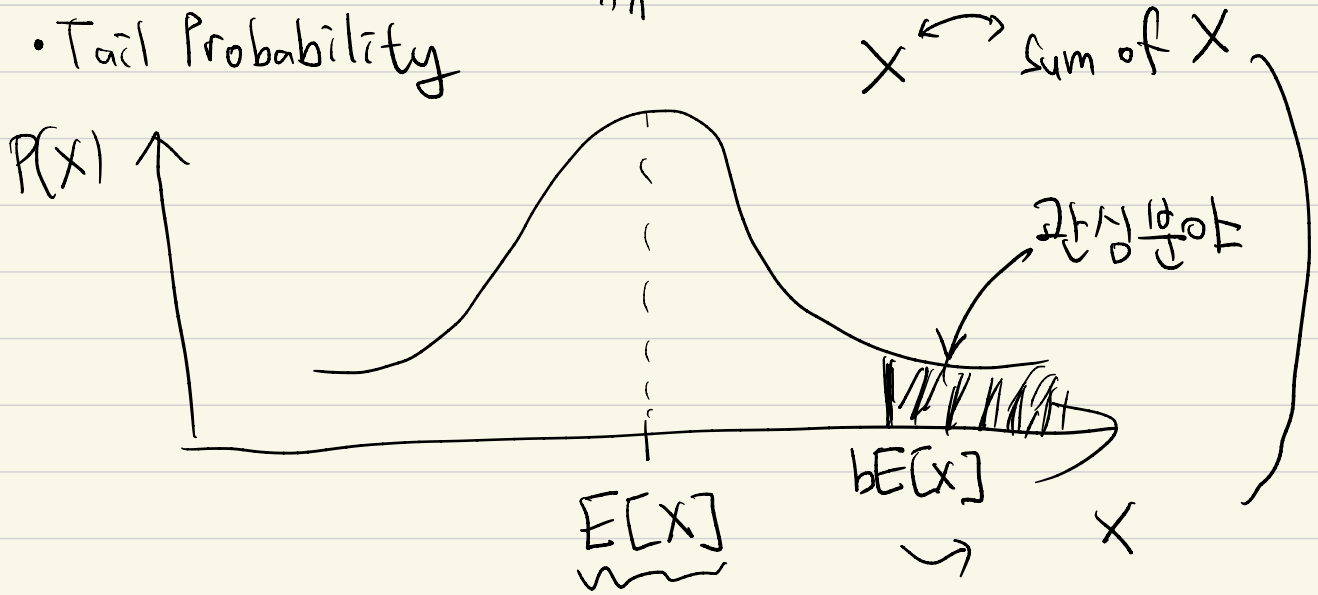
- Learning Theory 관련 내용 정리

210317

- Concentration Inequality ← Statistics

  - Tail Probability



- Concentration Property

  How the r.vs are concentrated around E[X] ?

Theorem.

- Markov Inequality

nonnegative X. any $a>0$

$$P[X \geq a] \leq \frac{E[X]}{a}$$

pf) $E[X] \geq E[X \mathbb{1}_{\{X \geq a\}}] \geq E[a \mathbb{1}_{\{X \geq a\}}] = aP(X \geq a)$

$X \geq a$

등호성립? $X=0$, a 인 ...

Corollary. $a = bE[X] \Rightarrow P(X \geq bE[X]) \leq \frac{1}{b}$ $(b>0)$

Rmk · 등호성립?

· nonnegative.

Corollary. Chebyshev's Inequality

$$P(|X - E[X]| \geq a) = P((X-E[X])^2 \geq a^2)$$

$$\leq \frac{E((X-E[X])^2)}{a^2} = \frac{Var(X)}{a^2}.$$

$a \leftarrow a\sigma$

$$P(|X-E[X]| \geq a\sigma) \leq \frac{Var(X)}{a^2\sigma^2} = \frac{1}{a^2}.$$

Theorem. Law of Large Numbers (Weak)

$S_n = X_1 + \cdots + X_n$

$$P\left(\left|\frac{1}{n}S_n - E[X]\right| \geq \varepsilon\right) \leq \frac{1}{n}\frac{Var(X)}{\varepsilon^2}$$

- Markov 사용법

X: RV, $t \in \mathbb{R}_+$ $(\Omega, \mathcal{F}, P)$: probability space

event

$E \in \mathcal{F}$, $E = \{X \geq t\} = \{w \in \Omega : X(w) \geq t\} = \{\phi(x) \geq \phi(t)\}$

under 특정 condition of ϕ.

