Lecture 2

Last time	**Today**
▷ Syllabus	▷ Chernoff Method
▷ Motivation	▷ Sub-Gaussians
▷ Overview	▷ Hoeffding's inequality

Chernoff method

Last time, we ended with the question of

Given $X_1, \ldots, X_n$ r.v., how can we prove a tail bound

$$\mathbb{P}\left(\left|\sum_i X_i\right| \geq t\right) \leq e^{-t^2/2} \,?$$

We start by proving a simple (suboptimal) bound.

Proposition (Markov's inequality) For any nonnegative X and $t > 0$, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}. \qquad \dashv$$

Proof:

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}X\mathbb{1}_{\{X\geq t\}} + \mathbb{E}X\mathbb{1}_{\{X\leq t\}} \\ &\geq t\,\mathbb{E}\mathbb{1}_{\{X\geq t\}} \\ &= t\,\mathbb{P}(X\geq t). \qquad \square\end{aligned}$$

Indicator

Corollary (Chebychev's inequality): For any r.v. X, we have

$$\mathbb{P}(|X-\mathbb{E}X|\geq t) \leq \frac{Var(X)}{t^2}$$

Proof: Apply Markov's to $\mathbb{P}(|X-\mathbb{E}X|^2 \geq t^2)$. $\square$

Although this is much weaker than what we want it introduces the key idea of mapping

$$\mathbb{P}(X-\mathbb{E}X \geq t) = \mathbb{P}(f(X-\mathbb{E}X)\geq f(t))$$

for any $f:\mathbb{R}\to\mathbb{R}_+$ monotonically increasing. In turn, it is a good idea to take $f(t)=e^{\lambda t}$,

Some tunable $\lambda \geq 0$

$$\mathbb{P}(X - \mathbb{E}X \geq t) = \mathbb{P}\left(e^{\lambda(X-\mathbb{E}X)} \geq e^{\lambda t}\right)$$

Moment generating function $M_{X-\mathbb{E}X}(\lambda)$

$$\leq \frac{\mathbb{E}\, e^{\lambda(X-\mathbb{E}X)}}{e^{\lambda t}}.$$

Taking logs and optimizing for λ

$$\log \mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda \geq 0} \left\{ \log \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} - \lambda t \right\}$$

$$= -\underbrace{\sup \left\{ \lambda t - \log \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} \right\}}.$$

What the heck is this thing?

It is a well-known object in convex analysis.

Def (Fenchel conjugate): Given $\varphi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, define

$$\varphi^*(t) = \sup_{\lambda} \left\{ t\lambda - \varphi(\lambda) \right\}.$$

$\lambda \leftarrow$ over all $\mathbb{R}$

Take

$$\varphi_X(\lambda) := \log \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)}.$$

For any $\lambda \in \mathbb{R}$

$$\varphi_X(\lambda) = \log \mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} \geq \mathbb{E} \log e^{\lambda(X-\mathbb{E}X)} = 0$$

Jensen's (−log is convex).

When $\lambda < 0$ and $t \geq 0$ we have

$$t\lambda - \varphi(\lambda) \leq 0 = t \cdot 0 - \varphi(0).$$

Thus, for $t \geq 0$ we get

$$\varphi^*(t) = -\sup_{\lambda \geq 0} \{t\lambda - \varphi(\lambda)\}.$$

We have proved the following.

Proposition (Chernoff Bound). Given X a r.v., we have

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \exp(-\varphi_X^*(t)). \quad \dashv$$

Example. Take $X \sim N(\mu, \sigma^2)$. Then, its moment generating function is

$$\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\frac{\sigma^2 \lambda^2}{2}} \qquad \forall \lambda \in \mathbb{R}.$$

Then,

Max of a quadratic

$$\varphi_X^*(t) = \sup_{\lambda} \left\{ t\lambda - \frac{\sigma^2\lambda^2}{2} \right\} = \frac{t^2}{2\sigma^2}.$$

Thus,

$$\mathbb{P}(X-\mu \geq t) \leq e^{-t^2/2\sigma^2} \quad \forall t \geq 0.$$

So, Gaussians exhibit very nice concentration, to get analogous bounds, we can consider r.vs that are Gaussian-like. A moment of thought reveals

$$\varphi_X \leq \varphi_Y \Rightarrow \varphi_X^* \geq \varphi_Y^*$$
$$\Rightarrow \exp(-\varphi_X^*(t)) \leq \exp(-\varphi_Y^*(t)).$$

This motivates the following.

