LECTURE 5

THM (Hoeffding's Inequality) Let $X_1, \dots, X_N$ be iid symmetric Bernoulli r.v's i.e. $P\{X_i = 1\} = P\{X_i = -1\} = 1/2$. Then

$$P\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \geq t \right\} \leq \exp\left(-\frac{t^2}{2}\right) \qquad \forall t \geq 0.$$

Proof: "The MGF method" (moment generating function)

- Let $\lambda \geq 0$ be a parameter TBD later ("to be determined").
- Multiply both sides of the inequality by $\lambda\sqrt{N}$ and exponentiate:

$$P\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \geq t \right\} = P\left\{ \exp\left(\lambda \sum_{i=1}^{N} X_i\right) \geq \exp(t\lambda\sqrt{N}) \right\}$$

$$\overset{\text{Markov's ineq.}}{\leq} \frac{E\exp\left(\lambda \sum_{i=1}^{N} X_i\right)}{\exp(\lambda t \sqrt{N})} = e^{-t\lambda\sqrt{N}} E \prod_{i=1}^{N} e^{\lambda X_i}$$

(independence)

$$= e^{-t\lambda\sqrt{N}} \prod_{i=1}^{N} E e^{\lambda X_i}$$

(all terms are the same (identical distr.))

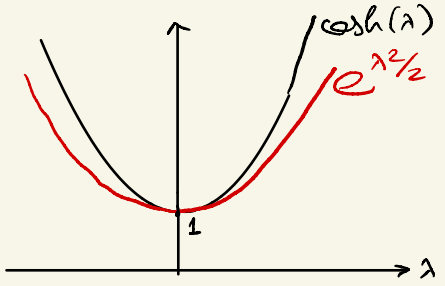
$$= e^{-t\lambda\sqrt{N}} \left(E e^{\lambda X}\right)^N \quad (\leq) \text{ ? where } X \sim \text{sym Bernoulli}$$

- Since X takes values 1 and -1 with probabilities $\frac{1}{2}$ each,

$$E e^{\lambda X} = \frac{1}{2} e^{\lambda} + \frac{1}{2} e^{-\lambda}$$

"Moment generating function" (MGF)

$$= \cosh(\lambda) \overset{\text{DIY Taylor}}{\leq} e^{\lambda^2/2} \quad \text{HW}$$



$$(\leq)\ e^{-t\lambda\sqrt{N}} \left(e^{\lambda^2/2}\right)^N = \exp\left(-t\lambda\sqrt{N} + \frac{\lambda^2 N}{2}\right) = \exp\left(-t\mu + \frac{\mu^2}{2}\right).$$

$-t\lambda\sqrt{N} = \mu$

- Minimize in $\mu \Rightarrow$ for $t = \mu$:

$$(=)\ e^{-t^2/2}. \qquad \text{QED}$$

Example: In N flips of a fair coin, $\mathbb{P}\{\text{at least } \frac{3}{4}N \text{ heads}\} \le \exp(-\frac{N}{8})$

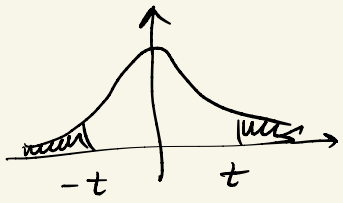
Proof: $\#\text{heads} = \sum_{i=1}^{N} Z_i$ where $Z_i = \begin{cases} 1 & \text{if } i\text{'th flip=head} \\ 0 & \text{if tail} \end{cases} \sim \text{Ber}(1/2)$

Create symmetric r.vs: $X_i = 2Z_i - 1 = \begin{cases} 1 & \text{if tail} \\ -1 & \text{if head} \end{cases} \sim \text{Sym Ber}$

$$\mathbb{P}\left\{\sum_{i=1}^{N} Z_i \ge \tfrac{3N}{4}\right\} = \mathbb{P}\left\{\tfrac{1}{\sqrt{N}}\sum_{i=1}^{N}(2Z_i-1) \ge \tfrac{\sqrt{N}}{2}\right\} \overset{\text{Hoeffding}}{\le} \exp\left(-\frac{(\sqrt{N}/2)^2}{2}\right) = \exp\left(-\tfrac{N}{8}\right). \quad \square$$

$\tfrac{\sqrt{N}}{2} \overset{!!}{=} t$

Remark (1) Two sided bound:



$$\mathbb{P}\left\{\left|\tfrac{1}{\sqrt{N}}\sum_{1}^{N} X_i\right| \ge t\right\} = \underbrace{\mathbb{P}\left\{\tfrac{1}{\sqrt{N}}\sum_{1}^{N} X_i > t\right\}}_{\text{Hoeffding} \Rightarrow e^{-t^2/2}} + \underbrace{\mathbb{P}\left\{\tfrac{1}{\sqrt{N}}\sum_{1}^{N} X_i < -t\right\}}_{\mathbb{P}\left\{\frac{1}{\sqrt{N}}\sum_{1}^{N}(-X_i) > t\right\}}$$

