Commun. Math. Res.
doi: 10.4208/cmr.2020-0041

Vol. **37**, No. 1, pp. 1-85
February 2021

REVIEW ARTICLE

Concentration Inequalities for Statistical Inference

Huiming Zhang[1,3,4] and Song Xi Chen[1,2,3,*]

[1] *School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China.*
[2] *Guanghua School of Management, Peking University, Beijing 100871, P.R. China.*
[3] *Center for Statistical Sciences, Peking University, Beijing 100871, P.R. China.*
[4] *Department of Mathematics, Faculty of Science and Technology, University of Macau, P.R. China.*

Received 4 November 2020; Accepted 12 December 2020

Abstract. This paper gives a review of concentration inequalities which are widely employed in non-asymptotical analyses of mathematical statistics in a wide range of settings, from distribution-free to distribution-dependent, from sub-Gaussian to sub-exponential, sub-Gamma, and sub-Weibull random variables, and from the mean to the maximum concentration. This review provides results in these settings with some fresh new results. Given the increasing popularity of high-dimensional data and inference, results in the context of high-dimensional linear and Poisson regressions are also provided. We aim to illustrate the concentration inequalities with known constants and to improve existing bounds with sharper constants.

AMS subject classifications: 60F10, 60G50, 62E17

Key words: Constants-specified inequalities, sub-Weibull random variables, heavy-tailed distributions, high-dimensional estimation and testing, finite-sample theory, random matrices.

*Corresponding author. *Email addresses:* `zhanghuiming@pku.edu.cn` (H. Zhang), `csx@gsm.pku.edu.cn` (S. X. Chen)

1 Introduction

In probability theory and statistical inference, researchers often need to bound the probability of a difference between a random quantity from its target, usually the error bound of estimation. Concentration inequalities (CIs) are tools for attaining such bounds, and play important roles in deriving theoretical results for various inferential situations in statistics and probability. The recent developments in high-dimensional (HD) statistical inference, and statistical and machine learning have generated renewed interests in the CIs, as reflected in [29, 47, 84, 86]. As the CIs are diverse in their forms and the underlying distributional requirements, and are scattered around in references, there is an increasing need for a review which collects existing results together with some new results (sharper and constants-specified CIs) from the authors for researchers and graduate students working in statistics and probability. This motivates the writing of this review.

CIs enable us to obtain non-asymptotic results for estimating, constructing confidence intervals, and doing hypothesis testing with a high-probability guarantee. For example, the first-order optimized condition for HD linear regressions should be held with a high probability to guarantee the well-behavior of the estimator. The concentration inequality for error distributions is to ensure the concentration from first-order optimized conditions to the estimator. Our review focuses on four types of CIs:

$$P(Z_n > \mathrm{E}Z_n + t), \quad P(Z_n < \mathrm{E}Z_n - t), \quad P(|Z_n - \mathrm{E}Z_n| > t), \quad \mathrm{E}(\max_{i=1,\ldots,n} |X_i|),$$

where $Z_n := f(X_1, \cdots, X_n)$ and $X_1, \cdots, X_n$ are random variables. We present two types of CIs: distribution-free and distribution-dependent. Distribution free CIs are free of distribution assumptions, while the distribution-dependent CIs are based on exponential moment conditions reflecting the tail property for the particular class of distributions. Concentration phenomenons for a sum of sub-Weibull random variables will lead to a mixture of two tails: sub-Gaussian for small deviations and sub-Weibull for large deviations from the mean, and it is closely related to Strong Law of Large Numbers, Central Limit Theorem, and Law of the Iterative Logarithm. We provide applications of the CIs to empirical processes and high-dimensional data settings. The latter includes the linear and Poisson regression with a diverging number of covariates. We organize the materials in the forms of lemmas, corollaries, propositions, and theorems. Lemmas and corollaries are on existing results usually without proof except for a few fundamental ones. Propositions are also for existing results but with sharper or more precise constants and sometimes come with proofs. Theorems are for new results. This review contains 26 lemmas, 21 corollaries, 14 propositions, and 4 theorems.

The review is organized as follows. Section 2 outlines distribution-free CIs. CIs for Sub-Gaussian, Sub-exponential, sub-Gamma, and sub-Weibull random variables are given in Sections 3, 4, 5, and 6, respectively. Section 7 reports concentration for the maximal of random variables and suprema of empirical processes. Applications for high dimensional linear and Poisson regression are outlined in Section 8. Section 9 discusses extensions to other settings.

2 Distribution-free concentration bounds

The purpose here is to introduce distribution-free CIs. We first review Markov's, Chebyshev's and Chernoff's tail probability bounds that constitute fundamental inequalities for deriving most of the concentration bounds, see [24, Chap. 1] or [34, Appendix B] for the proofs.

Lemma 2.1 (Markov's inequality). *Let $\varphi(x):\mathbb{R}\to\mathbb{R}^+$ be any non-decreasing positive function. For any real valued random variable (RV) X,*

$$P(X\geq a)\leq \mathrm{E}[\varphi(X)]\frac{1}{\varphi(a)},\quad \forall\, a\in\mathbb{R}.$$

By letting $\varphi(x)=x^2$, the following Chebyshev's inequality is merely an application of Markov's inequality for $|X-\mathrm{E}X|$.

Lemma 2.2 (Chebyshev's Inequality). *Let X be an RV with expectation $\mathrm{E}X$ and variance $\mathrm{Var}X$. Then, for any $a\in\mathbb{R}^+$*

$$P(|X-\mathrm{E}X|\geq a)\leq\frac{\mathrm{Var}X}{a^2}.$$

The Chebyshev's inequality prescribes a polynomial rate of convergence depending on the variance assumption. Another application of Markov's inequality is the Chernoff's bound which is sharper by optimizing the upper bounds.

Lemma 2.3 (Chernoff's inequality). *For an RV X with $\mathrm{E}e^{tX}<\infty$,*

$$P(X\geq a)\leq \inf_{t>0}\left\{e^{-ta}\mathrm{E}e^{tX}\right\}.$$

Proof. Lemma 2.1 with $\varphi(x)=e^{tx}$ implies $P(X\geq a)\leq e^{-ta}\mathrm{E}e^{tX}$ and minimize t on $t>0$. □

The Jensen's inequality and its truncated version [18, Lemma 14.6] are another powerful tool to derive useful inequalities by the convexity.

Lemma 2.4 (Jensen's inequality). *For any convex function $\varphi:\mathbb{R}^d\to\mathbb{R}$ and any RV X in $\mathbb{R}^d$, such that $\varphi(X)$ is integrable, we have $\varphi(\mathrm{E}X)\le\mathrm{E}[\varphi(X)]$.*

Lemma 2.5 (Truncated Jensen's inequality). *Let $g(\cdot)$ be an increasing function on $[0,\infty)$, which is concave on $[c,\infty)$ for some $c\ge 0$. Then*

$$\mathrm{E}g(|Z|)\le g\big[\mathrm{E}|Z|+cP(|Z|<c)\big]$$

for RV Z.

The Chebyshev's, Markov's, Chernoff's and Jensen's inequalities are also valid for conditional expectations [24, Chapter 4]. The Chernoff's bound typically lead to a tighter bound than Markov's inequality by optimization via an exponential $\varphi(x)$ function. A sharper bound for the sum of independent random variables (RVs) was attempted in [39]. The following is a slightly sharper bound from [13, Theorem 1.2].

Corollary 2.1 (Hoeffding's inequality). *Let $X_1,\cdots,X_n$ be independent RVs on $\mathbb{R}$ satisfying bound condition $a_i\le X_i\le b_i$ for $i=1,\ldots,n$. Then for $t,u>0$*

(a) ***Hoeffding's lemma:***

$$\mathrm{E}e^{u\sum_{i=1}^n(X_i-\mathrm{E}X_i)}\le e^{\frac{u^2}{8}\sum_{i=1}^n(b_i-a_i)^2},$$
$$\mathrm{E}e^{u|\sum_{i=1}^n(X_i-\mathrm{E}X_i)|}\le 2e^{\frac{u^2}{8}\sum_{i=1}^n(b_i-a_i)^2};$$

(b) ***Hoeffding's inequality:***

$$P\left(\left|\sum_{i=1}^n(X_i-\mathrm{E}X_i)\right|\ge t\right)\le 2e^{\frac{-2t^2}{\sum_{i=1}^n(b_i-a_i)^2}}.$$

Corollary 2.1 has a sharper bound than the Markov's inequality or Chebyshev's inequality with the requirement of first or moment condition on X. Hoeffding's inequality has many applications in statistics as shown in the next example.

The proof of Hoeffding's lemma. Without loss of generality, we assume $\mathrm{E}X_i=0$. This is from the fact that the concentration inequality is location shift-invariance. Since $f(x)=e^x$ is convex, for $u>0$, then

$$e^{ux}\le\frac{b_i-x}{b_i-a_i}e^{ua_i}+\frac{x-a_i}{b_i-a_i}e^{ub_i},\quad a_i\le x\le b_i.$$

Taking expectation, it gives by $\mathrm{E}X_i=0$

$$\mathrm{E}e^{uX_i}\le\frac{b_i}{b_i-a_i}e^{ua_i}-\frac{a_i}{b_i-a_i}e^{ub_i}=\left[1-s+se^{u(b_i-a_i)}\right]e^{-su(b_i-a_i)}\triangleq e^{f(r)},\tag{2.1}$$

where $r=u(b_i-a_i)$, $s=-\frac{a_i}{(b_i-a_i)}$ and $f(r)=-sr+\log(1-s+se^r)$. We can show that

$$f'(r)=-s+\frac{se^r}{1-s+se^r},\quad f''(r)=\frac{(1-s)se^r}{(1-s+se^r)^2}\le\frac{1}{4}\quad\text{for all}\quad r\ge 0.$$

Note that $f(0)=f'(0)=0$. Consider the Taylor's expansion of f, there exists $\xi\in[0,1]$ such that

$$f(r)=\frac{r^2f''(\xi r)}{2}\le\frac{r^2}{8}=\frac{u^2(b_i-a_i)^2}{8}.$$

Substitute it to (2.1), we get the Hoeffding's lemma.

The last assertion of Lemma 2.1(a) is by letting $Z=u\sum_{i=1}^n(X_i-\mathrm{E}X_i)$, so that

$$\mathrm{E}e^{|Z|}=\mathrm{E}e^{-Z}\cdot 1(Z\le 0)+\mathrm{E}e^{Z}\cdot 1(Z>0)\le 2e^{\frac{1}{8}u^2\sum_{i=1}^n(b_i-a_i)^2}.\tag{2.2}$$

The proof of Hoeffding's inequality. Let $S_n=\sum_{i=1}^n X_i$ and $c_i=a_i-b_i$. For any $t,u>0$,

$$\begin{aligned}
&P(S_n-\mathrm{E}S_n\ge t)=P\left(e^{u(S_n-\mathrm{E}S_n)}\ge e^{ut}\right)\\
&\le\inf_{u>0}e^{-ut}\prod_{i=1}^n\mathrm{E}e^{u(X_i-\mathrm{E}X_i)} &&:\text{[Chernoff's inequality]}\\
&\le\inf_{u>0}e^{-ut}\prod_{i=1}^n e^{\frac{u^2c_i^2}{8}} &&:\text{[Hoeffding's lemma]}\\
&=\inf_{u>0}e^{-ut+\frac{1}{8}u^2\sum_{i=1}^n c_i^2}=e^{\frac{-2t^2}{\sum_{i=1}^n c_i^2}}.
\end{aligned}\tag{2.3}$$

The smallest bound is attained at $u=\frac{4t}{\sum_{i=1}^n c_i^2}$ and

$$P\big(-[S_n-\mathrm{E}S_n]\ge t\big)\le e^{\frac{-2t^2}{\sum_{i=1}^n c_i^2}}$$

similarly. Hence, the Hoeffding's inequality is verified via

$$P\big(|S_n-\mathrm{E}S_n|\ge t\big)\le P\big(S_n-\mathrm{E}S_n\ge t\big)+P\big(-[S_n-\mathrm{E}S_n]\ge t\big)\le 2e^{\frac{-2t^2}{\sum_{i=1}^n c_i^2}}.$$

Corollary 2.1 has a sharper bound than the Markov's inequality or Chebyshev's inequality with the requirement of first or moment condition on X. A second approach for proving Hoeffding's lemma is given in [70, Lemma 1.8]. Hoeffding's inequality has many applications in statistics as shown in the next example.

