Lecture 6: Concentration Inequalities - Introduction to Martingales

John Sylvester Nicolás Rivera Luca Zanetti Thomas Sauerwald

Lent 2019

UNIVERSITY OF CAMBRIDGE

Outline

More Chernoff Bounds

Conditional Expectation

Chernoff Bounds

Remember the Chernoff Bounds from the previous lecture..

Chernoff Bounds: upper tails

Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter p_i. Let $X = X_1 + \ldots + X_n$ and $\mu = \mathbf{E}[X] = \sum p_i$. Then, for any $\delta > 0$ it holds that

$$\mathbf{P}[X \geq (1+\delta)\mu] \leq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}.$$

and for $t > \mu$ it holds that

$$\mathbf{P}[X \geq t] \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^{t},$$

Chernoff Bounds

.. and the lower tails..

Chernoff Bounds: Lower Tails

Suppose $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter p_i. Let $X = X_1 + \ldots + X_n$ and $\mu = \mathbf{E}[\, X \,] = \sum p_i$. Then, for any $\delta > 0$ it holds that

$$\mathbf{P}[\, X \leq (1-\delta)\mu \,] \leq \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{\mu}.$$

and for any $t < \mu$

$$\mathbf{P}[\, X \leq t \,] \leq e^{-\mu} \left(\frac{e\mu}{t} \right)^{t}.$$

Chernoff Bounds

..and the nicer version!

Nicer Chernoff Bounds

Suppose $X_1, \dots, X_n$ are independent Bernoulli random variables with parameter p_i. Let $X = X_1 + \dots + X_n$ and $\mu = \mathbf{E}[\,X\,] = \sum p_i$. Then,

- For all $t > 0$,

$$\mathbf{P}[\,X \geq \mathbf{E}[\,X\,] + t\,] \leq e^{-2t^2/n}$$

$$\mathbf{P}[\,X \leq \mathbf{E}[\,X\,] - t\,] \leq e^{-2t^2/n}$$

- For $0 < \delta < 1$,

$$\mathbf{P}[\,X \geq (1+\delta)\mathbf{E}[\,X\,]\,] \leq \exp\left(-\frac{\delta^2 \mathbf{E}[\,X\,]}{3}\right)$$

$$\mathbf{P}[\,X \leq (1-\delta)\mathbf{E}[\,X\,]\,] \leq \exp\left(-\frac{\delta^2 \mathbf{E}[\,X\,]}{2}\right)$$

Chernoff Bound: Extension to other Random Variables

- Most of the time we will use Chernoff Bounds for sum of independent Bernoulli random variables
- but not always
- it does not hurt to know how to derive similar bounds for other random variables

Remember the key steps:

Chernoff Bound recipe

1. Let $\lambda > 0$, then

$$\mathbf{P}[\, X \geq (1+\delta)\mu \,] \leq e^{-\lambda(1+\delta)\mu} \mathbf{E}\left[\, e^{\lambda X} \,\right]$$

2. Compute an upper bound for $\mathbf{E}\left[\, e^{\lambda X} \,\right]$
3. Optimise the value of $\lambda > 0$.

Exercise:

- Let X be a Poisson random variable of mean μ. Prove that

$$\mathbf{E}\left[e^{\lambda X} \right] = e^{\mu(e^{\lambda}-1)}$$

and deduce that for $t \geq \mu$

$$\mathbf{P}[\, X \geq t \,] \leq e^{-\mu} \left(\frac{e\lambda}{t} \right)^{t} \quad \text{and} \quad \mathbf{P}[\, X \geq (1+\delta)\mu \,] \leq e^{-\delta^2 \mu},$$

and the corresponding lower tails.

