**36-705: Intermediate Statistics** **Fall 2017**

# Lecture 2: August 30

*Lecturer: Siva Balakrishnan*

Recall in the last class we discussed that we would like to understand the behaviour of the average of independent random variables. Towards that goal let us begin by trying to understand the tail behaviour of a random variable.

## 2.1 Markov Inequality

The most elementary tail bound is Markov's inequality, which asserts that for a positive random variable $X \geq 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Intuitively, if the mean of a (positive) random variable is small then it is unlikely to be too large too often, i.e. the probability that it is large is small. While Markov on its own is fairly crude it will form the basis for much more refined tail bounds.

**Proof:** Fix an arbitrary $t > 0$. Define the indicator:

$$\mathbb{I}(t) = \begin{cases} 1 \text{ if } X \geq t \\ 0 \text{ if } X < t. \end{cases}$$

We have that,

$$t\mathbb{I}(t) \leq X,$$

so that

$$\mathbb{E}[X] \geq \mathbb{E}[t\mathbb{I}(t)] = t\mathbb{E}[\mathbb{I}(t)] = t\mathbb{P}(X \geq t).$$

## 2.2 Chebyshev Inequality

Chebyshev's inequality states that for a random variable $X$, with $\mathsf{Var}(X) = \sigma^2$:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2} \quad \forall k \geq 0.$$

Before we prove this lets look at a simple application. In the last lecture we saw that if we average i.i.d. random variables with mean $\mu$ and variance $\sigma^2$, we have that the average:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

has mean $\mu$ and variance $\sigma^2/n$. So applying Chebyshev's inequality to $\widehat{\mu}_n$ we obtain that,

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}.$$

Alternatively, with probability at least 0.99 (for instance) the average is within $10\sigma/\sqrt{n}$ from the its expectation. This is pretty neat and almost directly gives us something called the Weak Law of Large Numbers (but we will return to this).

We will study refinements of this inequality today, but in some sense it already has the correct "$1/\sqrt{n}$" behaviour. The refinements will mainly be to show that in many cases we can dramatically improve the constant 10.

**Proof:** Chebyshev's inequality is an immediate consequence of Markov's inequality.

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) &= \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq k^2\sigma^2) \\ &\leq \frac{\mathbb{E}(|X - \mathbb{E}[X]|^2)}{k^2\sigma^2} = \frac{1}{k^2}.\end{aligned}$$

## 2.3 Chernoff Method

There are several refinements to the Chebyshev inequality. One simple one that is sometimes useful is to observe that if the random variable $X$ has a finite $k$-th central moment then we have that,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}|X - \mathbb{E}[X]|^k}{t^k}.$$

For many random variables (we will see some examples today), the moment generating function will exist in a neighborhood around 0, i.e the mgf is finite for all $|t| \leq b$ where $b > 0$ is some constant. In these cases, we can use the mgf to produce a tail bound.

Define, $\mu = \mathbb{E}[X]$. For any $t > 0$, we have that,

$$\mathbb{P}((X - \mu) > u) = \mathbb{P}(\exp(t(X - \mu)) > \exp(tu)) \leq \frac{\mathbb{E}[\exp(t(X - \mu))]}{\exp(tu)}.$$

Now $t$ is a parameter we can choose to get a tight upper bound, i.e. we can write this bound as:

$$\mathbb{P}((X - \mu) > u) \leq \inf_{0 \leq t \leq b} \exp(-t(u + \mu))\mathbb{E}[\exp(tX)].$$

This bound is known as Chernoff's bound.

### 2.3.1 Gaussian Tail Bounds via Chernoff

Suppose that, $X \sim N(\mu, \sigma^2)$, then a simple calculation (see HW2) gives that the mgf of $X$ is:

$$M_X(t) = \mathbb{E}[\exp(tX)] = \exp(t\mu + t^2\sigma^2/2).$$

The mgf is defined for all $t$. To apply the Chernoff bound we then need to compute:

$$\inf_{t\geq 0} \exp(-t(u+\mu))\exp(t\mu + t^2\sigma^2/2) = \inf_{t\geq 0} \exp(-tu + t^2\sigma^2/2),$$

