a random experiment. Let the random experiment be modeled by a random variable X. Suppose the experiment is repeated n times. Denote $X_1, X_2, \cdots, X_n$ to be n copies of X (they have the same distribution). We also assume that the family of random variables $\{X_i\}_{i=1}^n$ is pairwise independent (any two random variables are independent).

The intuition is, as n gets bigger, the average value of $X_1, X_2, \cdots, X_n$ should be close to $E[X]$. So, define a new random variable,

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Hence, $\bar{X}$ is the average of n repetitions of X (as a random variable).

Exercise 3. What is the expectation of $\bar{X}$?

This is not very difficult, using linearity of expectation, $E[\bar{X}] = E[X]$. To use Chebyshev's inequality on $\bar{X}$, we need its variance (in terms of the variance of X). We first make an easier observation.

Exercise 4. If $Y = aX$, where X is a random variable, then $Var[Y] = a^2\, Var[X]$.

Our remaining task is to find variance of $\sum_{i=1}^n X_i$. In general, it is not possible to give bounds on the variance of a sum of random variables. Though, we also know that $\{X_i\}$'s are pairwise independent.

Lemma 1. *Let $\{X_i\}_{i=1}^n$ be a pairwise independent family of random variables. Then,*

$$Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i].$$

Proof. The proof follows by a straightforward calculation of variance.

$$\begin{aligned}
Var[\sum_{i=1}^{n} X_i] &= E[\left(\sum_{i=1}^{n} X_i\right)^2] - E[\sum_{i=1}^{n} X_i]^2 \\
&= \sum_{i,j=1}^{n} E[X_i X_j] - \left(\sum_{i=1}^{n} E[X_i]\right)^2 \\
&= \sum_{i=1}^{n} E[X_i^2] + \sum_{i \neq j} E[X_i]E[X_j] - \left(\sum_{i=1}^{n} E[X_i]\right)^2 \\
&= \sum_{i=1}^{n} E[X_i^2] + \sum_{i \neq j} E[X_i]E[X_j] - \sum_{i=1}^{n} E[X_i]^2 - \sum_{i \neq j} E[X_i]E[X_j] \\
&= \sum_{i=1}^{n} E[X_i^2] - \sum_{i=1}^{n} E[X_i]^2 \\
&= \sum_{i=1}^{n} Var[X_i]
\end{aligned} \tag{2}$$

The first equality used linearity of expectation. Where did we use pairwise independence? □

Getting back to the original question about variance of $\bar{X}$,

Exercise 5. What is $Var[\bar{X}]$?

From the lemma and the observation above, $Var[\bar{X}] = \frac{1}{n} Var[X]$. Armed with the expectation and variance of $\bar{X}$, Chebyshev gives us *the law of large numbers.*

Theorem 3 (Law of large numbers). *Define the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where each X_i has the same distribution as a random variable X and are pairwise independent. Then,*

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

Notice that the probability, of $\bar{X}$ being away from $E[X]$, goes down to 0 as n increases ($Var[X], a$ are fixed numbers). In other words, if we repeat X multiple times, the average value of the outcome will be close to the expectation with high probability.

Note 2. In literature, this is known as the weak law of large numbers. There are different notions of convergence of random variables, this theorem is weaker in the sese that it only shows weaker notion of *convergence in probability*). You are encouraged to look at the statement of strong law of large numbers.

1.3 Chernoff bound

While looking at the distribution of $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, we saw that a concentration inequality can be derived by controlling $E[\bar{X}]$ and $E[\bar{X}^2]$ (related to variance). We were able to say something about $E[\bar{X}^2]$ because the any two X_i's were independent. It suggests that we can get better concentration inequalities by looking at $E[\bar{X}^k]$ for higher values of k (these expectations are called *moments*). Though to control $E[\bar{X}^k]$, we need X_i's to be independent in a stronger sense.

