Note on Concentration

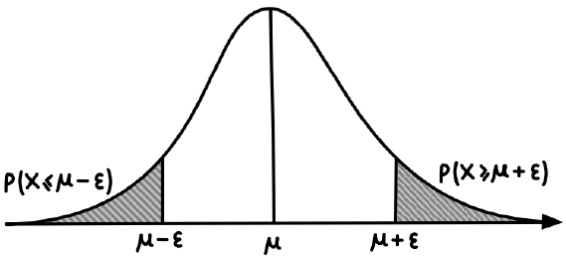
Shaojie Bai
white.shaojie@gmail.com

Abstract

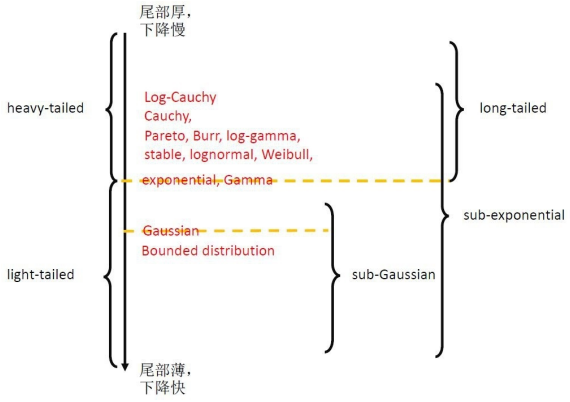
This note considers concentration bounds and some applications in learning theory.

1 Introduction

In a variety of settings, it is of interest to obtain bounds on the probability of tails of a random variable, or two-sided inequalities that guarantee that a random variable is close to its mean or median, e.g., in Figure 1(a). In this note, we explore some elementary techniques for obtaining both deviation and concentration inequalities, mainly refer to Wainwright (2019); Duchi (2021); Mohri et al. (2018), and their applications in learning theory, refer to Ma (2022). For different kinds of distributions, the property of the distribution tail varies. Some distributions have light tails and decay fast, while others may have heavy and long tails and decay slowly, refer to Figure 1(b). We will discuss some of them in the remaining sections.



(a) Tail Probability



(b) Distribution Tails

Figure 1: Concentration of Tails

We start with the famous concentration theorem in an asymptomatic way.

Theorem 1.1 (Weak Law of Large Number)**.** Let $\{X_n\}$ be a sequence of i.i.d. random variables, with expectation $\mathbb{E}X$ then

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

in probability.

However, we are more interested in how close the average is to its expectation given finite samples in practice, which is also the non-asymptomatic view.

2 Classical bounds

2.1 General Inqualities

Theorem 2.1 (Markov Inequality)**.** Given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \forall t > 0 \tag{2.1}$$

Proof. This is a simple instance of an upper tail bound.

$$\mathbb{E}[X] = \int_0^\infty x\mathbb{P}(X = x)dx \geq \int_t^\infty x\mathbb{P}(X = x)dx \geq t\int_t^\infty \mathbb{P}(X = x)dx = t\mathbb{P}[X \geq t] \tag{2.2}$$

□

Corollary 2.2 (High-order moments bound)**.** For a random variable X with $\mathbb{E}[X] = \mu$, which has a finite expectation of a central moment $\mathbb{E}\left[|X-\mu|^k\right]$ of order k,

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}\left[|X-\mu|^k\right]}{t^k} \quad \forall t > 0 \tag{2.3}$$

Specially, with finite variance, we have *Chebyshev Inequality*:

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathrm{var}(X)}{t^2} \quad \forall t > 0 \tag{2.4}$$

Proof. Apply markov inequality, $|X-\mu|^k$ is a non-negative random variable, with finite expectation of central moment, then for all $t \geq 0$

$$\mathbb{P}[|X - \mu| \geq t] = \mathbb{P}[|X-\mu|^k \geq t^k] \leq \frac{\mathbb{E}\left[|X-\mu|^k\right]}{t^k} \tag{2.5}$$

□

Theorem 2.3 (Chernoff Bound)**.** Suppose that random variable X has a moment generating function in a neighborhood of zero, meaning that there is some constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}\left[e^{\lambda(X-\mu)}\right]$ exists for all $\lambda \in [-b, b]$, then

$$\mathbb{P}[X - \mu \geq t] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \quad \forall t \geq 0, \lambda \in [0, b] \tag{2.6}$$

Optimizing the choice of λ so as to obtain the tightest result yields the *Chernoff bound*,

$$\mathbb{P}[X - \mu \geq t] \leq \inf_{\lambda \in [0,b]} \left\{\mathbb{E}\left[e^{\lambda(X-\mu)}\right] - e^{\lambda t}\right\} \quad \forall t \geq 0 \tag{2.7}$$

Remark 2.4. In general, Markov inequality and Chebyshev inequality are sharp in the sense that we can find some distribution for which the bound is tight. However, for small tail t, the bound derived from Markov (Chebyshev) inequality goes to infinite, which is terrible. In many cases (some kinds of distributions), we can improve the $O(1/t^k)$ rate to an $O(exp(poly(t)))$ rate, and we still have a tight bound for small tail t. An example is as follows.

Example 2.5 (Gaussian tail bounds)**.** Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a Gaussian random variable with mean μ and variance σ^2. We have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall \lambda \in \mathbb{R} \tag{2.8}$$

By optimizing the choice of λ defined in Chernoff bound, we obtain

$$\mathbb{P}[X - \mu \geq t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \forall t \geq 0 \tag{2.9}$$

Proof. By $\int_{-\infty}^{\infty} \exp(-x^2)dx = \sqrt{\pi}$, we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(X-\mu))] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\lambda x - \frac{1}{2\sigma^2}x^2\right)dx \\
&= \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\lambda\sigma^2\right)^2\right)dx \\
&= \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \sqrt{2\sigma^2}\exp(-y^2)dy \quad y = \frac{1}{\sqrt{2\sigma^2}}(x-\lambda\sigma^2) \\
&= \exp\left(\frac{\lambda^2\sigma^2}{2}\right)
\end{aligned} \tag{2.10}$$

Therefore, taking $\lambda = \frac{t}{\sigma^2}$ by Chernoff bound, we have

$$\begin{aligned}
\mathbb{P}[X-\mu \geq t] &\leq \inf_{\lambda\in\mathbb{R}}\left\{\mathbb{E}\left[e^{\lambda(X-\mu)}\right] - e^{\lambda t}\right\} = \exp\left\{\inf_{\lambda\in\mathbb{R}} \frac{\lambda^2\sigma^2}{2} + \lambda t\right\} \\
&= \exp\left\{-\frac{t^2}{2\sigma^2}\right\}
\end{aligned} \tag{2.11}$$

□

2.2 Sub-Guassian Variables and Bounds

Motivated by Chernoff Bound and Guassian example, we introduce the following class of random variables.

Definition 2.6 (sub-Guassian Variable)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-Gaussian* if there is a positive number σ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \forall\lambda \in \mathbb{R} \tag{2.12}$$

The definition requires that the central moments of X exist and grow mildly,

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(X-\mu)^k \tag{2.13}$$

Theorem 2.7 (sub-Gaussian Bound)**.** Any sub-Gaussian variables satisfy the concentration inequality for tails, for $\forall t > 0$

$$P(X-\mu \geq t) \vee P(X-\mu \leq -t) \leq e^{-\frac{t^2}{2\sigma^2}} \tag{2.14}$$

Proof. Similar to Chernoff technique in Example 2.5 and optimize the parameter λ. □

Example 2.8 (Bounded random variables)**.** Random variable supported on some interval $[a,b]$ is sub-Gaussian with $\sigma = \frac{b-a}{2}$.