which is minimized when $t = u/\sigma^2$ which in turn yields the tail bound,

$$\mathbb{P}(X - \mu > u) \leq \exp(-u^2/(2\sigma^2)).$$

This is often referred to as a one-sided or upper tail bound. We can use the fact that if $X$ has distribution $N(\mu, \sigma^2)$ then $-X$ has distribution $N(-\mu, \sigma^2)$ and repeat the above calculation to obtain the analogous lower tail bound,

$$\mathbb{P}(-X + \mu > u) \leq \exp(-u^2/(2\sigma^2)).$$

Putting these two pieces together, we have the two-sided Gaussian tail bound:

$$\mathbb{P}(|X - \mu| > u) \leq 2\exp(-u^2/(2\sigma^2)).$$

The main thing to observe is that this inequality is much sharper than Chebyshev's inequality. In particular, suppose we consider the average of i.i.d Gaussian random variables, i.e. we have $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and we construct the estimate:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Using the fact that the average of Gaussian RVs is Gaussian we obtain that $\widehat{\mu}$ has a $N(\mu, \sigma^2/n)$ distribution. In this case, using the Gaussian tail bound we derived we obtain that,

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq k\sigma/\sqrt{n}) \leq 2\exp(-k^2).$$

This is an example of an exponential tail inequality. Comparing with Chebyshev's inequality we should observe two things:

1. Both inequalities say roughly that the deviation of the average from the expected value goes down as $1/\sqrt{n}$.

2. However, the Gaussian tail bound says if the random variables are actually Gaussian then the chance that the deviation is much bigger than $\sigma/\sqrt{n}$ goes down *exponentially fast.* Let us look at a concrete example, we say previously that Chebyshev told us the average is within $10\sigma/\sqrt{n}$ with probability at least 0.99.

   On the other hand the exponential tail bound says that with probability 0.99 the average is within,

$$\sqrt{\ln(1/0.005)}\sigma/\sqrt{n} \approx 2.3\sigma/\sqrt{n}.$$

   More generally, Chebyshev tells us that with probability at least $1-\delta$,

$$|\widehat{\mu} - \mu| \leq \frac{\sigma}{\sqrt{n\delta}}$$

   while the exponential tail bound tells us that,

$$|\widehat{\mu} - \mu| \leq \sigma\sqrt{\frac{\ln(2/\delta)}{n}}.$$

   The first goes up polynomially as $\delta \to 0$, while the second more refined bound goes up only logarithmically.

### 2.3.2 Sub-Gaussian Random Variables

It turns out that the Gaussian tail inequality from the previous section is much more broadly applicable to a class of random variables called sub-Gaussian random variables. Roughly these are random variables whose tails decay faster than a Gaussian. Formally, a random variable $X$ with mean $\mu$ is *sub-Gaussian* if there exists a positive number $\sigma$ such that,

$$\mathbb{E}[\exp(t(X-\mu))] \leq \exp(\sigma^2 t^2/2),$$

for all $t \in \mathbb{R}$. Gaussian random variables with variance $\sigma^2$ satisfy the above condition with equality, so a $\sigma$-sub-Gaussian random variable basically just has an mgf that is dominated by a Gaussian with variance $\sigma$.

It is straightforward to go through the above Chernoff bound to conclude that for a sub-Gaussian random variable we have the same two-sided exponential tail bound,

$$\mathbb{P}(|X-\mu| > u) \leq 2\exp(-u^2/(2\sigma^2)).$$

Suppose we have $n$ i.i.d random variables $\sigma$ sub-Gaussian RVs $X_1, X_2, \ldots, X_n$, then by independence we obtain that,

$$\begin{aligned}\mathbb{E}[\exp(t(\widehat{\mu}-\mu))] &= \mathbb{E}[\exp(t/n\sum_{i=1}^{n}(X_i-\mu)] \\ &= \prod_{i=1}^{n}\mathbb{E}[\exp(t(X_i-\mu)/n)] \\ &\leq \exp(t^2\sigma^2/(2n)).\end{aligned}$$

