Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Nitish Joshi
Instructor: Purushottam Kar
Date: January 8, 2018

LECTURE

2

The Coin-toss Game

1 Introduction

We concluded the last lecture with a setting where Alice and Bob are playing a game of tossing a coin repeatedly and exchanging money depending on the outcome. Alice needs to choose which of the two available coins is to be used to play that game. In general there can be several coins to choose from but let us assume there are only two. Coins C_1 and C_2 have biases p_1 and p_2 respectively, which are not known to either of the two players beforehand. Note p_i denotes the probability of getting heads using coin C_i. As mentioned before, when playing the game, each time the outcome is heads, Alice pays ₹1 to Bob. If the outcome is tails, Bob pays ₹1 to Alice. Alice's goal being to maximize her winnings, the question that arises is how should Alice choose a coin in order to achieve her goal? It is obvious that it is advantageous for Alice to choose the coin with the smaller bias but how to identify which coin has small bias is not clear. In this lecture, we will try to answer this question. Note that it is better for Alice to choose a coin with a smaller bias. Even a small difference in the bias can cause a huge difference in the final payout if the game is played a large number of times.

2 Alice's strategy

One way Alice can choose a coin out of the two is by following the algorithm given below.

```
Algorithm 1: A Naive Algorithm

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

Intuitively, we expect that $\hat{p}_1$ and $\hat{p}_2$ would act as estimates of the biases p_1 and p_2 respectively. Note that the only variable in the algorithm above is 'n', the number of times we toss each coin. As per the weak law of large numbers, the more number of times we toss the coins, the more we are sure that $\hat{p}_1$ and $\hat{p}_2$ are close to p_1 and p_2 respectively. But since we don't want to toss the coin a large number of times, we would like to know how many times is enough and how confident we are with our final choice. More technically speaking, we would like to know the probability 'δ' of making an error 'ϵ' on the estimation of p_i given that we tossed the coin

n times. Ideally, we want to keep δ, ϵ and n small, but as we will see in section 4 and future lectures, there is always a tradeoff between the three terms – in general, all three cannot be reduced simultaneously. Typically one has to go for large samples i.e. large n if one wants high accuracy and confidence i.e. small ϵ and δ.

3 Analysing Alice's strategy

Without loss of generality, lets assume $p_1 < p_2$. Note that our algorithm fails to identify the advantageous coin precisely when $\hat{p}_1 > \hat{p}_2$. A simple step of adding and subtracting p_1 and p_2 on both sides tells us that this will happen if and only if

$$(\hat{p}_1 - p_1) + (p_2 - \hat{p}_2) > p_2 - p_1$$

Thus, Alice can ensure success by simply ensuring that

$$(\hat{p}_1 - p_1) + (p_2 - \hat{p}_2) \leq p_2 - p_1$$

Now, there are a lot of nice settings in which the above would happen. For instance, if we had

$$|\hat{p}_1 - p_1| \leq \left|\frac{p_1 - p_2}{2}\right|$$
$$|\hat{p}_2 - p_2| \leq \left|\frac{p_1 - p_2}{2}\right|,$$

then we would have ensured

$$|\hat{p}_1 - p_1| + |p_2 - \hat{p}_2| \leq p_2 - p_1,$$

which clearly ensures success. However, there may be quite bizarre situations where Alice still succeeds. For example, we may have $p_1 = 0.8, p_2 = 0.9$ but $\hat{p}_1 = 0.1, \hat{p}_2 = 0.2$. In this case, our estimates are way off but Alice still manages to succeed. Understandably, Alice is confused as to which route should she choose to ensure success.