Proof. We assume $\mathbb{E}X = \mu = 0$ for simplification. By definition, we aim to prove $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2(b-a)^2}{8}}$. By convexity and bounded assumpetion, we have

$$e^{\lambda x} \leq \frac{x-a}{b-a}e^{\lambda b} + \frac{b-x}{b-a}e^{\lambda a} \quad \forall x \in [a,b]$$

Take expectation on both sides,

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &\le \frac{-a}{b-a}e^{\lambda b} + \frac{b}{b-a}e^{\lambda a} \\ &= \gamma e^{(1-\gamma)u} + (1-\gamma)e^{-\gamma u} \qquad u = \lambda(b-a) \ge 0, \gamma = -\frac{a}{b-a} \\ &\triangleq e^{g(u)}\end{aligned} \tag{2.15}$$

Then we analyse the bound of $g(u)$.

$$\begin{aligned}g(u) &= -\gamma u + \ln(\gamma e^u + 1 - \gamma) \qquad g(0) = 0 \\ g'(u) &= -\gamma + \frac{\gamma e^u}{\gamma e^u + 1 - \gamma} \qquad g'(0) = 0 \\ g''(u) &= \frac{\gamma e^u (1-\gamma)}{(\gamma e^u + 1 - \gamma)^2} \le \frac{1}{4} \qquad (a+b)^2 \ge 4ab\end{aligned} \tag{2.16}$$

By Taylor's theorem and Lagrange remainder, we have

$$\begin{aligned}g(u) &= g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) \qquad \xi \in [0, u] \\ &= \frac{u^2}{2}g''(\xi) \le \frac{u^2}{8} = \frac{\lambda^2(b-a)^2}{8}\end{aligned} \tag{2.17}$$

Thus, we complete the proof by combining Equation 2.17 and 2.15. □

Property 1 (sub-Gaussian random variables are closed under linear combination)**.** 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 $\sigma^2 = \sum_{i=1}^n \sigma_i^2$

Proof. The property is easy to verify by definition.

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i - \mathbb{E}X_i)}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i-\mathbb{E}X_i)}\right] \le e^{\lambda^2 \frac{\sum_{i=1}^n \sigma_i^2}{2}} \tag{2.18}$$

□

As a consequence of the property 1, we obtain an important result applicable to sums of independent sub-Gaussian random variables, known as the *Hoeffding bound*.

Theorem 2.9 (Hoeffding Bound)**.** Suppose that the variables $X_i, i = 1, \ldots, n$ are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then,

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mu_i) \ge t\right] \vee \mathbb{P}\left[\sum_{i=1}^n (X_i - \mu_i) \le -t\right] \le \exp\left\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right\} \quad \forall t \ge 0 \tag{2.19}$$

In particular, if X_i is supported on $[a_i, b_i]$, we have

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \ge t\right) \vee \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \forall t \ge 0 \tag{2.20}$$

Corollary 2.10 (Uniform bound from Hoeffding bound)**.** Suppose that the variables $X_i, i = 1, \ldots, n$ are independent, and $\forall i, X_i$ has mean μ and sub-Gaussian parameter σ_i. Then, with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right| \le \sqrt{\frac{2\sum_{i=1}^n \sigma_i^2}{n^2} \log\left(\frac{2}{\delta}\right)} \tag{2.21}$$

In particular, if X_i is supported on $[a_i, b_i]$, then with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right| \le \sqrt{\frac{c}{2n}\log\left(\frac{2}{\delta}\right)} \tag{2.22}$$

where $c = \frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2$.

Proof. Refer to Theorem 2.9, let $\delta = 2\exp(-\frac{2nt^2}{c})$ and solve the solution of t. □

Theorem 2.11 (Equivalent characterizations of sub-Gaussian variables)**.** Given any zero-mean random variable X, the following properties are equivalent:

- There is a constant $\sigma \ge 0$ such that for all $\lambda \in \mathbb{R}$ $\mathbb{E}\left[e^{\lambda X}\right] \le e^{\frac{\lambda^2\sigma^2}{2}}$
- There is a constant $c \ge 0$ and Gaussian variable $Z \sim \mathcal{N}\left(0, \tau^2\right)$ such that for all $s \ge 0$ $\mathbb{P}[|X| \ge s] \le c\mathbb{P}[|Z| \ge s]$
- There is a constant $\theta \ge 0$ such that for all $k = 1, 2, \ldots$ $\mathbb{E}\left[X^{2k}\right] \le \frac{(2k)!}{2^k k!}\theta^{2k}$
- There is a constant $\sigma \ge 0$ such that for all $\lambda \in [0, 1)$ $\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \le \frac{1}{\sqrt{1-\lambda}}$

Proof. See Section 2.4 in Wainwright (2019) for the proof of these equivalences. □

2.3 Sub-Exponential Variables and Bounds

The notion of sub-Gaussianity is fairly restrictive, so we now turn to the class of sub-Exponential variables, which are defined by a slightly milder condition on the moment generating function, i.e., the moment generating function exists in a neighborhood of zero:

Definition 2.12 (sub-Exponential Variables)**.** A random variable X with mean $\mu = \mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν, b) such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\lambda^2\nu^2}{2}} \quad \forall|\lambda| < \frac{1}{b} \tag{2.23}$$

Theorem 2.13 (Sub-exponential tail bound)**.** Suppose that X is sub-exponential with parameters (ν, b). Then

$$\mathbb{P}[X-\mu \ge t] \vee \mathbb{P}[X-\mu \le -t] \le \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \le t \le \frac{\nu^2}{b} \\ e^{-\frac{t}{2b}} & \text{for } t > \frac{\nu^2}{b} \end{cases} = \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{b}\right\}\right) \quad t \ge 0 \tag{2.24}$$

Proof. We still assume $\mu = 0$ without generality. We follow the usual Chernoff-type approach and combine it with the definition of the sub-exponential variable:

$$\mathbb{P}(X \ge t) \le e^{-\lambda t}\mathbb{E}[e^{\lambda X}] \le \exp\underbrace{(-\lambda t + \frac{\lambda^2\nu^2}{2})}_{g(\lambda,t)} \quad \lambda \in [0, 1/b), t \ge 0$$

In order to complete the proof, it remains to compute the quantity $g^*(t) := \inf_{\lambda\in[0,1/b)} g(\lambda, t)$ for each fixed $t \ge 0$. Note that the unconstrained minimum of the function $g(\cdot, t)$ occurs at $\lambda^* = \frac{t}{\nu^2}$.

If $0 \le t < \frac{\nu^2}{b}$, the unconstrained optimum corresponds to the constrained minimum as well, so that $g^*(t) = -\frac{t^2}{2\nu^2}$ over this interval.

If $t \geq \frac{\nu^2}{b}$, $g(\cdot, t)$ is monotonocally decreasing in the interval $[0, \lambda^*]$, the constrained minimum is achieved at the boundary point $\lambda^* = 1/b$, and we have

$$g^*(t) = g(\lambda^*, t) = -\frac{t}{b} + \frac{\nu^2}{2b^2} \leq -\frac{t}{2b} \quad t \geq \frac{\nu^2}{b}$$

□

Example 2.14 (Laplace Variable). (μ, ϵ)–Laplace distribution with μ mean and $\epsilon > 0$ parameter: $f(x) = \frac{1}{2\epsilon} \exp \frac{-|x-\mu|}{\epsilon}$, which is a sub-exponential variable with parameter $(2\epsilon, \epsilon)$.

Proof.

$$\begin{aligned} \mathbb{E}[e^{\lambda(X-\mu)}] &= \int_{-\infty}^{\infty} e^{\lambda x} \cdot \frac{1}{2\epsilon} e^{-\frac{|x|}{\epsilon}} dx = \frac{1}{2\epsilon} \left\{ \int_{-\infty}^{0} e^{(\lambda + \frac{1}{\epsilon})x} dx + \int_{0}^{\infty} e^{(\lambda - \frac{1}{\epsilon})x} dx \right\} \\ &= \frac{1}{2\epsilon} \left(\frac{\epsilon}{\lambda\epsilon + 1} - \frac{\epsilon}{\lambda\epsilon - 1} \right) = \frac{1}{1 - \lambda^2\epsilon^2} \quad |\lambda| < \frac{1}{\epsilon} \\ &\leq \exp\left(\frac{\lambda^2 \cdot 4\epsilon^2}{2} \right) \quad |\lambda| < \frac{1}{\epsilon} \end{aligned} \tag{2.25}$$

Clearly, the moment generating function does not exist for $|\lambda| > 1/\epsilon$. The tail of this distribution does not decay as fast as the Gaussian variables, but we still can find useful bounds through Theorem 2.13. □

Example 2.15 (Exponential Variable). θ-exponential distribution defined in $X \in [0, \infty)$ with parameter $\theta > 0$: $f(x) = \theta e^{-\theta x}$, which is a sub-exponential variable with parameter $(\frac{2}{\theta}, \frac{2}{\theta})$.