Note 3. Mutual independence implies that any subset of random variables follow the product rule, $P[X_1 = x_1, X_2 = x_2, \cdots X_k = x_k] = P[X_1 = x_1]P[X_2 = x_2] \cdots P[X_k = x_k]$. We will cover the two notions, pairwise and mutual independence, and their differences in the next lecture. At this point, just notice that mutual independence is a stronger criteria than pairwise independence.

For now, let us just assume that our repetitions are mutually independent, this situation is encountered frequently in computer science.

Suppose an experiment succeeds with probability p and fails otherwise. The expected value of success is p. If we repeat the experiment n times then the expected number of successes is np (by linearity of expectation). Chernoff bound shows that if we repeat the experiment many times (say n), then the number of successes will be close to np with very high probability (exponentially small in n).

The assumptions on random variables are much stronger (mutual independence and not just pairwise independence), this allows us to prove very tight bounds. This mutual independence condition is encountered quite often in computer science, we use the stronger bound in that case.

Theorem 4 (Chernoff bound). *Let X be a random variable which takes value 1 with probability p and 0 otherwise. Let $X_1, X_2, \cdots, X_n$ be n copies of X, where the family $\{X_i\}_{i=1}^n$ is mutually independent. Define $S = \sum_{i=1}^n X_i$, then*

$$P(S < (1-\delta)nE[X]) \leq e^{\frac{-nE[X]\delta^2}{2}}.$$

Proof. This proof is taken from John Canny's lecture notes [1].

We used variance of the random variables (second moment, $E[X^2]$) to show bounds in laws of large number. We wanted to use higher moments to get better results. For Chernoff bounds, $E[e^{-tX}]$ is used, in some sense capturing all the moments.

The proof of Chernoff bound follows by looking at the random variable e^{-tS}, where t is a parameter and will be optimized later. Define $u := E[S] = nE[X]$, so

$$P(S < (1-\delta)u) = Pr(e^{-tS} > e^{-t(1-\delta)u}).$$

We can apply Markov's inequality for e^{-tS},

$$P(S < (1-\delta)u) \leq \frac{E[e^{-tS}]}{e^{-t(1-\delta)u}}.$$

But e^{-tS} is the product of e^{-tX_i}, where X_i are independent. So,

$$P(S < (1-\delta)u) \leq \frac{\Pi_{i=1}^n E[e^{-tX_i}]}{e^{-t(1-\delta)u}}. \tag{3}$$

Exercise 6. Show that $E[e^{-tX_i}] = 1 - p(1-e^{-t}) \leq e^{p(e^{-t}-1)}$.

Above exercise implies that $\Pi_{i=1}^n E[e^{-tX_i}] \leq e^{u(e^{-t}-1)}$. From Eq. 3, we get

$$P(S < (1-\delta)u) \leq e^{u(e^{-t}+t(1-\delta)-1)}$$

Exercise 7. Show that the bound on right is minimized for $t = \ln \frac{1}{1-\delta}$.

Putting the best t, we get

$$P(S < (1-\delta)u) \leq \left(\frac{e^{-\delta u}}{(1-\delta)^{u(1-\delta)}}\right).$$

Using the Taylor expansion of $\ln(1-\delta)$,

$$P(S < (1-\delta)u) \leq e^{\frac{-u\delta^2}{2}}.$$

Hence proved.

□

Exercise 8. Suppose we toss an unbiased coin 1000 times. What is the probability that we get less than 400 heads?

Exercise 9. Suppose we toss a biased coin 1000 times (probability of getting head is .6). What is the probability that we get less than 500 heads?

Exercise 10. We saw two different repetition theorems. What are the differences in these two settings, law of large number and Chernoff bound.

