

Sub-Gaussian Concentration

Junyi Liao

1 Sub-Gaussian random variables

1.1 Moment-generating function and Chernoff bound

In probability theory, the moment-generating function is an alternative characterization of probability distributions. The k -th moment of a distribution can be obtained by evaluating the k -th derivative of its moment-generating function at 0, as is implied by the nomenclature. In contrast to characteristic functions, the moment-generating function of a distribution does not necessarily exist. (As a counterexample, consider a standard Cauchy distribution with density $\frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$.)

Definition 1.1 (Moment-generating function, MGF). Let X be a real-valued random variable such that $\mathbb{E}[e^{tX}]$ exists in some neighborhood of 0, i.e. $\exists b > 0$ such that $\mathbb{E}[e^{tX}] < \infty$ for $t \in (-b, b)$. The moment-generating function (MGF) of X , denoted by M_X , is defined as

$$M_X(t) := \mathbb{E}[e^{tX}]. \quad (1.1)$$

We also define the centered MGF as

$$M_X^*(t) := \mathbb{E}[e^{t(X - \mathbb{E}X)}] = e^{-t\mathbb{E}X} M_X(t). \quad (1.2)$$

It can be verified that the existence of first-moment $\mathbb{E}X$ is ensured by the existence of MGF.

In practical situations, we may wonder if our sample properly depicts the population. In other words, we are interested in the probability that a variable falls in the tail of a distribution. Applying Markov's inequality to the integrand in MGF, we can attain the Chernoff bound:

Lemma 1.2 (Chernoff bound). Suppose that $M_X(t) < \infty$ for all $t \in \mathbb{R}$. Then for all $\epsilon \geq 0$, we have

$$\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq M_X^*(t) e^{-t\epsilon}, \quad \forall t \geq 0. \quad (1.3)$$

To obtain a tight bound, take the infimum of RHS:

$$\mathbb{P}(X - \mathbb{E}X \geq \epsilon) \leq \inf_{t \geq 0} M_X^*(t) e^{-t\epsilon}. \quad (1.4)$$

As an example, let's investigate the Chernoff bound of a Gaussian variable $Z \sim N(0, \sigma^2)$. The MGF of Z is

$$M_Z^*(t) = \frac{1}{\sqrt{2\pi}\sigma} \int e^{tz - \frac{z^2}{2\sigma^2}} dz = \exp\left(\frac{t^2\sigma^2}{2}\right). \quad (1.5)$$

And we get the bound

$$\mathbb{P}(Z \geq \epsilon) \leq \inf_{t \geq 0} \exp\left(\frac{t^2 \sigma^2}{2} - t\epsilon\right) = \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \quad (1.6)$$

1.2 Sub-Gaussian random variable

From the above discussion, we can conclude that if a random variable X satisfies $M_X^*(t) \leq \exp(t^2 \sigma^2 / 2)$ uniformly, then a decay rate in form of (1.6) can be obtained. This motivates the definition of sub-Gaussian random variables.

Definition 1.3 (Sub-Gaussian random variable). Let $\sigma > 0$. A random variable X with mean $\mu = \mathbb{E}X$ is said to sub-Gaussian with variance proxy σ^2 (or σ^2 -sub-Gaussian), if

$$M_X^*(t) = \mathbb{E}[e^{t(X-\mu)}] \leq \exp\left(\frac{t^2 \sigma^2}{2}\right), \quad \forall t \in \mathbb{R}. \quad (1.7)$$

By definition, we know that if a random variable is σ^2 -sub-Gaussian, then it is ρ^2 -gaussian for any $\rho > \sigma$. This definition generalizes Gaussian tail bounds to non-Gaussian variables on the MGF condition. Nevertheless, there are several equivalent characterizations of sub-Gaussianity. This is an exercise in Handel's book, chapter 3.

Theorem 1.4 (Characterizations of sub-Gaussian variables). Let X be a centered random variable, i.e., $\mathbb{E}X = 0$. Then the following statements are equivalent:

(i) (MGF condition). There is a constant $\sigma > 0$ such that

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2 \sigma^2}{2}\right), \quad \forall t \in \mathbb{R}. \quad (1.8)$$

(ii) (Tail bound condition). There is a constant $\rho > 0$ such that

$$\mathbb{P}(|X| \geq \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2\rho^2}\right), \quad \forall \epsilon > 0. \quad (1.9)$$

(iii) There is a constant $\nu > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{X^2}{2\nu^2}\right)\right] \leq 2. \quad (1.10)$$

(iv) (Moment condition) There is a constant $\theta > 0$ such that

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!} \theta^{2k}, \quad \forall k \in \mathbb{N}. \quad (1.11)$$

Proof. (i) \Rightarrow (ii): Fix $\epsilon > 0$. For any $t > 0$, we have

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(e^{tX} \geq e^{-t\epsilon}) \leq e^{t\epsilon} \mathbb{E}[e^{tX}] \leq \exp\left(\frac{1}{2}t^2 \sigma^2 - t\epsilon\right). \quad (1.12)$$

Setting $t = \epsilon/\sigma^2$ implies $\mathbb{P}(X \geq \epsilon) \leq \exp(-\epsilon^2/2\sigma^2)$. By applying similar calculation to $-X$, we can obtain $\mathbb{P}(X \leq -\epsilon) \leq \exp(-\epsilon^2/2\sigma^2)$, and the result (1.9) immediately follows for $\rho = \sigma$.

(ii) \Rightarrow (iii): Suppose (1.9) holds for $\rho > 0$. We will use the following fact, if Y is a random variable that is almost surely non-negative and has distribution F , and ϕ is a differentiable increasing function, then

$$\begin{aligned}\mathbb{E}[\phi(Y)] &= \int_0^\infty \phi(y) dF(y) = \int_0^\infty \left(\phi(0) + \int_0^y \phi'(\epsilon) d\epsilon \right) dF(y) \\ &= \phi(0) + \int_0^\infty \int_t^\infty \phi'(\epsilon) dF(y) d\epsilon = \phi(0) + \int_0^\infty \mathbb{P}(Y \geq \epsilon) \phi'(\epsilon) d\epsilon.\end{aligned}\quad (1.13)$$

Set $Y = X^2$. For any $\nu > \rho$, we have

$$\begin{aligned}\mathbb{E} \left[\exp \left(\frac{X^2}{2\nu^2} \right) \right] &= 1 + \int_0^\infty \frac{1}{2\nu^2} \exp \left(\frac{\epsilon}{2\nu^2} \right) \mathbb{P}(X^2 \geq \epsilon) d\epsilon \\ &\leq 1 + \frac{1}{\nu^2} \int_0^\infty \exp \left\{ \epsilon \left(\frac{1}{2\nu^2} - \frac{1}{2\rho^2} \right) \right\} d\epsilon = 1 + \frac{2\rho^2}{\nu^2 - \rho^2},\end{aligned}\quad (1.14)$$

where the inequality follows from (1.9). Then we can attain (1.10) by setting $\nu = \sqrt{3}\rho$ in (1.14).

(iii) \Rightarrow (iv): Note that $e^x \geq 1 + x^k/k!$ for $x \geq 0$ and $k \in \mathbb{N}$, we have

$$2 \geq \mathbb{E} \left[\exp \left(\frac{X^2}{2\nu^2} \right) \right] \geq 1 + \frac{\mathbb{E}[X^{2k}]}{2^k \nu^{2k} k!}.\quad (1.15)$$

Then

$$\mathbb{E}[X^{2k}] \leq (2k)!! \nu^{2k} \leq (2k+1)!! \nu^{2k} = \frac{(2k)!}{2^k k!} (2k+1) \nu^{2k},\quad (1.16)$$

and (1.11) follows for $\theta = \sqrt{3}\nu$.