Proof. The mean of exponential distribution is $\mathbb{E}(X) = \int_0^\infty x \cdot \theta e^{-\theta x} dx = -\left[e^{-\theta x} x \Big|_0^\infty - \int_0^\infty e^{-\theta x} dx \right] = \frac{1}{\theta}$. (Integration by Parts: $\int u dv = uv - \int v du$.)

$$\begin{aligned} \mathbb{E}[e^{\lambda(X - \frac{1}{\theta})}] &= \frac{\theta}{\theta - \lambda} e^{-\frac{\lambda}{\theta}} \qquad \lambda < \theta \\ &= \frac{1}{1-y} e^{-y} = \exp \underbrace{(-\log(1-y) - y)}_{g(y)} \qquad y = \frac{\lambda}{\theta} \in (0, 1) \\ &\leq \exp\left(\frac{y^2}{1-y} \right) \qquad g(y) \leq \frac{y^2}{1-y}, y \in (0, 1) \\ &\leq \exp(2y^2) \qquad y \in (0, \frac{1}{2}) \\ &= \exp\left(\frac{2\lambda^2}{\theta^2} \right) \qquad \lambda \in (0, \frac{\theta}{2}) \end{aligned} \tag{2.26}$$

□

Example 2.16 (χ^2 Variable with 1 degree of freedom). Let $Z \sim \mathcal{N}(0, 1)$, and consider the random variable $X = Z^2$, which is a sub-exponential variable with parameter $(2, 4)$.

Proof. The mean of X is $\mathbb{E}(X) = \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = -\frac{1}{\sqrt{2\pi}} \left(z e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz \right) = -\frac{1}{\sqrt{2\pi}} (0 - \sqrt{2\pi}) = 1$.

$$\begin{aligned} \mathbb{E}[e^{\lambda(X-1)}] &= \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1-2\lambda}{2} z^2} dz = \frac{e^{-\lambda}}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sqrt{1-2\lambda}} \int_{-\infty}^{\infty} e^{-y^2} dy \qquad y = \sqrt{\frac{1-2\lambda}{2}} z \\ &= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \qquad \lambda \in [0, \frac{1}{2}) \\ &\leq e^{2\lambda^2} = e^{\frac{4\lambda^2}{2}}, \qquad \lambda \in [0, \frac{1}{4}) \end{aligned} \tag{2.27}$$

□

Property 2 (sub-exponential random variables are closed under linear combination)**.** Consider an independent sequence of $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-exponential with parameter (ν_k, b_k). Then the variable $Z = \sum_{k=1}^n X_k$ is sub-exponential with parameters (ν_*, b_*), where

$$\nu_* := \sqrt{\sum_{k=1}^n \nu_k^2} \qquad b_* := \max_{k=1,\ldots,n} b_k \tag{2.28}$$

this observation leads directly to the upper tail bound (lower tail bound is the same)

$$\mathbb{P}\left[\sum_{i=1}^n (X_k - \mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{t^2}{2\nu_*^2}} & \text{for } 0 \leq t \leq \frac{\nu_*^2}{b_*} \\ e^{-\frac{t}{2b_*}} & \text{for } t > \frac{\nu_*^2}{b_*} \end{cases} \qquad \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n (X_k - \mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{n^2t^2}{2\nu_*^2}} & \text{for } 0 \leq t \leq \frac{\nu_*^2}{nb_*} \\ e^{-\frac{nt}{2b_*}} & \text{for } t > \frac{\nu_*^2}{nb_*} \end{cases} \tag{2.29}$$

Proof. By independence and sub-exponential property,

$$\mathbb{E}\left[e^{\lambda\sum_{k=1}^n (X_k-\mu_k)}\right] = \prod_{k=1}^n \mathbb{E}\left[e^{\lambda(X_k-\mu_k)}\right] \leq \prod_{k=1}^n e^{\lambda^2\nu_k^2/2} = \exp\left(\frac{\lambda^2}{2}\cdot\sum_{k=1}^n \nu_k^2\right), \quad |\lambda| < \frac{1}{\max_{k=1,\ldots,n} b_k} \tag{2.30}$$

The concentration bound can be directly obtained by Theorem 2.13. □

Example 2.17 (χ^2 variables)**.** A chi-squared χ^2 random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, can be represented as the sum $Y = \sum_{k=1}^n Z_k^2$ where $Z_k \sim \mathcal{N}(0,1)$ are i.i.d. variates. Consequently, the χ^2-variate Y is sub-exponential with parameters $(\nu, b) = (2\sqrt{n}, 4)$, and the preceding discussion yields the two-sided tail bound

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n Z_k^2 - 1\right| \geq t\right] \leq 2e^{-nt^2/8}, \quad \text{for all } t \in (0,1) \tag{2.31}$$

Proof. Together with Example 2.16 and Property 2, we know Y is a sub-exponential variable with parameter $(2\sqrt{n}, 4)$. Then by Theorem 2.13, we can derive the bound. □

2.3.1 Bernstein's type Bound

The sub-exponential property can be verified by explicitly computing or bounding the moment generating function. This direct calculation may be impracticable in many settings, so it is natural to seek alternative approaches. One method is based on control of the polynomial moments of X as follows:

Definition 2.18 (Bernstein's condition)**.** Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}[X-\mu]^2$, we say that *Bernstein condition* with parameter b holds if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \text{for } k = 2,3,4,\ldots \tag{2.32}$$

Indeed, if the random variables have a small variance, we would like to see it reflected in the exponential tail bound where the variance does not appear in Hoeffding's inequality.

Theorem 2.19 (Bernstein's type Bound)**.** For any random variable satisfying the Bernstein condition, we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \quad \text{for all } |\lambda| < \frac{1}{b} \tag{2.33}$$

and moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\left(\sigma^2+bt\right)}} \quad \text{for all } t \geq 0 \tag{2.34}$$

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

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) \geq t\right) \leq \exp\left(-\frac{nt^2}{2\left(\sigma^2+bt\right)}\right) \qquad P\left(\sum_{i=1}^{n}(X_i-\mu) \geq t\right) \leq \exp\left(\frac{-t^2}{2\left(n\sigma^2+bt\right)}\right) \tag{2.35}$$

Proof. W.o.g., we assume $\mu = 0$, and by Bernstein condition we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X)] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k EX^k}{k!} \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2} \qquad |\lambda| < \frac{1}{b} \\
&= 1 + \frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-|\lambda|b} \qquad \text{(by geometric series)} \\
&\leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-|\lambda|b)}\right) \qquad e^x \geq 1+x, \forall x \in \mathbb{R} \\
&\leq \exp\left(\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}\right) \qquad |\lambda| < \frac{1}{2b}
\end{aligned} \tag{2.36}$$

Thus, variable X satisfying Bernstein condition is a $(\sqrt{2\sigma}, 2b)$ sub-exponential variable. In addition to the concentration bound derived by the sub-exponential theorem, we can get a sharper one.

$$\begin{aligned}
P(X-\mu \geq t) = P\left(e^{\lambda(X-\mu)} \geq e^{\lambda t}\right) &\leq \mathbb{E}[e^{\lambda(X-\mu)}]\cdot e^{-\lambda t} \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b\lambda)} - \lambda t\right) \qquad \lambda \in [0, \frac{1}{b}) \\
&\leq \exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \quad \text{let } \lambda = \frac{t}{\sigma^2+bt} < \frac{1}{b}
\end{aligned} \tag{2.37}$$

