Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Prannay Khosla
Instructor: Purushottam Kar
Date: January 10, 2018

LECTURE
3

Concentration Inequalities - Chernoff Bound

1 Introduction

In this lecture we will present a proof of the *Chernoff bound* using the Cramèr-Chernoff method of moments, followed by an application of this inequality to help Alice choose the advantageous coin, our running problem for the previous two lectures. We will also see why Alice cannot do significantly better than the simple algorithm we looked at in the previous lecture.

Before we go any further, here is a nice exercise to help cement our concepts of independence and its use in proving tail bounds. Recall the notions of pairwise and mutually independent events, as well as pairwise and mutually independent[1] random variables from the previous lecture. Also recall that whereas mutual independence implies pairwise independence, the reverse is not true.

Exercise 3.1 (Exponential Tosses)**.** Let $Y_1, \ldots, Y_m$ be m independent copies of a fair Rademacher variable Y i.e. $Y \in \{-1, 1\}$ with $\mathbb{P}[Y_i = 1] = 0.5$. For every subset $S \subseteq [m]$, we define a new random variable

$$X_S = \mathbb{I}\left\{\prod_{i \in S} Y_i = 1\right\},$$

i.e. if the Rademacher variables in the set S have even parity, then the random variable X_S takes value 1 else 0.

1. Find $\mathbb{E}X_S$ for all $S \subseteq [m]$
2. Show that the family of random variables $\{X_s : S \subseteq [m]\}$ is pairwise independent
3. Show that the family of random variables $\{X_s : S \subseteq [m]\}$ is not mutually independent
4. Apply the Chebyshev's inequality to obtain a high confidence left and right tail bound on the quantity $\frac{1}{2^m}\sum_{S \subseteq [m]} X_S$.

For the last part, you will have to estimate the variance of the random variable $\frac{1}{2^m}\sum_{S \subseteq [m]} X_S$. Recall that Chebyshev's inequality may be applied even when random variables are merely pairwise independent and not mutually independent.

[1] When one says that a mutually of events or random variables is *independent* without any further qualification on the term, one usually means that the collection is mutually independent.

2 The Chernoff Bound

The Chernoff bound is a powerful and popularly used result on the concentration on the sums of (mutually) independent Bernoulli random variables. Although the result holds even if the Bernoulli random variables are not identical (i.e. they have different biases), we will study the result in a simpler setting where the random variables are independent and identical copies of the same (unknown) Bernoulli variable.

Theorem 3.1 (Chernoff Bound)**.** Let $X_1, \ldots, X_n$ be n independent copies of a Bernolli random variable X with bias p and denote $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then the following hold for any $\epsilon \in (0, 1)$:

1. Right/Upper Tail Bound: $\mathbb{P}\left[\bar{X}_n > (1+\epsilon)\cdot p\right] \leq \exp\left(\frac{-np\cdot\epsilon^2}{3}\right)$
2. Left/Lower Tail Bound: $\mathbb{P}\left[\bar{X}_n < (1-\epsilon)\cdot p\right] \leq \exp\left(\frac{-np\cdot\epsilon^2}{2}\right)$

To appreciate the above result, let us introduce some popular terminology

Definition 3.1 (Accuracy/Deviation)**.** In the above result, the parameter ϵ quantifies the degree of accuracy claimed by the bound. In previous results, such as the Markov's and Chebyshev's inequalities, the accuracies were quantified using a *deviation* term denoted by the variable t. All concentration inequalities use some such parameter to quantify how closely is concentration being claimed.

Definition 3.2 (Confidence)**.** The first and second statements in the Chernoff Bound are being claimed with an upper bound on the (failure) probability of $\exp(np\epsilon^2/3)$ and $\exp(np\epsilon^2/2)$ respectively. This failure probability is often denoted by the Greek letter δ and is called the *confidence* parameter of the concentration bound. A bound that assures small δ is said to be a *high-confidence* bound, i.e. the probability of the concentration result not holding true as claimed is very small[2].

