Algoritmos e Incerteza **(PUC-Rio INF2979, 2018.2)**

Lecture 09: An Introduction to Concentration Inequalities

09/10/2018

Lecturer: Marco Molinaro *Scribe: Guilherme Bodin*

In probability theory, concentration inequalities provide bounds on how a random variable deviates from some value (e.g. its expected value). This lecture presents Markov, Chebyshev and Chernoff inequalities and discusses their limits and applications. Bernstein inequality and Union Bound will also be presented.

1 Balls and bins

To illustrate concentration inequalities, an application on a classic probability problem, balls and bins, will be presented. The setup is the following, imagine we have m balls and n bins, the balls go to the bins under an uniform distribution.

Figure 1: Balls and bins example for $n = m = 7$

In theory we expect that after the n balls fall into the bins we will have 1 ball in each bin,but this is clearly not true all the time because the reality will suffer a deviation from the mean. In all cases we could ask ourselves some interesting questions on this problem, for example:

- What is the probability that bin number 3 contains 3 balls?
- What is the probability that one bin contains at least 3 balls?

These are the type of questions that we will try to answer using concentration inequalities.

2 Markov's inequality

Theorem 2.1. *(Markov's inequality) If Z is a non-negative random variable with $E(Z) > 0$ and $a > 0$ then the probability that Z is at least a is at most the expected value of Z divided by a:*

$$\mathbf{Pr}(Z \geq a) \leq \frac{E(Z)}{a}$$

Proof. We assume Z is discrete and taking finite values to simplify things

$$\begin{aligned} E(Z) &= \sum_{x \geq 0} x\mathbf{Pr}(Z = x) \\ &\geq \sum_{x \geq a} x\mathbf{Pr}(Z = x) \\ &\geq \sum_{x \geq a} t\mathbf{Pr}(Z = x) \\ &= a\sum_{x \geq a} \mathbf{Pr}(Z = x) \\ &= a\mathbf{Pr}(Z \geq a) \end{aligned}$$

Now we have

$$\begin{aligned} E(Z) &\geq a\mathbf{Pr}(Z \geq a) \\ \mathbf{Pr}(Z \geq a) &\leq \frac{E(Z)}{a} \end{aligned}$$

□

A very common way to use Markov inequality is by substituting $a = tE(Z)$, this gives an insight about the concentration of the distribution in terms of its mean. The theorem guarantees that

$$\mathbf{Pr}(Z \geq tE(Z)) \leq \frac{1}{t}$$

An interesting thing to notice about Markov's inequality is that it is insensitive to variance. Intuitively this means that even though it provides a bound for the concentration of the distribution it doesn't tell us much information about the shape of its pdf. This makes the inequality useful when we know nothing about its variance but weak when we know some information about it.

Let's discuss the bound created by Markov's inequality on the two following distributions:

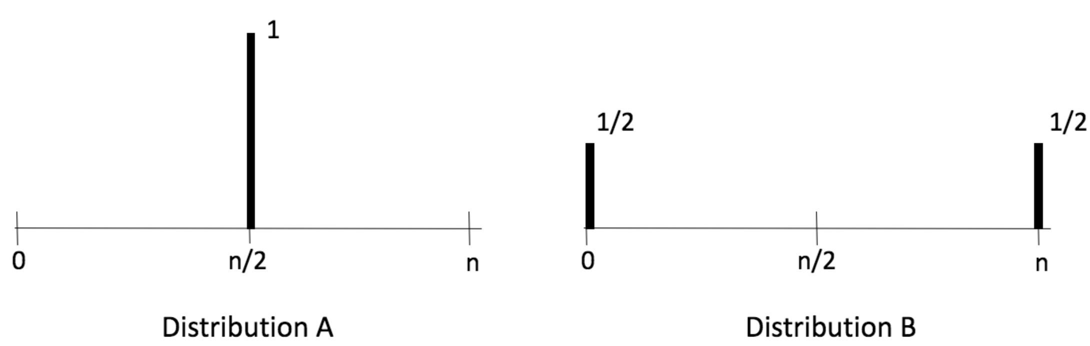

Figure 2: Distributions

If Z is a random variable that follows distribution A we can question ourselves what is the probability that Z is greater than n, $\mathbf{Pr}(Z \geq n)$. using Markov's inequality:

$$\mathbf{Pr}(Z \geq n) \leq \frac{n/2}{n} = \frac{1}{2}$$

Now if Z is a random variable that follows distribution B we can also question ourselves what is the probability that Z is greater than n, $\mathbf{Pr}(Z \geq n)$. using Markov's inequality:

$$\mathbf{Pr}(Z \geq n) \leq \frac{n/2}{n} = \frac{1}{2}$$

Notice that distributions have completely different pdf's but the bound given by Markov's inequality is exactly the same.

3 Chebyshev's inequality

Theorem 3.1. *(Chebyshev's inequality) Let Z be a random variable with finite expected value ($\mu < \infty$) and non-zero variance ($\sigma^2 > 0$). Then for any real number $a > 0$*

$$\mathbf{Pr}(|Z - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Proof. We can prove Chebyshev's inequality by applying Markov's inequality to $Y = |Z - \mu|$ Notice that

$$\mathbf{Pr}(|Z - \mu| \geq a) = \mathbf{Pr}((Z - \mu)^2 \geq a^2)$$

One way of understanding this is by thinking that if the difference between Z and its mean μ is greater that a, than the square of the the difference between Z and its mean μ must be greater that a^2.

We can rewrite our Markov's inequality application as

$$\mathbf{Pr}((Z - \mu)^2 \geq a^2) \leq \frac{E((Z - \mu)^2)}{a^2}$$

and notice that $E((Z - \mu)^2) = \sigma^2$, so

$$\mathbf{Pr}((Z - \mu)^2 \geq a^2) \leq \frac{\sigma^2}{a^2}$$

The last step is to going back from the squared difference of $Z - \mu$ to the absolute value of the difference. This way we can write

$$\mathbf{Pr}(|Z - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

□

A very common way of using Chebyshev's inequality is by substituting $a = t\sigma$ to give insight about how concentrated is the distribution in terms of its standard deviation. The theorem guarantees that

$$\mathbf{Pr}(|Z - \mu| \geq t\sigma) \leq \frac{1}{t^2}$$

Chebyshev's inequality is indeed variant to the variance which is good as it gives us more information about the shape of the pdf. It is stronger in terms of bound tightness than Markov's inequality but also requires more information on the distribution that we are working with.

Let's discuss the bound created by Markov's inequality on the two following distributions:

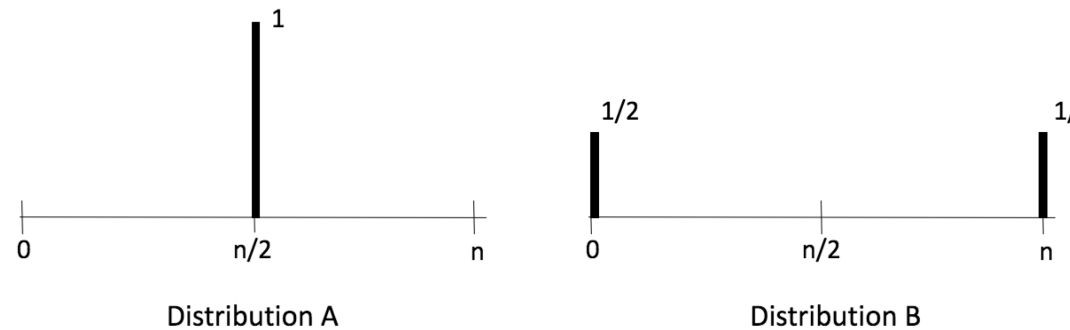

Figure 3: Distributions

If Z is a random variable that follows distribution A we can question ourselves what is the probability that Z is greater than n, $\mathbf{Pr}(Z \geq n)$. To use Chebyshev's inequality it let's think of A as the limiting distribution of putting all mass closer and closer to $\frac{n}{2}$, with $Var(Z) = \epsilon$, $\epsilon \to 0$:

$$\mathbf{Pr}(Z \geq n) = \mathbf{Pr}(Z - \frac{n}{2} \geq \frac{n}{2}) \leq \frac{\epsilon}{n^2/4} = \frac{4\epsilon}{n^2}$$