- Let X be a Normal random variable of mean μ and variance σ^2. Prove that

$$\mathbf{E}\left[e^{\lambda X} \right] = e^{\mu\lambda + \sigma^2\lambda^2/2},$$

and deduce that for $t > \mu$

$$\mathbf{P}[\, X \geq t \,] \leq e^{-(t-\mu)^2/2}.$$

Hoeffding's Extension

- Beside sums of independent Bernoulli Random variables, sums of independent and bounded random variables is very important in applications.
- Unfortunately the distribution of the X_i will be unknown or very hard to compute, thus it will be very hard to compute the moment-generating function of X_i.
- Hoeffding's Lemma helps us here

You can always consider $X' = X - \mathbf{E}[X]$

Hoeffding's Extension Lemma

Let X be a random variable with mean 0 such that $a \leq X \leq b$, then for all $\lambda \in \mathbb{R}$.

$$\mathbf{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{(b-a)^2\lambda^2}{8}\right)$$

Chernoff-Hoeffding Bounds

Chernoff-Hoeffding's Bounds

Let $X_1, \ldots, X_n$ be independent random variable with mean μ_i such that $a_i \leq X_i \leq b_i$. Let $X = X_1 + \ldots + X_n$, and let $\mu = \mathbf{E}[\, X\,] = \sum_{i=1}^{n} \mu_i$. Then for any $t > 0$

$$\mathbf{P}[\, X \geq \mu + t\,] \leq \exp\left[\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right]$$

and

$$\mathbf{P}[\, X \leq \mu - t\,] \leq \exp\left[\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right]$$

Proof:

- Let $X_i' = X_i - \mu_i$ and $X' = X_1' + \ldots, X_n'$, then $\mathbf{P}[\, X \geq \mu + t\,] = \mathbf{P}[\, X' \geq t\,]$
- $\mathbf{P}[\, X' \geq t\,] \leq e^{-\lambda t} \prod_{i=1}^{n} \mathbf{E}\left[\, e^{\lambda X_i'}\,\right] \leq \exp\left[-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^{n}(b_i - a_i)^2\right]$
- Choose $\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}$ to get the result.

This is not magic! you just need to optimise on λ

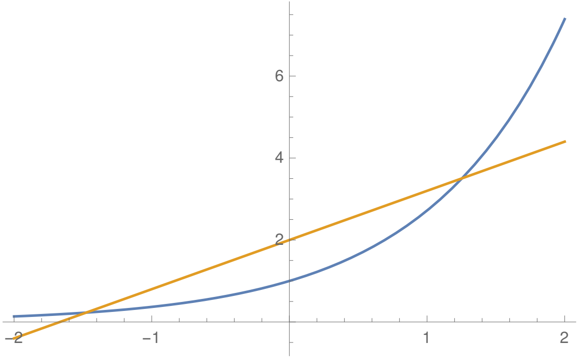
Hoeffding's Extension Lemma

Let X be a random variable with mean 0 such that $a \leq X \leq b$, then for all $\lambda \in \mathbb{R}$.

$$\mathbf{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{(b-a)^2\lambda^2}{8}\right)$$

Proof (for $\lambda \geq 0$):

- $f(x) = e^{\lambda x}$ is a convex function.



Hoeffding's Extension Lemma

Let X be a random variable with mean 0 such that $a \leq X \leq b$, then for all $\lambda \in \mathbb{R}$.

$$\mathbf{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{(b-a)^2\lambda^2}{8}\right)$$

Proof (for $\lambda \geq 0$):

1. $f(x) = e^{\lambda x}$ is a convex function.
2. As $a \leq X \leq b$, we consider the points $(a, e^{\lambda a})$ and $(b, e^{\lambda b})$
3. The straight line between those points is always above the graph of $e^{\lambda x}$
4. i.e.

$$e^{\lambda X} \leq \frac{b-X}{b-a}e^{\lambda a} + \frac{X-a}{b-a}e^{\lambda b}$$

5. Then

$$\mathbf{E}\left[e^{\lambda X}\right] \leq \frac{be^{\lambda a}}{b-a} - \frac{ae^{\lambda b}}{b-a}$$

