PRELIMINARIES

Concentration Inequalities (Foreword)

Two classical results to characterize the "many-sample" behavior of independent random variables:

① [Strong Law of Large Numbers]

$X_1, X_2, \dots, X_n$ independent r.v.s, $\mathbb{E}X_i = 0$, $\mathbb{E}[X_i^4] \leq K < \infty$, $\forall i \in \mathbb{N}$,

then $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} 0 \right) = 1.$

② [Central Limit Theorem]

$(X_n)_{n\in\mathbb{N}}$ is an iid sequence, $E[X_i] = 0$, $Var(X_i) = \sigma^2 < \infty$, $\forall i \in \mathbb{N}$.

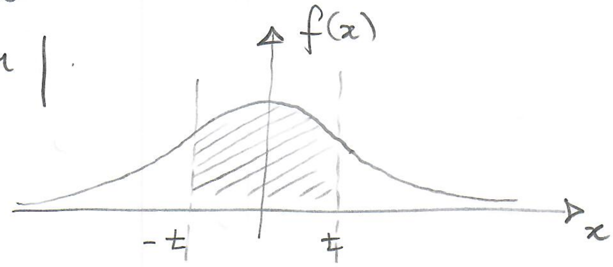
Set $G_n := \dfrac{\sum_{i=1}^n X_i}{\sigma\sqrt{n}}$

Then, $\forall x \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}(G_n \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$

<u>Observation</u> Consider independent and identically distributed r.v.'s $(X_n)_{n\in\mathbb{N}}$ with $\mathbb{E}[X_i] = \mu$. Then, $\frac{1}{n}\sum_{i=1}^n X_i$ converges asymptotically to μ, as $n \to \infty$.

<u>Question</u> Finite-sample guarantees for the deviation

$$\left| \frac{1}{n}\sum_{i=1}^n X_i - \mu \right|.$$



We are mainly interested in inequalities s.t.

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^n X_i - \mu \right| > t \right) \leq \delta(n,t),$$

where $\delta(n,t) \in (0,1)$ is s.t. $\lim_{n\to\infty} \delta(n,t) = 0$, $\forall t \geq 0$,

$\lim_{t\to\infty} \delta(n,t) = 0$, $\forall n \in \mathbb{N}$.

Of course, moments of X_i, i.e., $\mathbb{E}|X_i|^k$, $k \in \mathbb{N}$ will have a crucial importance in deriving these bounds.

The concentration behavior is much more general than just the concentration of the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ around the true mean.

Under very general conditions, as long as $(X_n)_n$ is an independent sequence of random variables,

$f(X_1, X_2, \dots, X_n)$ also shows concentration behavior around $\mathbb{E}f(X_1, X_2, \dots, X_n)$.

As such, the discussion will be as follows:

Characterize $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|$

Markov inequality,

$\downarrow$

Chebyshev's inequality

$\downarrow$

Chernoff bound $\rightarrow$ Hoeffding's inequality

Faster convergence

but

stronger moment assumptions

$\downarrow$

Characterize $|f(X_1,\dots,X_n) - \mathbb{E}f(X_1,\dots,X_n)|$

Azuma-Hoeffding inequality

$\downarrow$

$\boxed{\text{McDiarmid's inequality}}$

McDiarmid's inequality will be central to generalization bounds.

Hoeffding's inequality is useful for analyzing wide neural networks.

PRELIMINARIES #1 : CONCENTRATION INEQUALITIES

Basics : Markov, Chebyshev, Chernoff

Markov's Inequality Let $Y \in \mathbb{R}$ be a non-negative random variable with $\mathbb{E}|Y| < \infty$. Then, for any $t > 0$,

$$\mathbb{P}(Y > t) \leq \frac{\mathbb{E}Y}{t}.$$

Pf: Very easy. Since $Y \geq 0$ a.s., $t > 0$,

$$Y = Y(\mathbb{1}\{Y \geq t\} + \mathbb{1}\{Y < t\}) \geq Y\mathbb{1}\{Y \geq t\}$$
$$\geq t\,\mathbb{1}\{Y \geq t\}.$$

Taking expectation,

$$\mathbb{E}[Y] \geq \mathbb{E}[t\,\mathbb{1}\{Y \geq t\}] = t\,\mathbb{P}(Y \geq t).$$

Chebyshev's inequality Let $X \in \mathbb{R}$ be a random variable with $\mathbb{E}X^2 < \infty$. Then, for any $t > 0$,

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

Pf: Again, super easy. $Y := |X - \mathbb{E}X| \geq 0$ a.s. For any $t > 0$,

$$\mathbb{P}(Y > t) = \mathbb{P}(Y^2 > t^2) \leq \frac{\mathbb{E}Y^2}{t^2} = \frac{Var(X)}{t^2}.$$

How to use these inequalities?