Example 2.1 (Empirical distribution function, EDF)**.** Let $\{X_i\}_{i=1}^n \overset{\text{IID}}{\sim} F(x)$ for a distribution F. Let

$$\mathbb{F}_n(x) := \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \le x\}}(x), \quad x \in \mathbb{R}$$

be the empirical distribution. By Hoeffding's inequality ($a_i - b_i = \frac{1}{n}$),

$$P\big(|\mathbb{F}_n(x) - F(x)| > \varepsilon\big) \le 2e^{-2n\varepsilon^2}, \quad \forall \varepsilon > 0.$$

McDiarmid's inequality (also called bounded difference inequality, see [61]) is a concentration inequality for a multivariate function of random sequence $\{X_i\}_{i=1}^n$, says $f(X_1,\ldots,X_n)$. As a generalization of Hoeffding's inequality, it does not require any distribution assumptions about RVs and the $f(X_1,\ldots,X_n)$ may be dependent sum of RVs. The only requirement is the bounded difference condition by replacing X_j by X_j' meanwhile maintaining the others fixed in $f(X_1,\ldots,X_n)$.

Lemma 2.6 (McDiarmid's inequality)**.** *Suppose $X_1,\cdots,X_n$ are independent RVs all taking values in the set A, and assume $f: A^n \to \mathbb{R}$ satisfies the bounded difference condition*

$$\sup_{x_1,\cdots,x_n,x_k' \in A} \big|f(x_1,\cdots,x_n) - f(x_1,\cdots,x_{k-1},x_k',x_{k+1},\cdots,x_n)\big| \le c_k.$$

Then,

$$P\big(\big|f(X_1,\cdots,X_n) - \mathrm{E}\{f(X_1,\cdots,X_n)\}\big| \ge t\big) \le 2e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}, \quad \forall t > 0.$$

One method of proof is by the martingale argument, which needs to check the Azuma-Hoeffding's inequality below, see [86, Section 2.2.2]. Theorem 3.3.14 of [33] gives another proof based on the entropy method.

Lemma 2.7 (Azuma-Hoeffding's inequality)**.** *Let $\{X_n\}_{n=0}^\infty$ be a sequence of martingale (or supermartingale), adapted to an increasing filtration $\{\mathcal{F}_n\}_{n=0}^\infty$. Suppose $\{X_n\}_{n=0}^\infty$ satisfies the bounded difference condition $a_k \le X_k - X_{k-1} \le b_k$, a.s. for $k = 1,\ldots,n$. Then,*

$$P\big(|X_n - X_0| > t\big) \le 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_k - a_k)^2}}, \quad t \ge 0.$$

Two typical examples with bounded differences function are the concentration for U-statistics (a dependent summation) and the integral error of the kernel density estimation.

Example 2.2 (U-statistics)**.** Let $\{X_i\}_{i=1}^n$ be independent and identically distributed (IID) RVs and $g:\mathbb{R}^2\to\mathbb{R}$ be the bounded and symmetric function. Define a *U*-statistic of order 2 as

$$U_n=\binom{n}{2}^{-1}\sum_{i<j} g(X_i,X_j):=f(x_1,\dots,x_n).$$

Its bounded difference condition is

$$\begin{aligned}&\left|f(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots x_n)-f\left(x_1,\dots,x_{k-1},x_k',x_{k+1},\dots x_n\right)\right|\\ =&\frac{1}{\binom{n}{2}}\left|\sum_{j=1,j\neq k}^n\left[g(x_k,x_j)-g(x_k',x_j)\right]\right|\le\frac{2\cdot 2(n-1)\|g\|_\infty}{n(n-1)}=\frac{4\|g\|_\infty}{n}.\end{aligned}$$

So we have

$$P\big(|U_n-\mathrm{E}U_n|>t\big)\le 2e^{-\frac{nt^2}{8\|g\|_\infty^2}}.$$

Example 2.3 (L_1-error in kernel density estimation)**.** Let $\{X_i\}_{i=1}^n\overset{\text{IID}}{\sim}F(x)$ with density function $f(x)$. Define the kernel density estimator by

$$\hat{f}_{n,h}(x)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left(\frac{x-X_i}{h}\right),$$

where $K(\cdot)>1$ is the kernel function and $h>0$ is a smoothing parameter called the bandwidth. Usually, the kernel function $K(\cdot)$ is symmetric probability density and $h>0$ with $h\to 0$ and $nh\to\infty$. Define the L_1-error of $\hat{f}_{n,h}(x)$ by

$$Z_n=g(X_1,\dots,X_n)=\int\left|\hat{f}_{n,h}(x)-f(x)\right|dx.$$

By $\int K(u)du=1$, the McDiarmid's inequality with bound difference condition

$$\begin{aligned}&\left|g(x_1,\dots,x_n)-g(x_1,\dots,x_i',\dots,x_n)\right|\\ \le&\frac{1}{n}\int\left|K\left(\frac{x-x_i}{h}\right)-K\left(\frac{x-x_i'}{h}\right)\right|d\left(\frac{x}{h}\right)\le\frac{2}{n}\end{aligned}$$

gives

$$P\big(|Z_n-\mathrm{E}Z_n|\ge t\big)\le 2e^{-2t^2/n(\frac{2}{n})^2}=2e^{-\frac{nt^2}{2}},$$

which is free of the bandwidth.

3 Sub-Gaussian distributions

3.1 Motivations

In probability, there is a well-known inequality for bounding the Gaussian tail. If $X \sim N(0,1)$, [35] obtained for $x>0$

$$\left(\frac{1}{x}-\frac{1}{x^3}\right)\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}<\left(\frac{x}{x^2+1}\right)\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\le P(X\ge x)\le\frac{1}{x}\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}, \tag{3.1}$$

which is called Mills's inequality, relating to Mills's ratio [63]. The upper bound in (3.1) is mostly used to derive law of the iterated logarithm [24]. However, if x tends to zero the upper bound goes to $+\infty$ which makes it meaningless. So the Mill's inequality is useful only for larger x. We need a better inequality. In fact, the upper bound in (3.1) can be strengthened as in [34, Lemma B.3]: $P(|X|\ge x)\le e^{-\frac{x^2}{2}}$. We refer it as the sharper Mill's inequality.

In statistics, people want to study a general class of error distributions (beyond Gaussian) whose moment generating function (MGF): $\mathrm{E}e^{sX}$ have similar Gaussian properties with s in specific subset of $\mathbb{R}$. To derive sharper Mill's inequality, it is natural to define the class of sub-Gaussian RV as follows.

Definition 3.1 (Sub-Gaussian distribution)**.** *An RV $X\in\mathbb{R}$ with mean zero is sub-Gaussian with a variance proxy σ^2 (denoted $X\sim \mathrm{subG}(\sigma^2)$) if its MGF satisfies*

$$\mathrm{E}e^{sX}\le e^{\frac{\sigma^2 s^2}{2}}, \quad \forall s\in\mathbb{R}.$$

With Definition 3.1 and Chernoff's inequality, we will get the exponential decay of the tail as the alternative definition of sub-Gaussian:

$$P(X\ge t)\le\inf_{s>0}e^{-st}\mathrm{E}e^{sX}\le\inf_{s>0}e^{-st+\frac{\sigma^2 s^2}{2}}\xlongequal{s=t/\sigma^2}e^{-\frac{t^2}{2\sigma^2}}.$$

This argument is called Cramer-Chernoff method, and it is applied in proving Hoeffding's lemma for sum of independent variables. In general, let $Z_1,\dots,Z_n$ be n independent centralized RVs, and suppose there exists a convex function $g(t)$ and a domain D_0 containing $\{0\}$ such that

$$\mathrm{E}e^{t\sum_{i=1}^n Z_i}\le e^{ng(t)}, \quad \forall t\in D_0\subset\mathbb{R}.$$

Denote $g^*(s)=\sup_{t\in D_0}\{ts-g(t)\}$ as the convex conjugate function of g, therefore the Chernoff's inequality implies

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|>s\right)\le 2e^{-ng^*(s)}, \quad \forall s>0,$$

which has rich applications in high-dimensional statistics, machine learning, random matrix theory, and other fields on non-asymptotic results.

Note that $\mathrm{subG}(\sigma^2)$ denotes a class of distributions rather than a single distribution. Trivially, the Gaussian distribution is a special case of sub-Gaussian.

Example 3.1 (Normal distributions)**.** Consider the normal RV $X \sim N(\mu,\sigma^2)$. With the MGF of X:

$$\mathrm{E}e^{sX} := e^{\frac{\sigma^2 s^2}{2}}, \quad \forall s \in \mathbb{R}$$

it is sub-Gaussian with the variance proxy $\sigma^2 = \mathrm{Var}(X)$.

Example 3.2 (Bounded RVs)**.** By Hoeffding's lemma,

$$\mathrm{E}e^{sX} \le e^{\frac{1}{8}s^2(b-a)^2} \quad \text{for} \quad s>0$$

for the centralized bounded variable $X \in [a,b]$. So X is essentially sub-Gaussian with variance proxy $\sigma^2 = \frac{1}{4}(b-a)^2$. For Bernoulli variable $X \in \{0,1\}$, we have $X \sim \mathrm{subG}\left(\frac{1}{4}\right)$.

There are at least seven equivalent forms for sub-Gaussian as shown in the following.

Corollary 3.1 (Characterizations of sub-Gaussian)**.** *Let X be an RV in $\mathbb{R}$ with $\mathrm{E}X=0$. Then, the following are equivalent for finite positive constants $\{K_i\}_{i=1}^7$.*

(1) The MGF of X: $\mathrm{E}e^{sX} \le e^{K_1^2 s^2}$ for all $s \in \mathbb{R}$.

(2) The tail of X: $P\{|X| \ge t\} \le 2e^{-t^2/K_2^2}$ for all $t \ge 0$.

(3) The moments of X: $(\mathrm{E}|X|^k)^{1/k} \le K_3\sqrt{k}$ for all integer $k \ge 1$.

(4) The exponential moment of X^2 : $\mathrm{E}e^{X^2/K_4^2} \le 2$.

(5) The local MGF of X^2: $\mathrm{E}e^{l^2X^2} \le e^{K_5^2 l^2}$ for all l in a local set $|l| \le \frac{1}{K_5}$.

(6) There is a constant $\sigma \ge 0$ such that $\mathrm{E}e^{\lambda X^2/K_6^2} \le (1-\lambda)^{-1/2}$ for all $\lambda \in [0,1)$.

(7) Union bound condition: $\exists c>0$ s.t. $\mathrm{E}[\max\{|X_1|,\ldots,|X_n|\}] \le c\sqrt{\log n}$ for all $n \ge c$, where $\{X_i\}_{i=1}^n$ are IID copies of X.

Remark 3.1. The $\mathrm{E}X=0$ is for convenience as the zero mean is used in the proof of Corollary 3.1(1), see [83] for the details and the proof of the equivalences (1)-(5). The equivalences (6) is given in [86, Theorem 2.6] and the equivalences (7) is present in [75, p. 24]. The moment condition for integers k in (3) can be relaxed to even integers k by the symmetrization technique. By symmetry of X, let us consider a negative independent copy $-X'$ which is independent of X and has the same distribution as X. If (3) is true and $\mathrm{E}(-X')=0$, from Jensen's inequality $\mathrm{E}e^{\theta(-X')}\geq e^{\theta\mathrm{E}(-X')}=1$ since $-X'$ has zero mean. So we have by the independence of X' and X:

$$\begin{aligned}
\mathrm{E}e^{\theta X} &\leq \mathrm{E}e^{\theta X}\mathrm{E}e^{\theta(-X')}=\mathrm{E}e^{\theta(X-X')}=1+\sum_{k=1}^{n}\frac{\theta^{2k}\mathrm{E}(X-X')^{2k}}{(2k)!}\\
&\leq 1+\sum_{k=1}^{n}\frac{\theta^{2k}\mathrm{E}(|X|+|X'|)^{2k}}{(2k)!}\\
&\leq 1+\sum_{k=1}^{\infty}\frac{\theta^{2k}2^{2k}\mathrm{E}|X|^{2k}}{(2k)!}<1+\sum_{k=1}^{\infty}\frac{\left(2\theta K_3^2\sqrt{2k}\right)^{2k}}{k^k k!} \qquad \text{[By (3.7)]}\\
&=1+\sum_{k=1}^{\infty}\frac{\left(8\theta^2K_3^2\right)^{k}}{k!}=e^{8\theta^2K_3^2}, \quad \forall\theta\in\mathbb{R},
\end{aligned}$$

where the last inequality is due to $(2k)!>k^k\cdot k!$.