ϕ를 적절하게 잡으면 Markov Ineq. 추정해주는 Bound가 다 낫다!

non-negative, non-decreasing ϕ

$$P(X \geq t) = P(\phi(X) \geq \phi(t))$$

$$\leq \frac{E[\phi(X)]}{\phi(t)} = \phi(t)^{-1} E[\phi(X)]$$

cf) $\phi(x) = x^q$ $q > 0$

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{E[(X - E[X])^q]}{\varepsilon^q}$$

- MGF, Chernoff Bounds

$\phi(x) = e^{\theta x}$ 로 두기 $(\theta > 0)$.

→ $P(X \geq t) \leq e^{-\theta t} \mathbb{E}[e^{\theta X}]$, (Chernoff Bound)

θ 값에 무관하므로, $P(X \geq t) \leq \inf_{\theta > 0} e^{-\theta t} \mathbb{E}[e^{\theta X}]$

MGF

- Recall MGF. ✓

MGF $M_X(\theta) := \mathbb{E}[e^{\theta X}] \rightarrow \mathbb{E}[X^n] = \frac{d^n}{d\theta^n} M_X(\theta)\Big|_{\theta=0}$

테일러전개시 증명 상기능

$(M_X(\theta))' = \mathbb{E}[(e^{\theta X})'] = \mathbb{E}[X e^{\theta X}]$

- Cumulant Function ✓

$\psi_X(\theta) := \log \mathbb{E}[e^{\theta X}]$ (로그 모멘 MGF)

비스무리하게

$\psi_X'(\theta) = \frac{(M_X(\theta))'}{\mathbb{E}[e^{\theta X}]} = \frac{\mathbb{E}[X e^{\theta X}]}{\mathbb{E}[e^{\theta X}]}$

- Cumulant function is convex. $\psi_X(\theta) = \log \mathbb{E}[e^{\theta X}]$

Hölder Inequality

$f, g \quad p, q > 0, \quad \frac{1}{p} + \frac{1}{q} = 1$

$$\int fg \leq \left(\int |f|^p \right)^{1/p} \left(\int |g|^q \right)^{1/q}$$

$p = q = 2$ Cauchy-Schwartz.

$\frac{1}{p} + \frac{1}{q} = 1$

$f(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda f(\theta_1) + (1-\lambda) f(\theta_2)$

$f \Leftarrow \mathbb{E}[e^{\theta X}]$ $\quad p = \frac{1}{\lambda} \quad q = \frac{1}{1-\lambda}$

$\mathbb{E}[e^{(\lambda\theta_1 + (1-\lambda)\theta_2)}]$

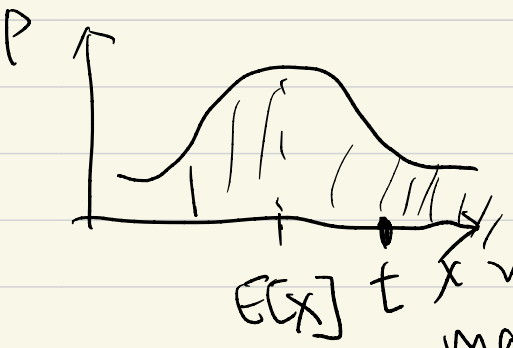
...