(iv) \Rightarrow (i): Let X' be an independent copy of X and $Y := X - X'$. Then Y is symmetric, and the odd moments vanish:

$$\mathbb{E}[e^{tY}] = \sum_{k=0}^\infty \frac{t^k \mathbb{E}[Y^k]}{k!} = \sum_{k=0}^\infty \frac{t^{2k} \mathbb{E}[Y^{2k}]}{(2k)!}.\quad (1.17)$$

For any $k \in \mathbb{N}$, we have the c_r -inequality:

$$Y^{2k} = (X - X')^{2k} = 2^{2k} \left(\frac{X}{2} + \frac{-X'}{2} \right)^{2k} \leq 2^{2k} \left(\frac{1}{2} X^{2k} + \frac{1}{2} (-X')^{2k} \right),\quad (1.18)$$

$$\mathbb{E}[Y^{2k}] \leq 2^{2k} \left(\frac{1}{2} \mathbb{E}[X^{2k}] + \frac{1}{2} \mathbb{E}[(X')^{2k}] \right) = 2^{2k} \mathbb{E}[X^{2k}].\quad (1.19)$$

Plug in to (1.17), we have

$$\mathbb{E}[e^{tY}] \leq \sum_{k=0}^\infty \frac{(2t)^{2k} \mathbb{E}[X^{2k}]}{(2k)!} \leq \sum_{k=0}^\infty \frac{2^k (t\theta)^{2k}}{k!} \leq \exp(2t^2\theta^2).\quad (1.20)$$

Since $\mathbb{E}X = 0$, we have

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{tX - t\mathbb{E}X'}] \leq \mathbb{E}[e^{t(X - X')}] = \mathbb{E}[e^{tY}] \leq \exp(2t^2\theta^2).\quad (1.21)$$

Then (1.8) holds for $\sigma = 2\theta$, and we finish the proof. \square

Proposition 1.5 (Sub-Gaussian vector). Suppose X_1, \dots, X_n are independent sub-Gaussian variables with variance proxy σ^2 . Then for any $u \in \mathbb{R}^n$ with $\|u\|_2 = 1$, $X^\top u$ is σ^2 -sub-Gaussian, where $X = (X_1, \dots, X_n)^\top$ is said to be a σ^2 -sub-Gaussian vector.

$$\text{Proof. } \mathbb{E} \left[e^{tX^\top u} \right] = \prod_{i=1}^n \mathbb{E} \left[e^{tu_i X_i} \right] \leq \prod_{i=1}^n \exp \left(\frac{t^2 u_i^2 \sigma^2}{2} \right) = \exp \left(\frac{t^2 \sigma^2 \|u\|_2^2}{2} \right) = \exp \left(\frac{t^2 \sigma^2}{2} \right). \quad \square$$

1.3 Illustrative examples

The sub-Gaussian family contains a wide range of random variables, such as Gaussian variables, Rademacher variables and bounded variables.

Proposition 1.6 (Rademacher variables are sub-Gaussian). Let X be a Rademacher random variable, i.e. $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Then X is 1-sub-Gaussian.

$$\text{Proof. } \text{For all } t \in \mathbb{R}, \text{ we have } \mathbb{E}[e^{tX}] = \frac{e^t + e^{-t}}{2} = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} = e^{t^2/2}. \quad \square$$

Lemma 1.7 (Hoeffding's lemma). Suppose X is a random variable such that $\mathbb{P}(X \in [a, b]) = 1$. Then X is a sub-Gaussian variable with variance proxy $(b - a)^2/4$.

Proof. This proof is adapted from Handel's notes. Without loss of generality, let $\mathbb{E}X = 0$. Use exponential tilting. Fix $t \in \mathbb{R}$. For any Borel set $B \subseteq \mathcal{B}(\mathbb{R})$, define $\mathbb{P}_t : \mathcal{B}(\mathbb{R}) \rightarrow \mathbb{R}$ as

$$\mathbb{P}_t(B) := \frac{\mathbb{E} \left[e^{tX} \mathbb{1}_{\{X \in B\}} \right]}{\mathbb{E}[e^{tX}]}. \quad (1.22)$$

It can be verified that \mathbb{P}_t is a valid probability measure on \mathbb{R} . Let random variable $U_t \sim \mathbb{P}_t$. Using simple approximation theorem, we have for any measurable function f that

$$\mathbb{E}[f(U_t)] = \frac{\mathbb{E}[e^{tX} f(X)]}{\mathbb{E}[e^{tX}]}. \quad (1.23)$$

Now we investigate the logarithmic MGF $\psi_X(t) = \log \mathbb{E}[e^{tX}]$. Using the interchangeability of derivative and integral, we have

$$\psi'_X(t) = \frac{\mathbb{E}[X e^{tX}]}{\mathbb{E}[e^{tX}]} = \mathbb{E}[U_t], \quad \psi''_X(t) = \frac{\mathbb{E}[X^2 e^{tX}]}{\mathbb{E}[e^{tX}]} - \left(\frac{\mathbb{E}[X e^{tX}]}{\mathbb{E}[e^{tX}]} \right)^2 = \text{Var}(U_t). \quad (1.24)$$

By definition, $\mathbb{P}(U_t \in [a, b]) = \mathbb{P}_t([a, b]) = 1$, hence

$$\text{Var}(U_t) = \mathbb{E}[(U_t - \mathbb{E}U_t)^2] = \inf_{c \in \mathbb{R}} \mathbb{E}[(U_t - c)^2] \leq \mathbb{E} \left[\left(U_t - \frac{a+b}{2} \right)^2 \right] \leq \left(\frac{b-a}{2} \right)^2. \quad (1.25)$$

Using (1.24) and (1.25), we can bound ψ_X as follows:

$$\psi_X(t) = \psi_X(0) + \int_0^t \left(\psi'_X(0) + \int_0^s \psi''_X(u) du \right) ds \leq \int_0^t \int_0^s \left(\frac{b-a}{2} \right)^2 du ds = \frac{t^2(b-a)^2}{8}. \quad (1.26)$$

Thus we complete the proof. \square

2 Gaussian concentration

2.1 Entropy and sub-Gaussianity

Definition 2.1 (Entropy). For a non-negative random variable Y , the entropy of Y is defined as

$$\text{Ent}(Y) = \mathbb{E}[Y \log Y] - \mathbb{E}Y \log(\mathbb{E}Y). \quad (2.1)$$

For a random variable X , the following lemma has established the connection between the entropy of e^{tX} and sub-Gaussianity.

Lemma 2.2 (Herbst). Suppose that random variable X satisfies

$$\text{Ent}(e^{tX}) = \mathbb{E}[tX e^{tX}] - \mathbb{E}[e^{tX}] \log \mathbb{E}[e^{tX}] \leq \frac{t^2 \sigma^2}{2} \mathbb{E}[e^{tX}], \quad \forall t \in \mathbb{R}. \quad (2.2)$$

Then X is σ^2 -sub-Gaussian. Conversely, if X is $\frac{\sigma^2}{4}$ -sub-Gaussian, then it satisfies (2.2).

Proof. (i) Let $\mu = \mathbb{E}X$, and define function $\varphi : \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R}, t \mapsto \frac{1}{t} \log \mathbb{E}[e^{t(X-\mu)}]$, then

$$\frac{d\varphi}{dt}(t) = \frac{1}{t} \frac{\mathbb{E}[(X-\mu)e^{t(X-\mu)}]}{\mathbb{E}[e^{t(X-\mu)}]} - \frac{1}{t^2} \log \mathbb{E}[e^{t(X-\mu)}] = \frac{1}{t} \frac{\mathbb{E}[X e^{tX}]}{\mathbb{E}[e^{tX}]} - \frac{1}{t^2} \log \mathbb{E}[e^{tX}] \leq \frac{\sigma^2}{2}. \quad (2.3)$$

We can complete φ on \mathbb{R} by redefining $\varphi(0) = \lim_{t \rightarrow 0} \varphi(t) = 0$. Then $\varphi(t) - t\sigma^2/2$ is non-increasing on \mathbb{R} , and X is σ^2 -sub-Gaussian:

$$\log \mathbb{E}[e^{tX}] - \frac{t^2 \sigma^2}{2} = t\varphi(t) - \frac{t^2 \sigma^2}{2} \leq t\varphi(0) = 0. \quad (2.4)$$