Now if Z is a random variable that follows distribution B we can also question ourselves what is the probability that Z is greater than n, $\mathbf{Pr}(Z \geq n)$. Using Chebyshev's inequality with $Var(Z) = \frac{n^2}{4}$:

$$\mathbf{Pr}(Z \geq n) = \mathbf{Pr}(Z - \frac{n}{2} \geq \frac{n}{2}) \leq \frac{n^2/4}{n^2/4} = 1$$

Notice that the figures have completely different pdf's and the bound given by Chebyshev's inequality shows that. Even though this case has better bound tightness, Chebyshev isn't tight in some cases, e.g, when we have a random variable that is the sum of various independent random variables.

4 Chernoff's inequality

There are many different forms of Chernoff bounds, each tuned to slightly different assumptions. We will use the one with the statement of the bound for the simple case of a sum of independent

Bernoulli trials, i.e. the case in which each random variable only takes the values 0 or 1

Theorem 4.1. *(Chernoff's inequality) Let $X_1 \dots X_n$ be independent random variables with value in $\{0, 1\}$ and $Z = X_1 + \cdots + X_n$. Then*

$$\mathbf{Pr}(Z \geq E(Z) + t\sqrt{n}) \leq e^{-t^2}.$$

Before starting the proof let's just do a little revision on exponential approximations. The first thing to remember is that the Taylor approximation of e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots;$$

so e^Z captures simultaneously "all moments" of Z (as opposed to Chebychev's inequality that only looked at the second moment). More technically, we will need the following fact:

$$1 + x \leq e^x \leq 1 + x + x^2 \quad \forall x \in [-1, 1]$$

as illustrated in Figure 4

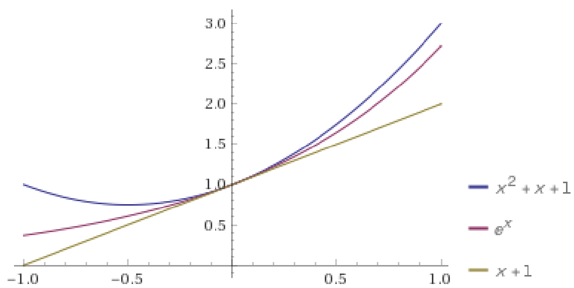

Figure 4: e^x bounds $\forall x \in [-1, 1]$

Proof. The main idea for this proof is to apply Markov's inequality to the exponential moment $e^{\lambda Z}$.

The first thing to do is

$$\mathbf{Pr}(Z - E(Z) \geq t\sqrt{n}) = \mathbf{Pr}(e^{\lambda(Z-E(Z))} \geq e^{\lambda(t\sqrt{n})}) \quad \forall \lambda \in [0, 1]$$

Now applying the Markov's inequality we have that

$$\mathbf{Pr}(e^{\lambda(Z-E(Z))} \geq e^{\lambda(t\sqrt{n})}) \leq \frac{E(e^{\lambda(Z-E(Z))})}{e^{\lambda(t\sqrt{n})}}$$

Notice that by independence:

$$\begin{aligned} e^{\lambda(Z-E(Z))} &= e^{\lambda(X_1-E(X_1))+\cdots+\lambda(X_n-E(X_n))} \\ &= e^{\lambda(X_1-E(X_1))} \times \cdots \times e^{\lambda(X_n-E(X_n))} \\ &= \prod_{i=1}^{n} e^{\lambda(X_i-E(X_i))} \end{aligned}$$

