A note on exponential inequality

Wei Xiong *

December 10, 2021

1 Introduction

We decide to (re)write this note because the author realizes that he is too vegetable to conduct some hard and rigorous analysis.

We start with the famous concentration results in asymptomatic analysis.

Theorem 1. *Weak Law of Large Number. Let $\{X_n\}$ be a sequence of i.i.d. random variables with expectation $\mathbb{E}X_1$. Then,*

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbb{E}X_1 \to 0, \tag{1.1}$$

in probability.

Actually, this may hold even though the expectation or the second moment of X does not exist. However, practically, we may concern the non-asymptomatic analysis which means that we are given only finitely many samples and n will not go to infinity.

2 Sub-Gaussian random variable

We start with the famous Markov's inequality.

Theorem 2. *Markov's inequality. Let X be a non-negative r.v. in the sense that $X \geq 0$ w.p. 1. Then,*

$$P(X \geq t) \leq \frac{\mathbb{E}X}{t} \tag{2.1}$$

We note that in general, Markov's inequality and Chebyshev's inequality are sharp in the sense that we can find some distribution for which the bound is tight. However, in many cases, we can improve the $O(\frac{1}{t})$ rate (or $O(\frac{1}{t^2})$ rate of Chebyshev's inequality) to an $\exp(-t)$ rate. For instance,

*The Hong Kong University of Science and Technology; email: wxiongae@connect.ust.hk.

the standard normal distribution:

$$\begin{aligned}
\int_x^\infty \phi(t)\mathrm{d}t &= \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\mathrm{d}t \\
&= \int_x^\infty \frac{1}{t}\frac{1}{\sqrt{2\pi}} t \cdot \exp(-t^2/2)\mathrm{d}t \\
&= -\frac{1}{t}\frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\Big|_x^\infty - \int_x^\infty \left(-\frac{1}{t^2}\right)\left(-\frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\right) \mathrm{d}t \\
&= \frac{\phi(x)}{x} - \int_x^\infty \frac{\phi(t)}{t^2}\mathrm{d}t \\
&\leq \frac{\phi(x)}{x},
\end{aligned}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. Therefore, some random variables can achieve an rate of $O(\exp(\text{poly(t)}))$. We want to find them beyond the normal distribution. We then motivate the sub-Gaussian random variables through a discussion for the moment generating function.

Theorem 3. *Inequality induced by generating function. If a random variable has a moment generating function $\phi(\lambda) = \mathbb{E}[e^{\lambda X}]$, for all $\lambda > 0$, we have*

$$P(X \geq t) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}} = \phi(\lambda) e^{-\lambda t} \tag{2.2}$$

Proof.

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda t)}$$

It also holds that

$$P(X - \mathbb{E}X \geq t) = P(e^{\lambda(X - \mathbb{E}X)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}\exp(\lambda(X - \mathbb{E}X))}{\exp(\lambda t)}.$$

Since we can replace $X - \mathbb{E}X = Y$, we only need to consider the mean-zero random variables. □

Remark 1. *This theorem is super important because it motivates us to consider random variables with an upper bound on the moment generating function. Moreover, the technique*

$$P(X - \mathbb{E}X > t) = P(\exp(\lambda(X - \mathbb{E}X) > \exp(\lambda t)) \leq \frac{\mathbb{E}\exp(\lambda(X - \mathbb{E}X))}{\exp(\lambda t)}$$

is standard and will be used throughout the rest of this note.

This result implies that an upper bound for $\mathbb{E}\lambda(X - \mathbb{E}X)$ leads to an upper bound for the tail

probability. For instance, for the normal distribution $N(0, \sigma^2)$, we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\lambda x - \frac{1}{2\sigma^2}x^2\right) dx \\ &= e^{\frac{\lambda^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \lambda\sigma^2 x\right)^2\right) dx \\ &= \exp(\frac{\lambda^2\sigma^2}{2})\end{aligned}$$

Therefore, $N(0, \sigma^2)$ achieves a rate of $O(\exp(-t^2))$ if we take $\lambda = \frac{t}{\sigma^2}$:

$$P(X \geq \mathbb{E}X + t) \leq \exp(-\frac{t^2}{2\sigma^2}).$$

This motivates us to consider the class of sub-Gaussian random variables whose moment generating functions are bounded.