$X_1, X_2, \dots, X_n$ independent $\mathbb{R}$-valued r.v.'s with $\mathbb{E}X_i = \mu_i$
$Var(X_i) \leq \sigma^2 < \infty$

Then, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) =: Z_n$, $\mathbb{E}[Z_n] = 0$,

$$Var(Z_n) = \frac{1}{n^2}\mathbb{E}\sum_{i,j}(X_i - \mu_i)(X_j - \mu_j) \leq \frac{\sigma^2}{n}.$$

$$\Rightarrow \quad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > t\right) \leq \frac{\sigma^2}{nt^2}$$

If $(X_i)_i$ are iid, then $\mu_i = \mu$, $Var(X_i) = \sigma^2$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right| > t\right) \leq \frac{\sigma^2}{nt^2}$$

Chernoff bound $X \in \mathbb{R}$, $\lambda > 0$, $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(X - \mathbb{E}X > t) &= \mathbb{P}(\lambda \cdot (X - \mathbb{E}X) > \lambda t) \\ &= \mathbb{P}\left(e^{\lambda(X-\mathbb{E}X)} > e^{\lambda t}\right) \\ &\leq \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] \cdot e^{-\lambda t}.\end{aligned}$$

Assumption: $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} < \infty$, $\lambda \in (-\lambda_0, \lambda_0)$ for some $\lambda_0 > 0$.

Then, let $\varphi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right]$.

$$\Rightarrow \quad \mathbb{P}(X - \mathbb{E}X > t) \leq e^{-[\lambda t - \varphi_X(\lambda)]}$$

Optimizing the r.h.s. of the above inequality,

$$\tilde{\varphi}_X^*(t) := \sup_{\lambda \in \mathbb{R}_+} \{\lambda t - \varphi_X(\lambda)\}$$

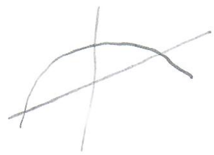
Thus,

$$\mathbb{P}(X - \mathbb{E}X > t) \leq e^{-\tilde{\varphi}_X^*(t)} \quad \text{for any } t \in \mathbb{R}$$

$\tilde{\varphi}_X^*(t)$ resembles <u>Legendre transform</u> in convex analysis, but the optimization is performed on $\mathbb{R}_+$ rather than $\mathbb{R}$.

Let $\varphi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \varphi_X(\lambda)\}$. Then, $\varphi_X^*(t) \geq \tilde{\varphi}_X^*(t)$.

However, we have $\varphi_X^*(t) = \tilde{\varphi}_X^*(t)$, $\underline{t \geq 0}$



<u>Pf</u> By Jensen's inequality,

$$\begin{aligned}\varphi_X(\lambda) &= \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] \\ &\geq \mathbb{E} \log e^{\lambda(X-\mu)} = 0\end{aligned}$$

Hence, $\varphi_X(\lambda) \geq 0$. If $t \geq 0$,

$$\sup_{\lambda \leq 0} \{\lambda t - \varphi_X(\lambda)\} \leq \sup_{\lambda \leq 0} \{\lambda t\} \leq 0.$$

Since $\{\lambda t - \varphi_X(\lambda)\}_{\lambda=0} = 0$, we have

$$\varphi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \varphi_X(\lambda)\} = \sup_{\lambda \in \mathbb{R}_+} \{\lambda t - \varphi_X(\lambda)\} = \tilde{\varphi}_X^*(t).$$

Example (Gaussian)

$$X \sim N(\mu, \sigma^2) \Rightarrow \varphi_X(\lambda) = \frac{\sigma^2\lambda^2}{2}, \ \forall \lambda \in \mathbb{R}.$$

Then,

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

for optimizing, $t - \sigma^2 \lambda_{opt} = 0 \Rightarrow \lambda_{opt} = \frac{t}{\sigma^2}$

$$\Rightarrow \varphi_X^*(t) = \frac{t^2}{2\sigma^2}$$

$$\Rightarrow \boxed{\mathbb{P}(X - \mu \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}}$$

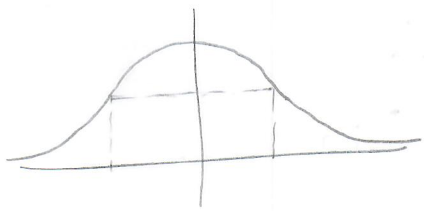
Sub-Gaussian Random Variables and Hoeffding's Lemma

DEF X is sub-Gaussian with parameter σ^2 if

(i) $\mathbb{E}X = 0$,

(ii) $\varphi_X(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$ for all $\lambda \in \mathbb{R}$.

$X \sim N(0, \sigma^2)$ is sub-Gaussian with σ^2 obviously.



Another important class of sub-Gaussian r.v.'s is bounded r.v.'s.

PROP If $X \in \mathbb{R}$ is sub-Gaussian with σ^2, then

$$\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}, \ \forall t > 0.$$

LEMMA $X \in [a, b]$ a.s. for some $a < b$, then X is sub-Gaussian with $\frac{(b-a)^2}{4}$.

Pf X has a density function $f(x)$.