Definition 3.3 (Left/Right Tail)**.** Concentration inequalities can guarantee that a certain random variable will not take a value that is too large, or too small, compared to a certain value (usually the expected value of that random variable). The former is called a *right tail* bound and the latter is called a *left tail* bound. The Chernoff bound does this for the random variable $\bar{X}_n$. The left and right tails are often called the lower and upper tails respectively. The bounds offered by the Chernoff bound are not the same for the left and the right tail and is called an *asymmetric tail* bound. The Chebyshev's inequality, on the other hand, was a *symmetric* tail bound since it claimed a symmetric deviation probability for the left and the right tail

$$\mathbb{P}\left[|X - \mathbb{E}X| \geq t\right] \leq \frac{\sigma^2}{t^2}$$

Definition 3.4 (Sample Complexity)**.** For concentration inequalities that guarantee the concentration of averages of (often identical but independent) random variables, such as the Chernoff bound which considers n copies $X_1, \ldots, X_n$, the number of copies of the random variables used in the average is often called the *sample complexity* of the bound since this is the number of times an experiment has to be repeated to obtain said independent and identical copies of the random variable. The letter n is often used to denote the sample complexity.

[2]The Chernoff bound is said to offer an *exponential* confidence bound since its failure probability falls of exponentially in terms of the deviation. Markov's and Chebyshev's inequalities offer only *polynomial* confidence since their failure probabilities merely go down as $1/t$ and $1/t^2$ respectively, in terms of the deviation

An admirable goal is to offer high accuracy bounds (for left and right tail) with high confidence, all the while using a small sample complexity. Thus, our focus of attention is on three parameters (n, ϵ, δ) and we wish to prove bounds which offer small values of all three, and that too simultaneously. Unfortunately, this is not possible in general[3].

One often has anti-concentration bounds which prevent all three quantities to be lowered simultaneously. This sets off a sort of tradeoff between the three quantities: to achieve high confidence and small deviation, one invests in larger sample sizes, or else makes peace with mediocre confidence and not-so-small deviation, if constrained by a budget on sample complexity.

Proof. (of Chernoff bound). We will prove the left tail bound by more or less following the Cramér-Chernoff protocol, and the upper tail bound can be proved similarly with the essentially same proof. We have the following series of elementary equalities.

$$\begin{aligned}
\mathbb{P}\left[\bar{X}_n < (1-\epsilon)\cdot p\right] &\overset{(a)}{=} \mathbb{P}\left[-\bar{X}_n > -(1-\epsilon)\cdot p\right] \\
&\overset{(b)}{=} \mathbb{P}\left[-s\bar{X}_n > -s(1-\epsilon)\cdot p\right] \\
&\overset{(c)}{=} \mathbb{P}\left[\exp(-s\bar{X}_n) > \exp(-s(1-\epsilon)\cdot p)\right] \\
&\overset{(d)}{\leq} \frac{\mathbb{E}\exp(-s\bar{X}_n)}{\exp(-s(1-\epsilon)\cdot p)},
\end{aligned}$$

where step (a) was performed in anticipation of an application of the Markov's inequality which only gives a right tail bound. This trick is popular in converting left tail bounds to look like right tail bounds. Step (b) holds for any positive scale parameter s (and we promise that we will never set s to a negative quantity later). Step (c) uses the fact that $\exp(\cdot) \uparrow$ as well as that the function is positive-valued everywhere, and step (d) applies the Markov's inequality, having satisfied all its preconditions in the previous steps.

Before we apply the Cramèr transform, we need to obtain a simplified upper bound on the MGF term $\mathbb{E}\exp(-s\bar{X}_n)$ for which we use the independence of the random variables $X_1, \ldots, X_n$. Note that this means that the Chernoff bound will not hold unless the collection is anything less than n-wise independent (i.e. mutually independent). Denote $\frac{s}{n} = t$ to avoid clutter.

$$\mathbb{E}\exp(-s\bar{X}_n) = \mathbb{E}\exp\left(-t\sum_{i=1}^{n} X_i\right) = \mathbb{E}\prod_{i=1}^{n}\exp\left(-tX_i\right) = \prod_{i=1}^{n}\mathbb{E}\exp\left(-tX_i\right) = \left(\mathbb{E}\exp\left(-tX\right)\right)^n.$$

The third equality holds due to independence and the fourth due to the random variables being identical[4]. Since X is a Bernoulli variable, we know exactly how this quantity behaves.

$$\mathbb{E}\exp\left(-tX\right) = p\cdot\exp\left(-t\cdot 1\right) + (1-p)\cdot\exp\left(-t\cdot 0\right) = 1 + p\left(e^{-t}-1\right)$$

