Hoeffding's Inequality:

suppose that $Z_1 \dots Z_n$ iid P, such that for each $Z_i \in [a, b]$, $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $\mu = \mathbb{E}[Z_i]$.

Then, for all $\varepsilon > 0$,

$$P(|\bar{Z} - \mu| > \varepsilon) \le 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$

$\mu-\varepsilon$ μ $\mu+\varepsilon$

Equivalently, $\forall \delta > 0$,

$$P\left(|\bar{Z} - \mu| > (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \le \delta.$$

Def: random variable X is said to be σ^2-subgaussian,

if $\forall \lambda \in \mathbb{R}$, $\mu = \mathbb{E}[X]$

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\sigma^2\lambda^2}{2}}.$$

(If $X \sim N(\mu, \sigma^2)$, $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\frac{\sigma^2\lambda^2}{2}}$)

1. linear transformation of Gaussian distribution is still a gaussian distribution

2. $X \cdot Y = A \cdot X + b$

$$\mathbb{E}[Y] = A \cdot \mathbb{E}[X] + b.$$

$$\mathrm{cov}(Y) \overset{?}{=} A\, \mathrm{cov}(X) \cdot A^T$$

$$= \mathbb{E}[(Y - \mathbb{E}Y) \cdot (Y - \mathbb{E}Y)^T]$$

$$= \mathbb{E}[(A \cdot (X - \mathbb{E}X)) \cdot (A \cdot (X - \mathbb{E}X))^T]$$

Lemma 1: If X takes values in $[a, b]$, then X is $\frac{(b-a)^2}{4}$ - subgaussian (SG)

Lemma 2: If $X_1 \sim X_n$ are independent, and for all i, X_i is σ_i^2 - subgaussian. then

$\sum_{i=1}^{n} a_i X_i$ is $\sum_{i=1}^{n} a_i^2 \sigma_i^2$ - SG for all $a_1 \sim a_n$

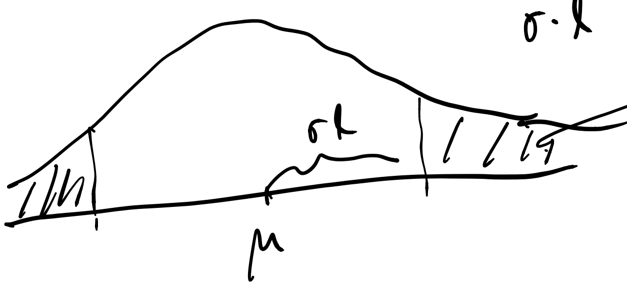
$a \cdot X_1 \quad a^2 \sigma_1^2$ - SG

$X_1 + X_2 \quad (\sigma_1^2 + \sigma_2^2)$ - SG

If X is $\underline{\sigma^2\text{-SG}}$ $\quad$ σ^2: variance proxy

Lemma 3 If X is σ^2-SG (SE), then $\forall t>0$

$$P(|X-\mu| \geq \underbrace{t}_{\sigma\cdot l}) \leq 2\cdot e^{-\frac{t^2}{2\sigma^2}}$$



$$\leq 2e^{-\frac{l^2}{2}}$$

the converse is almost true (up to constant scaling of σ).

Proof of Lemma 3

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

$$\Updownarrow$$

$$P(|X-\mu| \geq t)$$

$$= P(X-\mu \leq -t \text{ or } X-\mu \geq t)$$

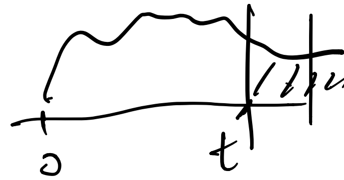
$$= P(X-\mu \leq -t) + \underbrace{P(X-\mu \geq t)}_{(*)}$$

$$(*) = P\left(e^{\lambda(X-\mu)} \geq e^{\lambda t} \right) \quad \text{for all } \lambda > 0.$$

$$\leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

Markov's ineq: $Y \geq 0$.

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



$$\leq e^{-\lambda t} \cdot e^{\frac{\sigma^2 \lambda^2}{2}}$$

$$= e^{-\lambda t + \frac{\sigma^2 \lambda^2}{2}} \quad (\circ)$$

choose $\lambda > 0$ to minimize the bound,

$$\sigma^2 \lambda - t = 0 \Rightarrow \lambda = \frac{t}{\sigma^2}$$

$$\Rightarrow (\circ) = e^{-\frac{t^2}{2\sigma^2}}$$

$$P(|X-\mu| \geq t) \leq 2 \cdot e^{-\frac{t^2}{2\sigma^2}}. \quad \boxtimes$$

proof of Lemma 2

$\mathbb{E}[X_1] = \mu_1$

① $a X_1$ $\mathbb{E}[aX_1] = a\mu_1$

$$\mathbb{E}\left[e^{\lambda (aX_1 - a\mu_1)} \right]$$

$$= \mathbb{E}\left[e^{\underline{\lambda a}(X_1 - \mu_1)} \right]$$

$$\le e^{\frac{(\lambda a)^2 \sigma_1^2}{2}} = e^{\frac{\lambda^2 (a^2 \sigma_1^2)}{2}}$$