Alternatively, the average of $n$ independent $\sigma$-sub Gaussian RVs is $\sigma/\sqrt{n}$-sub Gaussian. This yields the tail bound for the average of sub Gaussian RVs:

$$\mathbb{P}(|\widehat{\mu}-\mu| \geq k\sigma/\sqrt{n}) \leq 2\exp(-k^2).$$

### 2.3.3 Bounded Random Variables - Hoeffding's bound

We claimed in the previous section that many classes of RVs are sub-Gaussian. In this section, we show this for an important special case: *bounded random variables.*

**Example 1:** Let us first consider a simple case, of Rademacher random variables, i.e. random variables that take the values $\{+1,-1\}$ equiprobably. In this case we can see that,

$$\begin{aligned}\mathbb{E}[\exp(tX)] &= \frac{1}{2}\left[\exp(t)+\exp(-t)\right] \\ &= \frac{1}{2}\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{k!}+\sum_{k=0}^{\infty}\frac{t^k}{k!}\right] \\ &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{t^{2k}}{2^k k!} \\ &= \exp(t^2/2).\end{aligned}$$

This shows that Rademacher random variables are 1-sub Gaussian.

**Detour: Jensen's inequality:** Jensen's inequality states that for a convex function $g:\mathbb{R}\mapsto\mathbb{R}$ we have that,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

If $g$ is concave then the reverse inequality holds.

**Proof:** Let $\mu=\mathbb{E}[X]$ and let $L_\mu(x)=a+bx$ be the tangent line to the function $g$ at $\mu$, i.e. we have that $L_\mu(\mu)=g(\mu)$. By convexity we know that $g(x)\geq L_\mu(x)$ for every point $x$. Thus we have that,

$$\begin{aligned}\mathbb{E}[g(X)] \geq \mathbb{E}[L_\mu(X)] &= \mathbb{E}[a+bX] \\ &= a+b\mu = L_\mu(\mu) = g(\mu).\end{aligned}$$

**Example 2: Bounded Random Variables.** Let $X$ be a random variable with zero mean and with support on some bounded interval $[a, b]$.

You should convince yourself that the zero mean assumption does not matter in general (you can always subtract the mean, i.e. define a new random variable $Y = X - \mathbb{E}[X]$ and use $Y$ in the calculation below).

Let $X'$ denote an *independent* copy of $X$ then we have that,

$$\mathbb{E}_X[\exp(tX)] = \mathbb{E}_X[\exp(t(X - \mathbb{E}[X'])] \leq \mathbb{E}_{X,X'}[\exp(t(X - X')],$$

using Jensen's inequality, and the convexity of the function $g(x) = \exp(x)$.

Now, let $\epsilon$ be a Rademacher random variable. Then note that the distribution of $X - X'$ is identical to the distribution of $X' - X$ and more importantly of $\epsilon(X - X')$. So we obtain that,

$$\begin{aligned}\mathbb{E}_{X,X'}[\exp(t(X - X')] &= \mathbb{E}_{X,X'}[\mathbb{E}_\epsilon[\exp(t\epsilon(X - X')]] \\ &\leq \mathbb{E}_{X,X'}[\exp(t^2(X - X')^2/2],\end{aligned}$$

where we just use the result from Example 1, with $(X, X')$ fixed by conditioning. Now $(X - X')$ using boundedness is at most $(b - a)$ so we obtain that,

$$\mathbb{E}_X[\exp(tX)] \leq \exp(t^2(b - a)^2/2),$$

which in turn shows that bounded random variables are $(b - a)$-sub Gaussian.

This in turn yields Hoeffding's bound. Suppose that, $X_1, \ldots, X_n$ are independent identically distribution *bounded* random variables, with $a \leq X_i \leq b$ for all $i$ then,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{(b-a)^2}\right).$$

This is a two-sided exponential tail inequality for the averages of bounded random variables.

### 2.3.4 A simple generalization

It is worth noting that none of the exponential tail inequalities we proved required the random variables to be identically distributed. More generally, suppose that we have $X_1, \ldots, X_n$ which are each $\sigma_1, \ldots, \sigma_n$ sub Gaussian. Then using just independence you can verify that their average $\widehat{\mu}$ is $\sigma$-sub Gaussian, where,

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^n \sigma_i^2}$$

This in turn yields the exponential tail inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq \exp(-t^2/(2\sigma^2)).$$

Note that the random variables still need to be independent but no longer need to be identically distributed (i.e. they can for instance have different means and sub-Gaussian parameters).