Definition 1. *Sub-Gaussian random variable. A random variable is said to be sub-Gaussian with parameter σ^2 if*

$$\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(-\frac{\lambda^2\sigma^2}{2}), \forall\lambda \in \mathbb{R}. \tag{2.3}$$

The definition requires that the moments of X exist and grow mildly because we have

$$\mathbb{E}\exp(\lambda X) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}X^k.$$

Theorem 4. *Tail bound of Sub-Gaussian random variable.*

$$\begin{aligned}P(X \geq \mathbb{E}X + t) &\leq \exp(-\frac{t^2}{2\sigma^2}), \\ P(X \leq \mathbb{E}X - t) &\leq \exp(-\frac{t^2}{2\sigma^2}).\end{aligned} \tag{2.4}$$

Proof. Using similar Chernoff-type techniques as in theorem 3 and minimized w.r.t. λ as in the normal distribution case. □

One important thing is that sub-Gaussian random variables are closed under linear combination.

Theorem 5. *Linear combination of sub-Gaussian random variables.*

- *If $X_1, \cdots, X_n$ are independent sub-Gaussian with parameter $\sigma_1^2, \cdots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with parameter $\sum_{i=1}^n \sigma_i^2$;*
- *If X is sub-Gaussian with parameter σ^2, then cX is sub-Gaussian with parameter $c^2\sigma^2$.*

Consequently, we have

$$P(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}), \qquad P(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}) \tag{2.5}$$

Proof. The second property is easy to verify. The first one is because

$$\begin{aligned}\mathbb{E}\exp(\lambda(Z-\mathbb{E}Z)) &= \mathbb{E}\left[\prod_{i=1}^{n}\exp(\lambda(X_i-\mathbb{E}X_i)\right] \\ &= \prod_{i=1}^{n}\mathbb{E}\exp(\lambda(X_i-\mathbb{E}X_i) \\ &\leq \exp(\lambda^2\frac{\sum_{i=1}^{n}\sigma_i^2}{2}).\end{aligned}$$

□

Examples:

- $N(\mu,\sigma^2)$ is σ^2-sub-Gaussian;
- Bounded random variable on $[a,b]$ is $\frac{(b-a)^2}{4}$-sub-Gaussian;

The bounded random variable deserves a theorem!

Theorem 6. *Hoeffding's inequality. Let $X_1,\cdots,X_n$ be independent random variables s.t. X_i is supported on $[a_i,b_i]$. Then,*

$$P(\sum_{i=1}^{n}(X_i-\mathbb{E}X_i)\geq t)\leq\exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}),\qquad P(\sum_{i=1}^{n}(X_i-\mathbb{E}X_i)\leq -t)\}\leq\exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}) \tag{2.6}$$

Remark 2. *There are also results related to the concentration of functions of random variables (beyond the linear combination). For instance, the McDiarmid's inequality. However, since the authors do not know much about them, they are omitted here.*

3 Sub-Exponential random variable

A slightly weaker condition than sub-Gaussianity is sub-exponentiality, which, for a mean-zero random variable, means that its moment generating function exists in a neighborhood of zero. Indeed, if the random variables have small variance, we would like to see it reflected in the exponential tail bound where the variance does not appear in Hoeffding's inequality.

We begin with an example of Laplace distribution with parameter 1: $f(x)=\frac{1}{2}\exp(-|x|$.

$$\begin{aligned}&P(|X|>t)=\exp(-t), t\geq 0 \\ &\mathbb{E}\exp(sX)=\frac{1}{1-s^2}, |s|<1 \mathbb{E}\exp(sX) \quad =\exp(2s^2), |s|<\frac{1}{2}\end{aligned} \tag{3.1}$$

Clearly, it is not sub-Gaussian because its moment generating function does not exist for $|s|>1$. The tails of this distribution do not decay as fast as the Gaussian variables. However, we can still find some useful bound through its moment generating function using similar technique for sub-Gaussian random variables

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

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(\frac{\lambda^2\tau^2}{2}), \forall|\lambda| \leq \frac{1}{b}. \tag{3.2}$$

Therefore, σ^2-sub-Gaussian r.v. is $(\sigma^2, 0)$-sub-exponential.

Theorem 7. *Tail bound of Sub-exponential random variable. Suppose that X is a sub-exponential with parameters (τ^2, b), then it holds that*

$$\begin{aligned} P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t^2}{2\tau^2}), 0 \leq t \leq \frac{\tau^2}{b}; \\ P(X - \mathbb{E}X \geq t) &\leq \exp(-\frac{t}{2b}), t > \frac{\tau^2}{b}. \end{aligned} \tag{3.3}$$

Similar bounds hold for the left side.