Because $e^x \leq 1 + x + x^2$ and the variance of a Bernoulli distribution (which is our case for $X_1 \ldots X_n$) is given by $p(1-p), p \in [0,1]$ variance of X_1 is at most $\frac{1}{4}$ we have that

$$\begin{aligned} E[e^{\lambda(X_1 - E(X_1))}] &\leq 1 + E[\lambda(X_1 - E(X_1) + \lambda^2(X_1 - E(X_1))^2] \\ &= 1 + E[\lambda(X_1 - E(X_1)] + E[\lambda^2(X_1 - E(X_1))^2] \\ &= 1 + 0 + \frac{\lambda^2}{4} \end{aligned}$$

We can use the approximation $e^x \geq 1 + x$ to deduce

$$\prod_{i=1}^{n} E[e^{\lambda(X_i - E(X_i))}] \leq (1 + \frac{\lambda^2}{4})^n \leq e^{\frac{\lambda^2}{4}n}$$

Now setting $\lambda = \frac{2t}{\sqrt{n}}$ we finally get

$$\mathbf{Pr}(Z \geq E(Z) + t\sqrt{n}) = \mathbf{Pr}(e^{\lambda(Z - E(Z))} \geq e^{\lambda(t\sqrt{n})}) \leq \frac{E(e^{\lambda(Z-E(Z))})}{e^{\lambda(t\sqrt{n})}} \leq \frac{e^{\frac{\lambda^2}{4}n}}{e^{\lambda(t\sqrt{n})}} = e^{-t^2}$$

□

Notice that Chernoff's inequality also works in the other tail

$$\mathbf{Pr}(Z \leq E(Z) - t\sqrt{n}) \leq e^{-t^2}.$$

5 Union bound (Boole's inequality)

The union bound is a very simple yet useful tool in probability theory, the union bound states that

$$\mathbf{Pr}(A \textit{ or } B) = \mathbf{Pr}(A) + \mathbf{Pr}(B) - \mathbf{Pr}(A \cap B) \leq \mathbf{Pr}(A) + \mathbf{Pr}(B).$$

This is very intuitive when looking to Figure 5,

The most powerful part of the union bound is that is has no requirements whether how should be the expected value, the variance or any other possible parameter of a distribution, it works all cases.

6 Applying on the balls and bins problem

The setup is as in the first section. We have m balls and n bins, the balls go to the bins under an uniform distribution.

- $X_{i,j} = \begin{cases} 1 & \text{if ball } i \text{ falls into bin } j, \\ 0 & \text{otherwise} \end{cases}$

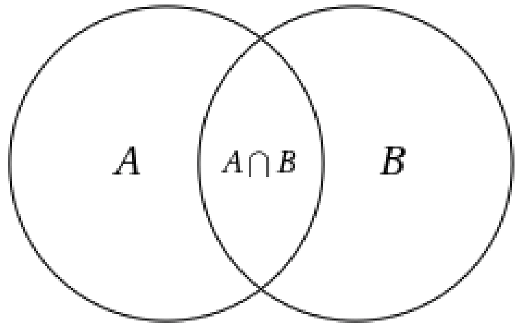

Figure 5: Venn diagram

- $E(X_{i,j}) = \frac{1}{n}$
- $Var(X_{i,j}) = \frac{1}{n} + \frac{1}{n^2}$
- $X_{1,j} + X_{2,j} + \cdots + X_{n,j} =$ number of balls in bin j

In each subsection we will try to create a bound to the question: For m balls and n bins what is the probability that one bin contains k balls?

Before starting it is necessary to use union bound to model properly the problem, one thing you might not have noticed is that it exists a dependency between variables as $X_{1,1} = 1 \rightarrow X_{1,2} = 0$. Let's say B_j it's a bad event, it is the case bad event is that bin j has more than k balls $\mathbf{Pr}(X_{1,j} + X_{2,j} + \cdots + X_{n,j} \geq k)$. Now we can use the union bound to say that

$$\mathbf{Pr}(B_1 \text{ or } B_2 \text{ or} \ldots \text{ or } B_n) \leq \mathbf{Pr}(B_1) + \mathbf{Pr}(B_2) + \cdots + \mathbf{Pr}(B_n)$$

Now we only need to bound the probability $\mathbf{Pr}(B_i)$ of having too many ball in bin i using concentration inequalities. [To be continued]