(ii) Suppose X is $\frac{\sigma^2}{4}$ -sub-Gaussian, and define $Z = e^{tX}/\mathbb{E}[e^{tX}]$. To prove (2.2), it suffices to show that

$$\mathbb{E}[Z \log Z] \leq \frac{t^2 \sigma^2}{2}. \quad (2.5)$$

Suppose $Z \sim F$. Since Z is non-negative and $\mathbb{E}Z = 1$, we can define a new probability measure G such that $dG(z) = z dF(z)$. Then by Jensen's inequality, we have

$$\mathbb{E}[Z \log Z] = \int z \log z dF(z) = \int \log z dG(z) \leq \log \left(\int z dG(z) \right) = \log \mathbb{E}[Z^2]. \quad (2.6)$$

Furthermore, note that $\mathbb{E}[e^{t(X-\mu)}] \geq e^{\mathbb{E}[t(X-\mu)]} = 1$, we have $Z \leq e^{t(X-\mu)}$, and

$$\mathbb{E}[Z \log Z] \leq \log \mathbb{E}[Z^2] \leq \log \mathbb{E}[e^{2t(X-\mu)}] \leq \frac{(2t)^2 \sigma^2}{8} = \frac{t^2 \sigma^2}{2}, \quad (2.7)$$

where the last equality follows from the sub-Gaussianity of X . Hence we conclude the proof. \square

2.2 Lipschitz function of Gaussian variables

Lemma 2.3 (Gaussian log-Sobolev inequality). Let $d\mu(z) = (2\pi)^{-n/2} \exp(-\frac{1}{2}\|z\|^2) dz$ be the standard Gaussian measure on \mathbb{R}^n . Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth function such that $\|f\|_{L^2(\mu)} := \int |f|^2 d\mu = 1$. Then

$$\int f^2 \log f^2 d\mu \leq 2 \int \|\nabla f\|_2^2 d\mu. \quad (2.8)$$

We do not cover the proof here since it is a bit complicated. Nonetheless, we can understand (2.8) from an information theory perspective. Define $g = f^2$ and $d\nu = g d\mu$, it can be verified that ν is also a probability measure on \mathbb{R}^n , and $g = d\nu/d\mu$ is the Radon-Nikodym derivative of ν with respect to μ . Moreover, (2.8) can be written as

$$\int g \log g d\mu \leq \frac{1}{2} \int \frac{\|\nabla g\|_2^2}{g} d\mu. \quad (2.9)$$

The LHS is the Kullback-Leibler divergence (or relative entropy) from μ to ν , and the RHS is half the relative Fisher information:

$$D_{\text{KL}}(\nu\|\mu) := \int \log \left(\frac{d\nu}{d\mu} \right) d\nu \leq \frac{1}{2} \int \|\nabla \log g\|_2^2 d\nu =: \frac{1}{2} \mathcal{I}(\nu\|\mu). \quad (2.10)$$

Therefore, this lemma gives an upper bound for the Kullback-Leibler divergence between ν and μ in terms of their relative entropy.

Theorem 2.4 (Gaussian concentration). Let $X \sim N(0, I_n)$, and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -Lipschitz continuous function. Then $f(X)$ is a sub-Gaussian variable with variance proxy L^2 .

Proof. Fix a smooth function $h \geq 0$ with $\|h\|_{L^1(\mu)} = \int h d\mu > 0$. Applying Theorem 2.3 to $f = \sqrt{\frac{h}{\|h\|_{L^1(\mu)}}}$, (2.9) can be written as

$$\int h \log h d\mu - \|h\|_{L^1(\mu)} \log \|h\|_{L^1(\mu)} \leq \frac{1}{2} \int \frac{\|\nabla h\|_2^2}{h} d\mu. \quad (2.11)$$

Suppose $f \in C_c^\infty(\mathbb{R}^n)$ and f is L -Lipschitz continuous. Fix $t \in \mathbb{R}$ and set $h = e^{tf}$. By (2.11), we have

$$\text{Ent} \left(e^{tf(X)} \right) \leq \frac{t^2}{2} \mathbb{E} \left[\|\nabla f(X)\|_2^2 e^{tf(X)} \right] \leq \frac{t^2 L^2}{2} \mathbb{E} \left[e^{tf(X)} \right], \quad (2.12)$$

where the last equality holds because f is L -Lipschitz continuous. By Lemma 2.2, $f(X)$ is L^2 -sub-Gaussian.

Now it remains to show that the conclusion holds for all L -Lipschitz f . (f is not necessarily differentiable.) Choose a non-negative $\psi \in C_c^\infty(\mathbb{R}^n)$ such that $\text{supp}(\psi) \subseteq \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ and $\int \psi(x) dx = 1$, and define $\psi_\epsilon(x) := \frac{1}{\epsilon} \psi\left(\frac{x}{\epsilon}\right)$ for $\epsilon > 0$. Then $\int \psi_\epsilon(x) dx = 1$.

Fix $\epsilon > 0$, and define $f_\epsilon = \psi_\epsilon * f : x \mapsto \int \psi_\epsilon(x - y) f(y) dy$. Then $f_\epsilon \in C_c^\infty(\mathbb{R}^n)$, and f_ϵ is L -Lipschitz:

$$\begin{aligned} |f_\epsilon(x) - f_\epsilon(x')| &\leq \int \psi_\epsilon(y) |f(x - y) - f(x' - y)| dy \\ &\leq \int \psi_\epsilon(y) L \|x - x'\|_2 dy \leq L \|x - x'\|_2. \end{aligned} \quad (2.13)$$

Moreover, f_ϵ converges uniformly to f as $\epsilon \rightarrow 0$:

$$\begin{aligned} \|f_\epsilon - f\|_\infty &:= \sup_{x \in \mathbb{R}^n} |f_\epsilon(x) - f(x)| = \sup_{x \in \mathbb{R}^n} \left| \int \psi_\epsilon(x - y) (f(y) - f(x)) dy \right| \\ &\leq \sup_{x \in \mathbb{R}^n} \int_{\|y - x\|_2 \leq \epsilon} \psi_\epsilon(x - y) L \|y - x\|_2 dy \\ &\leq \epsilon L \int_{\|y\|_2 \leq \epsilon} \psi_\epsilon(-y) dy = \epsilon L. \end{aligned} \quad (2.14)$$

Fix $t \in \mathbb{R}$. For any $\epsilon > 0$ and $x \in \mathbb{R}^n$, we have $e^{tf(x)} \leq e^{tf_\epsilon(x) + |t|\epsilon L}$. Moreover, $f_\epsilon \in C_c^\infty(\mathbb{R}^n)$ and f_ϵ is continuous, then $f_\epsilon(X)$ is L^2 -sub-Gaussian. Therefore

$$\mathbb{E} \left[e^{tf(X)} \right] \leq \inf_{\epsilon > 0} \mathbb{E} \left[e^{tf_\epsilon(X)} \right] e^{|t|\epsilon L} \leq \inf_{\epsilon > 0} \exp \left(\frac{t^2 L^2}{2} + |t|\epsilon L \right) = \exp \left(\frac{t^2 L^2}{2} \right), \quad (2.15)$$

which concludes the proof. \square

3 Tail bound for mean and maxima

3.1 Hoeffding bound

Proposition 3.1. Let X_1, \dots, X_n be independent sub-Gaussian variables with variance proxies $\sigma_1^2, \dots, \sigma_n^2$. Then

$$\mathbb{P} \left(\sum_{i=1}^n (X_i - \mathbb{E} X_i) \geq \epsilon \right) \leq \exp \left(-\frac{\epsilon^2}{2 \sum_{i=1}^n \sigma_i^2} \right). \quad (3.1)$$

Proof. It can be easily verified that $\sum_{i=1}^n (X_i - \mathbb{E} X_i)$ is a sub-Gaussian variable with mean 0 and variance proxy $\sum_{i=1}^n \sigma_i^2$. Then (3.1) immediately follows from (1.9) in Theorem 1.4. \square

Combining Lemma 1.7 and Theorem 3.1 gives the following Hoeffding's inequality:

Theorem 3.2 (Hoeffding). Let X_1, \dots, X_n be independent random variables such that $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ for $i = 1, \dots, n$. Then

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E} X_i) \geq \epsilon \right) \leq \exp \left\{ -\frac{2n^2 \epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right\}. \quad (3.2)$$

3.2 Maximum of sub-Gaussian variables

Suppose we have n centered independent sub-Gaussian variables with variance proxy σ^2 . A natural tail bound for the maximum can be attained from the fact that $\{\max_{1 \leq i \leq n} X_i \geq \epsilon\} = \bigcup_{i=1}^n \{X_i \geq \epsilon\}$:

$$\mathbb{P} \left(\max_{1 \leq i \leq n} X_i \geq \epsilon \right) \leq n \exp \left(-\frac{\epsilon^2}{2\sigma^2} \right). \quad (3.3)$$

We can also bound the expected value as follows.