Proof. We start with the same argument in the sub-Gaussian case. For all $|\lambda| \leq \frac{1}{b}$, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda t}} \leq \exp\left(\frac{\lambda^2\tau^2}{2} - \lambda t\right).$$

Denote $g(\lambda) = \frac{\lambda^2\tau^2}{2} - \lambda t$. It remains to minimize $g(\lambda)$. For each fixed $t > 0$, we know that $g(\lambda)$ attains minimum at $\frac{t}{\tau^2}$.

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: $t/\tau^2 \geq \frac{1}{b}$. In this case, since the function is monotonically decreasing in the interval $[0, \lambda^*)$, 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 $t/\tau^2 \geq \frac{1}{b}$. □

Again, we are concerning $\frac{X_1+\cdots X_n}{n}$ instead of X itself. The following result is useful.

Theorem 8. *Linear combination of sub-exponential random variables. Let $X_1, ..., X_n$ be independent mean-zero sub-exponential random variables, where X_i is (σ_i^2, b_i)-sub-exponential. Then for any vector $a_i \in R^n$, we have*

$$\mathbb{E}[\exp(\lambda \sum_{i=1}^n a_i X_i)] \leq \exp(\frac{\lambda^2 \sum_{i=1}^n a_i^2\sigma_i^2}{2}), |\lambda| \leq \frac{1}{b_*}, \tag{3.4}$$

where $b_* = \max_i b_i|a_i|$. *In other words,* $\sum_{i=1}^n a_i X_i$ *is* $(\sum_{i=1}^n a_i^2\sigma_i^2, b_*)$*-sub-exponential. Then, it holds that*

$$\begin{aligned} P(\sum_{i=1}^n a_i X_i \geq t) &\leq \exp(-\frac{1}{2}\frac{t^2}{\sum_{i=1}^n a_i^2\sigma_i^2}), 0 \leq t \leq \frac{\sum_{i=1}^n a_i^2\sigma_i^2}{b_*} \\ P(\sum_{i=1}^n a_i X_i \geq t) &\leq \exp(-\frac{1}{2}\frac{t}{b_*\|a\|_\infty}), 0 \leq t > \frac{\sum_{i=1}^n a_i^2\sigma_i^2}{b_*} \end{aligned} \tag{3.5}$$

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

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

□

3.1 Bernstein-type Bound

Using a moment condition, we can obtain so-called Bernstein-type Bound.

Definition 3. *Bernstein condition. A random variable* X *with mean* μ *and variance* σ^2 *is said to satisfy the Bernstein condition if*

$$\mathbb{E}(X-\mu)^k \leq \frac{k!}{2}\sigma^2 b^{k-2}, k \geq 2 \tag{3.6}$$

Theorem 9. *Bernstein-type Bound. For any random variable satisfying the Bernstein condition, we have*

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

and moreover, the concentration inequality

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

In particular, X *is* $(\sqrt{2}\sigma, 2b)$*-sub-exponential. Finally, if* $X_1, \cdots, X_n$ *are i.i.d. random variables satisfying Bernstein condition with* b*. Then, it holds that*

$$\begin{aligned} P(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t) &\leq \exp(-\frac{nt^2}{2(\sigma^2+bt)}) \\ P(\sum_{i=1}^n (X_i - \mathbb{E}X) \geq t) &\leq \exp(\frac{-t^2}{2(n\sigma^2+bt)}) \end{aligned} \tag{3.7}$$

Proof. We assume $\mu = 0$ W.L.O.G..

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

where we use the Bernstein condition in the last step. For any $|\lambda| < \frac{1}{b}$, we can sum the geometric series to obtain

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

where we use $e^x \geq 1 + x, \forall x \in \mathbb{R}$. To show that X is $(2\sigma^2, 2b)$-sub-exponential, we note that for $|\lambda| < \frac{1}{2b}$, we have

$$\mathbb{E}\exp(\lambda X) \leq \exp(\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}).$$

Now we give a proof for the tail probability inequality. Note that we will not directly use $\mathbb{E}\exp(\lambda X) \leq \exp(\frac{\lambda^2(\sqrt{2}\sigma)^2}{2})$ since we can get a sharper bound here.

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

where we use Markov's inequality and the above moment generating function bound in the last step. Setting $\lambda = \frac{t}{bt + \sigma^2} < \frac{1}{b}$ concludes the proof.