Fortunately, Alice need not worry. We will soon see that the number of times n Alice has to toss the coins to ensure the condition $|\hat{p}_i - p_i| \leq \left|\frac{p_1 - p_2}{2}\right|$ for $i = 1, 2$ is close (within constants) to the number of times she would have to toss the coins to ensure success using any other method as well. This result is quite interesting and proceeds by showing a result known as a *sample lower bound.* Basically, there is a lower bound on the number of times one must toss before one can distinguish between two coins with bias p and $p + \epsilon$.

If we believe the above, then this gives Alice a concrete goal to target (toss each coin enough number of times so that the inequalities $|\hat{p}_i - p_i| \leq \left|\frac{p_1 - p_2}{2}\right|$ hold for $i = 1, 2$. In order to ensure such results, we need to look at concentration inequalities.

4 Concentration Inequalities

Concentration Inequalities quantify the random fluctuations of random variables (or functions of random variables) by bounding the probability of the deviation of the random variables from some value (typically the expected value of the random variable) Boucheron et al. (2013). These inequalities fall in a loose hierarchy in terms of how generally do they apply and how strong a concentration do they guarantee Wikipedia (2017). We begin with one of the most basic inequalities, namely the Markov inequality.

4.1 Markov's Inequality

Markov's inequality is a basic and widely used concentration inequality. As we will see in the next subsection and future lectures, this inequality is used to derive other more sophisticated inequalities. Markov's inequality can only be applied if the following conditions are met:

- $X \geq 0$ a.s (almost surely)[1],
- Finite expectation $\mathbb{E}X < \infty$.

Notice that these conditions are very mild and one of the reasons why Markov's inequality forms the basis for proving other results.

Theorem 2.1 (Markov's Inequality). For any real-valued random variable X that takes non-negative values almost surely and has a finite expectation $\mathbb{E}X$, we have, for every $t > 0$,

$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}X}{t}$$

Proof. Since $\mathbb{I}\left\{X < t\right\} + \mathbb{I}\left\{X \geq t\right\} = 1$, we have

$$\begin{aligned} X &= X\left(\mathbb{I}\left\{X < t\right\} + \mathbb{I}\left\{X \geq t\right\}\right) \\ &= X \cdot \mathbb{I}\left\{X < t\right\} + X \cdot \mathbb{I}\left\{X \geq t\right\} \\ &\geq X \cdot \mathbb{I}\left\{X < t\right\} + t \cdot \mathbb{I}\left\{X \geq t\right\} \end{aligned}$$

Now, since $X \geq 0$, a.s. we have $\mathbb{E}\left[X \cdot \mathbb{I}\left\{X < t\right\}\right] \geq 0$. Taking expectations on both sides, using the fact that if $A \geq B$, then $\mathbb{E}A \geq \mathbb{E}B$, applying linearity of expectation, and using the fact that $\mathbb{E}\left[\mathbb{I}\left\{X \geq t\right\}\right] = \mathbb{P}\left[X \geq t\right]$ gives us

$$\mathbb{E}\left[X\right] \geq t \cdot \mathbb{P}\left[X \geq t\right] \qquad \square$$

Note that this inequality gives us a rather loose bound and the decay rate for the confidence bound is proportional to t, the deviation. One reason for this is that this bound merely looks at the first moment of the random variable. Following results will show stronger decay rates, indeed by considering higher moments of random variables.

Also notice that this inequality starts giving us non-trivial probability bounds only on the *right tail* of the distribution i.e. when $t > \mathbb{E}X$. For $t < \mathbb{E}X$, the right hand side of the inequality is a quantity greater than 1 and since all probability values are anyway upper bounded by 1, such results are vacuous. However, for say $t = 2\mathbb{E}X$, we get that $\mathbb{P}\left[X \geq 2\mathbb{E}X\right] \leq 0.5$ which is definitely a very non-trivial and useful statement.