Theorem 3.3. Let X_1, \dots, X_n be independent σ^2 -sub-Gaussian variables with mean zero. Then

$$\mathbb{E} \left[\max_{1 \leq i \leq n} X_i \right] \leq \sigma \sqrt{2 \log n}. \quad (3.4)$$

Proof. Fix $\epsilon > 0$. By Jensen's inequality, we have

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq i \leq n} X_i \right] &\leq \frac{1}{\epsilon} \log \mathbb{E} \left[\exp \left(\epsilon \max_{1 \leq i \leq n} X_i \right) \right] \leq \frac{1}{\epsilon} \log \mathbb{E} \left[\sum_{i=1}^n e^{\epsilon X_i} \right] \\ &\leq \frac{1}{\epsilon} \log \left\{ \sum_{i=1}^n \exp \left(\frac{\epsilon^2 \sigma^2}{2} \right) \right\} = \frac{\log n}{\epsilon} + \frac{\epsilon \sigma^2}{2}. \end{aligned} \quad (3.5)$$

Then we conclude the proof by setting $\epsilon = \sqrt{2 \log n} / \sigma$. \square

An immediate corollary of this theorem is the Massart's finite class lemma.

Lemma 3.4 (Massart). Let \mathcal{A} be a finite subset of \mathbb{R}^n and $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher variables. Denote by $r_{\mathcal{A}} = \max_{a \in \mathcal{A}} \|a\|_2$ the radius of \mathcal{A} . Then we have

$$\mathbb{E} \left[\max_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^m \epsilon_i a_i \right] \leq \frac{r_{\mathcal{A}} \sqrt{2 \log |\mathcal{A}|}}{n}. \quad (3.6)$$

4 Tail bound for quadratic forms

4.1 Gaussian quadratic forms

Lemma 4.1 (Hsu et al., 2012). Let Z_1, \dots, Z_m be independent standard Gaussian variables. Fix non-negative vector $\alpha \in \mathbb{R}_+^m$ and vector $\beta \in \mathbb{R}^m$. If $0 \leq t < \frac{1}{2\|\alpha\|_\infty}$, then

$$\mathbb{E} \left[\exp \left(t \sum_{i=1}^m \alpha_i Z_i^2 + \sum_{i=1}^m \beta_i Z_i \right) \right] \leq \exp \left(t \|\alpha\|_1 + \frac{t^2 \|\alpha\|_2^2 + \|\beta\|_2^2 / 2}{1 - 2t \|\alpha\|_\infty} \right). \quad (4.1)$$

Proof. Fix $0 \leq t < \frac{1}{2\|\alpha\|_\infty}$, and let $\eta_i = 1/\sqrt{1 - 2t\alpha_i} > 0$ for $i = 1, \dots, m$. Then

$$\begin{aligned} \mathbb{E} [\exp \{ t\alpha_i Z_i^2 + \beta_i Z_i \}] &= \frac{1}{\sqrt{2\pi}} \int \exp \left\{ -\left(\frac{1}{2} - t\alpha_i\right) z^2 + \beta_i z \right\} dz \\ &= \frac{1}{\sqrt{2\pi}} \int \exp \left\{ -\frac{1}{2} \left(\frac{z}{\eta_i} - \beta_i \eta_i \right)^2 + \frac{1}{2} \beta_i^2 \eta_i^2 \right\} dz \\ &= \eta_i \exp \left(\frac{1}{2} \beta_i^2 \eta_i^2 \right) = \exp \left\{ -\frac{1}{2} \log(1 - 2t\alpha_i) + \frac{\beta_i^2}{2(1 - 2t\alpha_i)} \right\}. \end{aligned} \quad (4.2)$$

To bound (4.2), note that

$$-\log(1 - 2t\alpha_i) = \sum_{k=1}^{\infty} \frac{(2t\alpha_i)^k}{k} \leq 2t\alpha_i + \sum_{k=2}^{\infty} \frac{(2t\alpha_i)^k}{2} = 2t\alpha_i + \frac{2t^2\alpha_i^2}{1 - 2t\alpha_i}. \quad (4.3)$$

Combining (4.2) and (4.3), we have

$$\mathbb{E} [\exp (t\alpha_i Z_i^2 + \beta_i Z_i)] \leq \exp \left(t\alpha_i + \frac{t^2\alpha_i^2 + \beta_i^2/2}{1 - 2t\alpha_i} \right) \leq \exp \left(t\alpha_i + \frac{t^2\alpha_i^2 + \beta_i^2/2}{1 - 2t\|\alpha\|_\infty} \right). \quad (4.4)$$

Summation of (4.4) from $i = 1$ to m immediately yields (4.1). \square

4.2 Quadratic forms of sub-Gaussian variables

Theorem 4.2 (Tail bound for quadratic form). Let X_1, \dots, X_n be independent sub-Gaussian variables with mean 0 and variance proxy σ^2 . Then for any positive definite matrix $\Sigma \in \mathbb{R}^{n \times n}$ and $t > 0$, we have

$$\mathbb{P} \left(X^\top \Sigma X \geq \sigma^2 \left\{ \text{tr}(\Sigma) + 2\|\Sigma\|_F \sqrt{t} + 2\|\Sigma\|_{2t} t \right\} \right) \leq e^{-t}, \quad (4.5)$$

where $X = (X_1, \dots, X_n)^\top$ is the vector of sub-Gaussian variables.

Proof. Since Σ is positive definite, it admits a spectral decomposition $\Sigma = Q^\top S Q$ where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $S = \text{diag}\{\rho_1, \dots, \rho_n\}$ with eigenvalues $\rho_1 \geq \dots \geq \rho_n > 0$. Let Z be a vector of n

independent standard Gaussian variables. Then for any $\alpha \in \mathbb{R}^n$ and $\epsilon > 0$, we have

$$\mathbb{E} \left[e^{Z^\top \alpha} \right] = e^{\|\alpha\|_2^2/2}. \quad (4.6)$$

Denote $A = Q^\top S^{1/2} Q$. For any $\epsilon > 0$, define set $E_\epsilon = \{x \in \mathbb{R}^n : x^\top \Sigma x \geq \epsilon\}$. Fix $\lambda > 0$, we have

$$\begin{aligned} \mathbb{E} [\exp(\lambda Z^\top A X)] &= \int_{\mathbb{R}^n} \mathbb{E} [\exp(\lambda Z^\top A X) | X = x] dF_X(x) \\ &\geq \int_{E_\epsilon} \mathbb{E} [\exp(\lambda Z^\top A X) | X = x] dF_X(x) \\ &= \int_{E_\epsilon} \exp\left(\frac{1}{2} \lambda^2 Z^\top \Sigma Z\right) dF_X(x) \\ &\geq \exp\left(\frac{1}{2} \lambda^2 \epsilon\right) \mathbb{P}(X^\top \Sigma X \geq \epsilon), \end{aligned} \quad (4.7)$$

where the second equality follows from (4.6). Moreover,

$$\mathbb{E} [\exp(\lambda Z^\top A X)] \leq \mathbb{E} \left[\exp\left(\frac{\lambda^2 \sigma^2}{2} Z^\top \Sigma Z\right) \right]. \quad (4.8)$$

Combining (4.7) and (4.8) yields

$$\mathbb{P}(X^\top \Sigma X \geq \epsilon) \leq \mathbb{E} \left[\exp\left(\frac{\lambda^2 \sigma^2}{2} Z^\top \Sigma Z - \frac{1}{2} \lambda^2 \epsilon\right) \right]. \quad (4.9)$$