Finally, we have

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

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

Example: we shall focus on the random variable satisfying $|X - \mathbb{E}X| \leq b$. Then, the Bernstein condition is satisfied with $\frac{b}{3}$. (the proof is tedious and is omitted. To prove this, it is sufficient to expand the $\exp(\cdot)$ and note that $\frac{k!}{2} \geq 3^{k-2}$ for $k \geq 2$.) Therefore, we have:

Theorem 10. *Let X_i be a sequence of i.i.d. random variables such that $|X_i - \mathbb{E}X_i| \leq b$. Then, it holds that*

$$P(\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t) \leq \exp(\frac{-t^2}{2n\sigma^2 + \frac{2}{3}bt}) \tag{3.8}$$

3.2 Discussion

Recall the Hoeffding-type inequality is of the form $P(\sum_{i=1}^n a_i X_i \geq t) \leq \exp(-\frac{t^2}{2\|a\|_2^2\sigma^2})$. For small t, Bernstein-type bound behaves similarly to sub-Gaussian tail ($\exp(-t^2)$). For large t, the bound is weaker ($\exp(t)$). However, we find that sub-exponential property sometimes can provide sharper inequality than that of sub-Gaussian because it uses the information of variance.

Example. Suppose that the random variables X_i are i.i.d., mean-zero and satisfy $X_i \in [-b, b]$ with probability 1, but have variance $\sigma^2 = E[X_i^2] \leq b^2$. Bernstein-type bound implies that

$$P\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{t^2}{\sigma^2\|a\|_2^2}, \frac{t}{2b\|a\|_\infty}\right\}\right)$$

With the fact that $5/12 > 1/3$ and taking $a_i = \frac{1}{n}$, we obtain that $\frac{1}{n}\sum_{i=1}^n X_i - 0 \leq t$ w.p. at least $1 - \exp(-n\min\{\frac{t^2}{3\sigma^3}, \frac{t}{4b}\})$. We can take $t = \max\{\sigma\sqrt{\frac{3\log\frac{1}{\delta}}{n}}, \frac{4b\log 1/\delta}{n}\}$ with probability $1-\delta$. On the contrary, the bound via Hoeffding-type bound (which only uses the information of range) is $b\frac{\sqrt{2\log\frac{1}{\delta}}}{\sqrt{n}}$.

We can also have a interpretation from another point of view similar to the discussion above. We consider a sequence of i.i.d. bounded random variables where $X_i \in [a, b]$ and $R = b - a$. Then, it holds that

$$\begin{aligned}
\text{Hoeffding}: &\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \leq \frac{R}{\sqrt{n}}\sqrt{\frac{\log(1/\delta)}{2}} = \tilde{\mathcal{O}}(\frac{R}{\sqrt{n}}) \\
\text{Bernstein}: &\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \leq \frac{2\sqrt{\sigma^2\log(1/\delta)}}{\sqrt{n}} + \frac{4b}{3}\frac{\log(1/\delta)}{n} = \tilde{\mathcal{O}}(\frac{\sigma}{\sqrt{n}} + \frac{b}{n})
\end{aligned} \tag{3.9}$$

Therefore, for random variable with a small variance compared to its range, Bernstein's inequality can give a sharper bound.

3.3 Ond-sided Bound

If we only have an upper bound for the range of X, it is still possible to derive one-sided bounds.

Theorem 11. *One-sided Bernstein inequality. If $X \leq b$ almost surely, then,*

$$\begin{aligned}
\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) &\leq \exp\left(\frac{\lambda^2\mathbb{E}X^2}{1 - \lambda b/3}\right), \forall\lambda \in [0, b/3) \\
P(\frac{1}{n}(\sum_{i=1}^n X_i - \mathbb{E}X_i) \geq t) &\leq \exp\left(\frac{-nt^2}{2(\frac{1}{n}(\sum_{i=1}^n \mathbb{E}X_i^2) + bt/3}\right)
\end{aligned} \tag{3.10}$$

Proof. See proposition 2.14 in [3]. □

4 Martingale

The inequalities provided so far are used for sum of independent variables. However, this is not true in general for multi-armed bandit and reinforcement learning. This is because the agent makes a decision based on the historic observations. For instance, the arm chosen in time 10 depends on $X_{a_1}, X_{a_2}, \cdots, X_{a_9}$ where X_{a_t} is the reward in time t and a_t is the distribution sampled in time t. However, we shall see that conditioned on the history and a_{10}, $X_{a_{10}}$ is sub-Gaussian if we assume it is bounded. We expect to have some concentration result in such a case.

