Basic Tail Inequality 2

Wei Xiong

Department of Mathematics
The Hong Kong University of Science and Technology

2022.2.23

1 Review and New

Review

- Tail inequality is implied by finite rate of M.G.F.;
- Sub-Gaussianity A random variable X is σ^2-sub-Gaussian if for all $\lambda \in R$ it holds that

$$E[\exp(\lambda(X - EX))] \leq \exp(\frac{\lambda^2\sigma^2}{2})$$

- Logarithmic moment generating function:

$$\Lambda_X(\lambda) = \ln \mathbb{E} e^{\lambda X}$$

- Rate function:

$$I_X(z) = \begin{cases} \sup_{\lambda > 0} [\lambda z - \Lambda_X(\lambda)] & z > \mu \\ 0 & z = \mu \\ \sup_{\lambda < 0} [\lambda z - \Lambda_X(\lambda)] & z < \mu \end{cases}$$

Asymptomatic Tightness of Bound from Rate Function

- Bound implied by the rate function:

$$\frac{1}{n} \ln \Pr\left(\bar{X}_n \geq \mu + \epsilon\right) \leq -I_{X_1}(\mu + \epsilon) = \inf_{\lambda > 0} \left[-\lambda\epsilon + \ln \mathbb{E} e^{\lambda(X_1 - \mu)}\right]$$
$$\frac{1}{n} \ln \Pr\left(\bar{X}_n \leq \mu - \epsilon\right) \leq -I_{X_1}(\mu - \epsilon) = \inf_{\lambda < 0} \left[\lambda\epsilon + \ln \mathbb{E} e^{\lambda(X_1 - \mu)}\right]$$

- The bound is asymptomatic optimal: for any $\epsilon' > \epsilon$, we have

$$\varliminf_{n \to \infty} \frac{1}{n} \ln \Pr\left(\bar{X}_n \geq \mu + \epsilon\right) \geq -I_{X_1}\left(\mu + \epsilon'\right).$$
$$\varliminf_{n \to \infty} \frac{1}{n} \ln \Pr\left(\bar{X}_n \leq \mu - \epsilon\right) \geq -I_{X_1}\left(\mu - \epsilon'\right).$$

Proof of Tightness

We define r.v.s with density: $d\Pr(X_i' \leq x) = e^{\lambda x - \Lambda_{X_1}(\lambda)} d\Pr(X_i \leq x)$. We have

$$\frac{d}{d\lambda}\Lambda_{X_1}(\lambda) = \frac{\int xe^{\lambda x} d\Pr(X_1' \leq x)}{\int e^{\lambda x} d\Pr(X_1' \leq x)} = \int x \frac{e^{\lambda x}}{\int e^{\lambda x} d\Pr(X_1' \leq x)} d\Pr(X_1' \leq x) = \mathbb{E}_{X_1'} X_1'$$

Here we take $\lambda = \arg\max_\lambda [\lambda(\mu + \epsilon') - \Lambda_{X_1}(\lambda)]$. By setting the derivative to 0, we have

$$\mathbb{E}_{X_1'} X_1' = \frac{d}{d\lambda}\Lambda_{X_1}(\lambda) = \mu + \epsilon'.$$

By LLN, for any $\epsilon'' > \epsilon' > \epsilon$, we have

$$\lim_{n\to\infty} \Pr\left(\bar{X}_n' - \mu \in \left[\epsilon, \epsilon''\right]\right) = 1.$$

Proof of Tightness

Recall $d\Pr(X_i' \le x) = e^{\lambda x - \Lambda_{X_1}(\lambda)} d\Pr(X_i \le x)$, we have

$$e^{-\lambda \sum_i x_i + n\Lambda_{X_1}(\lambda)} \prod_i d\Pr\left(X_i' \le x_i\right) = \prod_i d\Pr\left(X_i \le x_i\right)$$