$\Rightarrow a X_1$ is $a^2\sigma_1^2$ - SG

② X_1, X_2 $\mathbb{E}[X_1] = \mu_1$

$X_2 = \mu_2$

$$\mathbb{E}\left[e^{\lambda (X_1 + X_2 - \mu_1 - \mu_2)} \right] \overset{?}{\le} e^{\frac{\lambda^2 \sigma^2}{2}}$$

$$= \mathbb{E}\left[e^{\lambda (X_1 - \mu_1)} \cdot e^{\lambda (X_2 - \mu_2)} \right]$$

$$= \mathbb{E}\left[e^{\lambda(X_1-\mu_1)}\right] \cdot \mathbb{E}\left[e^{\lambda(X_2-\mu_2)}\right]$$

$$\le e^{\frac{\lambda^2\sigma_1^2}{2}} \cdot e^{\frac{\lambda^2\sigma_2^2}{2}}$$

$$= e^{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}}$$

$Z_1 \perp\!\!\!\perp Z_2$

$\mathbb{E}[Z_1 Z_2]$

$= \mathbb{E}[Z_1]\,\mathbb{E}[Z_2]$

$\le \sqrt{\mathbb{E}[Z_1^2]} \cdot \sqrt{\mathbb{E}[Z_2^2]}$

$X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$ - SG

③ $\sum_{i=1}^{n} a_i X_i$

$\downarrow$

$\sum_{i=1}^{n} a_i^2 \sigma_i^2$ - SG

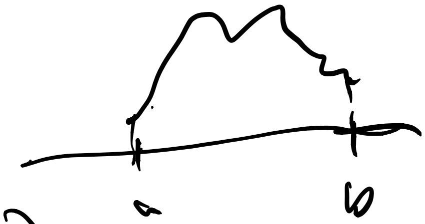
$\sigma^2 = \frac{(b-a)^2}{4}$

Lemma 1's pf:

want to show:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{(b-a)^2\lambda^2}{8}}$$

for X supported on $[a,b]$



suffices to show:

$$\underline{\psi(\lambda)} = \ln\left(\mathbb{E}\left[e^{\lambda \overset{Y}{(X-\mu)}}\right]\right)$$

$$\le \frac{(b-a)^2 \lambda^2}{8}$$

cumulant generating function of $X-\mu$.

b/w 0 and λ

$$\psi(\lambda) = \overset{=0}{\psi(0)} + \underline{\overset{=0}{\psi'(0)}}\cdot\lambda + \frac{\psi''(\xi)}{2}\cdot\lambda^2, \qquad \psi''(\xi) \le \frac{(b-a)^2}{4}$$

$$\psi(0) = 0$$

$$\psi'(\lambda) = \frac{\mathbb{E}\left[\frac{\partial}{\partial\lambda} e^{\lambda Y}\right]}{\mathbb{E}\left[e^{\lambda Y}\right]} = \frac{\mathbb{E}\left[Y \cdot e^{\lambda Y}\right]}{\mathbb{E}\left[e^{\lambda Y}\right]}$$

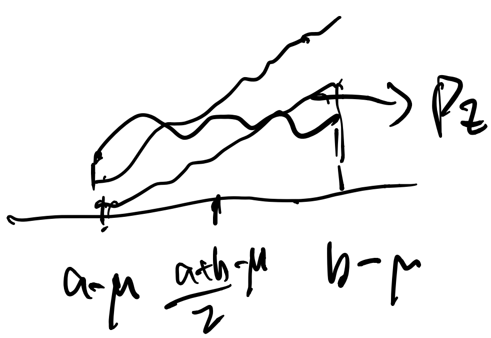
$$\psi'(0) = 0$$

$$\psi''(\lambda) = \frac{\mathbb{E}[e^{\lambda Y}\cdot Y^2]}{\mathbb{E}[e^{\lambda Y}]} - \left(\frac{\mathbb{E}[e^{\lambda Y}\cdot Y]}{\mathbb{E}[e^{\lambda Y}]}\right)^2$$

$$= \mathbb{E}[Z^2] - [\mathbb{E} Z]^2$$

where Z is a r.v. with distribution

pdf $$p_Z(y) = \frac{p_Y(y)\cdot e^{\lambda y}}{\int p_Y(y)\, e^{\lambda y}\, dy}$$



$$= \mathrm{Var}(Z)$$

$$= \mathbb{E}[Z - \mathbb{E} Z]^2$$

$$\le \mathbb{E}\left[Z - \left(\frac{a+b}{2} - \mu\right)\right]^2$$

$$\le \left(\frac{a-b}{2}\right)^2 = \frac{(b-a)^2}{4} \quad \boxtimes$$

Pf of Hoeffdings ineq:

$$X_i \quad \frac{(b-a)^2}{4}\text{-SG}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 \cdot \frac{(b-a)^2}{4} = \frac{(b-a)^2}{4n}\text{-SG}$$

Lemma 3: $\forall \varepsilon$.