4.1 Martingale

We consider $\{X_k\}_{k=1}^n$: a sequence of independent random variables and a function $f : \mathbb{R}^n \to \mathbb{R}$. We are interested in the deviations of $f(X) = f(X_1, c \ldots, X_n)$ from its mean. To analyze it, we define $Y_0 = \mathbb{E}[f(X)], Y_n = f(X)$, and

$$Y_k = \mathbb{E}[f(X)|X_1, \cdots, X_k], k = 1, 2, \cdots, n-1,$$

where we assume that all conditional expectations exist. We have a telescoping decomposition

$$f(X) - \mathbb{E}f(X) = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{D_k}$$

That is, $f(X) - \mathbb{E}f(X)$ is written as a sum of increments $\{D_k\}_{k=1}^n$. Here, $\{Y_k\}_{k=1}^n$ is a Doob martingale, whereas $\{D_k\}_{k=1}^n$ is a martingale difference. More generally, we consider a sequence of *σ-algebra* $\{\mathcal{F}_k\}_{k=1}^\infty$ s.t. $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \geq 1$. The sequence is known as a *filtration*. For instance, in the above example, we have $\sigma(X_1, \cdots, X_k) = \mathcal{F}_k$. We then have a sequence of random variables s.t. Y_k is measurable w.r.t. $\mathcal{F}_k$ (also referred as *adapted* to the filtration $\{\mathcal{F}_k\}_{k=1}^\infty$. Then, we have

Definition 4. *Martingale.* $\{(Y_k, \mathcal{F}_k)\}_{k=1}^\infty$ *is a martingale if for all* $k \geq 1$*, we have*

$$\mathbb{E}|Y_k| < \infty, \qquad \mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k. \tag{4.1}$$

Frequently we can see that $\mathcal{F}_k = \sigma(X_1, \cdots, X_k)$. Moreover, if $\mathcal{F}_k = \sigma(Y_1, \cdots, Y_k)$, we say $\{Y_k\}$ forms a martingale sequence. We find a somewhat easier interpretation and some examples will be provided later. Consider a sequence of random functionals $\xi_1(S_1), \cdots, \xi_n(S_n)$ where $S_n = (Z_1, \cdots, Z_n)$ and ξ_i is sub-Gaussian w.r.t. Z_i with parameter σ_i^2 which possibly depends on S_{i-1}. Then, we can still get some concentration result (Azuma-Hoeffding inequality which is provided later). For instance, consider the UCB-1 algorithm for the stochastic K-armed MAB with $Z_t = X_{a_t}$ and $\xi(S_n) = \frac{1}{\sum_{t=1}^n I(a_t=k)} \sum_{t=1}^n X_{a_t} I(a_t = k)$.

Example: Consider a sequence of i.i.d. random variables $\{X_k\}$. Let $\mathcal{F}_k = \sigma(X_1, \cdots, X_k)$ and $S_k = \sum_{j=1}^k X_j$. Then,

$$\mathbb{E}[S_{k+1}|\mathcal{F}_k] = \mathbb{E}[X_{k+1} + S_k|X_1, \cdots, X_k] = \mu + S_k.$$

Therefor, $\{S_k\}$ is not a martingale unless $\mu = 0$. We can take $Y_k = S_k - k\mu = \sum_{j=1}^{k}(X_j - \mu)$. Then, $\{Y_k\}$ is a martingale.

Example: Doob construction. Let $X_1^k = (X_1, \cdots, X_k)$ and suppose $\mathbb{E}|f(X)| < \infty$. Then,

$$\mathbb{E}|Y_k| = \mathbb{E}|\mathbb{E}[f(X)|X_1^k]| \leq \mathbb{E}\mathbb{E}|f(X)| \mid X_1^k$$

where we use $|\mathbb{E}[f(X)|X_1^k]| \leq \mathbb{E}|f(X)||X_1^k$ due to the convexity of $|\cdot|$. It also holds that,

$$\mathbb{E}[Y_{k+1}|X_1^k] = \mathbb{E}[\mathbb{E}[f(X)|X_1^{k+1}]|X_1^k] = \mathbb{E}[f(X)|X_1^k] = Y_k,$$

where the last step we use the tower property of conditional expectation: For sub-σ-algebras $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{F}$, we have

$$\mathbb{E}(\mathbb{E}(X|\mathcal{H}_2)|\mathcal{H}_1) = \mathbb{E}(X|\mathcal{H}_1),$$

where $\mathcal{H}_1 = \sigma(X_1^k)$ and $\mathcal{H}_2 = \sigma(X_1^{k+1})$ here.