We looked at the sum being smaller than expectation. Similar to above proof, you can also prove that the sum can't be much bigger than expectation. The bound turns out to be slightly different,

$$P(S > (1+\delta)nE[X]) \leq e^{\frac{-nE[X]\delta^2}{2+\delta}}.$$

In general, the exponential dependence on δ is important and not the constant in the denominator of the exponent. So, a safe bound to use for either side is,

$$e^{\frac{-nE[X]\delta^2}{3}}.$$

Applying the union bound,

$$P(|S - nE[X]| > \delta nE[X]) \leq 2e^{\frac{-nE[X]\delta^2}{3}}.$$

1.4 Calculating the value of π

Calculating the value of π is a natural question. We will see a *statistical* way to approximate the value of π (this example is taken from Mitzenmacher and Upfal [2]). Consider a square of area 4 and a circle inside it of radius 1 (look at Figure 1).

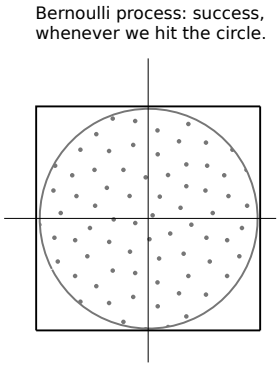

Fig. 1. Hitting a dart at a square.

Exercise 11. If we hit a dart uniformly randomly at the sqaure, what is the probability that it hits the circle?

This is not a difficult question to answer. The area of the square is 4, and the area of the circle is π. So, the probability that we land inside the circle is $\pi/4$. We can come up with a randomized strategy to compute the value of π.

Let X be the indicator random variable that a point lands inside the square, then we know $E[X] = \pi/4$. Let us pick n random points (uniformly) from the square. Denote the outcome by n independent copies of X, say $X_1, X_2, \cdots, X_n$.

Exercise 12. What do we expect $Y = \frac{1}{n}(\sum_{i=1}^{n} X_i)$ to be?

From linearity of expectation, $E[Y] = \pi/4$. The simple algorithm is: choose n points randomly in the square and say m of them hit the circle; our approximation of π is $4m/n$ (we equate the expectation of Y with our empirical estimate m/n).

Intuitively, as $n \to \infty$, we expect that our estimate will be close to the correct answer with high probability.

Exercise 13. How do we show this formally?

From the discussion above, our first guess should be a direct application of law of large numbers.

Exercise 14. What is the variance of X?

Since it is a Bernoulli random variable, the variance is $p(1-p)$, where $p = \pi/4$. Applying law of large numbers,

$$P[\Big|Y - \frac{\pi}{4}\Big| \geq a] \leq \frac{p(1-p)}{na^2}.$$

In other words, if we want an approximation error less than $4a$ (error of a in estimation of $\pi/4$) and failure probability at most δ, we must hit the square at least $n = \frac{p(1-p)}{\delta a^2}$ times. This is a linear dependence on $1/\delta$ (some might consider it to be too big a number).

Though, realize that the hits to the square were independent, we can improve our bounds on failure probability by using Chernoff bounds.

$$P(|nY - nE[X]| > \epsilon nE[X]) \leq 2e^{\frac{-nE[X]\epsilon^2}{3}}.$$

Cancelling n and using $E[X] = p$,

$$P(\Big|Y - \frac{\pi}{4}\Big| > \epsilon\frac{\pi}{4}) \leq 2e^{\frac{-np\epsilon^2}{3}}.$$

Like before, if we want an approximation error at most $4a$ (error of a in estimation of $\pi/4$) and failure probability at most δ, then

$$a = \epsilon\frac{\pi}{4} \text{ and } \delta = 2e^{\frac{-np\epsilon^2}{3}}.$$

Substituting ϵ and p,

$$\delta = 2e^{\frac{-4na^2}{3\pi}}.$$

This gives that we should repeat the experiment $n = \frac{3\pi}{4a^2}\ln\frac{2}{\delta}$ times. Notice that we should repeat the experiment only $O(\ln 1/\delta)$ times, instead of $O(1/\delta)$ times (as given by law of large numbers). The dependence on δ has reduced considerably (by application of Chernoff bound).

Note 4. This form of sampling, when we independently sample from a distribution to estimate its expectation, is called Monte Carlo sampling. You can read much more about it from the internet, Wikipedia and other resources.