1. $f(x) = e^{\lambda x}$ is a convex function.
2. As $a \leq X \leq b$, we consider the points $(a, e^{\lambda a})$ and $(b, e^{\lambda b})$
3. The straight line between those points is always above the graph of $e^{\lambda x}$
4. i.e.
$$e^{\lambda X} \leq \frac{b-X}{b-a} e^{\lambda a} + \frac{X-a}{b-a} e^{\lambda b}$$
5. Then
$$\mathbf{E}\left[e^{\lambda X} \right] \leq \frac{be^{\lambda a}}{b-a} - \frac{ae^{\lambda b}}{b-a}$$
6. Consider
$$\phi(\lambda) = \log\left(\frac{be^{\lambda a}}{b-a} - \frac{ae^{\lambda b}}{b-a} \right)$$
and check that (Exercise)
    - $\phi(0) = 0$
    - $\phi'(0) = 0$
    - $\phi''(t) \leq (b-a)^2/4$ for all $t \in \mathbb{R}$
7. For $t \geq 0$, use that $\phi'(t) = \int_0^t \phi''(x)dx \leq t(b-a)^2/4$
8. For $t \geq 0$, use that
$$\phi(t) = \int_0^t \phi'(x)dx \leq \int_0^t x(b-a)^2/4dx \leq t^2(b-a)^2/8$$
9. replace $t = \lambda$ for non-negative λ.

Chernoff-Bounds: Final Remarks

- There are several version of Chernoff-style Bounds that work for sum of independent random variables.
- The proof of all of them usually follows the same **recipe**
- Some bounds include more information about the random variables, e.g. the variance
- the limit is the amount of information we have about the random variables and our ability to manipulate/bound quantities.

Beyond sum of independent variables

Can we prove concentration of other type of random variables? Yes.. but

- There is no general tool to prove concentration beyond the basic **recipe**
- but in general it is very hard to compute moment generating functions
- It is worth trying to transform the problem into the setting of sum of independent random variable
- If everything fails.. There are a few other families of random variables for which proving concentration is doable One of them are the so-called **Martingales**

Outline

More Chernoff Bounds

Conditional Expectation

Conditional Expectation

Before talking about martingales, we need to review **conditional expectation.**

- Given two events A and B with $\mathbf{P}[\,A\,] > 0$ we define $\mathbf{P}[\,B|A\,] = \mathbf{P}[\,B \cap A\,]\,/\mathbf{P}[\,A\,]$.
- if $\mathbf{P}[\,A\,] = 0$, the usual convention is that $\mathbf{P}[\,B|A\,] = 0$.
- Given a discrete random variable Y, we define its conditional expectation with respect to the event A by

$$\mathbf{E}[\,Y|A\,] = \sum_b b\mathbf{P}[\,Y = b|A\,]$$

- a particular case is when the event $A = \{X = a\}$ where X is another discrete random variable. In such a case we define the function $f(a)$ by

$$f(a) = \mathbf{E}[\,Y|X = a\,]\,,$$

- We define the conditional expectation $\mathbf{E}[\,Y|X\,]$, as the **random variable** that takes the value $\mathbf{E}[\,Y|X = a\,]$ then $X = a$, i.e. $f(X)$.

Important Remarks

- The conditional expectation of Y w.r.t a event A, $\mathbf{E}[Y|A]$ is a deterministic number .
- The conditional expectation of Y w.r.t a random variable X, $\mathbf{E}[Y|X]$ is a random variable .
- X can be a random vector $(X_1, \ldots, X_N)$ in the definition of $\mathbf{E}[Y|X]$.
- There is a definition of conditional expectation with respect to general random variables[1], but most of the results in the discrete setting extend to the continuous setting.
- The conditional expectation $\mathbf{E}[Y|X]$ is always a function of X.
- Behind conditional expectation there is the notion of information [2]. The standard notion of expectation is like 'the best estimate of a random variable given no information of it', while the conditional expectation given X is like 'the best estimate of a random variable given the information of X'

[1] such a definition require the understanding of Measure theory
[2] Measure theory, again

Conditional Expectation: two dices

Suppose we independently roll two standard 6-sided dice. Let X_1 and X_2 the observed number in the first and second dice respectively. We compute a few conditional expectations.