We now use a very popular universal inequality $1 + x \leq e^x$.

$$\mathbb{E}\exp\left(-tX\right) \leq \exp\left(p\left(e^{-t}-1\right)\right)$$

The above gives us (using $s = tn$)

$$\mathbb{P}\left[\bar{X}_n < (1-\epsilon)\cdot p\right] \leq \left(\frac{\exp\left(p\left(e^{-t}-1\right)\right)}{\exp(-t(1-\epsilon)\cdot p)}\right)^n = \left(\exp(p(e^{-t}-1+t(1-\epsilon))\right)^n$$

[3]There may be cooked-up special cases where this is possible, but in any non trivial collection of random variables, such a feat is usually prohibited by information-theoretic lower bounds.

[4]Any temptations to apply results such as the Jensen's inequality to simplify things further, must be resisted at this point. If we do so, we will get $\left(\mathbb{E}\exp\left(-tX\right)\right)^n \leq \mathbb{E}\left[\left(\exp\left(-tX\right)\right)^n\right] = \mathbb{E}\exp(-sX)$ and thus lose any advantage having a large number of samples n could have given us, since now the bound is independent of n.

At this point we need to apply the Cramèr transform. Notice that the parameter t is the new scale parameter for us now. Using first order optimality to set $t = -\log(1-\epsilon)$ (note that it is indeed a positive value since $\epsilon \in (0,1)$ which is important since we promised we would set s to a positive value). This gives us the following result:

$$\mathbb{P}\left[\bar{X}_n < (1-\epsilon)\cdot p\right] \leq \exp(-np(\epsilon + (1-\epsilon)\log(1-\epsilon)))$$

The proof concludes with the observation that for all $\epsilon \in (0,1)$, the Taylor's expansion of the function $\log(1-\epsilon)$ tells us that $\log(1-\epsilon) > -\epsilon - \frac{\epsilon^2}{2}$. This gives us

$$\epsilon + (1-\epsilon)\log(1-\epsilon) > \epsilon + (1-\epsilon)\left(-\epsilon - \frac{\epsilon^2}{2}\right) = \frac{\epsilon^2}{2} + \frac{\epsilon^3}{2} > \frac{\epsilon^2}{2}$$

The right tail bound is left as an exercise. The reader may refer to Goemans and Markov (2015) for the complete proof, which essentially only differs in the last step where we approximate the $\log x$ differently to get a stronger bound. □

Remark 3.1. There exist more precise variants of the Chernoff's bound, notably in terms of the KL-divergence between the distributions induced by the actual coin with bias p and a coin with imprecisely estimated bias $p \pm \epsilon$. These variants offer confidence terms of the form $\exp\left(-KL(p+\epsilon\|p)\cdot n\right)$ and $\exp\left(-KL(p-\epsilon\|p)\cdot n\right)$ for the right and left tails respectively.

We can now start analyzing Alice's algorithm.

3 Alice's Strategy Revisited

Recall that we were interested in helping Alice identify, out of two given coins with biases p_1 and p_2, the coin with the smaller bias as it would give her an edge in a suspense-filled game she and Bob were playing where they would toss a coin repeatedly and upon every heads outcome Alice would pay ₹1 to Bob and upon every tails outcome, receive ₹1 from Bob. We had designed the following strategy for Alice.

Algorithm 1: A Naive Algorithm

1: Toss C_1 n times
2: Let $n_1 =$ # Heads // `Number of Heads for` C_1
3: Toss C_2 n times
4: Let $n_2 =$ # Heads // `Number of Heads for` C_2
5: Let $\hat{p}_1 = \frac{n_1}{n}$, $\hat{p}_2 = \frac{n_2}{n}$
6: Choose C_1 if $\hat{p}_1 \leq \hat{p}_2$, else choose C_2

The above strategy requires a parameter n to be specified. We first assume that Alice wishes to identify the more advantageous coin with probability at least $1-\delta$ and that she also knows $|p_1 - p_2|$. Then, if she sets $\epsilon_0 = \frac{|p_1 - p_2|}{2}$ and sets $n = \frac{3}{\epsilon_0^2}\log\frac{2}{\delta}$, then the following result holds.