Example: Likelihood ratio. Let f and g be two mutually absolutely continuous densities and let $\{X_k\}$ be a sequence of random variables i.i.d. from f. We define

$$Y_k = \prod_{\ell=1}^{k} \frac{g(X_\ell)}{f(X_\ell)}.$$

Then the sequence $\{Y_k\}$ is a martingale w.r.t. $\{X_k\}$ because

$$\mathbb{E}[Y_{n+1}|X_1, \cdots, X_n] = \mathbb{E}|\frac{g(X_{n+1})}{f(X_{n+1})}| \prod_{k=1}^{n} \frac{g(X_k)}{f(X_k)} = Y_n,$$

where the last step uses $\mathbb{E}\frac{g(X)}{f(X)} = \int f(x)\frac{g(x)}{f(x)}dx = \int g(x)dx = 1$ (note they are mutually absolutely continuous).

Example: Martingale difference.

$$\mathbb{E}|D_k| < \infty, \qquad , \mathbb{E}[D_{k+1}|\mathcal{F}_k] = 0.$$

Given a martingale $\{Y_k\}$, we can take $D_k = Y_k - Y_{k-1}$ to obtain

$$\mathbb{E}[D_{k+1}|\mathcal{F}_k] = \mathbb{E}[Y_{k+1}|\mathcal{F}_k] - \mathbb{E}[Y_k|\mathcal{F}_k] = 0.$$

We have

$$Y_n - Y_0 = \sum_{k=1}^{n} D_k.$$

4.2 Concentration bounds for martingale difference sequences

We derive bounds for difference $Y_n - Y_0$, or as bounds for $\sum_{k=1}^n D_k$ (sum of martingale difference). The main idea (from my understanding) is that we can replace the independence with

$$\mathbb{E}f(x_1, \cdots, x_n) = \mathbb{E}[\mathbb{E}[f(x_1, \cdots, x_n)|\mathcal{F}_{n-1}]].$$

Theorem 12. *Let $\{D_k, \mathcal{F}_k\}$ be a martingale difference and suppose that*

$$\mathbb{E}[\exp(\lambda D_k)|\mathcal{F}_{k-1}] \leq \exp(\frac{\lambda^2\tau_k^2}{2}), \forall|\lambda| < \frac{1}{b_k},$$

i.e., conditionally sub-exponential(τ_k^2, b_k). Then, it holds that $\sum_{k=1}^n D_k$ is sub-exponential $(\sum_{k=1}^n \tau_k^2, b_*)$ where $b_* = \max b_k$ and

$$P\left[\left|\sum_{k=1}^n D_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^n \tau_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \tau_k^2}{b_*} \\ 2e^{-\frac{t}{2b_*}} & \text{if } t > \frac{\sum_{k=1}^n \tau_k^2}{b_*} \end{cases} \tag{4.2}$$

Proof. The proof is standard: control the moment generating function, then apply Chernoff method. We have

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(\sum_{k=1}^n D_k)}\right] &= \mathbb{E}\left[e^{\lambda(\sum_{k=1}^{n-1} D_k)}\mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[e^{\lambda\sum_{k=1}^n D_k}\right] e^{\lambda^2\tau_n^2/2}. \end{aligned}$$

Iterating this procedure yileds that

$$\mathbb{E}\left[e^{\lambda(\sum_{k=1}^n D_k)}\right] \leq \exp(\lambda^2 \sum_{k=1}^n \frac{\tau_k^2}{2}), \forall|\lambda| < \frac{1}{b_*}.$$

□

We shall see that the above procedure also works for the conditionally sub-Gaussian case where $b_* = b_k = 0$. A special case is the bounded random variable and the corresponding famous Azuma-Hoeffding inequality.

Theorem 13. *Azuma-Hoeffding inequality. Let $\{D_k, \mathcal{F}_k\}$ be a martingale difference where there are constants $\{(a_k, b_k)\}$ s.t. $D_k \in [a_k, b_k]$ for all $k \geq 1$. Then, for all $t \geq 0$, we have*

$$P(|\sum_{k=1}^n D_k| \geq t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}) \tag{4.3}$$

Proof. Note that D_k conditioned on $\mathcal{F}_{k-1}$ is $\frac{(b_k-a_k)^2}{4}$-sub-Gaussian. □