1. $\mathbf{E}[X_1 + X_2 | X_1] = 3.5 + X_1$. Why? Because if $X_1 = a$ then

$$\begin{aligned}
\mathbf{E}[X_1 + X_2 | X_1 = a] &= \sum_{b=1}^{12} b\mathbf{P}[X_1 + X_2 = b | X_1 = a] \\
&= \sum_{b=1}^{12} b\mathbf{P}[X_1 + X_2 = b, X_1 = a] / \mathbf{P}[X_1 = a] \\
&= \sum_{b=1}^{12} b\mathbf{P}[X_2 = b - a, X_1 = a] / \mathbf{P}[X_1 = a] \\
\color{red}{X_1 \text{ indep } X_2} \quad &= \sum_{b=1}^{12} b\mathbf{P}[X_2 = b - a] \\
&= \sum_{c=1}^{6} (c + a)\mathbf{P}[X_2 = c] \\
&= 3.5 + a
\end{aligned}$$

Conditional Expectation: Properties

1. $\mathbf{E}[\,\mathbf{E}[\,Y|X\,]\,] = \mathbf{E}[\,Y\,]$.
2. $\mathbf{E}[\,1|X\,] = 1$
3. **Linearity :**
   - For any constant $c \in \mathbb{R}$, $\mathbf{E}[\,cY|X\,] = c\mathbf{E}[\,Y|X\,]$
   - $\mathbf{E}[\,Y + Z|X\,] = \mathbf{E}[\,Y|X\,] + \mathbf{E}[\,Z|X\,]$
4. If X is independent of Y, then $\mathbf{E}[\,Y|X\,] = \mathbf{E}[\,Y\,]$.
5. if Y is a function of X[3], i.e. $Y = f(X)$, then $\mathbf{E}[\,YZ|X\,] = Y\mathbf{E}[\,Z|X\,]$. Particularly, $\mathbf{E}[\,X|X\,] = X$
6. **Tower Property:**
   - $\mathbf{E}[\,\mathbf{E}[\,X|(Z,Y)\,]\,|Y\,] = \mathbf{E}[\,X|Y\,]$.
7. **Jensen Inequality:**
   - if f is a convex real function, then $f(\mathbf{E}[\,X|Y\,]) \leq \mathbf{E}[\,f(X)|Y\,]$.

By using this properties, everything becomes a bit easier, e.g., our two dices example

$$\mathbf{E}[\,X_1 + X_2|X_1\,] \overset{p3}{=} \mathbf{E}[\,X_1|X_1\,] + \mathbf{E}[\,X_2|X_1\,] \overset{p5,p4}{=} X_1 + \mathbf{E}[\,X_2\,] = X_1 + 3.5$$

[3] measurable function

Exercise: Prove the properties

1. $\mathbf{E}[\,\mathbf{E}[\,Y|X\,]\,] = \mathbf{E}[\,Y\,]$.
   Proof: e.g.

$$\sum_x \mathbf{E}[\,Y|X=x\,]\,\mathbf{P}[\,X=x\,] = \sum_x \sum_y y\mathbf{P}[\,Y=y|X=x\,]\,\mathbf{P}[\,X=x\,]$$
$$= \sum_x \sum_y y\mathbf{P}[\,Y=y, X=x\,]$$
$$= \sum_y y\mathbf{P}[\,Y=y\,]$$

Example: Expectation of a Geometric Random Variable

Suppose $X_1, X_2, \ldots,$ are an infinite sequence of independent Bernoulli (coins) random variables of parameter p, i.e. $\mathbf{P}[\, X_i = 1\,] = p$. Define

$G = \min\{k \geq 1 : X_k = 1\}$, which is the number of coins we have to observe until we get a head.