3.2 The variance proxy and sub-Gaussian norm

We show that the σ^2 in Definition 3.1 is indeed the upper bounds of variance of X. The σ^2 not only characterizes the speed of decay in the sub-Gaussian tail probability, but also bounds the variance of $n^{-\frac{1}{2}}\sum_{i=1}^{n}X_i$. The $\mathrm{Var}X\leq\sigma^2$ is because, by the sub-Gaussian MGF

$$\frac{\sigma^2s^2}{2}+o(s^2)=e^{\frac{\sigma^2s^2}{2}}-1\geq\mathrm{E}e^{sX}-1=s\mathrm{E}X+\frac{s^2}{2}\mathrm{E}X^2+\cdots=\frac{s^2}{2}\cdot\mathrm{Var}X+o(s^2). \tag{3.2}$$

Definition 3.2 (Sub-Gaussian norm)**.** *For a sub-Gaussian RV X, the sub-Gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined by*

$$\|X\|_{\psi_2}=\inf\left\{t>0:\mathrm{E}e^{\frac{X^2}{t^2}}\leq 2\right\}.$$

From Corollary 3.1(4), $\|X\|_{\psi_2}$ is the smallest K_4. An alternative definition of the sub-Gaussian norm is ([83])

$$\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-\frac{1}{2}} \left(\mathrm{E}|X|^p\right)^{\frac{1}{p}}.$$

The definition for sub-Gaussian norm makes Corollary 3.1 easily presented. In fact, if $\mathrm{E}e^{X^2/\|X\|_{\psi_2}^2} \leq 2$,

$$P\big(|X|\geq t\big) = P\left(e^{\frac{X^2}{\|X\|_{\psi_2}^2}} \geq e^{\frac{t^2}{\|X\|_{\psi_2}^2}}\right) \leq \mathrm{E}e^{\frac{X^2}{\|X\|_{\psi_2}^2}} / e^{\frac{t^2}{\|X\|_{\psi_2}^2}} \leq 2e^{-\frac{t^2}{\|X\|_{\psi_2}^2}}. \tag{3.3}$$

Example 3.3 (The sub-Gaussian norm of bounded RVs.)**.** Consider an RV $|X| \leq M < \infty$. Set

$$\mathrm{E}e^{\frac{X^2}{t^2}} \leq e^{\frac{M^2}{t^2}} \leq 2, \quad t \geq \frac{M}{\sqrt{\log 2}},$$

we have

$$\|X\|_{\psi_2} = \frac{M}{\sqrt{\log 2}}.$$

Example 3.4 (The sub-Gaussian norm of Gaussian RVs.)**.** For a $N(0,\sigma^2)$ and $t > \sqrt{2}\sigma$,

$$\mathrm{E}e^{\frac{X^2}{t^2}} = \int e^{\frac{x^2}{t^2}} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx = \frac{t}{(t^2-2\sigma^2)^{\frac{1}{2}}} \leq 2 \quad \Rightarrow \quad t \geq \sqrt{\frac{8}{3}}\sigma.$$

By the definition, $\|X\|_{\psi_2} = \sqrt{\frac{8}{3}}\sigma > \sqrt{2}\sigma$.

However, the neat notation for defining sub-Gaussian norm sometime leads to unknown constants in the CIs as shown next.

Corollary 3.2 ([84, Theorem 2.6.3])**.** *Let $\{X_i\}_{i=1}^n$ be independent mean-zero sub-Gaussian, $\forall t \geq 0$,*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right\} \leq 2e^{-\frac{C(nt)^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}}, \quad \forall t \geq 0$$

for a constant C.

The unknown constant C makes the above CIs cannot be used in constructing confidence bands for μ. To obtain more specific bounds (data dependent bounds as a statistics), we adopt the follow three propositions under sub-Gaussian.

Proposition 3.1 (Sub-Gaussian properties). *Let* $X \sim \mathrm{subG}(\sigma^2)$*, then for any* $t>0$*,*

(a) the tail satisfies

$$P\big(|X|>t\big) \leq 2e^{-\frac{t^2}{2\sigma^2}};$$

(b) (a) implies that moments

$$\mathrm{E}|X|^k \leq \big(2\sigma^2\big)^{\frac{k}{2}} k\Gamma\left(\frac{k}{2}\right), \quad \Big(\mathrm{E}\big(|X|^k\big)\Big)^{\frac{1}{k}} \leq \sigma e^{\frac{1}{e}}\sqrt{k}, \quad k\geq 2;$$

(c) if (a) holds and $\mathrm{E}X=0$*, then*

$$\mathrm{E}e^{sX} \leq e^{4\sigma^2 s^2} \quad \textit{for any} \quad s>0;$$

(d) if $X \sim \mathrm{subG}(\sigma^2)$*, then*

$$\|X\|_{\psi_2} \leq \frac{2\sqrt{2}}{\sqrt{\log 2}}\sigma,$$

conversely, if $\|X\|_{\psi_2}=\sigma$ *then* $X \sim \mathrm{subG}(4\sigma^2)$*.*

Proof. The proofs of (a)-(c) are in [70, Lemmas 1.4 and 1.5]. The proofs of (a, b) is similar to Proposition 3.2(a, b) below. For (d), note that

$$\begin{aligned}\mathrm{E}\exp\big(s^2X^2\big) &= 1+\sum_{k=1}^{\infty}\frac{s^{2k}\mathrm{E}X^{2k}}{k!} \overset{\text{(b)}}{\leq} 1+\sum_{k=1}^{\infty}\frac{2s^{2k}(2\sigma^2)^k k\Gamma(k)}{k!} \\ &= 1+4s^2\sigma^2\sum_{k=0}^{\infty}(2s^2\sigma^2)^k \xlongequal{\forall\ |2s^2\sigma^2|<1} 1+\frac{4s^2\sigma^2}{1-2s^2\sigma^2} \\ &\overset{\forall |s|\leq\frac{1}{2\sigma}}{\leq} 1+8s^2\sigma^2 \leq e^{8\sigma^2 s^2}. \end{aligned} \tag{3.4}$$

By (3.4), set

$$\mathrm{E}e^{s_0^2X^2} \leq e^{8{s_0}^2\sigma^2} \leq 2 \quad \text{for some} \quad s_0.$$

Then

$$|s_0| \leq \frac{\sqrt{\log 2}}{2\sqrt{2}\sigma} \leq \frac{1}{2\sigma}.$$

Put $|s_0| = \frac{\sqrt{\log 2}}{2\sqrt{2}\sigma}$ and the sub-Gaussian norm gives

$$\mathrm{E}e^{X^2/\left(\frac{2\sqrt{2}\sigma}{\sqrt{\log 2}}\right)^2} \leq 2 \quad \Rightarrow \quad \|X\|_{\varphi_2} \leq \frac{2\sqrt{2}\sigma}{\sqrt{\log 2}}.$$

Conversely, if $\|X\|_{\psi_2} = \sigma$ then the (3.3) gives

$$P(|X|>t) \le 2e^{-\frac{t^2}{\sigma^2}} = 2e^{-\frac{t^2}{2(\sigma/\sqrt{2})^2}}.$$

Then Proposition 3.1(c) concludes

$$\mathrm{E}e^{sX} \le e^{4(\frac{\sigma}{\sqrt{2}})^2 s^2} = e^{4\sigma^2 \frac{s^2}{2}}$$

for any $s>0$, and we have $X \sim \mathrm{subG}(4\sigma^2)$. □

Let $\{Y_i\}_{i=1}^n$ be a sequence of exponential family (EF) RVs, its density

$$f(y_i;\theta_i) = c(y_i)\exp\{y_i\theta_i - b(\theta_i)\} \tag{3.5}$$

with $\mathrm{E}Y_i = \dot{b}(\theta_i)$ and $\mathrm{Var}Y_i = \ddot{b}(\theta_i)$. We next introduce the sub-Gaussian CIs for the non-random weighted sum of EF RVs with compact parameter space, adapted from [69, Lemma 6.1] with more specific constants.

Proposition 3.2 (Concentration for weighted E-F summation)**.** *We assume* (3.5) *and*

- (E.1)*: Uniformly bounded variances condition: there exist a compact set Ω and some constant C_b such that $\sup_{\theta_i\in\Omega}\ddot{b}(\theta_i) \le C_b^2$ for all i.*

Let $\boldsymbol{w} := (w_1,\cdots,w_n)^T \in \mathbb{R}^n$ be a non-random vector and define $S_n^w =: \sum_{i=1}^n w_iY_i$. Then

(a) Closed under addition:

$$S_n^w - \mathrm{E}S_n^w \sim \mathrm{subG}\big(C_b^2\|\boldsymbol{w}\|_2^2\big), \quad P\big\{|S_n^w - \mathrm{E}S_n^w| > t\big\} \le 2e^{-\frac{t^2}{2C_b^2\|\boldsymbol{w}\|_2^2}}.$$

(b) Let $C_n := S_n^w - \mathrm{E}S_n^w$ and

$$\Gamma(t) := \int_0^\infty x^{t-1}e^{-x}dx$$

be the Gamma function. For all integer $k \ge 1$, we have moments bound:

$$\mathrm{E}|C_n|^k \le k\big(2C_b^2\big)^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)\|\boldsymbol{w}\|_2^k.$$

(c) The MGF of centralized $|C_n|^2$:

$$\mathrm{E}e^{s[|C_n|^2 - \mathrm{E}|C_n|^2]} \le e^{(s^2/2)(8\sqrt{2}C_b^2\|w\|_2^2)^2}, \quad \forall |s| \le \big(8C_b^2\|\boldsymbol{w}\|_2^2\big)^{-1}.$$

(d) *In this case, we do not assume* (3.5) *and* (E.1). *Suppose* $\{Y_i - \mathrm{E}Y_i\}_{i=1}^n$ *are independent distributed as* $\{\mathrm{subG}(\sigma_i^2)\}_{i=1}^n$ *with* $C_b^2 =: \max_{1\le i\le n}\sigma_i^2 > 0$, *then* (a)$-$(c) *also hold.*

Proof. Based on the MGF and uniformly bounded variances condition, the proof of (a) can be found in [69, Lemma 6.1]. In the proof of (c), we update the constant, and (d) is similar for the sub-Gaussian case.