Then,

$$\begin{aligned}
\Pr\left(\bar{X}_n \ge \mu + \epsilon\right) &\ge \Pr\left(\bar{X}_n - \mu \in \left[\epsilon, \epsilon''\right]\right) \\
&= \mathbb{E}_{X_1,\ldots,X_n} I\left(\bar{X}_n - \mu \in \left[\epsilon, \epsilon''\right]\right) \\
&= \mathbb{E}_{X_1',\ldots,X_n'} e^{-\lambda n \bar{X}_n' + n\Lambda(\lambda)} I\left(\bar{X}_n' - \mu \in \left[\epsilon, \epsilon''\right]\right) \\
&\ge e^{-\lambda n \epsilon'' + n\Lambda(\lambda)} \Pr\left(\bar{X}_n' - \mu \in \left[\epsilon, \epsilon''\right]\right)
\end{aligned}$$

Since the chosen λ is argmax, by the definition of rate function, we have $-\lambda n\epsilon' + n\Lambda(\lambda) = -nI\left(\mu + \epsilon'\right)$.This implies

$$\frac{1}{n} \ln \Pr\left(\bar{X}_n \ge \mu + \epsilon\right) \ge -I\left(\mu + \epsilon'\right) - \lambda\left(\epsilon'' - \epsilon'\right) + \frac{1}{n} \ln \Pr\left(\bar{X}_n' - \mu \in \left[\epsilon, \epsilon''\right]\right).$$

Letting $\epsilon'' \to \epsilon'$ and $n \to \infty$ concludes our proof.

2 Sub-Exponential Random Variable

Sub-Exponential Random Variable

- A random variable X is said to be sub-exponential with parameter (τ^2, b) if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(\frac{\lambda^2\tau^2}{2}), \forall|\lambda| \leq \frac{1}{b}. \tag{0.1}$$

- σ^2-sub-Gaussian r.v. is $(\sigma^2, 0)$-sub-exponential.
- M.G.F. condition holds in a neighbor of 0;
- Tail Inequality:

$$\begin{aligned} P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t^2}{2\tau^2}), 0 \leq t \leq \frac{\tau^2}{b}; \\ P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t}{2b}), t > \frac{\tau^2}{b}. \end{aligned} \tag{0.2}$$

- Given independent (τ_i^2, b_i) sub-exponential r.v.s. and $a \in \mathbb{R}^n$:

$$\mathbb{E}[\exp(\lambda\sum_{i=1}^{n} a_i X_i)] \leq \exp(\frac{\lambda^2 \sum_{i=1}^{n} a_i^2\sigma_i^2}{2}), |\lambda| \leq \frac{1}{\max|b_i a_i|}, \tag{0.3}$$

Proof of Tail Inequality

- (τ^2, b)-Sub-Gaussianity:$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(\frac{\lambda^2\tau^2}{2}), \forall|\lambda| \leq \frac{1}{b}$.
- Tail Inequality:

$$
\begin{aligned}
P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t^2}{2\tau^2}), 0 \leq t \leq \frac{\tau^2}{b};\\
P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t}{2b}), t > \frac{\tau^2}{b}.
\end{aligned}
\tag{0.4}
$$

We start with $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}} \leq \exp\left(\frac{\lambda^2\tau^2}{2} - \lambda t\right)$. It remains to minimize $g(\lambda) = \frac{\lambda^2\tau^2}{2} - \lambda t$. **Case** 1: $0 \leq t < \frac{\tau^2}{b}$, i.e., $\frac{t}{\tau^2} < \frac{1}{b}$. So, $\min_\lambda g(\lambda) = g(\frac{t}{\tau^2}) = -\frac{t^2}{2\tau^2}$. **Case** 2: $t/\tau^2 \geq \frac{1}{b}$. g is monotonically decreasing in $[0, \lambda^*)$, the constrained minimum occurs at $\frac{1}{b}$ and we have

$$
\min_\lambda g(\lambda) = -\frac{t}{b} + \frac{1}{2b}\frac{\tau^2}{b} \leq -\frac{t}{2b}
$$

where the last inequality uses the fact that $t/\tau^2 \geq \frac{1}{b}$.

Proof of Linear Combination Property

We aim to proof

$$\mathbb{E}[\exp(\lambda \sum_{i=1}^{n} a_i X_i)] \leq \exp(\frac{\lambda^2 \sum_{i=1}^{n} a_i^2 \sigma_i^2}{2}), |\lambda| \leq \frac{1}{\max |b_i a_i|}. \tag{0.5}$$

Proof.

