Lecture 5: Concentration inequalities

Rajat Mittal

IIT Kanpur

We learned about random variables and their expectation in previous lectures. 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. The statement above will be formalized in this lecture note.

1 Concentration inequalities

A concentration inequality allows us to obtain bounds on the distribution of the random variable given aggregate properties (like expectation) and 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. Next, we introduce another aggregate measure of a distribution, called *variance*, to quantify the divergence of distribution from its expectation. Using both, expectation and variance, we will derive more inequalities. Finally, *law of large numbers* and *Chernoff bound* will be shown, formalizing the interpretation of expectation.

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?

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. □

1.2 Variance

We have already seen one aggregate measure of a random variable, expectation. Clearly, expectation captures only a very small amount of information about the random variable. It was emphasized before, expectation does not imply that we get $E[X]$ with high probability. In other words, a random variable which always takes value 0 has same expectation as the one which takes value -2 and 2 with equal probability (even one which takes values 1000 and -1000 with same probability). Expectation does not distinguish between these cases.

This gives rise to another measure of interest, called the *variance* of a random variable. The idea is to measure how far X could be from $E[X]$.

Your first guess might be $E[X-E[X]]$, but we need to take care of the signs of random variable $X-E[X]$. The *variance* of a random variable X is defined as,

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

Exercise 3. Show that

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

You can easily calculate the variance of a random variable taking value $\alpha, -\alpha$ with equal probability. It increases rapidly as the value of α increases, even though the expectation remains same.

A very related quantity is called the *standard deviation* and is defined as the square root of the variance.

$$\sigma(X) := \sqrt{Var[X]} = \sqrt{E[X^2] - E[X]^2}.$$

Define a random variable X to be 0 for tails and 1 for head, where coin gets head with probability p. This is called a Bernoulli random variable with parameter p.

Exercise 4. What is the expectation of this random variable? What is the expectation of X^2?

From this exercise, it is easy to find the standard deviation of X,

$$\sigma(X) = \sqrt{p - p^2} = \sqrt{p(1-p)}.$$

Exercise 5. For what value of p is this maximum? does it agree with your intuition?

Using variance gives us another inequality, called *Chebyshev's inequality*, by looking at random variable $|X - E[X]|$.

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.

1.3 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.

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

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

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

Exercise 6. 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 7. 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? Hint: it is a question in the assignment. □

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

Exercise 8. 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. You are encouraged to look at the statement of strong law of large numbers.

1.4 Chernoff bound

We move to a more specific type of repetition, 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). In this case, we assume that the repetitions are mutually independent instead of just pairwise independent.

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 $\{X_i\}$ 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, http://www.cs.berkeley.edu/ jfc/cs174/lecs/lec10/lec10.pdf.

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 9. 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 10. 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 11. Suppose we toss an unbiased coin 1000 times. What is the probability that we get less than 400 heads?

Exercise 12. 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 13. 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.

2 Application: randomized algorithms

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 14. 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 15. 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 16. 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). In general, a 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. The randomized procedure above was a *one sided error* kind. When $x = y$, Jai and Veeru escape with certainty.

What we showed above is, if the failure probability of a one-sided error randomized algorithm is some constant ϵ. We can repeat the algorithm independently t times and failure probability reduces to ϵ^t. This is a sharp fall in the failure probability (exponential in t).

What if the randomized algorithm had two sided error?

Exercise 17. 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 $1/2 + \epsilon$, we can make the success probability as close to one as possible. You want to guess the tool we will use?

Exercise 18. How would you decrease the error of a two sided error algorithm?

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, 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 and running time will only change up to a constant.

3 Assignment

Exercise 19. Let X and Y be two independent random variables. Prove that,

$$E[XY] = E[X]E[Y].$$

Exercise 20. 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 21. 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 22. Read about central limit theorem.

Exercise 23. 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 24. 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. H. Tijms Understanding Probability. *Cambridge University Press*, 2012.
2. D. Stirzaker. Elementary Probability. *Cambridge University Press*, 2003.