Define $Y = QZ$, the orthogonality of Q implies that Y is also a vector of n independent standard Gaussian variables, and $Z^\top \Sigma Z = Y^\top S Y = \sum_{i=1}^n \rho_i Y_i^2$. Let $\rho = (\rho_1, \dots, \rho_n)^\top$ and $\gamma = \lambda^2 \sigma^2 / 2$. By Lemma 3.1, we have for $0 \leq \gamma < \frac{1}{2\|\rho\|_\infty}$ that

$$\mathbb{P}(X^\top \Sigma X \geq \epsilon) \leq \exp\left(-\frac{\gamma \epsilon}{\sigma^2} + \gamma \|\rho\|_1 + \frac{\gamma^2 \|\rho\|_2^2}{1 - 2\gamma \|\rho\|_\infty}\right). \quad (4.10)$$

Let $\delta = 1 - 2\gamma \|\rho\|_\infty$ with $0 < \delta \leq 1$. Then

$$\mathbb{P}(X^\top \Sigma X \geq \epsilon) \leq \exp\left\{\frac{1}{2\|\rho\|_\infty} \left[(1 - \delta) \left(\|\rho\|_1 - \frac{\epsilon}{\sigma^2}\right) + \frac{\|\rho\|_2^2}{2\|\rho\|_\infty} (\delta + \delta^{-1} - 2)\right]\right\}. \quad (4.11)$$

Let $\frac{\epsilon}{\sigma^2} - \|\rho\|_1 = \frac{\|\rho\|_2^2}{2\|\rho\|_\infty} (\delta^{-2} - 1)$, we have

$$\mathbb{P}\left(X^\top \Sigma X \geq \sigma^2 \left\{\|\rho\|_1 + \frac{\|\rho\|_2^2}{2\|\rho\|_\infty} \left(\frac{1}{\delta^2} - 1\right)\right\}\right) \leq \exp\left\{-\frac{\|\rho\|_2^2}{4\|\rho\|_\infty^2} \left(\frac{1}{\delta} - 1\right)^2\right\}. \quad (4.12)$$

Now let $t = \frac{\|\rho\|_2^2}{4\|\rho\|_\infty^2} (\delta^{-1} - 1)^2 \geq 0$, that is, $\delta^{-1} = 1 + \frac{2\|\rho\|_\infty}{\|\rho\|_2} \sqrt{t}$, then

$$\mathbb{P}\left(X^\top \Sigma X \geq \sigma^2 \left\{\|\rho\|_1 + 2\|\rho\|_2 \sqrt{t} + 2\|\rho\|_\infty t\right\}\right) \leq e^{-t}. \quad (4.13)$$

Recall that ρ_1, \dots, ρ_n are eigenvalues of Σ , we have $\|\rho\|_1 = \text{tr}(\Sigma)$, $\|\rho\|_2 = \|\Sigma\|_F$ and $\|\rho\|_\infty = \|\Sigma\|_2$, and the result (4.5) immediately follows from (4.13). \square

The following corollary immediately holds by setting Σ in Theorem 4.2 as the n -by- n identity matrix.

Corollary 4.3. Let X_1, \dots, X_n be independent sub-Gaussian variables with mean 0 and variance proxy σ^2 . Then for any $t > 0$, we have

$$\mathbb{P} \left(\sum_{i=1}^n X_i^2 \geq \sigma^2 (n + 2\sqrt{nt} + 2t) \right) \leq e^{-t}. \quad (4.14)$$

4.3 Application: Ordinary least square with a fixed design

We consider a fixed dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ from a linear model: $y_i = x_i^\top \beta^* + \epsilon_i$, where $\beta^* \in \mathbb{R}^d$ and $\{\epsilon_i\}_{i=1}^N$ are independent σ^2 -sub-Gaussian noises with $\mathbb{E}\epsilon_i = 0$. The solution to the ordinary least square (OLS) problem is

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^N (y_i - x_i^\top \beta)^2 = \Sigma^{-1} \left(\sum_{i=1}^N x_i y_i \right), \quad (4.15)$$

where $\Sigma = \sum_{i=1}^N x_i x_i^\top \in \mathbb{R}^{d \times d}$. In many cases, we are interested in the difference between our estimator $\hat{\beta}$ and the true parameter β^* .

Proposition 4.4 (OLS with a fixed design). Assume that Σ is invertible, and $0 < \delta < 1$. With probability at least $1 - \delta$, we have

$$\|\hat{\beta} - \beta^*\|_\Sigma^2 \leq \sigma^2 \left(d + 2\sqrt{d \log(1/\delta)} + 2 \log(1/\delta) \right). \quad (4.16)$$

Proof. Denote by $X = (x_1, \dots, x_N)^\top \in \mathbb{R}^{N \times d}$ the covariate matrix, $\epsilon = (\epsilon_1, \dots, \epsilon_N)^\top$ the noise vector, and $Y = (y_1, \dots, y_N)^\top = X\beta^* + \epsilon$ the response vector. Then we have $\Sigma = X^\top X$, and

$$\|\hat{\beta} - \beta^*\|_\Sigma^2 = (\Sigma^{-1} X^\top Y - \beta^*)^\top \Sigma (\Sigma^{-1} X^\top Y - \beta^*) = \epsilon^\top X \Sigma^{-1} X^\top \epsilon. \quad (4.17)$$

Note that ϵ is σ^2 -sub-Gaussian, $\operatorname{tr}(X \Sigma^{-1} X^\top) = d$, $\|X \Sigma^{-1} X^\top\|_F^2 = d$ and $\|X \Sigma^{-1} X^\top\|_2 = 1$, we can apply Theorem 4.2 to any $t > 0$:

$$\mathbb{P} \left(\epsilon^\top X \Sigma^{-1} X^\top \epsilon \geq \sigma^2 (d + 2\sqrt{dt} + 2t) \right) \leq e^{-t}. \quad (4.18)$$

Then the result immediately follows from (4.18) by setting $t = \log(1/\delta)$. \square

5 Empirical process

5.1 Dudley's entropy integral

Definition 5.1 (Sub-Gaussian process). Let $\{X_f : f \in \mathcal{F}\}$ be a collection of mean zero random variables indexed by $f \in \mathcal{F}$, and let d be a metric on the index set \mathcal{F} . Then $\{X_f : f \in \mathcal{F}\}$ is said to be a sub-Gaussian process with respect to d if

$$\mathbb{E} \left[e^{t(X_f - X_g)} \right] \leq \exp \left\{ \frac{t^2 d^2(f, g)}{2} \right\}, \quad \forall f, g \in \mathcal{F}. \quad (5.1)$$

That is, $X_f - X_g$ is sub-Gaussian with variance proxy $d^2(f, g)$.

Definition 5.2 (ϵ -covering number/metric entropy). Let (\mathcal{F}, d) be a metric space. For $\epsilon > 0$, a subset $\mathcal{N}_\epsilon \subseteq \mathcal{F}$ is called a ϵ -net of \mathcal{F} , if $\mathcal{F} \subseteq \bigcup_{f \in \mathcal{N}_\epsilon} B(f, \epsilon)$, where $B(f, \epsilon)$ is the open d -ball of radius ϵ centered at f . The ϵ -covering number of \mathcal{F} is the cardinality of the minimal ϵ -cover of \mathcal{F} , i.e.:

$$N(\epsilon, \mathcal{F}, d) = \min\{|\mathcal{N}_\epsilon|, \mathcal{N}_\epsilon \text{ is an } \epsilon\text{-net of } \mathcal{F}\}. \quad (5.2)$$

The following theorem can be seen as an extension of Theorem 3.3.

Theorem 5.3 (Dudley). Let (\mathcal{F}, d) be a metric space, and suppose that $D := \sup_{f, g \in \mathcal{F}} d(f, g) < \infty$. Let $\{X_f : f \in \mathcal{F}\}$ be a stochastic process such that

- (i) $\{X_f, f \in \mathcal{F}\}$ is sub-Gaussian with respect to d , and
- (ii) $\{X_f, f \in \mathcal{F}\}$ is sample-continuous, i.e., for each sequence $\{f_n\} \subset \mathcal{F}$ such that $\lim_{n \rightarrow \infty} d(f_n, f) = 0$ for some $f \in \mathcal{F}$, we have $X_{f_n} \rightarrow X_f$ almost surely.