(b) The proof relies on expectation formula for positive RV (in terms of integral of tail probability) which transforms tail bound to moment bound. For any integer $k\ge 1$,

$$\begin{aligned}\mathrm{E}|S_n^w - \mathrm{E}S_n^w|^k &= \int_0^\infty P\big(|S_n^w - \mathrm{E}S_n^w|^k \ge s\big)ds \\ &\overset{t=s^{1/k}}{=\!=\!=} \int_0^\infty kt^{k-1} P\big(|S_n^w - \mathrm{E}S_n^w| \ge t\big)dt. \end{aligned} \tag{3.6}$$

Applying tail bound in (a), we have by letting $D_{k,C} = k(2C_b^2)^{\frac{k}{2}}\Gamma(\frac{k}{2})$

$$\begin{aligned}\mathrm{E}|S_n^w - \mathrm{E}S_n^w|^k &\le 2k\int_0^\infty t^{k-1} e^{-\frac{t^2}{2C_b^2\|\boldsymbol{w}\|_2^2}}dt \\ &\overset{x=t^2/(2C_b^2\|\boldsymbol{w}\|_2^2)}{=\!=\!=\!=\!=} k\big(2C_b^2\big)^{\frac{k}{2}}\|\boldsymbol{w}\|_2^k \int_0^\infty x^{\frac{k}{2}-1}e^{-x}dx \\ &= D_{k,C}\|\boldsymbol{w}\|_2^k. \end{aligned}$$

(c) The proof will resort to $(\mathrm{E}|Z|)^k \le \mathrm{E}|Z|^k$ and Jensen's inequality

$$\left(\frac{|a|+|b|}{2}\right)^k \le \frac{1}{2}|a|^k + \frac{1}{2}|b|^k \quad \text{for integer} \quad k\ge 1. \tag{3.7}$$

From Taylor's expansion, (3.7) gives

$$\begin{aligned}\mathrm{E}e^{s[|C_n|^2 - \mathrm{E}|C_n|^2]} &= 1 + \sum_{k=2}^\infty \frac{s^k \mathrm{E}\big[|C_n|^2 - \mathrm{E}|C_n|^2\big]^k}{k!} \\ &\le 1 + \sum_{k=2}^\infty \frac{s^k 2^{k-1}\mathrm{E}\big\{|C_n|^{2k} + \big(\mathrm{E}|C_n|^2\big)^k\big\}}{k!} \\ &\le 1 + \sum_{k=2}^\infty \frac{s^k 2^{k-1}\mathrm{E}\big\{|C_n|^{2k} + \mathrm{E}|C_n|^{2k}\big\}}{k!} \qquad \big[\text{By } (\mathrm{E}|Z|)^k \le \mathrm{E}(|Z|^k)\big] \\ &\le 1 + \sum_{k=2}^\infty \frac{s^k 2^k \cdot 2k\big(2C_b^2\|\boldsymbol{w}\|_2^2\big)^k\Gamma(k)}{k!}, \end{aligned}$$

where the last inequality is by Proposition 3.2(b). Then, under $|4sC_b^2\|\boldsymbol{w}\|_2^2|<1$, we have

$$\begin{aligned}\mathrm{E}e^{s[|C_n|^2-\mathrm{E}|C_n|^2]}&=1+2\sum_{k=2}^{\infty}\left(4sC_b^2\|\boldsymbol{w}\|_2^2\right)^k=1+\frac{2\left(4sC_b^2\|\boldsymbol{w}\|_2^2\right)^2}{1-4|s|C_b^2\|\boldsymbol{w}\|_2^2}\\&\le 1+\frac{s^2(8\sqrt{2}C_b^2\|\boldsymbol{w}\|_2^2)^2}{2}\qquad\left[\left|4sC_b^2\|\boldsymbol{w}\|_2^2\right|\le\frac{1}{2}\Leftrightarrow|s|\le\frac{1}{8C_b^2\|\boldsymbol{w}\|_2^2}\right]\\&\le \mathrm{e}^{(s^2/2)(8\sqrt{2}C_b^2\|\boldsymbol{w}\|_2^2)^2}.\end{aligned}$$

(d) It follows by defining $C_b^2=:\max_{1\le i\le n}\sigma_i^2>0$ as the common variance proxy for $\{Y_i\}_{i=1}^n$. For $i=1,\ldots,n$, we have: $\mathrm{E}e^{sw_i(Y_i-\mathrm{E}Y_i)}\le e^{\sigma^2s^2w_i^2/2}$, $\forall s\in\mathbb{R}$. □

Proposition 3.2(a) yields the following results (the first result is in [70, Corollary 1.7]). The second sub-Gaussian CI below specifies the unknown constant in [84, Theorem 2.6.2].

Proposition 3.3. *Let $\{X_i\}_{i=1}^n$ be n independent* $\mathrm{subG}(\sigma_i^2)$. *Define $\sigma^2=\max_{1\le i\le n}\sigma_i^2$,*

$$P\left(\left|\sum_{i=1}^n w_iX_i\right|>t\right)\le 2e^{-\frac{t^2}{(2\sigma^2\|\boldsymbol{w}\|_2^2)}},$$

$$P\left(\left|\sum_{i=1}^n w_iX_i\right|>t\right)\le 2e^{-\frac{t^2}{(8\sum_{i=1}^n\|w_iX_i\|_{\psi_2}^2)}},\quad\forall t\ge 0$$

for any non-random vector $\boldsymbol{w}:=(w_1,\cdots,w_n)^T$.

Proof. To see the second CI, just use the Proposition 3.1(d) and the Proposition 3.2(d), by noticing that if $\|X_i\|_{\psi_2}<\infty$ then $w_iX_i\sim\mathrm{subG}(4\|w_iX_i\|_{\psi_2}^2)$. □

3.3 Randomly weighted sum of independent sub-Gaussian variables

In this part, we outline the sub-Gaussian type CIs for the randomly weighted sum of exponential family of RVs: $S_n^W=:\sum_{i=1}^n W_iY_i$, where $\{W_i\}_{i=1}^n$ are called the multipliers (or random weights) which are independent from $\{Y_i\}_{i=1}^n$. The normalized sum $\frac{1}{\sqrt{n}}(S_n^W-\mathrm{E}S_n^W)$ is also called multiplier empirical processes, and it serves for the multiplier Bootstrap inference where the multipliers $\{W_i\}$ are RVs independent from $\{Y_i\}_{i=1}^n$, see [82, Chapter 2.9]. To get sub-Gaussian concentration, some regularity conditions for the parameter space are required.

- (E.2): Let $\boldsymbol{W}:=(W_1,\cdots,W_n)^T\in\mathbb{R}^n$ be a random vector with some bounded components, i.e. $|W_i|\le w_i<\infty$ for a non-random vector $\boldsymbol{w}:=(w_1,\cdots,w_n)^T\in\mathbb{R}^n$.

Theorem 3.1 (Concentration inequalities for randomly weighted sum)**.** *Let* $\{Y_i\}_{i=1}^n$ *belong to the canonical exponential family* (3.5)*, and let* $\{W_i\}_{i=1}^n$ *be independent of* $\{Y_i\}_{i=1}^n$*. Define the randomly weighted sum* $S_n^W=:\sum_{i=1}^n W_iY_i$*, then under* (E.1) and (E.2)

$$P\left(\left|S_n^W-\mathrm{E}S_n^W\right|\ge t\right)\le 2e^{-\frac{t^2}{(2C_b^2\|\boldsymbol{w}\|_2^2)}}.$$

Proof. Let $Y_i=\dot{b}(\theta_i)+Z_i$, where $\{Z_i\}_{i=1}^n$ are centralized and independent E-F RVs. From $\mathrm{E}Y_i=\dot{b}(\theta_i)$ and the identity (3.8) for a dominating measure $\mu(\cdot)$

$$\int dF_{Y_i}(y)=1 \;\Leftrightarrow\; \int c(y)e^{y\theta_i}\mu(dy)=e^{b(\theta_i)}. \tag{3.8}$$

Let $\mathrm{E}_{\cdot|\boldsymbol{W}}(\cdot):=\mathrm{E}(\cdot|\boldsymbol{W})$ and s be in $(-\delta,\delta)$ (a neighbourhood of zero). Then

$$\begin{aligned}\mathrm{E}_{\cdot|\boldsymbol{W}}[\mathrm{e}^{sW_iY_i}]&=\int \mathrm{e}^{sW_iY_i}dF_{Y_i|\boldsymbol{W}}(y)=\int \mathrm{e}^{sW_iY_i}dF_{Y_i}(y)\quad\left[\text{by }\{W_i\}_{i=1}^n\perp\{Y_i\}_{i=1}^n\right]\\&=\int c(y)e^{y\theta_i-b(\theta_i)}e^{sW_iy}\mu(dy)\overset{(3.8)}{=\!=\!=}e^{b(\theta_i+sW_i)-b(\theta_i)}.\end{aligned}$$

It can be easily derived from (E.2) and Taylor's expansion,

$$\begin{aligned}\mathrm{E}_{\cdot|\boldsymbol{W}}[\mathrm{e}^{s(W_iY_i-\mathrm{E}_{\cdot|\boldsymbol{W}}(W_iY_i))}]&=e^{b(\theta_i+sW_i)-b(\theta_i)-\dot{b}(\theta_i)W_is}\\&\overset{\exists\tilde{\eta}_i\in[\theta_i,\theta_i+sW_i]}{=\!=\!=\!=\!=\!=\!=\!=}e^{\frac{s^2W_i^2}{2}\ddot{b}(\tilde{\eta}_i)}\le e^{\frac{s^2C_b^2W_i^2}{2}}.\end{aligned}\tag{3.9}$$

By the conditional independence for $\{W_iZ_i|\boldsymbol{W}\}$ and (3.9), it follows that when $s\in(-\delta,\delta)$

$$\begin{aligned}&\mathrm{E}_{\cdot|\boldsymbol{W}}[e^{s\sum_{i=1}^n[W_iZ_i-\mathrm{E}_{\cdot|\boldsymbol{W}}(W_iZ_i)]}]\\&=\prod_{i=1}^n\mathrm{E}_{\cdot|\boldsymbol{W}}e^{s[W_iZ_i-\mathrm{E}_{\cdot|\boldsymbol{W}}(W_iZ_i)]}\le\prod_{i=1}^n e^{\frac{s^2C_b^2W_i^2}{2}}\le 2e^{\frac{s^2C_b^2\|\boldsymbol{w}\|_2^2}{2}},\end{aligned}\tag{3.10}$$

where the last inequality is from $\{|W_i|\le w_i\}$ for a non-random vector $\boldsymbol{w}:=(w_1,\cdots,w_n)^T$.

By the conditional Markov's inequality and symmetry of Z_i, we have, as $s\in(-\delta,\delta)$

$$P\left(\left|\sum_{i=1}^n\left[W_iZ_i-\mathrm{E}_{\cdot|\boldsymbol{W}}(W_iZ_i)\right]\right|\ge t|\boldsymbol{W}\right)$$

$$
\begin{aligned}
&\leq \inf_{s>0}\left[e^{-st}\mathrm{E}_{\cdot|W}e^{s(\tilde{S}_n^W-\mathrm{E}_{\cdot|W}\tilde{S}_n^W)}+e^{-st}\mathrm{E}_{\cdot|W}e^{-s(\tilde{S}_n^W-\mathrm{E}_{\cdot|W}\tilde{S}_n^W)}\right]\\
&\leq 2\inf_{s>0}e^{\frac{s^2C_b^2\|\boldsymbol{w}\|_2^2}{2}-st}=2e^{-\frac{t^2}{2C_b^2\|\boldsymbol{w}\|_2^2}},
\end{aligned}
\tag{3.11}
$$

where the last equality is minimized by setting $s=\frac{t}{(C_b^2\|\boldsymbol{w}\|_2^2)}$. □

Note that [69, Lemma 6.1] is about the concentration for the non-random weighted sum of exponential family RVs. The assumption of compact parameter space for exponential family is vital for obtaining the sub-Gaussian type concentration. If we do not impose condition (E.2) and the assumption that $\{W_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ are dependent, a counterexample for sub-Gaussian concentration is $W_i=Y_i$. Thus, S_n^W is a quadratic form, and $S_n^W-\mathrm{E}S_n^W$ is sub-exponential by Lemma 4.2 below. If $\{W_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ are dependent but $\{W_i\}_{i=1}^n$ are still bounded, another counterexample is $W_i=\mathrm{sign}(\mathrm{Y_i})$. Therefore, $S_n^W=\sum_{i=1}^n|Y_i|$ is not zero-mean, and the concentration of $\sum_{i=1}^n|Y_i|$ fails.

3.4 Concentration for Lipschitz functions of random vectors

In the analyses of high-dimensional statistics by empirical processes, researches often resort to the CIs of Lipschitz functions for either bounded or strongly log-concave random vectors [86].