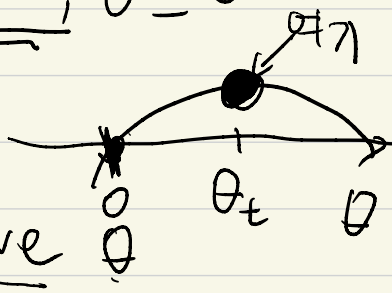
$\phi(x) = e^{\theta x}$

From Markov.. ineq

$$P(X \geq t) \leq e^{-\theta t} \cdot E[e^{\theta X}] = e^{-\theta t} \cdot e^{\log E[e^{\theta X}]}$$

$$= e^{-\theta t + \psi_X(\theta)}, \quad \theta \geq 0$$





$$\max_{\theta \geq 0} \left(t\theta - \psi_X(\theta) \right)$$ concave

$$\psi_X^*(t) := \sup_{\theta \geq 0} \{ t\theta - \psi_X(\theta) \}$$ Cramer-transform

$$P(X \geq t) \leq e^{-\psi_X^*(t)}$$

$\theta_t : t - \psi_X'(\theta) = 0$

$\psi_X'(\theta) = t$

$\psi_X'(\theta) = t$ iff $\theta \geq \theta_t$

$\theta = 0$: $t - \psi_X'(\theta) = t - E[X] > 0 \rightarrow \theta_t > 0$

$$\psi_X^*(t) = \sup_{\theta \in \mathbb{R}} \{ t\theta - \psi_X(\theta) \}$$

$\Updownarrow$

$\psi_X(\theta)$ & Fenchel-Legendre Dual. Conjugate

정리하면…

$$P(X \geq t) \leq e^{-t\theta + \psi_X(\theta)}$$

$$e^{-t\theta + \frac{1}{2}\theta^2\sigma^2}$$

① Gaussian $X \sim N(0, \sigma^2)$

$$\psi_X(\theta) = \log \mathbb{E}[e^{\theta X}] = \log\left(e^{\frac{1}{2}\theta^2\sigma^2}\right)$$

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

근데 정규분포는 두워서 사실 계산가능.

② 안 가우시안 시리즈

- Poisson $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$ X nonneg.

$$\mathbb{E}[e^{\theta X}] = e^{\lambda e^{\theta} - e\lambda - \lambda}$$

$$\psi_X^*(t) = \lambda h\left(\frac{t}{\lambda}\right) \quad \text{where } h(x) = (1+x)\log(1+x) - x\log x.$$

* Bernoulli $(p) \sim Y$

$Y = 1$ p (prob)
$= 0$ otherwise

- $\mathbb{E}[Y] = p$.

$X = Y - p$.

$X = 1-p$ p (prob)
$-p$ otherwise

$$\mathbb{E}[e^{\theta X}] = p \cdot e^{\theta(1-p)} + (1-p)e^{\theta(-p)} = e^{-p\theta}\left(pe^{\theta} + 1 - p\right)$$

$$\log \mathbb{E}[e^{\theta X}] = -p\theta + \log(pe^{\theta} + 1 - p) = \psi_X(\theta)$$

- $t\theta - \psi_X(\theta)$ 미분.

$$\psi_X(\theta) = \log(pe^\theta + 1-p) - p\theta$$

$$\frac{d}{d\theta}\left(t\theta + p\theta - \log(pe^\theta + 1-p)\right) = 0$$

$$\theta_E = \log \frac{(1-p)(p+t)}{p(1-p-t)} \qquad \underline{q := p+t}$$

$$= \log \frac{q\cdot(1-p)}{p\cdot(1-q)}$$

$$\Rightarrow \psi_X^*(t) = \underline{KL(P_q \| P_p)} \text{ 로 표현가능.}$$

$$= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$$

- 조건대 깨끗하게....

$$f(t) = (p+t)\log\frac{p+t}{p} + (1-p-t)\log\frac{1-p-t}{1-p}$$

$$f'(t) = \log\frac{p+t}{p} - \log\frac{1-p-t}{1-p}$$

$$\underline{f''(t)} = \underline{\frac{1}{(p+t)(1-p-t)}} = \frac{1}{a\cdot(1-a)} \geq \frac{4}{(a+(1-a))^2} = 4$$

$$\Rightarrow \underline{f(t)} = f(0) + f'(0)t + \frac{1}{2}f''(\xi)t^2 \geq \underline{2t^2}$$

$$P(X \geq t) \leq e^{-2t^2}.$$

- Independent sum. $\mathbb{E}[X_1]\mathbb{E}[X_2] = \mathbb{E}[X_1 X_2]$

$\mathbb{E}[X_i] = 0$, $X_1, X_2, \cdots$ (i.i.d), $S_n = X_1 + \cdots + X_n$

$$\mathbb{E}[e^{\theta S_n}] = \mathbb{E}[e^{\theta(X_1 + X_2 + \cdots + X_n)}] = \prod_{i=1}^{n} \mathbb{E}[e^{\theta X_i}]$$