Here we just choose an approximated minimizer λ to get a simplified expression in the Chernoff bound, rather than the exact minimizer λ^*. Besides, we omit the proof of the bounds for the sum of these independent variables, since it's easy to verify. □

Example 2.20 (Bounded variable)**.** Let X be a random variable with mean μ such that $|X-\mu| \leq b$, then we have variance $\sigma^2 = \mathbb{E}(X-\mu)^2 \leq b^2$. Then the Bernstein condition is satisfied with $b/3$, and it is a $(\sqrt{2\sigma}, 2b/3)$–sub-exponential variable. We have sub-exponential bound as

$$\mathbb{P}[X-\mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{4\sigma^2}} & \text{if } 0 \leq t \leq \frac{3\sigma^2}{b} \\ e^{-\frac{t}{4b/3}} & \text{for } t > \frac{3\sigma^2}{b} \end{cases} \tag{2.38}$$

and Bernstein-type bound as

$$\mathbb{P}[X-\mu \geq t] \leq e^{-\frac{t^2}{2\left(\sigma^2+\frac{b}{3}t\right)}} \quad \text{for all } t \geq 0 \tag{2.39}$$

and sub-Gaussian bound with parameter b as

$$\mathbb{P}[X-\mu \geq t] \leq e^{-\frac{t^2}{2b^2}} \quad \text{for all } t \geq 0 \tag{2.40}$$

Proof. Note that for $k \geq 2$, $\frac{k!}{2} \geq 3^{k-2}$ (can be verified by induction). Thus,

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \left|\mathbb{E}\left[(X-\mu)^2\right]\right| \cdot \left|\mathbb{E}\left[(X-\mu)^{k-2}\right]\right| \leq \sigma^2 b^{k-2} \leq \frac{1}{2}k!\sigma^2\left(\frac{b}{3}\right)^{k-2} \qquad \text{for } k = 2,3,4,\ldots \tag{2.41}$$

It means the Bernstein condition is satisfied with $b/3$, and applies the Theorem 2.13, 2.19 and 2.7. □

Corollary 2.21 (Uniform bound for Bernstein-type bound). Suppose that the variables $X_i, i = 1,\dots,n$ are independent, and $\forall i, X_i$ has mean μ and satisfies Bernstein condition with parameter (σ_i, b_i). Then, with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right| \le \frac{2b}{n}\log\frac{2}{\delta} + \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}} \tag{2.42}$$

In particular, if $|X_i-\mu| \le b$, then with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right| \le \frac{2b}{3n}\log\frac{2}{\delta} + \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}} \tag{2.43}$$

Proof. Refer to Theorem 2.19, let $\delta = 2\exp\left(-\frac{nt^2}{2(\sigma^2+bt)}\right)$. By solving the equation, we have

$$t = \frac{b}{n}\log\frac{2}{\delta} + \sqrt{\left(\frac{b}{n}\log\frac{2}{\delta}\right)^2 + \frac{2\sigma^2}{n}\log\frac{2}{\delta}} \le \frac{2b}{n}\log\frac{2}{\delta} + \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}} \tag{2.44}$$

The last inequality is due to $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}, \forall a \ge 0, b \ge 0$. □

Remark 2.22. For bounded variables, we can both apply sub-Gaussian bound, sub-exponential bound and Bernstein-type bound. Since $\sigma^2 = \mathbb{E}\left[(X-\mu)^2\right] \le b^2$, this sub-exponential bound and Bernstein-type bound sometimes can provide sharper inequality than the sub-Gaussian bound, because they use the information of variance. E.g., when $\sigma^2 \ll b^2$, as would be the case for a random variable that occasionally takes on large values, but has a relatively small variance. Such variance-based control frequently plays a key role in obtaining optimal rates in statistical problems.

In particular, we can take an interpretation from an example of a bounded variable. Referring to uniform bound 2.10 and 2.21, consider a sequence of i.i.d. variables where $|X_i-\mu| \le b$, we have event with probability at least $1-\delta$:

$$\begin{aligned}
&\text{Hoeffding: } \left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right| \le \sqrt{\frac{2b^2}{n}\log\frac{2}{\delta}} = \widetilde{\mathcal{O}}\left(\frac{b}{\sqrt{n}}\right) \\
&\text{Bernstein: } \left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right| \le \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}} + \frac{2b}{3n}\log\frac{2}{\delta} = \widetilde{\mathcal{O}}\left(\frac{\sigma}{\sqrt{n}} + \frac{b}{n}\right)
\end{aligned} \tag{2.45}$$

Therefore, for random variables with a small variance compared to their range, Bernstein-type inequality can give a sharper bound.

Theorem 2.23 (Bennett's type Bound). Bennett's inequality is a strengthening of Bernstein's inequality, which is at least as good as Bernstein's inequality:

- Consider a zero-mean random variable such that $|X_i| \le b$ for some $b > 0$, and $\sigma_i^2 = var(X_i)$. Then

$$\log\mathbb{E}\left[e^{\lambda X_i}\right] \le \sigma_i^2\lambda^2\left\{\frac{e^{\lambda b}-1-\lambda b}{(\lambda b)^2}\right\} \quad \text{for all } \lambda \in \mathbb{R} \tag{2.46}$$

- Given independent random variables $X_1,\dots,X_n$ satisfying the condition of part (1), let $\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$ be the average variance. Then

$$\mathbb{P}\left[\sum_{i=1}^{n}X_i \ge n\delta\right] \le \exp\left\{-\frac{n\sigma^2}{b^2}h\left(\frac{b\delta}{\sigma^2}\right)\right\} \tag{2.47}$$

  where $h(t) := (1+t)\log(1+t) - t$ for $t \ge 0$.

Proof. See proposition 3.19 in Duchi (2021). □

Theorem 2.24 (Equivalent characterizations of sub-exponential variables)**.** For a zero-mean random variable X, the following statements are equivalent:

- There are non-negative numbers (ν, b) such that for all $|\lambda| < \frac{1}{b}$ $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\nu^2\lambda^2}{2}}$
- There is a positive number $c_0 > 0$ such that $\mathbb{E}\left[e^{\lambda X}\right] < \infty$ for all $|\lambda| \leq c_0$.
- There are constants $c_1, c_2 > 0$ such that for all $t > 0$ $\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t}$
- The quantity $\gamma := \sup_{k\geq 2}\left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ is finite.

Proof. See Section 2.5 Appendix B in Wainwright (2019). □

2.4 One-side Bound

Theorem 2.25 (One-sided Bernstein's inequality)**.** If $X \leq b$ almost surely, then

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\frac{\lambda^2}{2}\mathbb{E}\left[X^2\right]}{1-\frac{b\lambda}{3}}\right) \quad \text{for all } \lambda \in [0, 3/b) \tag{2.48}$$

Consequently, given n independent random variables such that $X_i \leq b$ almost surely,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq n\delta\right] \leq \exp\left(-\frac{n\delta^2}{2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] + \frac{b\delta}{3}\right)}\right) \tag{2.49}$$

Proof. See Proposition 2.14 in Wainwright (2019). Since the random variable is only bounded from above, we can only derive bounds on its upper tail, rather than both sides. □

3 Martingale-based methods

So far, we introduce various types of bounds on sums of independent random variables. Many problems require bounds on more general functions of independent random variables, i.e., $f(X_1, \ldots, X_n)$. In this case, we would like to understand when $f(X_1, \ldots, X_n)$ concentrates on its expectation. One classical approach is based on martingale decompositions. There are several references introducing this, Duchi (2021) is one high-level treatment not requiring measure-theoretic knowledge, and Wainwright (2019) gives a more rigorous definition.