G has geometric distribution of parameter p. Indeed $\mathbf{P}[\, G = k\,] = p(1-p)^{k-1}$.

The expectation of G is given by the formula

$$\mathbf{E}[\, G\,] = \sum_{k=1}^{\infty} kp(1-p)^{k-1}$$

Let say that we forgot how to compute that type of sums...

We can compute $\mathbf{E}[\,G\,]$ by other means.

- $\mathbf{E}[\,G\,] \overset{p1}{=} \mathbf{E}[\,\mathbf{E}[\,G|X_1\,]\,]$
- $G = X_1 + (1 - X_1)(1 + G')$ where G' is the number of coins we need to wait to see a head after the first coin.
- $\mathbf{E}[\,X_1 + (1 - X_1)(1 + G')|X_1\,] \overset{p3,p5}{=} X_1 + (1 - X_1)\mathbf{E}[\,1 + G'|X_1\,]$
- G' has geometric distribution of parameter p and it is independent of X_1. Hence
$$\mathbf{E}\left[\,1 + G'|X_1\,\right] \overset{p4}{=} \mathbf{E}\left[\,1 + G'\,\right] = 1 + \mathbf{E}[\,G\,]$$
- Solve
$$\mathbf{E}[\,G\,] = p + (1 - p)(1 + \mathbf{E}[\,G\,])$$

Example: Balls into Bins

Suppose we have n bins but a random number of balls, say M. Suppose M has finite expectation. What is the expected number of balls in the first bin?.

1. Recall balls are assigned to bins uniformly at random and independent of everything
2. Let $X_i = 1$ if the ball i falls in bin 1
3. The total number of balls in bin 1 is $\sum_{i=1}^{M} X_i$ (recall M is a random variable, and M is independent of X_i)

4. $\mathbf{E}\left[\sum_{i=1}^{M} X_i\right] \overset{p1}{=} \mathbf{E}\left[\mathbf{E}\left[\sum_{i=1}^{M} X_i \,\middle|\, M\right]\right]$

5. $\mathbf{E}\left[\sum_{i=1}^{M} X_i \,\middle|\, M\right] = \mathbf{E}\left[\sum_{i=1}^{\infty} X_i \mathbf{1}_{\{i \leq M\}} \,\middle|\, M\right] \overset{p3}{=}{}^{4} \sum_{i=1}^{\infty} \mathbf{E}\left[X_i \mathbf{1}_{\{i \leq M\}} \,\middle|\, M\right]$

6. $\mathbf{E}\left[X_i \mathbf{1}_{\{i \leq M\}} \,\middle|\, M\right] \overset{p5}{=} \mathbf{1}_{\{i \leq M\}} \mathbf{E}\left[X_i \,\middle|\, M\right] \overset{p4}{=} \mathbf{1}_{\{i \leq M\}} \mathbf{E}[X_i] = \mathbf{1}_{\{i \leq M\}} \cdot (1/n)$

7. Replacing 6 in 5: $\mathbf{E}\left[\sum_{i=1}^{M} X_i \,\middle|\, M\right] = \sum_{i=1}^{\infty} (1/n) \cdot \mathbf{1}_{\{i \leq M\}}$

8. replacing 7 into 4: $\mathbf{E}\left[\sum_{i=1}^{M} X_i\right] = \sum_{i=1}^{\infty} (1/n) \cdot \mathbf{P}[i \leq M] =^{5} (1/n) \cdot \mathbf{E}[M]$

---

[4] Technically linearity works for a finite sum, but in most cases it can be done for infinite case. We need measure theory to justify that

[5] See Q2 of the Homework Assessment