Theorem 3.2. If Alice knows $|p_1 - p_2|$ and sets n as shown above, then with probability at least $1-\delta$, she will identify the advantageous coin.

Proof. Without loss of generality, assume $p_1 < p_2$ so that C_1 is advantageous. Applying the Chernoff Bound (right tail) to the tosses of coin C_1 with the accuracy parameter set to $\epsilon = \frac{\epsilon_0}{p_1}$ tells us that

$$\mathbb{P}\left[\hat{p}_1 > p_1 + \epsilon_0\right] \leq \exp\left(\frac{-n\epsilon_0^2}{3p_1}\right) \leq \exp\left(\frac{-n\epsilon_0^2}{3}\right),$$

since $p_1 \leq 1$. Applying the Chernoff Bound (left tail) to the tosses of coin C_2 with the accuracy parameter set to $\epsilon = \frac{\epsilon_0}{p_2}$ tells us that

$$\mathbb{P}\left[\hat{p}_2 < p_2 - \epsilon_0\right] \leq \exp\left(\frac{-n\epsilon_0^2}{2p_2}\right) \leq \exp\left(\frac{-n\epsilon_0^2}{3}\right).$$

Since we set $\epsilon_0 = \frac{|p_1 - p_2|}{2}$, we know that

$$\{\hat{p}_1 < p_1 + \epsilon_0\} \cap \{\hat{p}_2 > p_2 - \epsilon_0\} \Rightarrow \{\hat{p}_1 < \hat{p}_2\}$$

Thus, the contrapositive tells us that

$$\begin{aligned}
\mathbb{P}\left[\hat{p}_1 \not< \hat{p}_1\right] &\leq \mathbb{P}\left[\{\hat{p}_1 > p_1 + \epsilon_0\} \cup \{\hat{p}_2 < p_2 - \epsilon_0\}\right] \\
&\leq \mathbb{P}\left[\hat{p}_1 > p_1 + \epsilon_0\right] + \mathbb{P}\left[\hat{p}_2 < p_2 - \epsilon_0\right] \\
&\leq 2\exp\left(\frac{-n\epsilon_0^2}{3}\right) = \delta,
\end{aligned}$$

since we set $n = \frac{3}{\epsilon_0^2}\log\frac{2}{\delta}$. In the second step above, we used the union bound. Thus, Alice fails to identify C_1 as the advatageous coin with probability at most δ. □

Exercise 3.2. What should Alice do if she does not know $|p_1 - p_2|$? Show that if Alice keeps tossing long enough so that $|p_1 - p_2| \leq 2\sqrt{\frac{3\log 2/\delta}{n}}$, even then she will be able to recover the advantageous coin with probability at least $1 - \delta$.

4 Can Alice do any Better?

The above strategy is quite simple and the proof that Alice will recover the advantageous coin, is also quite straightforward. However, this begs the question whether this is the only/best possible strategy. While this is not the only strategy, and there indeed exist online strategies that can allow Alice to identify the advantageous coin with a lesser number of tosses, no strategy can allow Alice to discover the advantageous coin with an *asymptotically* smaller number of tosses.

Suppose $|p_1 - p_2| \triangleq \epsilon_0$. Then the strategy above requires Alice to perform $\tilde{\Theta}\left(\frac{1}{\epsilon_0^2}\right)$ coin tosses. This is asymptotically optimal, i.e. no strategy can guarantee successful recovery with, say $\mathcal{O}\left(\frac{1}{\epsilon_0}\right)$ coin tosses. This is because with such small number of coin tosses, the two coins look near-identical and it becomes impossible to make a high confidence claim about which coin is advantageous.

Notice that there do exist some special cases where Alice can get away with far fewer coin tosses. For instance, if she knows that $|p_1 - p_2| = 1$ i.e. one of the coins always lands heads and the other always tails, then she can just toss each coin once, i.e., $n = 1$ and know which is which. However, in general, there is no way to identify the more advantageous coin with an asymptotically lesser number of coin tosses.