3.1 Martingale

Definition 3.1 (Martingale)**.** Given a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a martingale if, for all $k \geq 1$,

$$\mathbb{E}[|Y_k|] < \infty \quad \text{and} \quad \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = Y_k \tag{3.1}$$

It is frequently the case that the filtration is defined by a second sequence of random variables $\{X_k\}_{k=1}^{\infty}$ via the canonical σ–fields $\mathcal{F}_k := \sigma(X_1, \ldots, X_k)$. In this case, we say that $\{Y_k\}_{k=1}^{\infty}$ is a martingale sequence with respect to $\{X_k\}_{k=1}^{\infty}$.

Definition 3.2 (Martingale Difference)**.** Let $D_1, D_2, \ldots$ be a sequence of random variables. They form a martingale difference sequence if $Y_n := \sum_{i=1}^{n} D_i$ is a martingale. In particular, for $k \geq 1$, we have

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0.$$

Example 3.3 (Doob Martingales)**.** Given a sequence of independent random variables $\{X_k\}_{k=1}^{\infty}$, define $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ for $k = 1, \ldots, n-1$, and $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$ for some function $f : \mathcal{X}^n \to \mathbb{R}$. Suppose that $\mathbb{E}[|f(X)|] < \infty$, then $\{Y_k\}_{k=1}^{\infty}$ is a martingale sequence with respect to $\{X_k\}_{k=1}^{\infty}$, associated with function f.

Proof. In term of the shorthand $X_1^k = (X_1, \ldots, X_k)$, for the first property we have

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

where the bound is due to the convexity of $|\cdot|$, that is $\left|\mathbb{E}_{X_{k+1:n}}[f(X)|X_1^k]\right| \leq \mathbb{E}_{X_{k+1:n}}[|f(X)|]\big|X_1^k$. Turning to the second property, we have

$$\mathbb{E}\left[Y_{k+1} \mid X_1^k\right] = \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1^{k+1}\right] \mid X_1^k\right] \overset{\text{(i)}}{=} \mathbb{E}\left[f(X) \mid X_1^k\right] = Y_k$$

where we have used the tower property of conditional expectation in step (i). □

Remark 3.4. By the definition of Doob martingale, we see that Y_0 is a constant, and the random variable Y_k will tend to exhibit more fluctuations as we move along the sequence from Y_0 to Y_n. Thus, the martingale approach can result in the tail bound based on the telescoping decomposition,

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{D_k} \tag{3.2}$$

where the sequence $\{D_k\}_{k=1}^{n}$ is an example of a *martingale difference sequence*, which captures exactly the difference between f and its expectation, and plays an important role in the development of concentration inequalities.

Example 3.5 (Partial sums as martingales)**.** Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. random variables with mean μ, and define the partial sums $S_k := \sum_{j=1}^{k} X_j$. Defining $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the variable S_k is measurable with respect to $\mathcal{F}_k$, then $Y_k := S_k - k\mu$ for $k \geq 1$ is a martingale sequence with respect to $\{X_k\}_{k=1}^{\infty}$.

Proof. The recentered partial sums of an i.i.d. sequence is the simplest instance of a martingale.

$$\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[S_{K+1} - (K+1)\mu \mid X_1, \ldots, X_k] = \mathbb{E}[X_{k+1} - \mu] + S_k - k\mu = Y_k.$$

□

3.2 Concentration bounds for martingale difference sequences

With these motivating ideas introduced, we turn to provide generalizations of our concentration inequalities from sub-Gaussian, sub-exponential sums to sub-Gaussian, sub-exponential martingales.

Definition 3.6 (Sub-Gaussian martingale difference)**.** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, Then D_k is a σ_k^2-sub-Gaussian martingale difference if

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\frac{\lambda^2 \sigma_k^2}{2}} \qquad \forall k \text{ and } \lambda \in \mathbb{R} \tag{3.3}$$

Theorem 3.7 (Azuma–Hoeffding Inequality)**.** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a σ_k^2-sub-Gaussian martingale difference sequence. Then $M_n = \sum_{k=1}^{n} D_k$ is $\sum_{k=1}^{n} \sigma_k^2$-sub-Gaussian, and,

$$\mathbb{P}[M_n \geq t] \vee \mathbb{P}[M_n \leq -t] \leq \exp\left(-\frac{nt^2}{2\sum_{k=1}^{n} \sigma_k^2}\right) \qquad \forall t \geq 0 \tag{3.4}$$

Corollary 3.8 (McDiarmid Inequality)**.** Let X_i be independent random variables, and $f : \mathcal{X}^n \to \mathbb{R}$ satisfy bounded difference with constants c_i. That is, $\forall i = 1, \ldots, n, x_1^n \in \mathcal{X}^n$ and $x_i' \in \mathcal{X}$, we have

$$\left| f\left(x_1^{i-1}, x_i, x_{i+1}^n\right) - f\left(x_1^{i-1}, x_i', x_{i+1}^n\right) \right| \le c_i. \tag{3.5}$$

It implies that martingale difference $|D_i| \le c_i$. Then, $M_n = f(X) - \mathbb{E}[f(X)]$ is $\frac{1}{4}\sum_{i=1}^n c_i^2$-sub-Gaussian, and

$$\mathbb{P}[M_n \ge t] \vee \mathbb{P}[M_n \le -t] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \qquad \forall t \ge 0 \tag{3.6}$$

Proof. This condition also means that f is c-Lipschitz with respect to the Hamming norm $d_H(x, y) = \sum_{i=1}^n \mathbb{I}[x_i \ne y_i]$, which counts the number of positions in which x and y differ, then the bounded difference inequality holds with parameter L uniformly across all coordinates. □

Definition 3.9 (Sub-exponential martingale difference)**.** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ be a martingale difference sequence, Then $\{D_k\}$ is a (ν_k, b_k)-sub-exponential martingale difference if

$$\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \le e^{\lambda^2 \nu_k^2/2} \quad \forall k \text{ and } |\lambda| < 1/b_k \tag{3.7}$$

Theorem 3.10. Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ be a (ν_k, b_k)-sub-exponential martingale difference sequence. Then $M_n = \sum_{k=1}^n D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^n \nu_k^2},\ \alpha_*\right)$ where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$, and

$$\mathbb{P}[M_n \ge t] \vee \mathbb{P}[M_n \le -t] \le \begin{cases} e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \le t \le \frac{\sum_{k=1}^n \nu_k^2}{\alpha_k} \\ e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}. \end{cases} \tag{3.8}$$

Theorem 3.11 (Azuma–Bernstein Inequality)**.** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ be a martingale difference sequence, and $|D_k| \le C$, $\mathrm{Var}(X_k|\mathcal{F}_{k-1}) \le \sigma_i^2$. Then for $M_n = \sum_{k=1}^n D_k$,

$$\mathbb{P}[M_n \ge t] \vee \mathbb{P}[M_n \le -t] \le \exp\left(-\frac{t^2}{2\left(\sum_{k=1}^n \sigma_i^2 + \frac{1}{3}Ct\right)}\right) \tag{3.9}$$

Example 3.12 (Rademacher Complexity[1])**.** Let $\mathcal{X}$ be some space, and let $\mathcal{F}$ be some collection of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\varepsilon_i \in \{-1, 1\}$ be a collection of independent random sign vectors. Then the *empirical Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F} \mid x_1^n) := \frac{1}{n}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(x_i)\right], \tag{3.10}$$

where the expectation is over only the random signs ε_i. (In some cases, depending on context and convenience, one takes the absolute value $\left|\sum_i \varepsilon_i f(x_i)\right|$.) The Rademacher complexity of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}\left[R_n(\mathcal{F} \mid X_1^n)\right], \tag{3.11}$$

the expectation of the empirical Rademacher complexities.