Lemma 3.1 ([86, Theorem 2.26]). *Let $\boldsymbol{N}\sim N(\boldsymbol{0},\mathrm{I}_p)$. Let $f:\mathbb{R}^n\to\mathbb{R}$ be L-Lipschitz with respect to (with respect to) the Euclidean norm, i.e.,*

$$|f(\boldsymbol{a})-f(\boldsymbol{b})|\leq L\|\boldsymbol{a}-\boldsymbol{b}\|_2 \quad \textit{for any} \quad \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n.$$

Then,

$$P\big(|f(\boldsymbol{N})-\mathrm{E}f(\boldsymbol{N})|\geq t\big)\leq 2e^{-\frac{t^2}{(2L^2)}},\quad \forall t>0.$$

A non-negative function $f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$ is log-concave if for any $\lambda\in[0,1]$ and any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$,

$$\log f\big(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\big)\geq\lambda\log f(\boldsymbol{x})+(1-\lambda)\log f(\boldsymbol{y}). \tag{3.12}$$

A function $\psi(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$ is γ-strongly concave if there is $\gamma>0$ s.t. for any $\lambda\in[0,1]$ and any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$

$$\lambda\psi(\boldsymbol{x})+(1-\lambda)\psi(\boldsymbol{y})-\psi\big(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}\big)\leq\frac{\gamma}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

A continuous probability density $f(\boldsymbol{x})$ and the corresponding RV are log-concave (or strongly log-concave) if $f(\boldsymbol{x})$ is a log-concave function (or strongly log-concave function), see [73] for a review of the log-concavity in statistics.

Lemma 3.2 ([86, Theorem 3.16]). *Let $\mathbb{P}$ be any γ-strongly log-concave distribution on $\mathbb{R}^n$ with parameter $\gamma>0$. Then for any function $f:\mathbb{R}^n\to\mathbb{R}$ that is L-Lipschitz with respect to the Euclidean norm, we have*

$$P\big[f(X)-\mathrm{E}f(X)\geq t\big]\leq e^{-\frac{\gamma t^2}{4L^2}}\quad \textit{for}\quad X\sim\mathbb{P},\quad t\geq 0.$$

The standard Gaussian random vector is 1-strongly log-concave distributed. However, Lemma 3.1 has the sharper constant $2L^2$ than the Gaussian case of Lemma 3.2 with constant $4L^2$. Beyond Gaussian and strongly log-concave, it is possible to establish concentration for distributions involving bounded RVs. A function $f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$ is said to be separately convex if, the univariate function $y_k\mapsto f(x_1,x_2,\dots,x_{k-1},y_k,x_{k+1},\dots,x_n)$ for each index $k\in\{1,\dots,n\}$, is convex for each fixed vector $(x_1,x_2,\dots,x_{k-1},x_{k+1},\dots,x_n)\in\mathbb{R}^{n-1}$.

Lemma 3.3 ([86, Theorem 3.4]). *Let $\{X_i\}_{i=1}^n$ be independent RVs, each supported on the interval $[a,b]$. Let $f:\mathbb{R}^n\to\mathbb{R}$ be separately convex, and L-Lipschitz with respect to the Euclidean norm. Then*

$$P\big[f(X)-\mathrm{E}f(X)\geq t\big]\leq e^{-\frac{t^2}{4L^2(b-a)^2}}\quad \textit{for}\quad X\sim\mathbb{P},\quad t\geq 0.$$

Example 3.5 (Order Statistics). From Lemmas 3.2 and 3.3, suppose that $\{X_i\}_{i=1}^n$ are independent RVs which are γ-strongly log-concave distributed satisfying

$$P\big[f(X)-\mathrm{E}f(X)\geq t\big]\leq e^{-\frac{\gamma t^2}{4L^2}}$$

for any function $f:\mathbb{R}^n\to\mathbb{R}$ that is L-Lipschitz with respect to the Euclidean norm. Let $X_{(k)}$ be the k-th order statistic of $X_1,\dots,X_n$, it can be shown that

$$P\big(|X_{(k)}-\mathrm{E}X_{(k)}|>\delta\big)\leq 2e^{-\frac{\delta^2}{2}}$$

by checking $|X_{(k)}-Y_{(k)}|\leq\|X-Y\|_2$, i.e. $L=1$. Indeed, we have

$$X_{(k)}-Y_{(k)}\leq|X_l-Y_l|\leq\|X-Y\|_2\quad\text{for some}\quad l\in\{1,\dots,n\}.$$

More results of the tail bounds for the order statistics of IID RVs are reported in [15].

4 Sub-exponential distributions

4.1 Characterizations

The requirement in definition of sub-Gaussian RV $\mathrm{E}e^{sX} \le e^{\frac{\sigma^2 s^2}{2}}, \forall s \in \mathbb{R}$ is too strong. We consider the MGF of exponential distributions.

Example 4.1 (MGF of exponential distributions)**.** Consider the exponential RV $X \sim \mathrm{Exp}(\mu)$ with $\mathrm{E}X = \mu > 0$. The MGF of $X - \mu$ satisfies

$$\begin{aligned} \mathrm{E}e^{s(X-\mu)} &= \frac{e^{-s\mu}}{1-s\mu} = \left(e^{-\frac{s\mu}{2}}(1-s\mu)^{-\frac{1}{2}}\right)^2 \\ &\le e^{2(\frac{s\mu}{2})^2} < e^{\frac{s^2(2\mu)^2}{2}}, \quad \forall |s| \le (2\mu)^{-1}, \end{aligned} \tag{4.1}$$

where the second last inequality is by

$$\frac{e^{-t}}{\sqrt{1-2t}} \le e^{2t^2} \quad \text{for} \quad |t| \le \frac{1}{4}.$$

In (4.1), the MGF of the exponential RV is divergent on $s = \frac{1}{\mu}$ and it cannot be bounded by a Gaussian MGF of s in $\mathbb{R}$, and the exponential MGF is bounded by Gaussian MGF for $|s| \le \frac{1}{2\mu}$ via inequality (4.1). Motivated by Example 4.1, the first definition of sub-exponential distribution (4.2) below is exactly the locally sub-Gaussian property.

Definition 4.1 (Sub-exponential distributions)**.** *A RV* $X \in \mathbb{R}$ *with* $\mathrm{E}X = 0$ *is sub-exponential with parameter* λ *(denoted* $X \sim \mathrm{subE}(\lambda)$*) if its MGF satisfies*

$$\mathrm{E}e^{sX} \le e^{\frac{s^2\lambda^2}{2}} \quad \textit{for all} \quad |s| < \frac{1}{\lambda}. \tag{4.2}$$

In [86], *sub-exponential distributions are generally defined by two positive parameters* (λ, α) *(denoted* $X \sim \mathrm{subE}(\lambda,\alpha)$*)*

$$\mathrm{E}e^{sX} \le e^{\frac{s^2\lambda^2}{2}} \quad \textit{for all} \quad |s| < \frac{1}{\alpha}.$$

The λ^2 in (4.2) is treated as a variance proxy and α is seen as locally sub-Gaussian factor, see Remark 4.1 later. Specifically, $\mathrm{subE}(\lambda) = \mathrm{subE}(\lambda,\lambda)$. Sub-Gaussian RVs are sub-exponential by definition, but not vice verse. In Corollary 3.1, one equivalence of sub-Gaussian RVs is that the survival function is bounded by the Gaussian-like survival function up to a constant. Similarly, the sub-exponential RV has a characterization that the survival function is bounded by that of a exponential distribution. Similar to sub-Gaussian characterizations, there are at least six equivalent forms for sub-exponential distributions which are useful for checking the sub-exponential distribution.

Corollary 4.1 (Characterizations of sub-exponential)**.** *Let X be an RV in $\mathbb{R}$ with* $\mathrm{E}X=0$*. Then the following properties are equivalent, where $\{K_i\}_{i=1}^6$ are positive constants.*

(1) The tails of X satisfy $P\{|X|\geq t\}\leq 2e^{-t/K_1}$ for all $t\geq 0$.

(2) The MGF of X satisfies $\mathrm{E}e^{lX}\leq e^{K_2^2 l^2}$ for all $|l|\leq \frac{1}{K_2}$.

(3) The moments of X satisfy $(\mathrm{E}|X|^p)^{1/p}\leq K_3 p$ for integer $p\geq 1$.

(4) The MGF of $|X|$ satisfies $\mathrm{E}e^{l|X|}\leq e^{K_4 l}$ for all $0\leq l\leq \frac{1}{K_4}$.

(5) The MGF of $|X|$ is bounded at some point: $\mathrm{E}e^{|X|/K_5}\leq 2$.

(6) Bounded MGF of X in a compact set: $\mathrm{E}e^{tX}<\infty$, $\forall |t|<\frac{1}{K_6}$.

The zero mean is only used in the proof of (2) of Corollary 4.1. The equivalence among (1)-(5) is proved in [84] and that between (5) and (6) can be found in [65, Lemma 5]. The (6) is the called Cramer's condition which is an essential characterization, it signifies that: All RVs. are sub-exponential if their MGF exist in a neighborhood of zero. [67] names the property (6) as the exponentially integrable RV.

Example 4.2 (Moment of exponential distributions)**.** The

$$P(X-\mu\geq t)=e^{-\frac{(t+\mu)}{\mu}}\leq e^{-\frac{t}{\mu}}$$

and the symmetry of $X-\mu$ implies $K_1=\mu$ in Corollary 4.1. Continue to Example 4.1, the "$\leq$" in (4.1) implies

$$\mathrm{E}e^{s(X-\mu)}\leq e^{(\frac{s\mu}{\sqrt{2}})^2}\leq e^{(2s\mu)^2}, \quad \forall |s|<(2\mu)^{-1}.$$

So $K_2=\frac{\mu}{\sqrt{2}}$ and $K_6=2\mu$ in Corollary 4.1. Next, we evaluate the moment of X for any $p\geq 1$,

$$\mathrm{E}|X|^p=\int_0^\infty x^p\cdot\mu^{-1}e^{-\mu^{-1}x}dx \xlongequal{y=\mu^{-1}x} \mu^p\int_0^\infty y^p e^{-y}dy=\Gamma(p+1)\mu^p.$$

By $\Gamma(p+1)\leq p^p$ for $p\geq 1$, it gives:

$$\big(\mathrm{E}|X|^p\big)^{\frac{1}{p}}=\big(\Gamma(p+1)\big)^{\frac{1}{p}}\mu\leq p\mu.$$

Via (4.4) shows that $(\mathrm{E}|X-\mu|^p)^{\frac{1}{p}}\leq 2p\mu$ and thus $K_3=2\mu$ in Corollary 4.1. Assume $\mathrm{E}X=0$, then by Stirling's approximation $p!\geq(\frac{p}{e})^p$

$$\begin{aligned}
\mathrm{E}e^{\lambda|X|}&=1+\sum_{p=2}^{\infty}\frac{\lambda^p\mathrm{E}|X|^p}{p!}\leq 1+\sum_{p=2}^{\infty}\frac{(\lambda K_3 p)^p}{(p/e)^p}\\
&=1+\sum_{p=2}^{\infty}(eK_3\lambda)^p=1+\frac{(eK_3\lambda)^2}{1-eK_3\lambda}\qquad \forall|eK_3\lambda|<1\\
&\leq 1+2(eK_2\lambda)^2\leq e^{2(eK_3\lambda)^2}\qquad\qquad \left(\text{Restrict } eK_3\lambda\leq\frac{1}{2}\right)\\
&\leq e^{eK_3\lambda}\leq e^{2eK_3\lambda}\quad \forall\lambda\leq\frac{1}{2eK_3}.
\end{aligned}\tag{4.3}$$

Thus $K_4=2eK_3=4e\mu$. That $\mathrm{E}e^{\lambda|X|}\leq e^{eK_3\lambda}$ for $0<\lambda\leq\frac{1}{2eK_2}$ in (4.3) implies $\mathrm{E}e^{\frac{|X|}{2eK_3}}<e^{\frac{1}{2}}<2$. Hence $K_5=K_3$.