2 Application: randomized algorithms

Let us start with an example to illustrate the idea of a randomized algorithm.

Assume that Jai and Veeru are caught by Gabbar. Given the mathematical inclinations of Gabbar (I bet you didn't know this about Gabbar), he proposes a condition to Jai and Veeru for their release.

Gabbar will give two non-zero strings $x, y \in \{0,1\}^n$ (one each) to Jai and Veeru after putting them into their respective rooms. That means, x will be given to Veeru and y will be given to Jai.

Jai and Veeru have a simple task, they need to find if $x = y$? If they give the correct answer, they are released, otherwise they are killed. The only problem is they cannot communicate a whole lot between these two rooms but the strings x, y are pretty big.

They can discuss the strategy now, but the strings will be given to them only after they are back to their rooms. Unfortunately, it is very difficult to communicate between these two rooms and they can only communicate very small number of bits to each other (after getting the strings). For the sake of this example, suppose $n = 10000$ and they are only allowed to communicate 10 bits.

What should Jai and Veeru do? Clearly transferring the entire string to each other is not possible. What if Jai and Veeru are ready to take some risk? They start thinking of a randomized strategy, such that, whatever be x, y, they are released with high probability.

One suggestion could be, pick a random small subset of $[n]$ and check whether x and y are equal on those co-ordinates.

Exercise 15. Show that if x, y are very close to each other, this strategy will fail.

Another strategy could be, Jai and Veeru agree on a randomly uniform string z of length n while discussing the strategy. After receiving the strings, Jai sends the inner product modulo 2 between x, z,

$$x^T z \mod 2 = \sum_{i=1}^{n} x_i z_i \mod 2$$

to Veeru. Veeru also calculates $y^T z \mod 2$, if both values agree then they say that x, y are equal.

If $x = y$ then $x^T z = y^T z \mod 2$ and they will be released.

Exercise 16. Show that if $x \neq y$ then Jai and Veeru are released with probability $1/2$.

Jai and Veeru just transferred 1 bit and got saved with $1/2$ probability. Is it possible to communicate some more bits and increase the probability of success?

The answer is not very difficult. They choose multiple strings $z_1, z_2, \cdots, z_{10}$ randomly (uniform and independent) and Jai sends the corresponding inner products $\{x^T z_i \mod 2\}$ to Veeru. Veeru matches the answer and says x, y are equal iff all inner products match.

Exercise 17. What is the probability that they succeed now?

You can show that if they share t random strings, their failure probability is $1/2^t$. In other words, just by 10 bits of communication, their error probability is $1/1024$, very small.

The algorithm (strategy) adopted by Jai and Veeru is a randomized strategy (picking z_i's).

2.1 Definition of randomized algorithms

Most of you have only seen *deterministic* algorithms till now, they come up with a correct answer all the time. This requirement is too strict; many a times, it is enough to correctly answer *almost all the time*. It means that our algorithm can make some randomized choices and it should output the correct answer with high probability over these choices. You have already seen an example in the last subsection.

More concretely, let F be a decision problem (output is 0 or 1) which takes $x \in X$ as input and denote the correct answer by $F(x) \in \{0,1\}$. A randomized algorithm A takes x and a random string r as input and outputs $A(x, r)$. There are different conditions under which randomized algorithm is considered *successful*.

- Las Vegas: $A(x, r) = F(x)$ for all x, r. In this case the expected running time of A is considered (we would want it to be better than a deterministic algorithm).
- Monte Carlo two-sided: For each x, $P_r[A(x, r) = F(x)] \geq 2/3$. Notice that the probability is over r (and not the input x). In this case, the worst case running time (over the inputs) of A is considered.
- Monte Carlo one-sided: For each x such that $F(x) = 0$, we want $P_r[A(x, r) = F(x)] = 1$. For each x such that $F(x) = 1$, we want $P_r[A(x, r) = F(x)] \geq 1/2$. The role of output being 0 and 1 can be reversed. Again, the probability is over r (and not the input x) and the worst case running time is considered.