If $f : \mathcal{X} \to [b_0, b_1]$ for all $f \in \mathcal{F}$, then the Rademacher complexity satisfies bounded differences, because for any two sequences x_1^n and z_1^n differing in only element j, we have

$$n\left|R_n(\mathcal{F} \mid x_1^n) - R_n(\mathcal{F} \mid z_1^n)\right| \le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i (f(x_i) - f(z_i))\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \varepsilon_i \left(f\left(x_j\right) - f\left(z_j\right)\right)\right] \le b_1 - b_0. \tag{3.12}$$

Consequently, the empirical Rademacher complexity satisfies $R_n\left(\mathcal{F} \mid X_1^n\right) - R_n(\mathcal{F})$ is $\frac{(b_1-b_0)^2}{4n}$ sub-Gaussian by Theorem 3.24.

[1] This quantity captures the richness of a family of functions by measuring the degree to which a hypothesis set can fit random noise, and it plays an important role in generalization error analysis.

4 Uniformity, basic generalization bounds, and complexity classes

In this section, we will apply concentration bounds to learning problems, and provide the non-asymptotic guarantee on the sub-optimality of the learned model by empirical risk-minimizing, mainly based on Ma (2022); Mohri et al. (2018).

4.1 Setup

4.1.1 Supervised Learning

In supervised learning, we have a dataset where each data point is associated with a label, and we aim to learn from the data a function that maps data points to their labels. Formally, suppose we draw a set of n i.i.d. data points $D = \{x^{(i)}, y^{(i)}\}_{i=1}^n$ from a specific joint probability distribution P over $\mathcal{X} \times \mathcal{Y}$, the goal is to learn a function $h : \mathcal{X} \to \mathcal{Y}$ using the training data, i.e., $h(x) = \hat{y}$, which is also called a hypothesis or model or predictor. We define a loss function to measure how good a model is, $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and assume $l(\hat{y}, y) \geq 0$, i.e., the difference between the prediction made by h and the true label.

Precisely, we aims to find a model h that minimizes the *expected loss* (or *population risk*, *expected risk*, also well-known as generalization error Mohri et al. (2018)[2]):

$$L(h) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p} [\ell(h(x), y)] \tag{4.1}$$

In practice, we don't have a way of optimizing over arbitrary functions. Instead, we work within a more constrained set of functions $\mathcal{H}$, which is called *hypothesis family* or *hypothesis class*, e.g., linear functions or neural networks. Given one particular function $h \in \mathcal{H}$, we define the excess risk of h with respect to $\mathcal{H}$:

$$E(h) \triangleq L(h) - \inf_{g\in\mathcal{H}} L(g). \tag{4.2}$$

Generally, we need more assumptions about a specific problem and hypothesis class to bound absolute population risk, hence we focus on bounding the excess risk.

Usually, the family we choose to work with can be parameterized by a vector of parameters $\theta \in \Theta$. In that case, we can refer to an element of H by h_θ, making that explicit.

4.1.2 Empirical risk minimization

Our ultimate goal is to minimize population risk. However, in practice we do not have access to the entire population: we only have a training set of n data points, drawn from the same distribution as the entire population. We can compute empirical risk, the loss over the training set, and try to minimize that and find the minimizer of L, i.e., $\hat{\theta} \triangleq \operatorname{argmin}_{\theta\in\Theta} \widehat{L}(h_\theta)$. The paradigm is known as empirical risk minimization (ERM):

$$\min_{\theta\in\Theta} \widehat{L}(h_\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \ell\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = \frac{1}{n}\sum_{i=1}^n \ell\left(\left(x^{(i)}, y^{(i)}\right), \theta\right) \tag{4.3}$$

Besides, we know that empirical risk and population risk are equal in expectation (over the randomness of the training dataset). This is one reason why it makes sense to use empirical risk: it is an unbiased estimator of the population risk.

The key question that we seek to answer is: **what guarantees do we have on the excess risk for the parameters learned by ERM?** The hope with ERM is that minimizing the training error will lead to small testing errors. One way to make this rigorous is by showing that the ERM minimizer's excess risk is bounded.

[2] They define generalization error on binary classification under 0/1-binary loss, which can be applied more generally.

4.2 Uniform Convergence and Generalization

Uniform convergence is a key tool for proving non-asymptotic guarantees on excess risk, where we have bounds of the following form:

$$\mathbb{P}\left[\sup_{\theta\in\Theta}|\hat{L}(\theta)-L(\theta)|\leq\epsilon\right]\geq 1-\delta \tag{4.4}$$

Let's look at a motivating example on why this type of bounds is useful. Assume we have a bounded loss function, e.g., $l((x,y),\theta)\in[0,1]$. First, we can decompose the excess risk into three terms via telescoping sums:

$$L(\hat{\theta})-L(\theta^*)=\underbrace{L(\hat{\theta})-\hat{L}(\hat{\theta})}_{(1)}+\underbrace{\hat{L}(\hat{\theta})-\hat{L}(\theta^*)}_{(2)}+\underbrace{\hat{L}(\theta^*)-L(\theta^*)}_{(3)}. \tag{4.5}$$

we know the second term is non-positive since $\hat{\theta}$ is a minimizer of $\hat{L}$. Thus,

$$\begin{aligned}L(\hat{\theta})-L(\theta^*)&\leq|L(\hat{\theta})-\hat{L}(\hat{\theta})|+\hat{L}(\hat{\theta})-\hat{L}(\theta^*)+\left|\hat{L}(\theta^*)-L(\theta^*)\right|\\&\leq|L(\hat{\theta})-\hat{L}(\hat{\theta})|+0+\left|\hat{L}(\theta^*)-L(\theta^*)\right|\\&\leq 2\sup_{\theta\in\Theta}|L(\theta)-\hat{L}(\theta)|.\end{aligned} \tag{4.6}$$

This result tells us that if $\sup_{\theta\in\Theta}|L(\theta)-\hat{L}(\theta)|\leq\varepsilon/2$, then excess risk $L(\hat{\theta})-L(\theta^*)\leq\varepsilon$, which is exactly in the form of the uniform convergence.

Let us try to apply our knowledge of concentration inequalities to this problem. Earlier we assumed that $\ell((x,y);\theta)$ is bounded, so we can bound (3) by $\widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$ via Hoeffding's inequality 2.9. However, we cannot apply the same concentration inequality to (1): since $\hat{\theta}$ is data-dependent by definition, the i.i.d. assumption no longer holds. (To see this, note that $\hat{\theta}$ depends on the training dataset $\left\{\left(x^{(i)},y^{(i)}\right)\right\}$, so the terms in $\widehat{L}(\hat{\theta}),\ell\left(\left(x^{(i)},y^{(i)}\right);\hat{\theta}\right)$, all depend on the training dataset too.) This is concerning: it is certainly possible that $L(\hat{\theta})-\hat{L}(\hat{\theta})$ is large. You've probably encountered this yourself when a model exhibits low training loss, but high validation/testing loss. That's why we aim to construct uniform convergence, i.e., $\forall\theta\in\Theta,\hat{L}(\theta)$ converges.

4.2.1 Derive uniform convergence bound

The high-level idea is as follows:

- Suppose we have a bound of the form $\Pr\left[|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right]\leq\delta'$ for some single, fixed choice of θ.
- If we know all possible values of θ in advance, we can use the above bound to create a more general bound over all values of θ.