Example 4.3 (Geometric distributions)**.** The geometric distribution $X\sim\mathrm{Geo}(q)$ for RV X is defined by

$$P(X=k)=(1-q)q^{k-1},\quad q\in(0,1),\quad k=1,2,\ldots.$$

The mean and the variance of $\mathrm{Geo}(q)$ are $\frac{1-q}{q}$ and $\frac{1-q}{q^2}$, respectively. Apply [38, Lemma 4.3], we get $(\mathrm{E}|X|^k)^{\frac{1}{k}}<\frac{-2k}{\log(1-q)}$. It follows from the Minkowski's inequality and Jensen's inequality $(\mathrm{E}|Z|)^k\leq\mathrm{E}|Z|^k$ for integer $k\geq 1$ that

$$\big(\mathrm{E}|X-\mathrm{E}X|^k\big)^{\frac{1}{k}}\leq\big(\mathrm{E}|X|^k\big)^{\frac{1}{k}}+|\mathrm{E}X|\leq 2\big(\mathrm{E}|X|^k\big)^{\frac{1}{k}}\leq\frac{-4k}{\log(1-q)}\tag{4.4}$$

and Corollary 4.1(3) implies the centralized $\mathrm{Geo}(q)$ is sub-exponential with $K_3=\frac{-4}{\log(1-q)}$.

Example 4.4 (Discrete Laplace RVs)**.** An RV $X\sim\mathrm{DL}(q)$ obeys the discrete Laplace distribution if

$$f_q(k)=\mathbb{P}(X=k)=\frac{1-q}{1+q}q^{|k|},\quad k\in\mathbb{Z}=\{0,\pm1,\pm2,\ldots\}$$

with parameter $q\in(0,1)$. The discrete Laplace RV is the difference of two IID Geo(q). The geometric distribution is sub-exponential, thus Corollary 4.2(a) mentioned later implies that the discrete Laplace is also sub-exponential distributed. In differential privacy of network models, the noises are assumed following the discrete Laplace distribution, see [30] and references therein.

The next result shows that a sum of independent sub-exponential RVs has two tails with difference convergence rate, which is slightly different from Hoeffding's inequality. Deviating from the mean, it tells us that the tail of the sum of sub-exponential RVs behaves like a combination of a Gaussian tail and a exponential tail.

Corollary 4.2 (Concentration for weighted sub-exponential sums)**.** *Let* $\{X_i\}_{i=1}^n$ *be independent* $\{\mathrm{subE}(\lambda_i)\}_{i=1}^n$ *distributed with zero mean. Define* $\lambda=\max_{1\le i\le n}\lambda_i>0$ *and the non-random vector* $\boldsymbol{w}:=(w_1,\cdots,w_n)^T\in\mathbb{R}^n$ *with* $w=\max_{1\le i\le n}|w_i|>0$*, we have*

(a) Closed under addition: $\sum_{i=1}^n w_iX_i\sim\mathrm{subE}(\|\boldsymbol{w}\|_2\lambda)$.

(b) $$P\left(\left|\sum_{i=1}^n w_iX_i\right|\ge t\right)\le 2e^{-\frac{1}{2}\left(\frac{t^2}{\|\boldsymbol{w}\|_2^2\lambda^2}\wedge\frac{t}{w\lambda}\right)}=\begin{cases}2e^{-\frac{t^2}{2\|\boldsymbol{w}\|_2^2\lambda^2}}, & 0\le t\le\frac{\|\boldsymbol{w}\|_2^2\lambda}{w},\\ 2e^{-\frac{t}{2w\lambda}}, & t>\frac{\|\boldsymbol{w}\|_2^2\lambda}{w}.\end{cases}$$

(c) Let $\{X_i\}_{i=1}^n$ *be independent zero-mean* $\{\mathrm{subE}(\lambda_i,\alpha_i)\}_{i=1}^n$ *distributed. Define*

$$\alpha:=\max_{1\le i\le n}\alpha_i>0,\quad \|\boldsymbol{\lambda}\|_2:=\left(\sum_{i=1}^n\lambda_i^2\right)^{\frac{1}{2}},\quad \bar{\lambda}:=\left(\frac{1}{n}\sum_{i=1}^n\lambda_i^2\right)^{\frac{1}{2}}.$$

Then $\sum_{i=1}^n X_i\sim\mathrm{subE}(\|\boldsymbol{\lambda}\|_2,\alpha)$ *and*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right|\ge t\right)\le 2e^{-\frac{1}{2}\left(\frac{nt^2}{\bar{\lambda}^2}\wedge\frac{nt}{\alpha}\right)}=\begin{cases}2e^{-\frac{nt^2}{2\bar{\lambda}^2}}, & 0\le t\le\frac{\bar{\lambda}^2}{\alpha},\\ 2e^{-\frac{nt}{2\alpha}}, & t>\frac{\bar{\lambda}^2}{\alpha},\end{cases}\quad \forall t\ge 0. \tag{4.5}$$

Remark 4.1. The $(\frac{nt^2}{\bar{\lambda}^2}\wedge\frac{nt}{\alpha})$ in (4.5) reveals that the smaller α (locally sub-Gaussian factor) leads to sharper sub-exponential concentration. The sub-exponential concentration tends to the sub-Gaussian concentration with variance proxy $\bar{\lambda}^2$ when $\alpha\to 0$, which coincides the locally sub-Gaussian definition for sub-exponential distribution in Definition 4.1.

Proof. (a) By definition of sub-exponential RVs,

$$\mathrm{E}e^{sw_iX_i} \le e^{\frac{s^2w_i^2\lambda_i^2}{2}}, \quad \forall |s| \le \frac{1}{|w_i|\lambda_i}, \quad i=1,\dots,n,$$

and it implies

$$\mathrm{E}e^{sw_iX_i} \le e^{\frac{s^2w_i^2\lambda_i^2}{2}}, \quad |s| \le \frac{1}{w\lambda} \quad \text{for all} \quad i.$$

By the independence among $\{X_i\}_{i=1}^n$,

$$\mathrm{E}\exp\left\{s\sum_{i=1}^n w_iX_i\right\} = \prod_{i=1}^n \mathrm{E}e^{sw_iX_i} \le \exp\left\{s^2\sum_{i=1}^n \frac{w_i^2\lambda_i^2}{2}\right\} \le e^{\frac{s^2\|w\|_2^2\lambda^2}{2}}, \quad |s| \le \frac{1}{w\lambda}.$$

(b) The proof can be found in [70, Theorem 1.13].
(c) The proof is similar to (b), see [86, p. 29]. □

Corollary 4.2(b) is due to Petrov, and it is also called Petrov's exponential inequalities, see [57]. Although Corollary 4.2(b,c) are non-asymptotically valid for any number of summands. Nevertheless, it also has asymptotical merit, which implies: Strong Law of Large Numbers (SLNN), Central Limit Theorem (CLT), and Law of the Iterated Logarithm (LIL) for sub-exponential sums, as discussed below.

(1) **SLNN**. Let $w_i = \frac{1}{n}$. Consider the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for IID $\{\mathrm{subE}(\lambda_i)\}_{i=1}^n$ data $\{X_i\}_{i=1}^n$ with population mean μ, and we can use Corollary 4.2(b) to prove that $\bar{X}_n \xrightarrow{a.s.} \mu$. We verify the Borel-Cantelli lemma by observing that

$$\sum_{n=1}^{\infty} P\big(|\bar{X}_n - \mu| > \varepsilon\big) \le \sum_{n=1}^{\infty} 2e^{-\frac{n}{2}\left(\frac{\varepsilon^2}{\lambda^2} \wedge \frac{\varepsilon}{\lambda}\right)} < \infty,$$

which shows the strong convergence: $\bar{X}_n \xrightarrow{a.s.} \mu$. Corollary 4.2(b) also implies the rate of convergence for sample mean for all n with a high probability. It is easy to see that the sample mean $\bar{X}_n$ has the non-asymptotic error bounds by

$$|\bar{X}_n - \mu| \le \sqrt{\frac{2\lambda^2 t}{n}} \vee \frac{2\lambda t}{n} = \begin{cases} \sqrt{\dfrac{2\lambda^2 t}{n}}, & n \ge 2t \text{ (slow global rate)}, \\ \dfrac{2\lambda t}{n}, & n < 2t \text{ (fast local rate)}, \end{cases} \tag{4.6}$$

$\forall t > 0$ with the probability at least $1 - 2e^{-t}$.

(2) **CLT**. To study the convergence rate of CLT, we standardize the sum by letting $w_i = \frac{1}{\sqrt{n}}$ and apply Corollary 4.2(b) to

$$P\left(|\sqrt{n}\bar{X}_n| \geq t\right) \leq 2\exp\left\{-\frac{1}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda/\sqrt{n}}\right)\right\} = \begin{cases} 2e^{-\frac{ct^2}{\lambda^2}}, & t \leq \lambda\sqrt{n}, \\ 2e^{-\frac{t\sqrt{n}}{\lambda}}, & t > \lambda\sqrt{n}. \end{cases}$$

The above deviation inequality is powerful as it indicates the phase transition about the tail behavior of $\sqrt{n}\bar{X}_n$:

Small Deviation Regime. In the regime $t \leq \lambda\sqrt{n}$, we have a sub-Gaussian tail bound with variance proxy λ^2 as if the sum had the normal distribution with a constant variance. Note that the domain $t \leq \lambda\sqrt{n}$ widens as n increases and then the central limit theorem becomes more powerful.

Large Deviation Regime. In the regime $t \geq \lambda\sqrt{n}$, the sum has a heavier tail. The sub-exponential tail bound is affected from the extreme variable among $\{\text{subE}(\lambda_i)\}_{i=1}^n$ with parameter $\frac{\lambda}{\sqrt{n}}$.

(3) **LIL**. Let $w_i = \frac{1}{n}$ and

$$t = \frac{R\sqrt{\log\log n}}{\sqrt{n}} \leq \frac{\|\boldsymbol{w}\|_2^2 \lambda}{w} = \lambda$$

for some positive constant R. Corollary 4.2(b) claims

$$\begin{aligned} P\left(|\bar{X}_n| \geq \frac{R\sqrt{\log\log n}}{\sqrt{n}}\right) &\leq 2e^{-\frac{t^2}{2}\|\boldsymbol{w}\|_2^2\lambda^2} = 2\exp\left\{-\frac{n}{2\lambda^2} \cdot \frac{R^2 \log\log n}{n}\right\} \\ &= 2\exp\left\{\log(\log n)^{-\frac{R^2}{2\lambda^2}}\right\} = \frac{2}{(\log n)^{\frac{R^2}{2\lambda^2}}}. \end{aligned}$$

Therefore, with probability $1 - \frac{2}{(\log n)^{R^2/2\lambda^2}}$ we have

$$|\bar{X}_n| \leq \frac{R\sqrt{\log\log n}}{\sqrt{n}}.$$

Although some researchers claims that LIL is useless, we clarify that there are still some meaningful applications of LIL, see [43, 91] for the statistical and machine learning applications of the LIL.

4.2 Sub-exponential norm

Recall the Corollary 4.1(5). The absolute value of sub-exponential RV $|X|$ has a bound MGF at point $K_5^{-1}: \phi_{|X|}(K_5^{-1}) := \mathrm{E}e^{\frac{|X|}{K_5}} \le 2$. Similar to the definition of sub-Gaussian norm, we define the sub-exponential norm.

Definition 4.2 (sub-exponential norm)**.** *The sub-exponential norm of X is defined as*

$$\|X\|_{\psi_1} = \inf\left\{ t > 0 : \mathrm{E}\exp\left(\frac{|X|}{t}\right) \le 2 \right\}. \tag{4.7}$$

An alternative definition of the sub-exponential norm is

$$\|X\|_{\psi_1} := \sup_{p \ge 1} p^{-1} \left(\mathrm{E}|X|^p\right)^{\frac{1}{p}}$$

as in [83]. The sub-exponential RV X satisfies the equivalent properties in Corollary 4.1 (Characterizations of sub-exponential). Next, we present a useful lemma below which is to determine the sub-exponential parameter in the Definition 4.1 by its MGF if we adopt Definition 4.2 of the sub-exponential norm.