Showing such a lower bound or impossibility result rigorously will require use of several tools we have not introduced. However, the following argument attempts to give a hopefully convincing argument for the same. Note that the following argument is not a proper lower bound. Recall the lower bound of Canetti et al. (1995) which shows that the Chernoff's bound is tight and cannot be improved, save constant factors. Specifically it is shown that, if $\hat{p}$ is the empirical bias of a coin with true bias p, after n tosses, then we must have

$$\mathbb{P}\left[|\hat{p} - p| > \epsilon \cdot p\right] \geq \frac{1}{8e\sqrt{\pi}}\exp(-4n\epsilon^2)$$

We will use the above lower bound to show an effective argument that Alice could not have done much better than tossing the coins $n = \frac{3}{\epsilon_0^2} \log \frac{2}{\delta}$ times each. At a high level, what we will do is the following, we will construct a virtual coin, each of whose tosses can be simulated by $\mathcal{O}(1)$ tosses of the two coins C_1 and C_2.

However, we will show that the bias of this virtual coin is very close to 0.5. We will also show that figuring out whether the bias of this virutal coin is greater or lesser than 0.5 is equivalent to whether C_1 or C_2 is the advantageous coin. Since Canetti et al. (1995) assure us that unless the virtual coin is tossed a certain number of times, its bias cannot be estimated accurately, we will deduce that it should not be possible for Alice to do much better either.

4.1 Rejection Sampling

We will construct a virtual coin D by tossing C_1 and C_2 once each. Depending on the outcomes, we will decide on an outcome for the virtual coin $\tilde{C}$. The following table summarizes this – in table $\perp$ indicates a failed experiment or *rejection* and in this case we repeat the experiment (toss C_1 and C_2 once each) again.

C_1	C_2	$\tilde{C}$
H	T	H
T	H	T
H	H	$\perp$
T	T	$\perp$

Such a sampling strategy, as we used to design the virtual coin $\tilde{C}$ is called a *rejection sampling* strategy since there are situations where the sampler rejects and requests a re-sampling. It is easy to show the following results for the above sampler.

1. $\mathbb{P}\left[\tilde{C} = 1\right] =: q > 0.5 \Leftrightarrow p_1 > p_2$
2. If $p_1.p_2 \in \left[\frac{1}{4}, \frac{3}{4}\right]$ then $\left|q - \frac{1}{2}\right| \leq \frac{8}{3} \cdot |p_1 - p_2|$

The first result ensures that the task of figuring out whether the virtual coin is biased towards heads or tails is equivalent to discovering the advantageous coin. The second result indicates that the closer p_1 and p_2 get to each other (arguably making Alice's job of distinguishing them harder) the closer the bias of $\tilde{C}$ i.e. q gets to 0.5 and the harder it becomes to figure out if its bias is above 0.5 or below 0.5.

Now, Canetti et al's result applies to $\tilde{C}$ since it is a Boolean random variable and shows us that estimating its bias to an accuracy ξ, i.e. ensure $|\hat{q} - q| \leq \xi$, with confidence δ is impossible unless the virtual coin is tossed at least $n \geq \frac{1}{4\xi^2} \log \frac{1}{8e\sqrt{\pi}\delta}$ times. However, since we need $\xi \leq \frac{8}{3} \cdot |p_1 - p_2|$ to find out whether the bias of the virtual coin is smaller than 0.5 or not (since $\left|q - \frac{1}{2}\right| \leq \frac{8}{3} \cdot |p_1 - p_2|$), we must need

$$n \geq \frac{9}{256(p_1 - p_2)^2} \log \frac{1}{8e\sqrt{\pi}\delta}$$

However, since each toss of the virtual coin requires one toss each of the coins C_1 and C_2, we conclude that Alice should require at least $n \geq \Omega\left(\frac{1}{(p_1-p_2)^2}\right)$ coins tosses to ensure that she chooses the advantageous coin. Since our strategy required Alice to toss the coins $n = \frac{12}{(p_1-p_2)^2} \log \frac{2}{\delta}$ times, this shows that Alice's strategy is optional upto constants. However, we note that what we have shown is not a rigorous lower bound on Alice's sampling complexity. Among other reasons, this is because we have not shown that tossing coins C_1 and C_2 a few times each is the only way to simulate the virtual coin $\tilde{C}$.

References

Ran Canetti, Guy Even, and Oded Goldreich. Lower bounds for sampling algorithms for estimating the average. *Information Processing Letters*, 53(1):17–25, 1995.

Michel Goemans and Apply Markov. Chernoff bounds , and some applications. pages 1–6, 2015.