$-X_i \sim$ Sym Bernoulli, too

Hoeffding $\Rightarrow e^{-t^2/2}$

$$\le \boxed{2e^{-t^2/2}}$$

Remark (2) Same MGF method $\Rightarrow$ HW

THM (General Hoeffding) Let $X_1, \dots, X_n$ be independent r.v's such that $X_i \in [a_i, b_i]$ $\forall i$. Then $S_N = \sum_{i=1}^{N} X_i$ satisfies

$$\mathbb{P}\{S_N - \mathbb{E}S_N \ge t\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(b_i-a_i)^2}\right)$$

- This contains Hoeffding we proved earlier (check!)
- And it works for a biased coin, i.e. for $X_i \sim \text{Ber}(p) \Rightarrow S_N = \text{Binom}(N,p)$

Ex. (a) $\mathbb{P}\{\frac{1}{4}\text{-biased coin comes up} \ge \frac{N}{2} \text{ heads}\} \le \exp(-\frac{N}{8})$

(b) More generally, $S_N \sim \text{Binom}(N,p)$ satisfies $\mathbb{P}\{|S_N - pN| \ge \delta N\} \le 2\exp(-2\delta^2 N)$ $\forall \delta \ge 0$ (HW)

APPLICATION: MEAN ESTIMATION

- **Problem:** Estimate the mean μ of a distribution from a sample $X_1, \dots, X_n$, which is drawn independently from this distribution.

- **Classical estimator:** the sample mean $\hat{\mu} := \frac{1}{N} \sum_{i=1}^{N} X_i$

- $\mathbb{E}\hat{\mu} = \mu$, i.e. unbiased. Mean-squared error:

$$\mathbb{E}(\hat{\mu} - \mu)^2 = \mathrm{Var}(\hat{\mu}) = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(X_i) = \frac{\sigma^2}{N}.$$

$\Rightarrow$ the error is $\boxed{\frac{\sigma}{\sqrt{N}}}$ on average

- Confidence interval? Chebyshev's inequality:

$$\mathbb{P}\left\{ |\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}} \right\} \leq \frac{\sigma^2/N}{(t\sigma/\sqrt{N})^2} = \boxed{\frac{1}{t^2}}$$ Meh.

- For Gaussian distribution, we can get a much higher confidence:

if $X_i \sim N(\mu, \sigma^2) \Rightarrow \hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$.

Gaussian tail $\Rightarrow$ $\mathbb{P}\left\{ |\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}} \right\} \leq \boxed{\exp(-t^2/2)}$ ☺

↑ Gaussian tail (Lec. 4)

- Can we get a similar confidence for general distr.?

SURPRISINGLY, YES:

Thm $\forall\ 0 \le t \le \sqrt{N}$, $\exists$ estimator $\tilde{\mu} = \tilde{\mu}_t(X_1, \dots, X_N)$ that satisfies

$$P\left\{ |\tilde{\mu} - \mu| \ge \frac{t\sigma}{\sqrt{N}} \right\} \le 2\exp(-ct^2)$$

if X_i's are drawn from a distribution with mean μ and variance σ^2.
Here $c > 0$ is an absolute constant.

- With only finite variance assumption, Gaussian power is surprising!

Proof $\tilde{\mu}$ = "Median-of-Means" (MoM):

- Partition the sample into K blocks of equal size M $\quad N = MK$

(For simplicity, assume N is divisible by M. General case → HW)

$$\underbrace{X_1 X_2 \dots X_M}_{B_1}\ \underbrace{X_{M+1} \dots X_{2M}}_{B_2}\ \dots\ \underbrace{\dots X_N}_{B_K}$$

- Take the mean of each block; then take the median of those means:

$$(*)\quad \hat{\mu}_j := \frac{1}{M} \sum_{i \in B_j} X_i \ ; \qquad \tilde{\mu} := \mathrm{Med}(\hat{\mu}_1, \dots, \hat{\mu}_K)$$

(i.e. at least $\frac{K}{2}$ of $\hat{\mu}_j$ are $\le \tilde{\mu}$ and at least $\frac{K}{2}$ are $\ge \tilde{\mu}$)

- Accuracy of each $\hat{\mu}_j$? Mean μ, variance σ^2/M $\Rightarrow$

$$P\left\{ \hat{\mu}_j \ge \mu + \frac{t\sigma}{\sqrt{N}} \right\} \le \frac{\sigma^2/M}{(t\sigma/\sqrt{N})^2} = \frac{N}{t^2 M} = \frac{K}{t^2} = \frac{1}{4}$$

(Chebyshev)

if we set $K := t^2/4$; note that $K \le N$ by assumption

(For simplicity, assume $t/4$ is an integer. General case → HW)

- Therefore

$$P\left\{ \tilde{\mu} \ge \mu + \frac{t\sigma}{\sqrt{N}} \right\} \overset{\text{def of median}}{=} P\left\{ \text{at least } \frac{K}{2} \text{ of } \hat{\mu}_j\text{'s are} \ge \mu + \frac{t\sigma}{\sqrt{N}} \right\}$$

$$= P\{\text{at least a half of } \tfrac{1}{4}\text{-biased coins come up heads}\}$$

$$\le \exp(-K/8) = \exp(-t^2/32)$$

(Ex. on the bottom of p. 2) QED.

Remarks: ① The restriction $t \le \sqrt{n}$ is necessary. (→ HW)

② Can one construct the same estimator $\tilde{\mu}$ for all t simultaneously?
Not always, see [Gabor Lugosi: Median-of-means tournaments].