Proposition 4.1 (Properties of sub-exponential norm)**.** *If* $\mathrm{E}\exp(|X|/\|X\|_{\psi_1}) \le 2$, *then*

(a) Tail bounds $P(|X| > t) \le 2e^{-\frac{t}{\|X\|_{\psi_1}}}$ *for all* $t \ge 0$.

(b) Moment bounds $\mathrm{E}|X|^k \le 2\|X\|_{\psi_1}^k k!$ *for all integer* $k \ge 1$.

(c) If $\mathrm{E}X = 0$, *the MGF bounds* $\mathrm{E}e^{sX} \le e^{(2\|X\|_{\psi_1})^2 s^2}$ *for all* $|s| < \frac{1}{2\|X\|_{\psi_1}}$, *i.e.* $X \sim \mathrm{subE}(2\|X\|_{\psi_1})$.

Proof. (a). To verify (a), using exponential Markov's inequality, we have

$$P(|X| \ge t) = P\left(e^{\left|\frac{X}{\|X\|_{\psi_1}}\right|} \ge e^{\frac{t}{\|X\|_{\psi_1}}}\right) \le e^{-\frac{t}{\|X\|_{\psi_1}}} \mathrm{E}e^{\left|\frac{X}{\|X\|_{\psi_1}}\right|} \le 2e^{-\frac{t}{\|X\|_{\psi_1}}}$$

by Definition 4.2.
(b). Similar to the proof of Theorem 5.1 (b), we get from (a)

$$\begin{aligned}
\mathrm{E}|X|^k &= \int_0^\infty P(|X| \ge t) k t^{k-1} dt \le 2k \int_0^\infty e^{-\frac{t}{\|X\|_{\psi_1}}} t^{k-1} dt \\
&= 2k \int_0^\infty e^{-s} \left(s\|X\|_{\psi_1}\right)^{k-1} \|X\|_{\psi_1} ds \qquad \left[\text{let } s = \frac{t}{\|X\|_{\psi_1}}\right] \\
&= 2\|X\|_{\psi_1}^k k\Gamma(k-1) = 2\|X\|_{\psi_1}^k k!.
\end{aligned}$$

(c). Applying Taylor's expansion to MGF, we have

$$
\begin{aligned}
\mathrm{E}\exp(sX) &= 1+\sum_{k=2}^{\infty}\frac{s^k\mathrm{E}X^k}{k!} \overset{\text{(b)}}{\le} 1+2\sum_{k=2}^{\infty}(s\|X\|_{\psi_1})^k \\
&= 1+\frac{2(s\|X\|_{\psi_1})^2}{1-s\|X\|_{\psi_1}} \qquad \big(|s\|X\|_{\psi_1}|<1\big) \\
&\le 1+4\big(s\|X\|_{\psi_1}\big)^2 \le e^{(2\|X\|_{\psi_1})^2s^2}, \quad \text{if } |s|<\frac{1}{(2\|X\|_{\psi_1})}.
\end{aligned}
$$

Therefore, $X\sim \mathrm{subE}(2\|X\|_{\psi_1})$. □

Lemma 4.1(c) implies the following user-friendly concentration inequality which contains all known constant. One should note that [84, Theorem 2.8.1] includes an un-specific constant, which makes it is inefficacious when constructing non-asymptotic confident intervals for sub-exponential sample mean.

Proposition 4.2 (Concentration for RV with sub-exponential sum)**.** *Let $\{X_i\}_{i=1}^n$ be zero mean independent sub-exponential distributed with $\|X_i\|_{\psi_1}\le\infty$. Then for every $t\ge 0$,*

$$
P\left(\left|\sum_{i=1}^n X_i\right|\ge t\right)\le 2\exp\left\{-\frac{1}{4}\left(\frac{t^2}{\sum_{i=1}^n 2\|X_i\|_{\psi_1}^2}\wedge\frac{t}{\max_{1\le i\le n}\|X_i\|_{\psi_1}}\right)\right\}.
$$

Proof. If $\mathrm{E}\exp(|X|/\|X\|_{\psi_1})\le 2$, then $X\sim\mathrm{subE}(2\|X\|_{\psi_1})$ by using Lemma 4.1(c). The result follows by employing Corollary 4.2(b). □

[36] mentions an explicitly calculation the sub-exponential norm with example of Poisson distributions. Therefore, it is convenient to apply Proposition 4.2 to get the concentration of sub-exponential summation.

Lemma 4.1. *If $\|X\|_{\psi_1}$ exists, then $\|X\|_{\psi_1}=1/\phi_{|X|}^{-1}(2)$ for the MGF $\phi_X(t):=\mathrm{E}e^{tX}$.*

Proof. Note that $\|\cdot\|_{\psi_1}$ is the smallest t such that $\mathrm{E}e^{\frac{|X|}{t}}=\phi_{|X|}(t^{-1})\le 2$, so $t^{-1}\le\phi_{|X|}^{-1}(2)$ and $t\ge 1/\phi_{|X|}^{-1}(2)$. By the definition of $\|\cdot\|_{\psi_1}$ again, we have $\|X\|_{\psi_1}=1/\phi_{|X|}^{-1}(2)$. □

Example 4.5 (The sub-exponential norm of bounded RV)**.** Consider a RV $|X|\le M<\infty$. Set $\mathrm{E}e^{\frac{|X|}{t}}\le e^{\frac{M}{t}}\le 2$ and $t\ge\frac{M}{\log 2}$. By the definition of $\|X\|_{\psi_1}$, we have $\|X\|_{\psi_1}=\frac{M}{\log 2}$.

Example 4.6 (The sub-exponential norm of Poisson RV)**.** Poisson RV X has the probability mass function

$$P(X=k)=\frac{\lambda^k}{k!}e^{-k},\quad k=1,\ldots,n,\quad \lambda>0.$$

We denote it as $X\sim \text{Poisson}(\lambda)$. The MGF of the Poisson$(\lambda)$ is $\phi_X(t):=e^{\lambda(e^t-1)}$. We have $\|X\|_{\psi_1}=[\log(\log(2)\lambda^{-1}+1)]^{-1}$, and the triangle inequality shows

$$\begin{aligned}\|X-\mathrm{E}X\|_{\psi_1}&\le\|X\|_{\psi_1}+\|\mathrm{E}X\|_{\psi_1}=\|X\|_{\psi_1}+\frac{\lambda}{\log 2}\\&\le\Big[\log\big(\log(2)\lambda^{-1}+1\big)\Big]^{-1}+\frac{\lambda}{\log 2}\propto\lambda,\end{aligned}$$

where we used inequality $\|\mathrm{E}X\|_{\psi_1}=\frac{|\mathrm{E}X|}{\log 2}$ by Example 4.5.

Corollary 4.2 is useful in the next subsection for the concentration for quadratic forms.

4.3 Concentration for quadratic forms and norm of random vectors

All concentration results in the above sections are about the mean. The inference for the variance and covariance in high-dimensional models is an important problem, see [86, Section 6]. It is connected with squares of RVs. The sample variance is a quadratic form (with shift term) of the data. The data are often postulated as sub-Gaussian. For the square of a sub-Gaussian RV, it is natural to ask what is the behavior of the tail (or the exponential moment). The answer is sub-exponential by using (5) in Corollary 3.1.

A simple example that the quadratic form of Gaussian is χ^2 distributed, and the χ^2-distribution of 2 degrees of freedom is exponentially distributed with mean 2. Let us look the χ^2-concentration below:

Example 4.7 (Chi-squared RVs)**.** If $\{X_i\}_{i=1}^n\overset{\text{IID}}{\sim}N(0,1)$, then we say $Y_n:=\sum_{i=1}^n X_i^2$ follows χ^2-distribution with n-degree of freedom, denoted as $Y_n\sim\chi^2(n)$. The density function is

$$f(y)=\Gamma^{-1}\left(\frac{n}{2}\right)\left(\frac{1}{2}\right)^{\frac{n}{2}}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}\cdot 1,\quad y>0.$$

As $s<\frac{1}{2}$, the MGF of X_i^2-1 is

$$\mathrm{E}e^{s(X_i^2-1)}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{s(x^2-1)}e^{-\frac{x^2}{2}}dx$$
$$=\frac{e^{-s}}{\sqrt{1-2s}}\le e^{2s^2}=e^{\frac{(2s)^2}{2}}\quad \text{for all}\quad |s|<\frac{1}{4},$$

where the second last inequality is due to $\frac{e^{-t}}{\sqrt{1-2t}}\le e^{2t^2}$ for $|t|<\frac{1}{4}$. Then $X_i^2\sim$ subE(2,4). Applying Corollary 4.2(c), we have $Y_n\sim\text{subE}(2\sqrt{n},4)$, therefore

$$P\left(\left|\frac{Y_n-n}{n}\right|\ge t\right)\le 2e^{-\frac{n}{8}(t^2\wedge t)}.$$

Similar sub-exponential results also hold for independent sum of square of sub-Gaussian RVs. The following two lemmas in [84, p. 31] confirm this simple example to the general situation.

Lemma 4.2 (Square and product of sub-Gaussian are sub-exponential)**.** *(a) A RV X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover, $\|X^2\|_{\psi_1}=\|X\|_{\psi_2}^2$.*

(b) Let X and Y be sub-Gaussian RVs. Then XY is sub-exponential and $\|XY\|_{\psi_1}\le\|X\|_{\psi_2}\|Y\|_{\psi_2}$.

For Lemma 4.2(a), it follows from $\|X^2\|_{\psi_1}=\|X\|_{\psi_2}^2$ and Lemma 4.1 that Corollary 4.2 coincides Proposition 3.3 as $\max_{1\le i\le n}\|X_i\|_{\psi_1}\to 0$, i.e. the sub-exponential RV degenerates to the sub-Gaussian RV the next proposition gives the accurately sub-exponential parameter for the square of sub-Gaussian RV in Definition 4.1, and it improves the constant in [70, Lemma 1.12] (from $\text{subE}(16\sigma^2)$ to $\text{subE}(8\sqrt{2}\sigma^2)$).

Proposition 4.3. *Let $X\sim\text{subG}(\sigma^2)$, then $Z:=X^2-\mathrm{E}X^2\sim\text{subE}(8\sqrt{2}\sigma^2)$ or $\sim\text{subE}(8\sqrt{2}\sigma^2,8\sigma^2)$.*

Proof. The proof is immediately follows from Proposition 3.2(c) by letting $\boldsymbol{w}:=(1,0,\cdots,0)^T$. ☐

In below, we deal with a sharper Hanson-Wright inequality in [8]. The Hanson-Wright (HW) inequality is a general concentration result for quadratic forms of sub-Gaussian RVs, which was first studied in [37]. Let $\mathbf{A}=(a_{ij})\in\mathbb{R}^{n\times n}$ be a real matrix and the $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^T$ be a centered random vector with independent components. Define the Frobenius norm (Hilbert-Schmidt norm)

$$\|\mathbf{A}\|_{\mathrm{F}}:=\sqrt{\operatorname{tr}(\mathbf{A}^T\mathbf{A})}=\sqrt{\sum_{i,j}A_{i,j}^2}$$

and the spectral norm (operator norm) $\|\mathbf{A}\|_2:=\sup_{\|\boldsymbol{u}\|_2\le 1}\|\mathbf{A}\boldsymbol{u}\|_2$. As an extension of χ^2 RVs, it is of interest to study the concentration behavior of $\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}]$. Under the setting above, [14, Example 2.12] gives the Gaussian chaos concentration.

Corollary 4.3 (Gaussian chaos of order 2)**.** *Let $\xi_1,\ldots,\xi_n$ be zero-mean Gaussian with $\mathrm{E}\xi_i^2=\sigma_i^2$. Define $\mathbf{D}_\sigma=\operatorname{diag}(\sigma_1,\ldots,\sigma_n)$, then for any $x>0$*

$$P\Big(\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}\big[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\big]\ge 2\|\mathbf{D}_\sigma\mathbf{A}\mathbf{D}_\sigma\|_{\mathrm{F}}\sqrt{x}+2\|\mathbf{D}_\sigma\mathbf{A}\mathbf{D}_\sigma\|_2 x\Big)\le e^{-x}. \tag{4.8}$$

The similar concentration phenomenon is also available for sub-Gaussian RVs. which is named as the HW inequality. [72] gives a modern proof by the so-called decoupling argument attributed to [16].