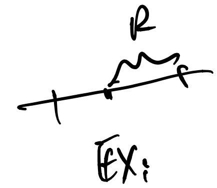
$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \varepsilon\right)$$

$$\le 2 \cdot \exp\left\{-\frac{\varepsilon^2}{2 \cdot \frac{(b-a)^2}{4n}}\right\}$$

$$= 2\exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\} \quad \boxtimes$$

Bernstein's inequality:

Let $X_1 \cdots X_n$ be iid. r.v's. $\forall i$. $|X_i - \mathbb{E}X_i| \le R$



let $\mu = \mathbb{E}[X_i]$

$\sigma^2 = \mathrm{Var}(X_i)$. Then,

for all $\varepsilon \ge 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \varepsilon\right) \le 2\exp\left\{-\frac{n\varepsilon^2}{2\sigma^2+\frac{2}{3}R\cdot\varepsilon}\right\} \quad (*)$$

$2(b-a)^2$

σ^2 can be $\ll (b-a)^2$

Set a small value of ε so that (*) is at most δ.

$$2\exp\left\{-\frac{n\varepsilon^2}{2\sigma^2+\frac{2}{3}R\varepsilon}\right\}\le\delta$$

$$\Leftarrow \quad n\varepsilon^2 \geqslant \left(2\sigma^2+\frac{2}{3}R\varepsilon\right)\ln\frac{2}{\delta}$$

$$\Leftarrow \quad n\varepsilon^2\geqslant 4\sigma^2\ln\frac{2}{\delta}, \text{ and } n\varepsilon^2\geqslant\frac{4}{3}R\varepsilon\ln\frac{2}{\delta}$$

$$\Leftrightarrow \quad \varepsilon\geqslant\sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}} \text{ and } \varepsilon\geqslant\frac{4R\ln\frac{2}{\delta}}{3n}.$$

choosing $\varepsilon=\sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}}+\frac{4R\ln\frac{2}{\delta}}{3n}$,

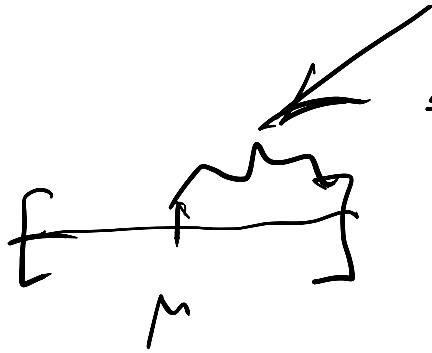
$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i-\mu\right|\geqslant\varepsilon\right)\le\delta.$$

$\Rightarrow$ w.p. $1-\delta$:

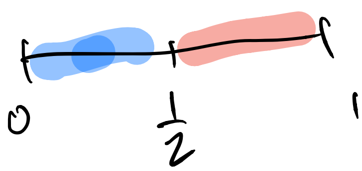
$$\left|\frac{1}{n}\sum_{i=1}^n X_i-\mu\right|\le\sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}}+\frac{4R\ln\frac{2}{\delta}}{3n}$$

$\delta = 10^{-5}$.

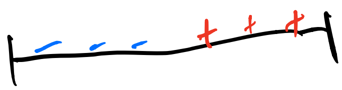
$\leq (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{n}}$.



Pf uses subexponential r.v's.

Ex D: 

$X \sim$ uniform$([0,1])$



Algorithm: (memorization) $\leftarrow \mathcal{H}$

given S. returns $\hat{h}$ such that

$$\hat{h}(x) = \begin{cases} y_i , & x = x_i \text{ for some } i \\ +1 & \text{otherwise} \end{cases}$$

① $\mathrm{err}(\hat{h}, S) = 0$

② Is it correct ~~that~~ w.p. $1-\delta$: $|\mathcal{H}|$ ↗

$$|\,err(\hat{h},S) - err(\hat{h},D)\,| \le \sqrt{\frac{\ln 1/\delta}{2m}} \;?$$

No?

③ $err(\hat{h},D) = \frac{1}{2}$

$$err(\hat{h},S) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{I(\hat{h}(x_i) \ne y_i)}_{\sim \text{Bernoulli}(err(\hat{h},D))}$$

$$\text{Bernoulli}(err(\hat{h},D)) \ne \frac{1}{2}$$

(*) $\hat{h}$ generated after seeing the training example then

$$m\, err(\hat{h},S) \not\sim \text{Bin}(m, err(\hat{h},D))$$

$\hat{h}$ generated before ...

--

then the above is true.