$$\psi_{S_n}(\theta) = \sum_{i=1}^{n} \psi_{X_i}(\theta) = n\psi_{X_1}(\theta)$$

$$\psi_{S_n}^*(t) = \sup_\theta \{t\theta - n\psi_{X_1}(\theta)\}$$

$$= n \cdot \sup_\theta \left\{\left(\frac{t}{n}\right)\theta - \psi_{X_1}(\theta)\right\}$$

$$= n\psi_{X_1}^*\left(\frac{t}{n}\right)$$

$$P(S_n \geq t) \leq e^{-n\psi_{X_1}^*\left(\frac{t}{n}\right)}$$

- Bernoulli 인 경우.

$X_i \sim \text{Bernoulli}(p) \Rightarrow S_n \sim \text{Binomial}(n, p)$

$$P(S_n - np \geq t) \leq e^{-n \cdot \psi_{X_1}^*\left(\frac{t}{n}\right)}$$

$$\leq e^{-2n\left(\frac{t}{n}\right)^2} = e^{\frac{-2t^2}{n}} \quad \checkmark$$

$t = n \cdot \ell$

$$= e^{-2n\ell^2}$$

• 다는 부분

- Sub Gaussian RV. $(sG(\sigma))$

Theorem 2.1.1. *Let X be a centered random variable on $\mathbb{R}$, each statement bellow implies the next (we take $\sigma^2 > 0$ in the first definition as a variance proxy).*

- *Laplace transform: for any $s \in \mathbb{R}$, $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$.*
- *Concentration: for any $t > 0$, $\max\{\mathbb{P}(X \geq t), \mathbb{P}(X \leq -t)\} \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)$.*
- *Moment condition: for any $q \in \mathbb{N}^*$, $\mathbb{E}\left[X^{2q}\right] \leq q!(4\sigma^2)^q$.*
- *Orlicz condition: $\mathbb{E}\left[\exp\left(\frac{X^2}{8\sigma^2}\right)\right] \leq 2$.*
- *Laplace transform: for any $t \in \mathbb{R}$, $\mathbb{E}[\exp(tX)] \leq \exp\left(\frac{24\sigma^2 t^2}{2}\right)$.*

$\mathbb{E}[e^{\theta X}] \leq e^{\frac{1}{2}\theta^2\sigma^2}$, $\psi_X(\theta) \leq \frac{1}{2}\theta^2\sigma^2$ 인 분포

sub-gaussian!

$\psi_X^*(t) = \sup_\theta \{t\theta - \psi_X(\theta)\} \geq \sup_\theta \{t\theta - \frac{1}{2}\theta^2\sigma^2\} = \frac{t^2}{2\sigma^2}$

이것으로 $\mathbb{P}(X \geq t) \leq e^{-\psi_X^*(t)} \leq e^{-\frac{t^2}{2\sigma^2}}$ (Thm 2.1)

- Sum of i.i.d vars

$\mathbb{E}[e^{\theta S_n}] \leq e^{\frac{1}{2}\theta^2 \sum_{i=1}^n \sigma_i^2}$, $\mathbb{P}(S_n \geq t) \leq e^{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}}$

- Bounded RVs. $X_i \in [a,b]$

(Claim) Bounded RVs are sub-Gaussian RV

(Hoeffding Lemma)

$Y \in [a,b]$, $\mathbb{E}[Y]=0$ $\quad \frac{1}{2}\theta^2\sigma^2$

$$\psi_Y(\theta) \le \frac{1}{8}(b-a)^2\theta^2, \ \frac{2}{1} \quad Y \sim SG\left(\tfrac{1}{2}(b-a)\right)$$

pf) $\left|Y - \frac{a+b}{2}\right| \le \frac{b-a}{2}$.

$$\mathrm{Var}(Y) = \mathrm{Var}\left(Y - \frac{b+a}{2}\right) \le \mathbb{E}\left[\left(Y-\frac{a+b}{2}\right)^2\right] \le \left(\frac{b-a}{2}\right)^2$$