4.2 Chebyshev's (Tchebysheff's) Inequality

This inequality is stronger than Markov's and can handle random variables that take positive and negative values. To apply this inequality, a random variable X must satisfy the following

- Finite expectation $\mathbb{E}X < \infty$,
- Finite variance $\sigma^2 = \mathrm{Var}\left[X\right] := \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] < \infty$.

[1] The qualifier "almost surely" is used to describe a property that is violated with probability 0. For example, if $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space underlying the random variable X, then the statement $X \geq 0$ a.s. implies that $\mathbb{P}\left[\omega \in \Omega : X(\omega) < 0\right] = 0$. Note that we must have $\{\omega \in \Omega : X(\omega) < 0\} \in \mathcal{F}$ due to properties of random variables and σ-algebras.

Theorem 2.2 (Chebyshev's Inequality)**.** For any real-valued random variable X with a finite expectation $\mathbb{E}X$ and finite variance σ^2, we have, for every $t > 0$,

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

Note that unlike Markov's inequality, this inequality requires the variance of X to be finite as well. However, it offers a much faster rate of decay, that of t^2 for a deviation of t, as well.

Exercise 2.1. Prove Chebyshev's inequality. *Hint*: apply Markov's inequality to the random variable $Y = (X - \mathbb{E}X)^2$.

Exercise 2.2. Let X be a Bernoulli random variable with bias p i.e. $X \in \{0, 1\}$ and $\mathbb{E}X = p$. Also let $X_1, \ldots, X_n$ be n identical and independent copies of X (think of n i.i.d. coin tosses) and let $\hat{p} = \frac{1}{n} X_i$. Use the Markov's and Chebyshev's inequalities to obtain a probabilistic bound on the deviation $|p - \hat{p}|$. Your bound should be of the form $\mathbb{P}\left[|p - \hat{p}| \geq \epsilon\right] \leq \delta$

Exercise 2.3. Show that Markov's inequality may be applied to the above problem even if the variables X_i are not independent at all! What happens to $\mathbb{E}\left[\hat{p}\right]$? What bound can you show for the deviation $|p - \hat{p}|$ if the random variables $X_1, \ldots, X_n$ are not guaranteed to be independent? Is it worse than what you could show when the variables were independent?

Exercise 2.4. Show that Chebyshev's inequality may be applied to the above problem of bounding the deviation $|p - \hat{p}|$ even if the variables X_i are merely pairwise independent and not mutually independent. What happens to $E\hat{p}$ and $\text{Var}\left[\hat{p}\right]$? Is the new bound worse than what you could show when the variables were mutually independent?

Exercise 2.5. Suppose we have a random variable X that has a finite expectation as well as a finite fourth central moment (related to the so-called *kurtosis* of the random variable) i.e. suppose that $\kappa \triangleq \mathbb{E}\left[X - \mathbb{E}X\right]^4 < \infty$ as well[2]. Apply Markovs inequality to the random variable $Z = (X - \mathbb{E}X)^4$ to obtain a new concentration inequality. Is it more powerful than Markov and Chebyshev?

Exercise 2.6. Use the new concentration inequality you have derived above to obtain a new bound on the deviation $|p - \hat{p}|$ assuming X_i are mutually independent. What is the value of κ for the random variable $\hat{p}$? Do the random variables X_i need to be mutually independent in order for you to apply your new concentration inequality?

Exercise 2.7. Can you keep going like this and consider the 6-th, 8-th moments and so on and keep getting more and more powerful inequalities? Is there a limit to this? What if n is even (recall that n is the number of copies of X we have) and we look at the n-th moment, assuming that the copies are mutually independent?