Then the expected supremum of $\{X_f, f \in \mathcal{F}\}$ can be bounded with Dudley's entropy integral:

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} X_f \right] \leq 12 \int_0^{D/2} \sqrt{\log N(\epsilon, \mathcal{F}, d)} \, d\epsilon. \quad (5.3)$$

Proof. This proof uses Dudley's chaining rule. Choose an arbitrary $f_0 \in \mathcal{F}$ and set $\epsilon_0 = D$, then $\mathcal{N}_{\epsilon_0} = \{f_0\}$ is a ϵ_0 -net of \mathcal{F} . Now we choose a sequence of minimal ϵ -nets $\{\mathcal{N}_{\epsilon_j}\}$ by setting $\epsilon_j := 2^{-j} \epsilon_0$ for $j = 1, 2, \dots$. For brevity, write $\mathcal{N}_j = \mathcal{N}_{\epsilon_j}$. By definition of ϵ -net, $\forall f \in \mathcal{F}$, we can find $f_j \in \mathcal{N}_j$ such that $d(f, f_j) \leq \epsilon_j$ for all $j \in \mathbb{N}$. Fixing $m \in \mathbb{N}$, we have

$$X_f = (X_f - X_{f_m}) + \sum_{j=1}^m (X_{f_j} - X_{f_{j-1}}) + X_{f_0}. \quad (5.4)$$

Note that both f_j and f_{j-1} are close to f , we have $d(f_j, f_{j-1}) = d(f_j, f) + d(f, f_{j-1}) \leq 3\epsilon_j$. Define a new class $\mathcal{H}_j = \{(g_{j-1}, g_j) \in \mathcal{N}_{j-1} \times \mathcal{N}_j : d(g_j, g_{j-1}) \leq 3\epsilon_j\}$, $j \in \mathbb{N}$. We have $|\mathcal{H}_j| \leq |\mathcal{N}_{j-1}| |\mathcal{N}_j| \leq |\mathcal{N}_j|^2$.

Revisiting (4.4), we have

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} X_f \right] &\leq \mathbb{E} \left[\sup_{g, g' \in \mathcal{F}, d(g, g') \leq \epsilon_m} (X_g - X_{g'}) + \sum_{j=1}^m \max_{(g_{j-1}, g_j) \in \mathcal{H}_j} (X_{g_j} - X_{g_{j-1}}) + X_{f_0} \right] \\ &= \mathbb{E} \left[\sup_{g, g' \in \mathcal{F}, d(g, g') \leq \epsilon_m} (X_g - X_{g'}) \right] + \sum_{j=1}^m \mathbb{E} \left[\max_{(g_{j-1}, g_j) \in \mathcal{H}_j} (X_{g_j} - X_{g_{j-1}}) \right], \end{aligned} \quad (5.5)$$

where the equality follows from $\mathbb{E}[X_{f_0}] = 0$. Since $\{X_{g_j} - X_{g_{j-1}}, (g_{j-1}, g_j) \in \mathcal{H}_j\}$ are sub-Gaussian with variance proxy $9\epsilon_j^2$, it follows from applying Theorem 3.3 that

$$\mathbb{E} \left[\max_{(g_{j-1}, g_j) \in \mathcal{H}_j} (X_{g_j} - X_{g_{j-1}}) \right] \leq 3\epsilon_j \sqrt{2 \log |\mathcal{H}_j|} \leq 6\epsilon_j \sqrt{\log |\mathcal{N}_j|} = 12(\epsilon_j - \epsilon_{j+1}) \sqrt{\log N(\epsilon_j, \mathcal{F}, d)}. \quad (5.6)$$

Plug in (5.6) to (5.5), we have

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} X_f \right] &\leq \mathbb{E} \left[\sup_{g, g' \in \mathcal{F}, d(g, g') \leq \epsilon_m} (X_g - X_{g'}) \right] + 12 \sum_{j=1}^m \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\log N(\epsilon, \mathcal{F}, d)} \, d\epsilon \\ &\leq \mathbb{E} \left[\sup_{g, g' \in \mathcal{F}, d(g, g') \leq \epsilon_m} (X_g - X_{g'}) \right] + 12 \sum_{j=1}^m \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{\log N(\epsilon, \mathcal{F}, d)} \, d\epsilon \\ &= \mathbb{E} \left[\sup_{g, g' \in \mathcal{F}, d(g, g') \leq \epsilon_m} (X_g - X_{g'}) \right] + 12 \int_{\epsilon_{m+1}}^{\epsilon_1} \sqrt{\log N(\epsilon, \mathcal{F}, d)} \, d\epsilon. \end{aligned} \quad (5.7)$$

Let $m \rightarrow \infty$, then $\epsilon_m \rightarrow 0$, and the first term in (5.7) converges to 0 by sample-continuity. Thus we obtain the bound in (5.3) provided that Dudley's entropy integral exists. \square

Remark. (Absolute values in suprema). In some cases, we may be interested in the supremum of absolute value. Note that

$$\begin{aligned}
\sup_{f \in \mathcal{F}} |X_f| &= \sup_{f \in \mathcal{F}} X_f + \sup_{f \in \mathcal{F}} (-X_f) - \sup_{f \in \mathcal{F}} X_f \wedge \sup_{f \in \mathcal{F}} (-X_f) \\
&= \sup_{f \in \mathcal{F}} X_f + \sup_{f \in \mathcal{F}} (-X_f) + \inf_{f \in \mathcal{F}} X_f \vee \inf_{f \in \mathcal{F}} (-X_f) \\
&\leq \sup_{f \in \mathcal{F}} X_f + \sup_{f \in \mathcal{F}} (-X_f) + \inf_{f \in \mathcal{F}} (X_f \vee (-X_f)) = \sup_{f \in \mathcal{F}} X_f + \sup_{f \in \mathcal{F}} (-X_f) + \inf_{f \in \mathcal{F}} |X_f|, \tag{5.8}
\end{aligned}$$

then by applying Theorem 5.3 to both X_f and $-X_f$, we have

$$\begin{aligned}
\mathbb{E} \left[\sup_{f \in \mathcal{F}} |X_f| \right] &\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} X_f \right] + \mathbb{E} \left[\sup_{f \in \mathcal{F}} (-X_f) \right] + \inf_{f \in \mathcal{F}} \mathbb{E} |X_f| \\
&\leq 24 \int_0^{D/2} \sqrt{\log N(\epsilon, \mathcal{F}, d)} d\epsilon + \inf_{f \in \mathcal{F}} \mathbb{E} |X_f|. \tag{5.9}
\end{aligned}$$

We can also use the chaining rule to construct a tail bound for the supremum of a sub-Gaussian process.

Lemma 5.4 (Adapted from Lemma 3.2 of van de Geer 2000). Suppose (\mathcal{F}, d) and $\{X_f, f \in \mathcal{F}\}$ are the metric space and the stochastic process proposed in Theorem 5.3, the entropy integral in the RHS of (5.3) exists, and $\exists f_0 \in \mathcal{F}$ such that $X_{f_0} = 0$. Then there exist constants $C_0, C_1 > 0$ depending only on \mathcal{F} , such that for all $t > C_0 D$, we have

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} X_f \geq t \right) \leq C_1 \exp \left(-\frac{t^2}{C_1^2 D^2} \right). \tag{5.10}$$

Proof. We inherit the definition of $\epsilon_j = D2^{-j}, \mathcal{N}_j$ and \mathcal{H}_j from Theorem 3, with the crudest ϵ_0 -net being $\mathcal{N}_0 = \{f_0\}$. Take C_0 sufficiently large such that

$$t \geq \left(12 \sum_{j=1}^{\infty} \epsilon_j \sqrt{2 \log |\mathcal{N}_j|} \right) \vee 6D \geq 24D \sqrt{\log \frac{24}{23}}. \tag{5.11}$$