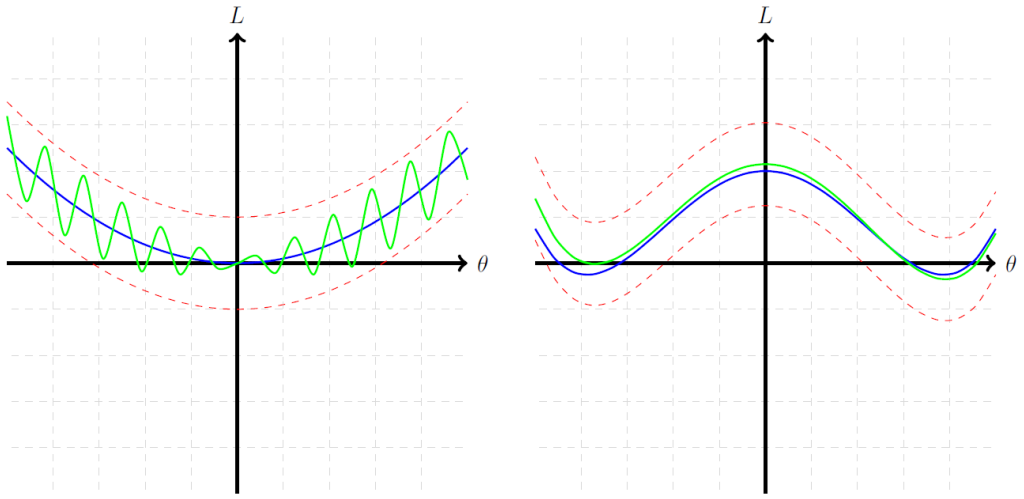
Thus, we use the union-bound inequality to create the general bound:

$$\Pr\left[\forall\theta\in\Theta,|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right]\leq\sum_{\theta\in\Theta}\Pr\left[|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right]. \tag{4.7}$$

We can then use Hoeffding's inequality to deal with the summands as θ there is no longer data-dependent.

4.2.2 Intuitive interpretation of uniform convergence

Since uniform convergence implies generalization, if we know that population risk and empirical risk are always "close," then the excess risk is "small" as well (Figure 2(a)). In fact, it is possible to show that not only is $L(\theta)$ "close" to $\hat{L}(\theta)$ for sufficiently large data, but that the "shape" of $\hat{L}$ is "close" to the shape of L as well (Figure 2(b)). This holds for the convex case; furthermore, there are conditions under which this holds in the non-convex case.



(a) (b)

Figure 2: These curves demonstrate how we apply uniform convergence to bound the population risk. The blue curves are the unobserved population risk we aim to bound. The green curves denote the empirical risk we observe. Though this curve is often depicted as the fluctuating curve used in Figure 2(a), it is more often a smooth curve whose shape mimics that of the population risk (Figure 2(b)). Uniform convergence allows us to construct additive error bounds for the excess risk, which are depicted using the red, dashed lines.

4.2.3 Finite hypothesis class

Theorem 4.1 (Uniform Convergence for Finite Hypothesis Class)**.** Suppose that the hypothesis class $\mathcal{H}$ is finite and that the loss function l is bounded in $[0,1]$, i.e. $0 \le l((x,y),h) \le 1$. Then $\forall \delta$ s.t. $0 < \delta < \frac{1}{2}$, with probability at least $1-\delta$, we have bound for uniform convergence

$$|L(h) - \hat{L}(h)| \le \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}} \quad \forall h \in \mathcal{H}. \tag{4.8}$$

As a corollary, we also have bound for excess risk

$$L(\hat{h}) - L(h^*) \le \sqrt{\frac{2(\ln|\mathcal{H}| + \ln(2/\delta))}{n}}. \tag{4.9}$$

Proof. The proof is in two steps:

1. Use concentration inequalities to prove the bound for a fixed $h \in \mathcal{H}$, then
2. Let $E_h = \{|\hat{L}(h) - L(h)| \ge \epsilon\}$:, and use a union bound across all the h. (Recall that if $E_1, \ldots, E_k$ are a finite set of events, then the union bound states that $\Pr(E_1 \cup \cdots \cup E_k) \le \sum_{i=1}^{k} \Pr(E_i)$.)

$$\mathbb{P}(\exists h \text{ s.t. } |\hat{L}(h) - L(h)| \ge \epsilon) \le \sum_{h \in \mathcal{H}} \Pr(|\hat{L}(h) - L(h)| \ge \epsilon) \le 2|\mathcal{H}| \exp\left(-2n\epsilon^2\right). \tag{4.10}$$

Take $\delta = 2|\mathcal{H}| \exp\left(-2n\epsilon^2\right)$, we obtain the bound. □

Compared with standard concentration inequality, we have the such bound depending on each h,

$$\forall h \in \mathcal{H}, \quad \text{w.h.p.} \quad |\hat{L}(h) - L(h)| \le \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{4.11}$$

In contrast, the uniform convergence bound we obtain from Theorem 4.1 is uniform over all $h \in \mathcal{H}$:

$$\text{w.h.p.} \quad \forall h \in \mathcal{H}, \quad |\hat{L}(h) - L(h)| \leq \widetilde{O}\left(\frac{\ln|\mathcal{H}|}{\sqrt{n}}\right) \tag{4.12}$$

Hence, the extra $\ln|\mathcal{H}|$ term that depends on the size of the finite hypothesis family $\mathcal{H}$ can be viewed as a trade-off in order to make the bound uniform.

4.2.4 Infinite hypothesis class

We can't generalize the results from the previous section directly to the case where the hypothesis class is infinite, which is uncountable. However, if we consider a *bounded* and *continuous* parameterized space of $\mathcal{H}$, we can obtain a similar uniform bound by applying *brute-force discretization*.

Assume that the infinite hypothesis class $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, \|\theta\|_2 \leq B\}$. The intuition behind brute-force discretization is as follows: Let $E_\theta = \{|\widehat{L}(\theta) - L(\theta)| \geq \epsilon\}$ be the "bad" events. We want the bound the probability of any one of these bad events happening (i.e. $\bigcup_\theta E_\theta$). The union bound does not work as we end up with an infinite sum. However, the union bound is very loose: these events can overlap with each other significantly. Instead, we can try to find "prototypical" bad events $E_{\theta_1}, \ldots, E_{\theta_N}$ that are somewhat disjoint so that $\bigcup_\theta E_\theta \approx \bigcup_{i=1}^N E_{\theta_i}$. We can then use the union bound on $\bigcup_{i=1}^N E_{\theta_i}$ to get a non-vacuous upper bound.

Theorem 4.2 (Uniform Convergence for Infinite Hypothesis Class)**.** Suppose $\ell((x,y),\theta) \in [0,1]$, and $\ell((x,y),\theta)$ is κ-Lipschitz in θ with respect to the ℓ_2-norm for all (x,y). Then, with probability at least $1 - O(\exp(-\Omega(p)))$, we have

$$\forall \theta, \quad |\hat{L}(\theta) - L(\theta)| \leq O\left(\sqrt{\frac{p \max(\ln(\kappa B n), 1)}{n}}\right). \tag{4.13}$$

Proof. See Section 4.3.2 in Ma (2022). □

5 Other Content

5.1 Understanding Generalization Error: Bounds and Decompositions

The quantity analyzed in the previous section $L(\theta)$ is also known as *Generalization Error*, which is quite important for learning theory. We have described the simplest type – uniform convergence bound based on the capacity of the function class searched by an algorithm. Here we will discuss the generalization error in a deeper view, refer to Mohri et al. (2018); Agarwal (2018).

In many learning algorithms, they may choose a sufficiently complex model to achieve low training error $\hat{L}(\theta)$, while the generation error would not always decrease together with training error. For example, a high-degree polynomial kernel, a neural network with a large number of hidden nodes, and so on. We observed that models of low complexity tend to underfit the data, while models of high complexity tend to overfit the data, and they both suffer over the generalization error. The challenge is that the "right" model complexity depends on the unknown data distribution, and so must also be estimated from the data itself. This is known as the *model selection* problem.

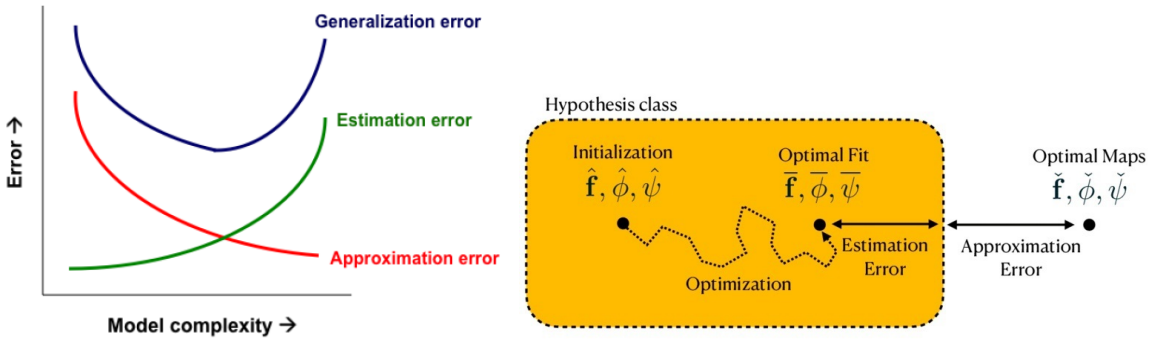
There are two different views for understanding this issue. One is *Estimation-Approximation Error Decomposition*, and the other one is *Bias-Variance Decomposition*.