$\mathbb{E}[e^{\theta Y}]$ well-defined, $\psi_Y(\theta)$ differentiable

$$\psi_Y'(\theta) = \frac{\mathbb{E}[Ye^{\theta Y}]}{\mathbb{E}[e^{\theta Y}]} = \mathbb{E}\left[Y e^{\theta Y - \psi_Y(\theta)}\right]$$

$\mathbb{E}[e^{\theta Y}] = e^{\log \mathbb{E}[e^{\theta Y}]} = e^{\psi_Y(\theta)}$

- Brilliant idea: $Q(A) := \mathbb{E}\left[\mathbb{1}_A \cdot e^{\theta Y - \psi_Y(\theta)}\right]$ event A.

$\Rightarrow$ Q becomes probability measure!!!!!!!!

$$P(A) = \mathbb{E}[\mathbb{1}_A]$$

$$Q(\Omega) = \mathbb{E}[\ ] = 1$$

Ω, F, P

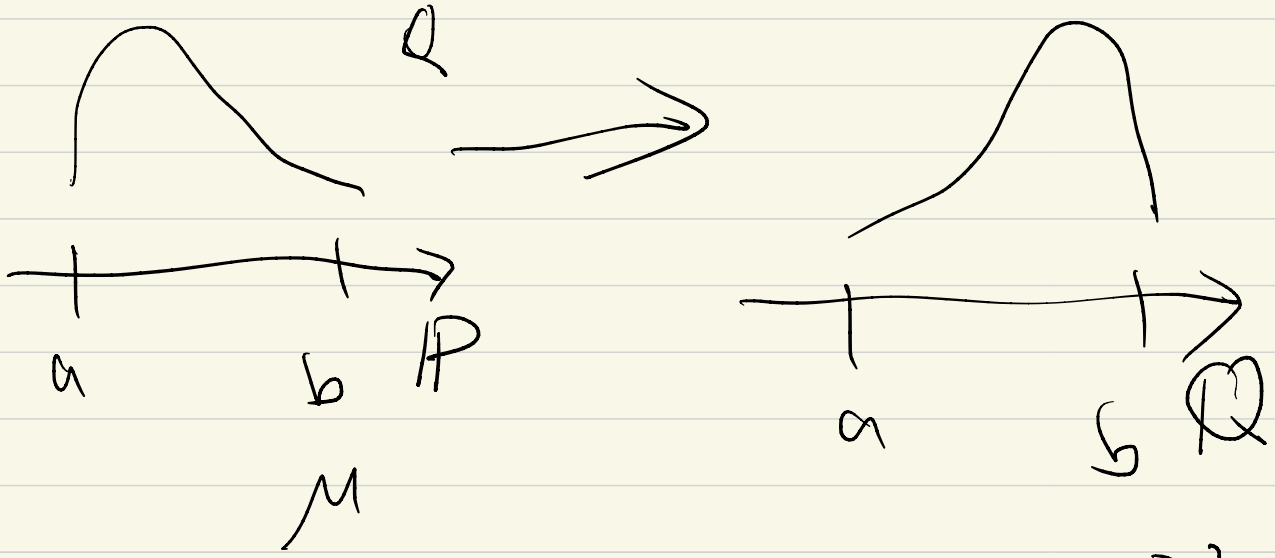
RmK.

Radon-Nikodym Theorem.

measurable space (X, Σ). absolutely continuous

$\nu << \mu$. ($\mu(A)=0$ implies $\nu(A)=0$)

$\nu(A) = \int_A f \, d\mu \Rightarrow f := \boxed{\frac{d\nu}{d\mu}} \geq 0$



$\mu(A)=0 \rightarrow \nu(A)=0$

$$\nu(A) = \int_A f \, d\mu = \int_A \frac{d\nu}{d\mu} \, d\mu$$

$P \Rightarrow Q$ measure change $dQ = e^{\theta Y - \psi_Y(\theta)} dP$.