The above exercises should have shown you that more and more powerful concentration inequalities can be derived if we consider higher and higher moments of the random variables in question. However, that implicitly restricts the application of these techniques to random variables for which these moments are indeed finite. Below we take this process to the limit and derive a very powerful concentration inequality in the process.

[2] The decoration $\triangle$ on top of the equality sign denotes a definition. Other popular notations for denoting a definition include $:=$ and $\overset{\text{def}}{=}$.

4.3 Chernoff's Bound

Below we will prove the Cheroff's bound for Bernoulli random variables. Let $X_1 ... X_n$ be i.i.d. copies of a Bernoulli random variable X with bias p i.e. $X \in \{0, 1\}$ and $\mathbb{E}X = p$. This exactly models the n coin tosses Alice would be making in her quest to discover the advantageous coin. Also let $\bar{X}_n := \hat{p} = \frac{1}{n} X_i$ denote the empirical mean of the random variable (in other words the fraction of heads we get in the tosses). Our goal in this section would be to prove the following result:

Theorem 2.3 (Chernoff's Bound)**.** Let $X_1 ... X_n$ be i.i.d. copies of a Bernoulli random variable X with bias p. Then for any $\epsilon \in (0, 1)$, we have

$$\mathbb{P}\left[\bar{X}_n \geq (1+\epsilon)p\right] \leq \exp\left(\frac{-np \cdot \epsilon^2}{3}\right)$$

$$\mathbb{P}\left[\bar{X}_n \leq (1-\epsilon)p\right] \leq \exp\left(\frac{-np \cdot \epsilon^2}{2}\right)$$

Combining the two inequalities above and applying a union bound gives us

$$\mathbb{P}\left[\left|\bar{X}_n - p\right| \geq \epsilon \cdot p\right] \leq 2\exp\left(\frac{-np \cdot \epsilon^2}{3}\right)$$

Note that the Chernoff bound offers an exponential decay in terms of the deviation. However, the result, at least in the form given above, holds only for Bernoulli random variables. It also requires the random copies X_is to be mutually independent, a condition required neither by Markov's nor Chebyshev's inequality.

Proving the above result will require us to analyze higher moments of random variables. Rather than fiddling with small finite moments as we did in the exercises above, we jump straight ahead to the concept of a *moment generating function* that encodes all finite moments of a random variable. The full proof of Chernoff's bound will be given in the next lecture.

4.4 Moments

The i^{th} moment of a random variable X is defined as $\mathbb{E}\left[X^i\right]$. The i^{th} *central* moment of X is defined as $\mathbb{E}\left[(X - \mathbb{E}X)^i\right]$. Note that the variance of a random variable is simply its 2^{nd} central moment. The third and fourth central moments of a random variable are related to its *skewness* and *kurtosis* respectively and for unit variance random variables, the third moment is equal to the skewness and the fourth to the kurtosis.

4.5 Moment Generating Function

The moment generating function of a random variable X is defined, for a *scale parameter* s, as

$$M_X(s) \triangleq \mathbb{E}\left[e^{sX}\right] = \sum_{i=1}^{\infty} \frac{s^i \mathbb{E}\left[X^i\right]}{i!}$$

The reason for this name is because if the function takes finite values in an interval around zero then the k^{th} (non-central) moment of X may be found simply by evaluating the k^{th} derivative of the MGF at $s = 0$ i.e.

$$\mathbb{E}\left[X^k\right] = M_X^{(k)}(0) = \left.\frac{d^k M_X(s)}{ds^k}\right|_{s=0}$$

The proof of Chernoff's bound will not only crucially use the MGF of the random variable $\hat{p} = \bar{X}_n$, but also introduce us to a fairly generic and powerful proof technique that we will use to prove other bounds such as Hoeffding's inequality. This proof technique is outlined below for a generic (not necessarily Bernoulli) random variable X.