Inductively applying the condition for each random variable is sufficient. We first note that

$$\mathbb{E}\left[\exp\left(\lambda a_i X_i\right)\right] \leq \exp\left(\frac{\lambda^2 a_i^2 \sigma_i^2}{2}\right), |\lambda a_i| \leq \frac{1}{b_i}.$$

Then, it holds that

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} a_i X_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda a_i X_i\right)\right] \leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 a_i^2 \sigma_i^2}{2}\right), \forall |\lambda| \leq \frac{1}{\max |b_i a_i|}.$$

□

Sub-Exponential: Bernstein Condition

- A random variable X with mean μ and variance σ^2 is said to satisfy the Bernstein condition if

$$\mathbb{E}(X-\mu)^k \leq \frac{k!}{2}\sigma^2 b^{k-2}, k \geq 2 \tag{0.6}$$

- Berinstein r.v. is $(\sqrt{2}\sigma, 2b)$-sub-exponential.
- Along the line, we have

$$\begin{aligned} \mathbb{E}\exp(\lambda(X-\mu)) &\leq \exp(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}), \forall|\lambda| < \frac{1}{b} \\ P(|X-\mu| \geq t) &\leq 2\exp(-\frac{t^2}{2(\sigma^2+bt)}), \forall t \geq 0 \\ P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t) &\leq \exp(-\frac{nt^2}{2(\sigma^2+bt)}) \end{aligned} \tag{0.7}$$

Proof of Bernstein-Type Bound.

$$\mathbb{E}\exp(\lambda(X-\mu)) \leq \exp(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}), \forall|\lambda| < \frac{1}{b}$$

$$P(|X-\mu| \geq t) \leq 2\exp(-\frac{t^2}{2(\sigma^2+bt)}), \forall t \geq 0$$

Proof.

W.L.O.G., we prove for $\mu = 0$. By $e^2 \geq 1+x$,

$$\mathbb{E}[\exp(\lambda X)] = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k EX^k}{k!} \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}$$

$$\leq 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-|\lambda|b} \leq \exp(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)})$$

The tail inequality:

$$P(X-\mu \geq t) = P(\exp(\lambda(X-\mu)) \geq e^{\lambda t}) \leq \exp(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} - \lambda t), \forall|\lambda| < \frac{1}{b},$$

Setting $\lambda = \frac{t}{bt+\sigma^2} < \frac{1}{b}$ concludes the proof. $\square$

Proof of Bernstein-Type Bound.

Berinstein r.v. is $(\sqrt{2}\sigma, 2b)$-sub-exponential and

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t) \leq \exp(-\frac{nt^2}{2(\sigma^2 + bt)})$$

Proof.

For $|\lambda| < \frac{1}{2b}$, we have

$$\mathbb{E}[\exp(\lambda X)] = \leq \exp(\frac{\lambda^2\sigma^2}{2(1 - b|\lambda|)}) \leq \exp(\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}).$$

Finally, we have

$$\mathbb{E}\exp(\lambda(\frac{1}{n}\sum_{i=1}^{n} X_i)) \leq \prod_{i=1}^{n} \exp(\frac{\lambda^2 \frac{\sigma^2}{n^2}}{2(1 - b|\lambda|/n)}) = \exp(\frac{\mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i)\lambda^2}{2(1 - (b/n)|\lambda|)})$$

Therefore, $\frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies Bernstein condition with $\frac{b}{n}$ and $\frac{1}{n}\sigma^2$. $\square$

Discussion.

- Bounded r.v. $|X - \mu| \leq b$ satisfies Bernstein condition with $\sigma^2 = Var(X)$.
- Let $X_i \in [a, b]$ and $R = b - a$:

$$\begin{aligned} \text{Hoeffding}: &\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \leq \frac{R}{\sqrt{n}}\sqrt{\frac{\log(1/\delta)}{2}} = \tilde{\mathcal{O}}(\frac{R}{\sqrt{n}}) \\ \text{Bernstein}: &\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \leq \frac{2\sqrt{\sigma^2 \log(1/\delta)}}{\sqrt{n}} + \frac{4R}{3}\frac{\log(1/\delta)}{n} = \tilde{\mathcal{O}}(\frac{\sigma}{\sqrt{n}} + \frac{R}{n}) \end{aligned} \tag{0.8}$$

- Bernstein's inequality is superior if the variance is small.