$$\psi_Y'(\theta) = E[Y e^{\theta Y - \psi_Y(\theta)}] = E_Q[Y]$$

$$\psi_Y''(\theta) = \left(\frac{E[Y e^{\theta Y}]}{E[e^{\theta Y}]} \right)' = \frac{E[Y^2 e^{\theta Y}] E[e^{\theta Y}] - E[Y e^{\theta Y}]^2}{E[e^{\theta Y}]^2}$$

$$= \frac{E[Y^2 e^{\theta Y}]}{E[e^{\theta Y}]} - \left(\frac{E[Y e^{\theta Y}]}{E[e^{\theta Y}]} \right)^2$$

$$= E_Q[Y^2] - E_Q[Y]^2 = \mathrm{Var}_Q[Y]$$

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

$$\psi_Y(\theta) = \underbrace{\psi_Y(0)}_{0} + \underbrace{\psi_Y'(0)}_{0}\theta + \frac{1}{2}\psi_Y''(\xi)\theta^2 \leq \frac{(b-a)^2}{8}\theta^2$$

$\frac{b-a}{2}$ とは？

Radamacher distribution

$+1$	$\frac{1}{2}$
-1	$\frac{1}{2}$

■

Thm (Hoeffding)

$Y_i \in [a_i, b_i]$ $\quad Y_i - \mathbb{E}[Y_i] \in [a_i - \mathbb{E}[Y_i],\ b_i - \mathbb{E}[Y_i]]$

$$P\left(\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) \geq t\right) \leq e^{-\frac{t^2}{2\sum_{i=1}^{n}\left(\frac{b_i-a_i}{2}\right)^2}}$$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

- Bernoulli 인 경우 위에서 했던 대로 하고 $a_i = 0,\ b_i = 1$ 해주면

- 사실 variance of S_n $\sum_{i=1}^{n}(b_i-a_i)^2$ 보다 작은경우가 많다!

$\Rightarrow$ Bernstein, Bennett (Boucheron, (2.10))

MGF에 더 tighter bound 주는 것이 가능함 .....

- Martingale 마팅게일, Hoeffding (1963)

$$\mathbb{E}[X_{n+1} \mid \underbrace{X_0, X_1, \cdots, X_n}_{\sigma(X_0, \cdots, X_n) = \mathcal{F}_n \text{ (information ...)}}] = X_n \qquad \text{ex) 승식!}$$

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n.$$

• Azuma-Hoeffding

$X_0, X_1, \cdots$ martingale $\quad a_i \le X_i - X_{i-1} \le b_i$ 이고 $\forall i$.

$$Y_i := X_i - X_{i-1}$$

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[X_{k+1} - X_k \mid \mathcal{F}_k] = \mathbb{E}[X_{k+1} \mid \mathcal{F}_k] - X_k$$

$$S_k := \sum_{i=1}^{k} Y_i \qquad = 0$$

Hoeffding's Lemma: $\mathbb{E}[e^{\theta Y_{k+1}} \mid \mathcal{F}_k] \le e^{\theta^2 (b_k - a_k)^2 / 8}$

$$\mathbb{E}[e^{\theta S_k}] = \mathbb{E}\left[\mathbb{E}[e^{\theta S_k} \mid \mathcal{F}_{k-1}]\right] \quad \text{(Tower Rule)}$$

$$= \mathbb{E}[e^{\theta S_{k-1}} \cdot \mathbb{E}[e^{\theta Y_k} \mid \mathcal{F}_{k-1}]] \le \mathbb{E}[e^{\theta S_{k-1}}] e^{\frac{\theta^2 (b_k - a_k)^2}{8}}$$

$$\cdots\cdots \le \exp\left(\frac{1}{2}\theta^2 \sum_{i=1}^{n} \left(\frac{b_i - a_i}{2}\right)^2\right)$$

$$\therefore S_k = \sum_{i=1}^{k} Y_i = \sum_{i=1}^{k} (X_i - X_{i-1}) = X_k - X_0,$$

$$X_k - X_0 \sim SG\left(\sqrt{\frac{1}{4}\sum_{i=1}^{m_k}(b_i - a_i)^2}\right)$$

나머진 Hoeffding의 증명과정과 동일.

- McDiarmid Inequality.

- Reference
  - Subgaussian random variables: An expository note
    Omar Rivasplata
  - Boucheron, Concentration Inequalities, Oxford 2012
  - course AI603 KAIST, course notes
    MAS480