Let Z be a r.v. with density $\frac{f(x)e^{\lambda x}}{e^{\varphi_X(\lambda)}}$. For sanity, $\frac{\int_a^b f(x) e^{\lambda x} dx}{e^{\varphi_X(\lambda)}} = 1$

$$E[Z] = \int x f(x) e^{\lambda x} dx \cdot e^{-\varphi_X(\lambda)} = \mathbb{E}[X e^{\lambda X}] e^{-\varphi_X(\lambda)}$$

$$\mathbb{E}[Z^2] = \mathbb{E}[X^2 e^{\lambda X}] e^{-\varphi_X(\lambda)}$$

$$\varphi_X(\lambda) = \varphi_X(0) + \lambda . \varphi_X'(0) + \frac{\lambda^2}{2} . \varphi_X''(\lambda_0)$$

$$\varphi_X(0) = 0 \, , \quad \varphi_X'(\lambda) = \frac{\mathbb{E}[(X-\mu)e^{\lambda(X-\mu)}]}{\mathbb{E}[e^{\lambda(X-\mu)}]} = 0 \text{ if } \lambda = 0.$$

$$\varphi_X''(\lambda) = \frac{\mathbb{E}[(X-\mu)^2 e^{\lambda(X-\mu)}] e^{\varphi_X(\lambda)} - \mathbb{E}^2[(X-\mu)e^{\lambda(X-\mu)}]}{e^{2\varphi_X(\lambda)}} \, , \forall \lambda \in \mathbb{R}$$

$$= \frac{\mathbb{E}[X^2 e^{\lambda X}]}{e^{\varphi_X(\lambda)}} - \left(\frac{\mathbb{E}[X e^{\lambda X}]}{e^{\varphi_X(\lambda)}} \right)^2$$

$$= \mathbb{E}[Z_\lambda^2] - \mathbb{E}^2[Z_\lambda] = \operatorname{Var}(Z_\lambda)$$

Hence,

$$\varphi_X(\lambda) = \frac{\lambda^2}{2} . \operatorname{Var}(Z_{\lambda_0}) \quad \text{for some } \lambda_0.$$

Note that : $\underset{t \in \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}[|Z - t|^2] = \mathbb{E}Z$

and $\left| Z - \frac{a+b}{2} \right| \leq \frac{b-a}{2}$ since $Z \in [a,b]$.

Then,

$$\operatorname{Var}(Z_{\lambda_0}) \leq \mathbb{E}\left[\left| Z_{\lambda_0} - \frac{a+b}{2} \right|^2 \right] \leq \frac{(b-a)^2}{4}$$

$$\Rightarrow \quad \varphi_X(\lambda) \leq \frac{(b-a)^2}{4} . \frac{\lambda^2}{2} \quad \Rightarrow \quad X \text{ is sub-Gaussian with } (b-a)^2/4$$

Hoeffding's Lemma at work:

$X \in [a,b]$, $\mathbb{E}X = 0$ $\Rightarrow$ X is sub-Gaussian with $\frac{(b-a)^2}{4}$:

$$\varphi_X(\lambda) \leq \frac{\lambda^2}{8} \cdot (b-a)^2$$

$$\sup_{\lambda \in \mathbb{R}} \{ \lambda t - \varphi_X(\lambda) \} \leq \sup_{\lambda \in \mathbb{R}} \{ \lambda t - \frac{\lambda^2 (b-a)^2}{8} \}$$

$$t - \frac{\lambda^* \cdot (b-a)^2}{4} = 0 \Rightarrow \lambda^* = \frac{4t}{(b-a)^2} \quad \text{and}$$

$$\sup_{\lambda} \{ \lambda t - \frac{\lambda^2 (b-a)^2}{8} \} = \frac{2t^2}{(b-a)^2}$$

Hence,

$$\mathbb{P}(X > t) \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

<u>LEMMA</u> | X, Y sub-Gaussian with σ_X^2, σ_Y^2, they are independent

$\Rightarrow$ $X+Y$ is sub-Gaussian with $\sigma_X^2 + \sigma_Y^2$.

Also, for any $\alpha \in \mathbb{R}$, αX is sub-Gaussian with $\alpha^2 \sigma_X^2$.

<u>Pf</u>:

$$\mathbb{E}[e^{\lambda(X+Y)}] \underset{\substack{\uparrow \\ \text{independence}}}{=} \mathbb{E}[e^{\lambda X}] \cdot \mathbb{E}[e^{\lambda Y}] \leq e^{\lambda^2(\sigma_X^2 + \sigma_Y^2)/2}$$

$$\mathbb{E}[e^{\lambda \alpha X}] \leq e^{(\alpha\lambda)^2/2 \, \sigma_X^2} = e^{\alpha^2 \sigma_X^2 \cdot \frac{\lambda^2}{2}}, \quad \forall \lambda, \alpha \in \mathbb{R}$$

Then,

consider independent $X_i \in [a_i, b_i]$ with $\mathbb{E}X_i = 0$. Then,

$\frac{1}{n}\sum_{i=1}^n X_i$ is sub-Gaussian with $\frac{1}{n^2} \sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$.

$$\Rightarrow \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \leq \exp\left(-\frac{2t^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i < -t\right) = \mathbb{P}\left(\frac{1}{n}\sum_i (-X_i) > t\right) \leq \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Alternative form:

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^n X_i \right| > \sqrt{\frac{1}{2n^2} \sum_{i=1}^n (b_i - a_i)^2 \log(2/\delta)} \right) \leq \delta.$$