3 Martingale Concentration

Motivation

- Empirical Risk Minimization (ERM):

$$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n, \quad \mathcal{F} = \{f_i : i \in \mathcal{I}\}, \quad \ell(y', Y_i) \in [0, 1];$$

  - $\hat{f} := \mathrm{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$;
  - Bound $L(\hat{f}) - L(f^*) \leq c(\sup_{f \in \mathcal{F}} \hat{L}(f) - L(f))$.
- Online Learning: K distributions supported on $[0, 1]$ with means μ_i
  - For $t = 1, \cdots, T$, select $a(t) \in [K]$ and observe $r(t) \sim P_{a(t)}$;
  - We assume $r(t)$ are independent across $t \in [T]$;
  - Sample mean estimator: with $N_k(t) = \sum_{i=1}^t I(a(i) = k)$,

$$\bar{X}_k(t) := \frac{1}{N_k(t)} \sum_{i=1}^t r(i) I(a(i) = k)$$

  - Problem: $a(t)$ depends on $\{a(1), r(1), \cdots, a(t-1), r(t-1)\}$.

Martingale Difference

- $\{D_t = X_t - \mathbb{E}X_t\}_{t=1}^T$ is a Martingale Difference;
- if D_t is conditionally sub-Gaussian/Exponential:

$$\mathbb{E}[\exp(\lambda D_t)|\mathcal{F}_{t-1}] \leq \exp(\frac{\lambda^2\sigma_t^2}{2}),$$

Then,

$$\mathbb{E}\left[e^{\lambda\left(\sum_{t=1}^n D_t\right)}\right] = \mathbb{E}\left[e^{\lambda\left(\sum_{t=1}^{n-1} D_t\right)}\mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \leq \mathbb{E}\left[e^{\lambda \sum_{t=1}^n D_k}\right] e^{\lambda^2\sigma_{n-1}^2/2} \leq \cdots$$

Azuma-Hoeffding

- Let $\{D_k, \mathcal{F}_k\}$ be a martingale difference $D_k \in [a_k, b_k]$ for all $k \geq 1$:

$$P(|\sum_{k=1}^{n} D_k| \geq t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2})$$

- If $X_i - \mathbb{E}X_i \leq R$ and $var(D_k|\mathcal{F}_{k-1}) \leq \sigma_k^2$:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2 + \frac{2}{3}Rt}\right)$$

- Let $\mathcal{F}_{t-1}$ be the filtration induced by $\{a(1), r(1), \cdots, a(t)\}$. Then,

$$\mathbb{E}[r(t) - \mathbb{E}r_{a(t)}(t)|\mathcal{F}_{t-1}] = 0$$

because the expectation captures the randomness only over $P_{a(t)}$;

Condition for Online Learning

- Let $S_t = \{Z_1, \cdots, Z_t\}$ and define random functionals $\xi_i(S_i)$. Then,

$$\mathbf{E}_{S_n} \exp\left(\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \ln \mathbf{E}_{Z_i} e^{\xi_i}\right) = 1$$

- e.g. $\ln \mathbf{E}_{Z_i} e^{\lambda \xi_i} \leq \lambda \mathbf{E}_{Z_i} \xi_i + \frac{\lambda^2 \sigma_i^2}{2}$ implies Azuma-Hoeffding where σ_i can depend on S_{i-1}.

4 Functions Beyong Linear Combination

Lipschitz functions of Gaussian variables

- Let f be L-Lipshitz: $\|f(x) - f(y)\| \leq L\|x - y\|$;
- Let X_i be i.i.d. standard normal r.v.s;
- Then $f(X) - \mathbb{E}f(X)$ is sub-Gaussian with parameter at most L:

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}}$$

- Any L-Lipschitz function of a standard Gaussian random vector behaves like $N(\mu_0, L^2)$;