```
Algorithm 2: The Cramér-Chernoff Method of Moments

Input: Random variable $X$ with finite expectation whose MGF exists
Output: A concentration bound for $X$ around its expectation $\mathbb{E}X$
                                                            //Right Tail
$\mathbb{P}[X > \mathbb{E}X + t] = \mathbb{P}[sX > s(\mathbb{E}X + t)]$ for $s > 0$
$\mathbb{P}[sX > s(\mathbb{E}X + t)] = \mathbb{P}[\exp(sX) > \exp(s(\mathbb{E}X + t))]$ since $\exp(\cdot) \uparrow$
Apply Markov's inequality to get
    $\mathbb{P}[\exp(sX) > \exp(s(\mathbb{E}X + t))] \leq \frac{\mathbb{E}\left[e^{sX}\right]}{\exp(s(\mathbb{E}X + t))} = \frac{M_X(s)}{\exp(s(\mathbb{E}X + t))}$
Obtain a good bound on the MGF $M_X(s)$ and substitute it in RHS
Find a value of $s$ that minimizes RHS (first order optimality), and simplify
//This effectively performs the so-called Cramér transform
                                                            //Left Tail
$\mathbb{P}[X < \mathbb{E}X - t] = \mathbb{P}[-X > -\mathbb{E}X + t] = \mathbb{P}[-sX > -s(\mathbb{E}X - t)]$ for $s > 0$
$\mathbb{P}[-sX > -s(\mathbb{E}X - t)] = \mathbb{P}[\exp(-sX) > \exp(-s(\mathbb{E}X - t))]$ since $\exp(\cdot) \uparrow$
Apply Markov's inequality to get
    $\mathbb{P}[\exp(-sX) > \exp(-s(\mathbb{E}X - t))] \leq \frac{M_X(-s)}{\exp(-s(\mathbb{E}X - t))}$
Proceed as in right tail analysis
```

Note that we are able to apply the Markov's inequality only because the exponential function always takes positive values.

5 Anti-Concentration Inequalities

Concentration inequalities basically give results that show that random variables are highly likely to concentrate around a value, typically the expectation of the random variable itself. Thus, these results are always of the form

$$\mathbb{P}[|X - v| > \epsilon] \leq \delta,$$

where v is typically $\mathbb{E}X$. However, anti concentration inequalities (see Krishnapur, for example) show that there is a limit to this behavior and that beyond a point, random variables will refuse to concentrate. Anti-concentration inequalities come in two flavors.

The first kind of anti-concentration inequalities show that certain random variables refuse to concentrate around their expectation beyond a point. The work of Canetti et al. (1995) gives such a lower bound which actually shows that the result offered by Chernoff's bound is actually tight and cannot be improved, save constant factors. Specifically they show that

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

We will use this inequality in the next lecture to show that Algorithm 1 is pretty much the best Alice can do in terms of choosing the advantageous coin with the least number of trial tosses.

The second kind of anti-concentration inequalities are much more powerful. They show that certain random variables refuse to concentrate *anywhere*, let alone at the expectation. The powerful *central limit theorems* are examples of such results. A simple result of this kind can be derived for the standard Gaussian random variable.

Let $X \sim \mathcal{N}(0,1)$. Then, since the density of this random variable is upper bounded by $\frac{1}{\sqrt{2\pi}}$ everywhere, it is easy to see that $\mathbb{P}\left[X \in [a,b]\right] \leq \frac{|b-a|}{\sqrt{2\pi}}$. This means that the random variable refuses to squeeze into tiny intervals, wherever they might be. In particular, if $|b-a| \leq 2\epsilon$ then the above shows that $\mathbb{P}\left[X \notin [a,b]\right] \geq 1-\epsilon$ i.e. with high probability, the random variable will take values outside this interval. This has the following interesting corollary

$$\sup_{v \in \mathbb{R}} \mathbb{P}\left[X \in [v-\epsilon, v+\epsilon]\right] \leq \epsilon,$$

i.e. the random variable refuses to concentrate around any point $v \in \mathbb{R}$. Another way of writing the above statement is write it in the negative.

$$\inf_{v \in \mathbb{R}} \mathbb{P}\left[X \notin [v-\epsilon, v+\epsilon]\right] \geq 1-\epsilon,$$

which makes an equivalent statement that the best concentration one can hope for around any point is upto a (really small) confidence of ϵ and no more. The random variable escapes every region with high probability.

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

Manjunath Krishnapur. Anti-concentration inequalities. URL `http://math.iisc.ernet.in/~manju/anti-concentration.pdf`. [Online; accessed 12-January-2018].

Wikipedia. Concentration inequality — wikipedia, the free encyclopedia, 2017. URL `https://en.wikipedia.org/w/index.php?title=Concentration_inequality&oldid=793477684`. [Online; accessed 12-January-2018].