Def (sub-Gaussians): A r.v. X with mean $\mu = \mathbb{E}X$ is σ-sub-Gaussian if

$$\log \mathbb{E}\, e^{\lambda(X-\mu)} \leq \frac{\sigma^2 \lambda^2}{2} \quad \forall \lambda \in \mathbb{R}.$$

(why?)

If X is sub-Gaussian so is $-X$,

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq 2\exp\left(-t^2/2\sigma^2\right)$$

Example (Bounded rv) Let X be a r.v. supported on $[a,b]$. Then, X is $\frac{(b-a)}{2}$ - sub-Gaussian.

Proof: Take $Y = X - \mu$ and take $\psi_Y(\lambda) = \log \mathbb{E}\, e^{\lambda Y}$. Then,

$$\underbrace{\psi_Y'(\lambda) = \frac{\mathbb{E}\, Y e^{\lambda Y}}{\mathbb{E}\, e^{\lambda Y}} \text{ and } \psi_Y''(\lambda) = \frac{\mathbb{E}\, Y^2 e^{\lambda Y}}{\mathbb{E}\, e^{\lambda Y}} - \left(\frac{\mathbb{E}\, Y e^{\lambda Y}}{\mathbb{E}\, e^{\lambda Y}}\right)^2.}_{\text{Check!}}$$

By Taylor's theorem we have that

$$\psi_Y(\lambda) \leq \psi_Y(0) + \psi_Y'(0)\lambda + \frac{L}{2}\lambda^2 \qquad (\heartsuit)$$

where L is any uniform upper bound on $\psi_Y''(\lambda)$. It is easy to see that $\psi_Y(0) = \psi_Y'(0) = 0$. Thus, we only need to bound $\varphi_y''(\lambda)$. We use the following

Fact (Exercise): For any r.v. X

$$\mathrm{Var}(X) = \inf_{\omega} \mathbb{E}(X-\omega)^2.$$

Define the r.v. z with density

$$f_z(x) = \frac{e^{\lambda x} f_y(x)}{\mathbb{E}_y e^{\lambda y}}.$$

density/measure of y.

It is easy to see that

$$\begin{aligned}
\varphi_y''(\lambda) &= \mathrm{Var}(z) \\
&= \inf_{\omega} \mathbb{E}(z-\omega)^2 \\
&\leq \mathbb{E}\left(z - \frac{a+b}{2} + \mu\right)^2
\end{aligned}$$

By construction $z+\mu$ is supported on $[a,b]$, thus,

$$\varphi_y''(\lambda) \leq (b-a)^2/4 \qquad \forall \lambda$$

So by ($\heartsuit$)

$$\varphi_y(\lambda) \leq \left(\frac{b-a}{2}\right)^2 \frac{\lambda^2}{2}. \qquad \square$$

We started by wishing to prove concentration of sums. Do we still have nice tails after summing?

Lemma (Sum Rule): Suppose X_i are independent, σ_i-sub-Gaussian r.vs.
Then, $\sum_{i=1}^{n} X_i$ is $\sqrt{\sum_{i=1}^{n} \sigma_i^2}$-sub-Gaussian.

$\|\sigma\|_2$

Proof:

$$\log \mathbb{E} \exp\left(\lambda \sum_{i=1}^{n} X_i\right) \overset{\text{independence}}{=} \sum_{i=1}^{n} \log \mathbb{E} e^{\lambda X_i} \overset{\text{Sub-Gaussianity}}{\leq} \left(\sum_{i=1}^{n} \sigma_i^2\right) \frac{\lambda^2}{2}. \qquad \square$$

This establishes our main result

Theorem (Hoeffding Inequality):
Suppose $X_1, \ldots, X_n$ are ind. with $\mathbb{E}X_i = \mu_i$ and X_i σ_i-sub-Gaussian.
Then,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i-\mu_i) \geq t\|\sigma\|_2\right) \leq e^{-t^2/2}. \quad \dashv$$

In particular, if $\mu_i = \mu$, $\sigma_i = \sigma$.

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i-\mu_i) \geq t\sigma\sqrt{n}\right) \leq e^{-t^2/2}.$$

Back to our coin flipping example:

Let $c_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$ be iid.

Radamacher

Then, $\mathbb{E} c_i = 0$, $\sigma_i = 1$, then

$$\mathbb{P}\left(\frac{1}{n}\sum c_i \geq \frac{1}{2}\right) \leq e^{-n/8}.$$

We only have exponential decay!

$\dashv$