4.3 One-sided results

Theorem 14. *Azuma-Hoeffding, one side. Let $X_i \in \mathcal{F}_i$ and $\mathcal{F}_{k-1} \subset \mathcal{F}_k$. If it holds that*

$$\mathbb{E}[X_i - \mathbb{E}[X_i]|\mathcal{F}_{i-1}] = 0, \qquad X_i \leq \mathbb{E}X_i + R_i,$$

then it holds that

$$P(\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) \geq t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^{n} R_i^2}) \tag{4.4}$$

Note that it means that $\{X_k - \mathbb{E}X_k\}$ is a martingale difference. Similarly, we have

Theorem 15. *Azuma-Bernstein, one side. Let $X_i \in \mathcal{F}_i$ and $\mathcal{F}_{k-1} \subset \mathcal{F}_k$. If it holds that*

$$\mathbb{E}[X_i - \mathbb{E}[X_i]|\mathcal{F}_{i-1}] = 0, \qquad X_i \leq \mathbb{E}X_i + R, \qquad \mathbb{V}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$$

then it holds that

$$P(\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) \geq t) \leq 2\exp(-\frac{t^2}{2\sum_{k=1}^{n}\sigma_i^2 + 2/3Rt}) \tag{4.5}$$

5 Uniform Convergence

Consider a supervised learning problem with hypothesis space $\mathcal{H}$ where $|\mathcal{H}| < \infty$. We assume (X, Y) is sampled from some unknown distribution $P(X, Y)$. For a fixed a hypothesis $f \in \mathcal{H}$, the *population risk* is given by

$$L(f) = \mathbb{E}_{(X,Y)\sim P}\ell(f(X), Y),$$

where $\ell(\cdot,\cdot)$ is a loss function, e.g., $\ell(f(x), y) = \frac{1}{2}(f(x)-y)^2$. Given a data set $\mathcal{D} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, we also define the *empirical risk* as

$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i).$$

Clearly, we have

$$\mathbb{E}_P\hat{L}(f) = L(f), \qquad \hat{L}(f) \to L(f),$$

as n tends to infinity (under some mild condition of Law of Large Number). Actually, we can find that $\sqrt{n}(\hat{\theta} - \theta^*)$ is asymptomatic normal under some regularization condition where we assume that $\mathcal{H} = \{f_\theta : \theta \in \Theta\}$.

What we care about? Suppose that we find a $\hat{f} \in \mathcal{H}$ by minimizing $\hat{L}(f), f \in \mathcal{H}$ (e.g. we run SGD/Adam) and assume that the minimizer of $L(f)$ is $f^* \in \mathcal{H}$. Here we assume $0 \leq \ell(\cdot,\cdot) \leq 1$ for simplicity (sub-Gaussian assumption).

$$L(\hat{f}) - L(f^*) = \underbrace{\left(L(\hat{f}) - \hat{L}(\hat{f})\right)}_{A} + \underbrace{\left(\hat{L}(\hat{f}) - \hat{L}(f^*)\right)}_{B} + \underbrace{\left(\hat{L}(f^*) - L(f^*)\right)}_{C}$$

We first note that $B \leq 0$ because $\hat{f}$ minimizes $\hat{L}(\cdot)$. Then, we have

$$\begin{aligned} L(\hat{f}) - L(f^*) &\leq \underbrace{\left(L(\hat{f}) - \hat{L}(\hat{f})\right)}_{A} + \underbrace{\left(\hat{L}(f^*) - L(f^*)\right)}_{C} \\ &\leq \sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)| + \left(\hat{L}(f^*) - L(f^*)\right) \\ &\leq 2 \sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)|. \end{aligned}$$

Clearly, we shall expect that $|\hat{L}(f) - L(f)|$ is small according to concentration inequality for each fixed $f \in \mathcal{H}$ because they are sample mean and true expectation. However, we cannot obtain such a result because

$$\{\ell(\hat{f}(X_i, Y_i) : i = 1, 2, \cdots, n\}$$

are not independent since $\hat{f}(\cdot)$ is obtained via a minimization problem over the data set $\mathcal{D}$ and consequently the i.i.d. assumption over $\mathcal{D}$ does not hold for $\ell(X_i, Y_i)$. This is the reason why we need a uniform convergence for all $f \in \mathcal{H}$ and clearly such a bound directly gives a generalization bound over $L(\hat{f}) - L(f^*)$.

5.1 Technique to obtain uniform convergence bounds

We first note that for a sequence of i.i.d. random variables $\{X_k\}_{k=1}^n$, $\{f(X_k)\}_{k=1}^n$ are also i.i.d. for a fixed function. In particular, the function $f(\cdot)$ does not relies on $\{X_k\}_{k=1}^n$.