5.1.1 Estimation-Approximation Error Decomposition

The Generalization error is decomposed as follows:

$$L(\theta) = \underbrace{\left(L(\theta) - \inf_{\theta' \in \Theta} L(\theta')\right)}_{\text{Estimation Error in } \Theta} + \underbrace{\left(\inf_{\theta' \in \Theta} L(\theta') - L(\theta^*)\right)}_{\text{Approximation Error in } \Theta} + \underbrace{L(\theta^*)}_{\text{Irreducible Bayes error}} \tag{5.1}$$

Recall that the Bayes error is the smallest possible generalization error over all possible mapping; it is an irreducible error associated with the distribution P, sometimes also called the "noise" intrinsic to P. The approximation error of $\mathcal{H}$ measures how far the best model in $\mathcal{H}$ is from the Bayes optimal classifier; it is a property of the function class $\mathcal{H}$. The estimation error measures how far the learned classifier h_S is from the best model in $\mathcal{H}$; this is a property of the learning algorithm, and depends on the training sample D (for a good learning algorithm, one would expect that the estimation error would become smaller with increasing sample size n). Figure 3 also demonstrates this idea, a tread-off between the estimation error and approximation error.



(a) Complexity and Errors (b) Decomposition

Figure 3: For a fixed sample size, as model complexity increases, the approximation error decreases, while the estimation error increases. A high value of either contributes to a high generalization error: high approximation error is associated with underfitting; high estimation error is associated with overfitting.

Based on such decomposition, one can using *structual risk minimization* to allow the function class $\mathcal{H}$ to grow with sample size n (so that the approximation error goes to zero), but does slowly that can estimate a good function in the class (so that the estimation error also goes to zero). See Section 4.3 in Mohri et al. (2018).

5.1.2 Bias-Variance Decomposition

The bias-variance decomposition aims to understand the *average* generalization error of the model h_D over any dataset with a fixed size, if we trained an algorithm on several different training datasets D. Thus, our goal is to understand the behavior of average (expected) generalization error[3],

$$\mathbb{E}_D[\hat{L}(h_D)] = \sum_{D'} \mathbb{P}(D = D')\hat{L}(h_{D'}) = \sum_{D'} \mathbb{P}(D = D') \cdot \frac{1}{n} \sum_{i=1}^{n} (h_{D'}(x_i) - y_i^{D'})^2 \tag{5.2}$$

Define an "average" model $\bar{h}(x) = \mathbb{E}_D[h_D(x)]$, then the average (expected) generalization error can be decomposed as follows[4], refer to Agarwal (2018),

$$\begin{aligned} \mathbb{E}_D[\hat{L}(h_D)] &= \mathbb{E}_D\left[(h_D(x) - y^D)^2\right] = \mathbb{E}_D\left[(h_D(x) - \bar{h}(x) + \bar{h}(x) - y^D)^2\right] \\ &= \mathbb{E}_D\left[(h_D(x) - \bar{h}(x))^2\right] + \mathbb{E}_D\left[(\bar{h}(x) - y + y - y^D)^2\right] \\ &= \underbrace{\mathbb{E}_D\left[(h_D(x) - \bar{h}(x))^2\right]}_{\text{Variance}} + \underbrace{(\bar{h}(x) - y)^2}_{\text{Bias}} + \underbrace{\mathbb{E}_D\left[(y - y^D)^2\right]}_{\text{Irreducible Error}} \end{aligned} \tag{5.3}$$

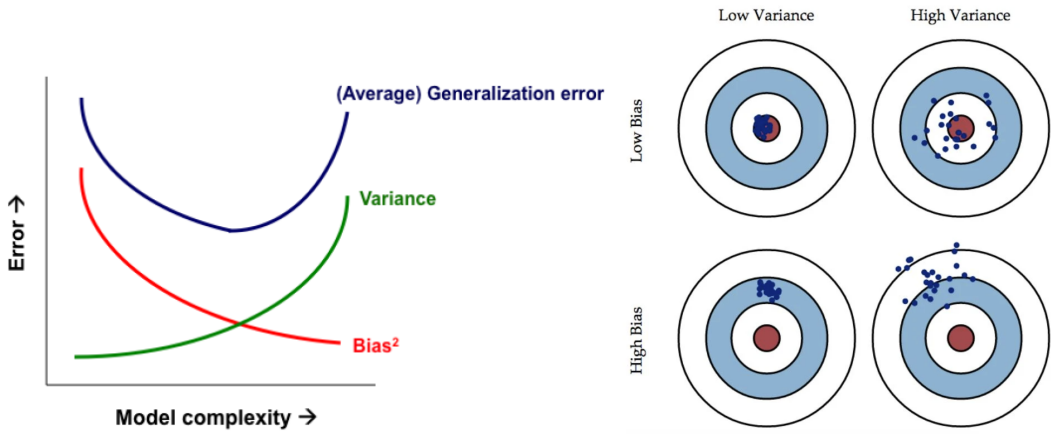
[3]This decomposition is most natural in the context of regression under squared loss, and the sampled label y^D of x in the dataset D may not equal to its real label y.

[4]Note that here, the notions of bias and variance apply to an algorithm, not necessarily to a function class.

The variance term is related to the "stability" of an algorithm: an algorithm with high variance has low stability, in the sense that changing the training sample D a little can produce a very different model h_D. The practice of bootstrap aggregation (bagging), where one creates multiple randomly selected bootstrap samples from a given training sample D and aggregates (averages) the models learned from the various bootstrap samples, can be viewed as a practice aimed at reducing variance. This is especially useful in reducing the error of algorithms that otherwise have high variance, such as decision tree learning algorithms.

The bias term is related to the "accuracy" of an algorithm, and low bias means being closer to the correct label in expectation, referring to Figure 4(b) and Fortmann-Roe (2012).

Looking at Figure 4, as more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls. For example, as more polynomial terms are added to linear regression, the greater the resulting model's complexity will be. In other words, bias has a negative first-order derivative in response to model complexity while variance has a positive slope.



(a) Complexity and Errors

(b) Decomposition Example

Figure 4: For a fixed sample size, as model complexity increases, the bias typically decreases, while the variance typically increases. A high value of either contributes to a high (average) generalization error: high bias is associated with underfitting; high variance is associated with overfitting.

5.2 More examples

Similar issues also arise in the setting of bandits and RL. In particular, the uniform convergence is fundamental for the class of UCB algorithms, which can be obtained through Hoeffding or Bernstein-type inequality.

References

Agarwal, S. (2018). Lecture11 notes of cis 520 machine learning: Understanding generalization error: Bounds and decompositions. *URL : http : //www.shivani – agarwal.net/Teaching/CIS – 520/Spring – 2018/Lectures/Reading/error – bounds – decompositions.pdf*.

Duchi, J. (2021). Lecture notes for statistics 311/electrical engineering 377: Information theory and statistics. *URL: https://web.stanford.edu/class/stats311/lecture-notes.pdf*.

Fortmann-Roe, S. (2012). Understanding the bias-variance tradeoff. *URL : http : //scott.fortmann – roe.com/docs/BiasVariance.html*.

Ma, T. (2022). Lecture notes for machine learning theory: Cs229m / stats214. *URL : https : //raw.githubusercontent.com/tengyuma/cs229m_notes/main/master.pdf*.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.