Inspired by (5.4) and (5.5), we choose a sequence $\{\eta_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty} \eta_j \leq 1$. Then

$$\begin{aligned}
\mathbb{P} \left(\sup_{f \in \mathcal{F}} X_f \geq t \right) &\leq \mathbb{P} \left(\sum_{j=1}^{\infty} \max_{(g_{j-1}, g_j) \in \mathcal{H}_j} (X_{g_j} - X_{g_{j-1}}) \geq t \sum_{j=1}^{\infty} \eta_j \right) \\
&\leq \sum_{j=1}^{\infty} \mathbb{P} \left(\max_{(g_{j-1}, g_j) \in \mathcal{H}_j} (X_{g_j} - X_{g_{j-1}}) \geq \eta_j t \right) \leq \sum_{j=1}^{\infty} \exp \left(2 \log |\mathcal{N}_j| - \frac{\eta_j^2 t^2}{18 \epsilon_j^2} \right), \tag{5.12}
\end{aligned}$$

where the last inequality follows from (3.3). Now take

$$\eta_j = \frac{6 \epsilon_j \sqrt{2 \log |\mathcal{N}_j|}}{t} \vee \frac{2^{-j} \sqrt{j}}{4}, \tag{5.13}$$

then by (5.11) we have

$$\sum_{j=1}^{\infty} \eta_j \leq \frac{6\sqrt{2}}{t} \sum_{j=1}^{\infty} \epsilon_j \sqrt{\log |\mathcal{N}_j|} + \frac{1}{4} \sum_{j=1}^{\infty} 2^{-j} \sqrt{j} \leq \frac{1}{2} + \frac{1}{2} = 1. \quad (5.14)$$

Here we use the bound

$$\sum_{j=1}^{\infty} 2^{-j} \sqrt{j} \leq \sum_{j=1}^{\infty} 2^{-j} j = \frac{2^{-1}}{(1 - 2^{-1})^2} = 2. \quad (5.15)$$

By (5.13), we have that $\log |\mathcal{N}_j| \leq \frac{\eta_j^2 t^2}{72\epsilon_j}$ and $\eta_j \geq \frac{2^{-j}\sqrt{j}}{4} = \frac{\epsilon_j \sqrt{j}}{4D}$, hence

$$\begin{aligned} \mathbb{P} \left(\sup_{f \in \mathcal{F}} X_f \geq t \right) &= \sum_{j=1}^{\infty} \exp \left(2 \log |\mathcal{N}_j| - \frac{\eta_j^2 t^2}{18\epsilon_j^2} \right) \leq \sum_{j=1}^{\infty} \exp \left(-\frac{\eta_j^2 t^2}{36\epsilon_j^2} \right) \leq \sum_{j=1}^{\infty} \exp \left(-\frac{jt^2}{576D^2} \right) \\ &= \left[1 - \exp \left(-\frac{t^2}{576D^2} \right) \right]^{-1} \exp \left(-\frac{t^2}{576D^2} \right) \leq 24 \exp \left(-\frac{t^2}{576D^2} \right), \end{aligned} \quad (5.16)$$

where the last inequality uses (5.11). Plug in (5.16) to (5.12), then (5.10) holds for $C_1 = 24$, which concludes the proof. \square

5.2 Rademacher complexity

Definition 5.5 (Empirical Rademacher complexity). The empirical Rademacher complexity of a function class \mathcal{F} based on a sample $\{x_i\}_{i=1}^n$ is defined as the expected supremum of inner product with independent Rademacher variables $\{\epsilon_j\}_{j=1}^n$:

$$\mathcal{R}(\mathcal{F}, x_{1:n}) := \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right]. \quad (5.17)$$

Denote by P_n the empirical distribution of $\{x_1, \dots, x_n\}$. Then we can define norm and inner product on $L^2(P_n)$ space:

$$\|f\|_{P_n} = \left(\frac{1}{n} \sum_{i=1}^n f(x_i)^2 \right)^{1/2}, \quad \langle f, g \rangle_{P_n} = \frac{1}{n} \sum_{i=1}^n f(x_i)g(x_i). \quad (5.18)$$

Now we define the process

$$Z_f := \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(x_i), \quad f \in \mathcal{F}. \quad (5.19)$$

By proposition 1.5, $(Z_f - Z_g)$ is sub-Gaussian with variance proxy $\|f - g\|_{P_n}^2$ for any $f, g \in \mathcal{F}$, namely, $\{Z_f, f \in \mathcal{F}\}$ is sub-Gaussian with respect to $\|\cdot\|_{P_n}$. It is worth noting that $\|\cdot\|_{P_n}$ is possibly a pseudo-metric on \mathcal{F} , which means that $\|f\|_{P_n} = 0$ does not necessarily imply $f = 0$. Nonetheless, this slight change does not impact our conclusion, and you can verify that $\{Z_f, f \in \mathcal{F}\}$ is sample-continuous. Using Theorem 5.3, we can establish the connection between Dudley's entropy integral and Rademacher complexity.

Definition 5.6 (Localized empirical Rademacher complexity and critical radius). Suppose we have a function class $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}$ that is uniformly bounded by b , i.e. $\forall f \in \mathcal{F}, \|f\|_{\infty} \leq b$. The localized empirical Rademacher

complexity of a function class \mathcal{F} based on a sample $\{x_i\}_{i=1}^n$ is defined as

$$\mathcal{R}_{\text{loc}}(\delta, \mathcal{F}, x_{1:n}) := \mathbb{E} \left[\sup_{f \in \mathcal{F}: \|f\|_{P_n} \leq \delta} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right] = \mathcal{R}(\mathcal{F} \cap B_n(\delta), x_{1:n}), \quad (5.20)$$

where $\{\epsilon_i\}_{i=1}^n$ are independent Rademacher variables and $B_n(\delta)$ is the closed ball in the norm $\|\cdot\|_{P_n}$ of radius $\delta > 0$ centered at the origin. The empirical critical radius of \mathcal{F} on dataset $\{x_i\}_{i=1}^n$ is defined as the minimum solution smallest positive solution to $\mathcal{R}_{\text{loc}}(\delta, \mathcal{F}, x_{1:n}) \leq \delta^2/b$:

$$\hat{\delta}_n = \min \left\{ \delta > 0 : \mathcal{R}_{\text{loc}}(\delta, \mathcal{F}, x_{1:n}) \leq \frac{\delta^2}{b} \right\}. \quad (5.21)$$

Proposition 5.7. Denote by $B_n(\rho)$ the closed ball in the norm $\|\cdot\|_{P_n}$ of radius $\rho > 0$ centered at the origin. Then the empirical Rademacher complexity of \mathcal{F} satisfies

$$\mathcal{R}_{\text{loc}}(\rho, \mathcal{F}, x_{1:n}) \leq \frac{12}{\sqrt{n}} \int_0^\rho \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|_{P_n})} d\epsilon. \quad (5.22)$$

Proof. Applying Theorem 5.3 to $\{Z_f := n^{-1/2} \sum_{i=1}^n \epsilon_i f(x_i), f \in \mathcal{F}\}$ immediately concludes the proof. \square

5.3 Sub-Gaussian complexity

Motivation. Let's consider a penalized least square problem. Suppose we have data $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ collected from

$$y_i = f^*(x_i) + \epsilon_i, \quad i = 1, \dots, n. \quad (5.23)$$

Given a vector space \mathcal{F} of mappings from \mathcal{X} to \mathcal{Y} with $f^* \in \mathcal{F}$, and let J be seminorm on \mathcal{F} . We construct an estimator of f^* from this class by minimizing the regularized risk for some tuning parameter $\lambda \geq 0$:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J(f) \right\}. \quad (5.24)$$

Recall that we denote by P_n the empirical distribution of $\{x_1, \dots, x_n\}$. We also abuse the notation $\langle \cdot, \cdot \rangle_{P_n}$ by defining $\langle \cdot, \cdot \rangle_{P_n} : \mathbb{R}^n \times \mathcal{F} \rightarrow \mathbb{R}, (z, f) \mapsto \frac{1}{n} \sum_{i=1}^n z_i f(x_i)$. Let $Y = (y_1, \dots, y_n)^\top$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ be the response and noise vectors. Then (5.15) implies

$$\|Y - \hat{f}\|_{P_n}^2 + \lambda J(\hat{f}) \leq \|Y - f^*\|_{P_n}^2 + \lambda J(f^*), \quad (5.25)$$

and we have the basic inequality for \hat{f} :

$$\begin{aligned} \|\hat{f} - f^*\|_{P_n}^2 &\leq 2\langle \epsilon, \hat{f} - f^* \rangle_{P_n} + \lambda(J(f^*) - J(\hat{f})) \\ &\leq 2(J(f^*) + J(\hat{f})) \left\langle \epsilon, \frac{\hat{f} - f^*}{J(\hat{f}) + J(f^*)} \right\rangle_{P_n} + \lambda(J(f^*) - J(\hat{f})) \\ &\leq 2(J(f^*) + J(\hat{f})) \sup_{J(g) \leq 1} \langle \epsilon, g \rangle_{P_n} + \lambda(J(f^*) - J(\hat{f})). \end{aligned} \quad (5.26)$$

Then we can bound the empirical estimation error by controlling the supremum of an empirical process $\{Z_g := \langle \epsilon, g \rangle_{P_n}\}$ indexed by g . Generally, for a function class \mathcal{F} , we call $\sup_{f \in \mathcal{F}} |\langle \epsilon, f \rangle_{P_n}|$ the sub-Gaussian complexity associated with \mathcal{F} .