Finite $\mathcal{H}$. We take $\delta_f = \frac{\delta}{|\mathcal{H}|}$ to construct a concentration bound for each $f \in \mathcal{H}$. Then, applying a union bound over all $f \in \mathcal{H}$:

$$\begin{aligned} P\left(\sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)| > \sqrt{\frac{1}{2n} \log \frac{2}{\delta/|\mathcal{H}|}}\right) &\leq \sum_{f \in \mathcal{H}} P\left(|L(f) - \hat{L}(f)| > \sqrt{\frac{1}{2n} \log \frac{2}{\delta/|\mathcal{H}|}}\right) \\ &\leq |\mathcal{H}| \times \frac{\delta}{|\mathcal{H}|} = \delta, \end{aligned} \tag{5.1}$$

where we use boundedness assumption in the second inequality to apply the Hoeffding's inequality. We can apply the Hoeffding's inequality here because f in $P\left(|L(f) - \hat{L}(f)| > \sqrt{\frac{1}{2n} \log \frac{2}{\delta/|\mathcal{H}|}}\right)$ is a fixed function instead of a function obtained via a minimization problem over $\mathcal{D}$.

Infinite $\mathcal{H}$. We can first find a finite covering $\mathcal{H}_\epsilon$ of $\mathcal{H}$ s.t. for all $f \in \mathcal{H}$, we can find some $f_\epsilon \in \mathcal{H}_\epsilon$ and

$$\sup_x |f(x) - f_\epsilon(x)| < \epsilon.$$

Then, we derive a uniform convergence result on $\mathcal{H}_\epsilon$ via union bound in the finite $\mathcal{H}$ case and obtain

that for all $f \in \mathcal{H}$, we have

$$\begin{aligned} |L(f) - \hat{L}(f)| &= |\frac{1}{n}\sum_{i=1}^{n}\left((f_\epsilon - \mathbb{E}f_\epsilon) + (f - f_\epsilon) + \mathbb{E}(f_\epsilon - f)\right)(X_i)| \\ &\leq |\frac{1}{n}\sum_{i=1}^{n}(f_\epsilon - \mathbb{E}f_\epsilon)| + |\frac{1}{n}\sum_{i=1}^{n}(f - f_\epsilon)| + |\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f_\epsilon - f)| \\ &\leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}_\epsilon|}} + 2\epsilon, \end{aligned} \tag{5.2}$$

where $|\mathcal{H}|_\epsilon$ is referred to ϵ-covering number. Clearly, we can tolerate an exponential covering number, e.g., $\exp(\frac{c}{\epsilon}), \exp(\frac{c}{\epsilon^2})$ which will contribute to the generalization bound with $\text{poly}(1/\epsilon)$. The author is too vegetable to be aware of the general case but knows an example.

Example: Let $B, \epsilon > 0$ and consider $S = \{x \in \mathbb{R}^p : \|X\|_2 \leq B\}$. Then, we can find a ϵ-covering w.r.t. ℓ_2-norm with at most $(\frac{3B}{\epsilon})^p$ elements and we have

$$\log(\frac{3B}{\epsilon})^p = p\log(\frac{3B}{\epsilon}).$$

Roughly speaking, we have

$$L(\hat{f}) - L(f^*) \leq \tilde{O}(\frac{\log|\mathcal{H}_\epsilon|}{\sqrt{n}}).$$

6 Other contents.

- McDiarmid' inequality (also referred as Bounded differences inequality, Lipschitz w.r.t. Hamming norm);
- Concentration of functions of Gaussian random variables;
- X is sub-Gaussian, then X^2 behaves in a sub-Gaussian way;
  - [3] exercise 2.6;
  - $X^2 - \mathbb{E}X^2$ is sub-exponential$(16\sigma^2)$, see note of MIT.
- Maximum of sub-Gaussian;

6.1 More Examples

Similar issues arise in the setting of bandit and RL. In particular, the uniform convergence is required for the class of UCB algorithms. See the uniform concentration result (step 1) in [2]; lemma 39 of [1], this example also demonstrates that we may get a sharper bound via Bernstein-type inequality; unstable issue in [4].

7 Reference

This note is largely based on

- Note on high-dimensional statistics by Philippe Rigollet and Jan-Christian Hutter;
- Note on hign-dimensional probability by Ramon van Handel;
- The book on high-dimensional statistics [3].

References

[1] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

[2] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[3] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[4] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*, 2020.