The previous algorithm by Jai and Veeru was a one-sided Monte Carlo algorithm (when $x = y$, Jai and Veeru escape with certainty).

In general, a Monte Carlo randomized algorithm uses some randomness and outputs the correct answer with high probability on *every possible input*. Notice that the probability is over the randomness of the algorithm and not on picking the input.

Increasing the success probability: You might wonder, what is the significance of $2/3$ and $1/2$ in the definition of Monte Carlo algorithms? It turns out, they are not that important.

We have already seen how to boost the probability of success for the case of one sided error. If the failure probability of a one-sided error randomized algorithm is some constant ϵ (does not depend on input size), it can be repeated independently t times and failure probability reduces to ϵ^t. This is a sharp fall in the failure probability (exponential in t).

In other words, we can decrease the error probability to any small constant by repeating the algorithm constant many times (here constant means, it does not depend on input size). This shows that we could have chosen any constant between 0 and 1 instead of $1/2$ and still got the same definition (but with a constant factor increase in running time).

What if the randomized algorithm had two sided error?

Exercise 18. Convince yourself that if the algorithm succeeds with probability less than half on both sides, it is useless.

We will show that even if the two sided error randomized algorithm succeeds with probability slightly more than $1/2$, say $1/2 + \epsilon$ (again, ϵ is a constant), we can make the success probability as close to one as possible. In other words, for the definition of two-sided Monte Carlo, any constant between $1/2$ and 1 would have been fine. You want to guess the tool we will use to increase the success probability?

Exercise 19. How would you decrease the error of a two sided error algorithm?

Exercise 20. Suppose you want to decide whether a coin is biased towards heads or tails. Can you decide that with high probability by tossing it multiple times?

The natural answer seems to be, toss the coin multiple times and take the majority output as answer. What is the probability that the majority output is incorrect? How would you analyze such a situation?

This is exactly the situation where we toss a *biased* coin (towards heads) multiple times and ask the probability of getting less than half heads.

In this case, we will repeat the algorithm k times and take the majority vote to decide the output. Suppose the original algorithm (the one we are repeating) gives the correct answer with probability more than $\frac{1}{2} + \epsilon$. We assume that ϵ is a constant. Then after repeating it k times, using Chernoff bound (Thm. 4), the probability that we get the wrong answer is less than

$$e^{-\epsilon^2 k/2}.$$

You will show this in the assignment. Notice, that the failure probability reduces exponentially (though not as fast as the one sided case).

The number of times we need to repeat the algorithm, k, in bounded error or one sided error case is independent of size of input and only depends on probability we want to achieve. In other words, if we want to reduce error probability from one constant to another, it will only take constant many iterations.

Since we are mostly interested in asymptotic running time of a randomized algorithm, we can fix our favorite failure probability (we fixed it to $2/3$) and running time will only change up to a constant.

3 Assignment

Exercise 21. Let X be a random variable with $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Find $E[X_1 X_2 \cdots X_n]$ where $X_1, X_2, \cdots X_n$ are identical and independent copies of X.

Exercise 22. Prove Chebyshev's inequality.

$$P(|X - E[X]| \geq a) \leq \frac{Var(X)}{a^2}$$

Where $Var(X) = E[(X - E[X])^2]$.

Exercise 23. Read about central limit theorem.

Exercise 24. What bound will you get using law of large number in the setting of Chernoff bound (assume variance to be some constant). Is it better or worse?

Exercise 25. Suppose an algorithm with two sided error gives the correct answer with probability more than $\frac{1}{2} + \epsilon$. In this case, we will repeat the algorithm k times and take the majority vote to decide the output (assume that ϵ is a constant). Then, after repeating it k times, show that the probability of getting the wrong answer is less than $e^{-\epsilon^2 k/2}$

References

1. J. Canny. Cs174: Chernoff bounds. `http://www.cs.berkeley.edu/~jfc/cs174/lecs/lec10/lec10.pdf`.
2. M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2017.