Lemma 5.7 (Adapted from Lemma 8.4 of van de Geer 2000). Let $\{\epsilon_i, i = 1, \dots, n\}$ denote independent sub-Gaussian random variables, each having mean zero and variance proxy σ^2 . Assume that there exist constants $0 < w < 2$ and $C > 0$ such that for some fixed x_1, \dots, x_n (which define the empirical norm $\|\cdot\|_{P_n}$),

$$\log N(\eta, \mathcal{F}, \|\cdot\|_{P_n}) \leq C\eta^{-w} \quad (5.27)$$

for sufficiently small $\eta > 0$. Then for any fixed $\rho > 0$, there exists constants $c, c' > 0$, depending only on σ, ρ, C, w such that for all $\gamma > c'$,

$$\sup_{f \in \mathcal{F} \cap B_n(\rho)} \frac{\langle \epsilon, f \rangle_{P_n}}{\|f\|_{P_n}^{1-w/2}} \leq \frac{\gamma}{\sqrt{n}} \quad (5.28)$$

with probability at least $1 - c \exp\left(-\frac{\gamma^2}{c^2}\right)$.

Proof. Note that $\frac{\langle \epsilon, f \rangle_{P_n}}{\sqrt{\sigma^2/n}}$ is a sub-Gaussian process with respect to $\|\cdot\|_{P_n}$. For any $0 < \delta < \rho$, the Dudley's entropy integral satisfies

$$\int_0^\delta \sqrt{\log N(\eta, \mathcal{F}, \|\cdot\|_{P_n})} d\eta \leq c_0 \delta^{1-w/2} \quad (5.29)$$

for some constant $c_0 > 0$, hence is bounded. By Lemma 5.4, there exists $c_1, c_2 > 0$ depending only on σ, ρ, C, w such that for all $T \geq \frac{c_1 \delta}{\sqrt{n}}$, we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{F} \cap B_n(\delta)} \langle \epsilon, f \rangle_{P_n} \geq T\right) \leq c_2 \exp\left(-\frac{nT^2}{c_2^2 \sigma^2 \delta^2}\right). \quad (5.30)$$

Then for any $T \geq \frac{c_1}{\sqrt{n}} 2^{1-w/2} \rho^{w/2}$, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F} \cap B_n(\rho)} \frac{\langle \epsilon, f \rangle_{P_n}}{\|f\|_{P_n}^{1-w/2}} \geq T\right) &= \mathbb{P}\left(\bigcup_{j=1}^{\infty} \left\{\sup_{f \in \mathcal{F} \cap (B_n(2^{1-j}\rho) \setminus B_n(2^{-j}\rho))} \frac{\langle \epsilon, f \rangle_{P_n}}{\|f\|_{P_n}^{1-w/2}} \geq T\right\}\right) \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{f \in \mathcal{F} \cap B_n(2^{1-j}\rho)} \langle \epsilon, f \rangle_{P_n} \geq T(2^{-j}\rho)^{1-w/2}\right) \\ &\leq \sum_{j=1}^{\infty} c_2 \exp\left(-\frac{nT^2(2^{-j}\rho)^{2-w}}{c_2^2 \sigma^2 (2^{1-j}\rho)^2}\right) = \sum_{j=1}^{\infty} c_2 \exp\left(-\frac{nT^2 2^{jw}}{4c_2^2 \sigma^2 \rho^w}\right) \\ &\leq \sum_{j=1}^{\infty} c_2 \exp\left(-\frac{nT^2(1+jw \log 2)}{4c_2^2 \sigma^2 \rho^w}\right) \\ &= c_2 \exp\left(-\frac{nT^2}{4c_2^2 \sigma^2 \rho^w}\right) \frac{\exp\left(-\frac{nT^2 w \log 2}{4c_2^2 \sigma^2 \rho^w}\right)}{1 - \exp\left(-\frac{nT^2 w \log 2}{4c_2^2 \sigma^2 \rho^w}\right)} \\ &\leq c_2 \exp\left(-\frac{nT^2}{4c_2^2 \sigma^2 \rho^w}\right) \frac{\exp\left(-\frac{c_1^2 w \log 2}{c_2^2 2^w \sigma^2}\right)}{1 - \exp\left(-\frac{c_1^2 w \log 2}{c_2^2 2^w \sigma^2}\right)} \leq c \exp\left(-\frac{nT^2}{c^2}\right) \end{aligned} \quad (5.31)$$

for some $c > 0$. Set $\gamma = \frac{T}{\sqrt{n}}$, then there exists $c' > 0$ depending only on ρ, σ, C, w such that for any $\gamma \geq c'$,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F} \cap B_n(\rho)} \frac{\langle \epsilon, f \rangle_{P_n}}{\|f\|_{P_n}^{1-w/2}} \geq \frac{\gamma}{\sqrt{n}}\right) \leq c \exp\left(-\frac{\gamma^2}{c^2}\right). \quad (5.32)$$

Thus we complete the proof. \square

This Lemma gives a bound of the empirical process $\{\langle \epsilon, f \rangle_{P_n}, f \in \mathcal{F}\}$. Suppose that $\|f\|_{P_n}$ decays with a rate of $n^{-1/(2+w)}$, then with high probability, the following inequality holds uniformly for all $f \in \mathcal{F}$:

$$\langle \epsilon, f \rangle_{P_n} \lesssim \mathcal{O} \left(\frac{\|f\|_{P_n}^{1-w/2}}{\sqrt{n}} \right) \approx \mathcal{O} \left(n^{-\frac{2}{2+w}} \right). \quad (5.33)$$

Lemma 5.8. Assume that \mathcal{F} satisfies the entropy bound (5.27) for fixed x_1, \dots, x_n , where $0 < w < 2$ and $C > 0$ are constants. Then the empirical critical radius of \mathcal{F} satisfies $\hat{\delta}_n \leq c_1 n^{-1/(2+w)}$ for a constant c_1 .

Proof. By (5.22) and (5.29), we have

$$\mathcal{R}_{\text{loc}}(\delta, \mathcal{F}, x_{1:n}) \leq \frac{c_0}{\sqrt{n}} \delta^{1-w/2} \quad (5.34)$$

for some constant $c_0 > 0$. Then the smallest solution $\hat{\delta}_n$ to $\mathcal{R}_{\text{loc}}(\delta, \mathcal{F}, x_{1:n}) \leq \delta^2/b$ can be upper bounded by

$$\delta^2/b = \frac{c_0}{\sqrt{n}} \delta^{1-w/2} \Leftrightarrow \delta^{1+w/2} = \frac{c_0 b}{\sqrt{n}}, \quad (5.35)$$

which gives $\hat{\delta}_n \leq c_1 n^{-1/(2+w)}$ for some constants c_1 . □

References

- [1] Ryan Tibshirani. *Empirical Process Theory for Nonparametric Analysis*. Notes for Advanced Topics in Statistical Learning, Spring 2023.
- [2] Ramon van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University. December 21, 2016.
- [3] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.
- [4] John Lafferty, Han Liu, and Larry Wasserman. *Concentration of Measure*. Carnegie Mellon University.
- [5] Alekh Agarwal, Nan Jiang, Sham M. Kakade and Wen Sun. *Reinforcement Learning: Theory and Algorithms* (draft of January 31, 2022).
- [6] Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.