Corollary 4.4 (R-V's HW inequality)**.** *Let $n\ge 1$ and $\boldsymbol{\xi}:=(\xi_1,\ldots,\xi_n)^T$ be an independent zero-mean sub-Gaussian RVs with $\max_{i=1,\ldots,n}\|\xi_i\|_{\psi_2}\le K$ for $K>0$. Let A be any $n\times n$ real matrix. Then there exists a constant $c>0$ such that*

$$P\Big(\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}\big[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\big]>t\Big)\le e^{-c\left(\frac{t^2}{K^4\|\mathbf{A}\|_{\mathrm{F}}^2}\wedge\frac{t}{K^2\|\mathbf{A}\|_2}\right)},\quad t\ge 0. \tag{4.9}$$

Furthermore, for any $x>0$

$$P\Big(\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}\big[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\big]\le cK^2\big(\|\mathbf{A}\|_2 x+\|\mathbf{A}\|_{\mathrm{F}}\sqrt{x}\big)\Big)\ge 1-e^{-x}.$$

Intuitively, the term $K^2\|\mathbf{A}\|_{\mathrm{F}}$ is seen as the "variance term". When $\mathbf{A}$ is diagonal-free (i.e. the $\mathbf{A}$ matrix has zeros down its diagonal: $a_{ii}=0$), the RV $\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}$ is zero-mean. [68] shortens the proof without unknown constant.

Corollary 4.5 (Diagonal-free Hanson-Wright inequality)**.** *Let $\xi_1,\ldots,\xi_n$ be independent, centered sub-Gaussian RVs with $\max_{i=1,\ldots,n}\|\xi_i\|_{\psi_2}\le K<\infty$. Let $\mathbf{A}$ be an $n\times n$ matrix of real numbers with $a_{ii}=0$ for each i. Then*

$$P\big(\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\ge t\big)\le e^{-\left(\frac{t^2}{64K^4\|\mathbf{A}\|_{\mathrm{F}}}\wedge\frac{t}{8\sqrt{2}K^2\|\mathbf{A}\|_2}\right)}\quad \text{for}\quad t\ge 0.$$

Under assumptions on the moments of $\xi_1,\ldots,\xi_n$ (do not need sub-Gaussian assumption), the next corollary provides a concentration inequality for quadratic forms of independent RVs satisfying Bernstein's moment condition (discussed in the next subsection).

Corollary 4.6 (Quadratic forms concentration with moment conditions). *Assume that the RV $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^T$ satisfies the condition on independent variables $\xi_1^2,\ldots,\xi_n^2$*

$$\mathrm{E}|\xi_i|^{2p}\le\frac{1}{2}p!\sigma_i^2\kappa^{2p-2},\quad\forall p\ge1$$

for some $\kappa>0$. Let A be any $n\times n$ real matrix. Then for all $t\ge0$

$$P\left(\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}\left[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\right]>t\right)\le e^{-\left(\frac{t^2}{192\kappa^2\|\mathbf{A}\mathbf{D}_\sigma\|_{\mathrm{F}}^2}\wedge\frac{t}{256\kappa^2\|\mathbf{A}\|_2}\right)},\tag{4.10}$$

where $\mathbf{D}_\sigma:=\mathrm{diag}(\sigma_1,\ldots,\sigma_n)$. Furthermore, with probability greater than $1-e^{-x}$

$$\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}-\mathrm{E}\left[\boldsymbol{\xi}^T\mathbf{A}\boldsymbol{\xi}\right]\le256\kappa^2\|\mathbf{A}\|_2x+8\sqrt{3}\kappa\|\mathbf{A}\mathbf{D}_\sigma\|_{\mathrm{F}}\sqrt{x},\quad\forall x\ge0.\tag{4.11}$$

The bound in (4.10) is exactly $\exp(-\frac{t^2}{192\kappa^2\|\mathbf{A}\mathbf{D}_\sigma\|_{\mathrm{F}}^2})$ if t is small, and while the $\exp(-c\frac{t^2}{K^4\|\mathbf{A}\|_{\mathrm{F}}^2})$ in right hand side of the R-V's HW inequality (4.9) has an unspecific constant $c>0$.

We finish this subsection with an exponential inequality for quadratic forms of a sub-Gaussian random vector. Consider the n-dimensional unit sphere $S^{n-1}:=\{\boldsymbol{x}\in\mathbb{R}^n:\|\boldsymbol{x}\|_2=1\}$. Early in [32], a random vector $\boldsymbol{X}$ in $\mathbb{R}^n$ is called sub-Gaussian (sub-exponential) if the one-dimensional marginals $\langle\boldsymbol{X},\boldsymbol{x}\rangle$ are sub-Gaussian (sub-exponential) RVs for all $\boldsymbol{x}\in\mathbb{R}^n$. Naturally, the sub-Gaussian (sub-exponential) norm of $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\|_{\psi_2}:=\sup_{\boldsymbol{x}\in S^{n-1}}\|\langle\boldsymbol{X},\boldsymbol{x}\rangle\|_{\psi_2},\quad(\|\boldsymbol{X}\|_{\psi_1}:=\sup_{\boldsymbol{x}\in S^{n-1}}\|\langle\boldsymbol{X},\boldsymbol{x}\rangle\|_{\psi_1}).$$

For the sub-Gaussian, [32] definition is equivalent to [86, Chapter 6.3], a random vector $\boldsymbol{X}\in\mathbb{R}^n$ with parameter $\sigma\in\mathbb{R}$ is sub-Gaussian (denote subGV(σ^2)) so that

$$\mathrm{E}e^{\lambda(\boldsymbol{v},\boldsymbol{X}-\mathrm{E}\boldsymbol{X})}\le e^{\frac{\lambda\sigma^2}{2}},\ \forall\lambda\in\mathbb{R}^n,\ \boldsymbol{v}\in S^{n-1}\ \Leftrightarrow\ \mathrm{E}e^{\boldsymbol{\alpha}^T(\boldsymbol{X}-\mathrm{E}\boldsymbol{X})}\le e^{\|\boldsymbol{\alpha}\|^2\frac{\sigma^2}{2}},\ \forall\boldsymbol{\alpha}\in\mathbb{R}^n.\tag{4.12}$$

In [64], the subG random vector with parameter $v_0\ge1$ is defined by

$$P\left(|\langle\boldsymbol{u},\boldsymbol{X}\rangle|\ge v_0\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}\cdot t\right)\le2e^{-\frac{t^2}{2}}$$

for all $\boldsymbol{u}\in\mathbb{R}^n$ and $t\ge0$, where $\boldsymbol{\Sigma}=\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)$ and $\|\boldsymbol{u}\|_{\mathbf{A}}=\|\mathbf{A}^{\frac{1}{2}}\boldsymbol{u}\|_2$ is the norm indexed by $\mathbf{A}$. For Definition (4.12), [40] obtains a tail bound for subG random vectors.

Corollary 4.7 (Tail inequality for quadratic forms of sub-Gaussian vectors). *Let $\mathbf{\Sigma}=\mathbf{A}^T\mathbf{A}$ for $p\times n$ matrix $\mathbf{A}$. Consider a sub-Gaussian random vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^T\sim$ $\mathrm{subGV}(\sigma^2)$ with independent components for $\boldsymbol{\mu}=\mathrm{E}\boldsymbol{\xi}$. Then, for any $t\geq 0$*

$$P\left\{\|\mathbf{A}\boldsymbol{\xi}\|^2>\sigma^2\left[\mathrm{tr}(\mathbf{\Sigma})+2\mathrm{tr}(\mathbf{\Sigma}^2t)^{\frac{1}{2}}+2\|\mathbf{\Sigma}\|_2t\right]+\mathrm{tr}(\mathbf{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^T)\left(1+2\sqrt{\frac{\|\mathbf{\Sigma}\|_2^2}{\mathrm{tr}(\mathbf{\Sigma}^2)}t}\right)\right\}\leq e^{-t}.$$

Conditioning on a divergence number of non-random covariates, an application of Corollary 4.7 for the prediction error (8.4) in regressions with sub-Gaussian noise is given in Section 8.1. The concentration bounds of sub-Gaussian random vectors depend on the parameter σ: the smaller σ, the tighter concentration bounds. Eq. (4.12) requires the distribution of subG random vectors to be isotropic, and the random vectors have an exponential tail, but the sub-Gaussian parameter σ may be large, which leads to loose bounds for constructing confidence bands. To establish tighter bounds, [44] define a different and general class of sub-Gaussian distributions in $\mathbb{R}^n$, called norm-subGaussian random vectors as follows.

Definition 4.3 (Norm-subGaussian). *A random vector $\boldsymbol{X}\in\mathbb{R}^n$ is norm-subG (denoted $\mathrm{nsubG}(\sigma^2)$), if $\exists\sigma$ so that*

$$P\big(\|\boldsymbol{X}-\mathrm{E}\boldsymbol{X}\|\geq t\big)\leq 2e^{-\frac{t^2}{2\sigma^2}},\quad \forall t\in\mathbb{R}^+.$$

The Definition 4.3 only requires the tail probability estimate has sub-Gaussian tail under l_2-norm, avoiding the uniform condition in (4.12). If $\mathrm{E}\boldsymbol{X}=\mathbf{0}$ and

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

from $\mathrm{nsubG}(\sigma^2)$, we get

$$P\big(|\langle\boldsymbol{u},\boldsymbol{X}\rangle|\geq t\big)\leq P\big(\|\boldsymbol{X}\|\geq t\big)\leq 2e^{-\frac{t^2}{2\sigma^2}},\quad \boldsymbol{u}\in S^{n-1}$$

by Cauchy's inequality. Thus, $\mathrm{nsubG}(\sigma^2)$ implies (4.12), and this verifies that the norm-subG is more general. [44] show that if $\boldsymbol{X}\in\mathbb{R}^n$ is $\mathrm{subGV}(\frac{\sigma^2}{n})$, then $\boldsymbol{X}\sim$ $\mathrm{nsubG}(8\sigma^2)$.

5 Sub-Gamma distributions and Bernstein's inequality

5.1 Sub-Gamma distributions

Comparing to the classical Chebyshev's inequality, Bernstein-type inequalities have more precise concentration, it originally is an extension of the Hoeffding's inequality with bounded assumption (see [9,10]). As mentioned by [68], the proof of Hoeffding's inequality with endpoints of the interval $[a,b]$ in Lemma 2.1 (with $n=1$) crudely depends on the variance bound

$$\mathrm{Var}X = \mathrm{E}(X-\mathrm{E}X)^2 \le \mathrm{E}\left[X-\frac{(b-a)}{2}\right]^2 \le \left[\frac{(b-a)}{2}\right]^2, \quad \text{if} \quad a\le X\le b. \tag{5.1}$$

The following tail bound for the sum $S_n := \sum_{i=1}^n X_i$ needs extra variance information.

Corollary 5.1 (Bernstein's inequality with the bounded condition)**.** *Let $X_1,\ldots,X_n$ be centralized independent variables such that $|X_i|\le M$ a.s. for all i. Then, $\forall t>0$*

$$P\big(|S_n|\ge t\big) \le 2e^{-\frac{t^2/2}{\sum_{i=1}^n \mathrm{Var}X_i + Mt/3}},$$
$$P\left\{|S_n| \ge \left(2t\sum_{i=1}^n \mathrm{Var}X_i\right)^{\frac{1}{2}} + \frac{Mt}{3}\right\} \le 2e^{-t}.$$

The next example illustrates a sharp confidence interval for sample mean if we known that the variance is sufficient small.

Example 5.1 (Non-asymptotic confidence intervals)**.** Let $\{X_i\}_{i=1}^n \overset{\mathrm{IID}}{\sim} X$ with the support $[-c,c]$ and the mean μ. Hoeffding's and Bernstein's inequalities show for $\bar{X} := n^{-1}\sum_{i=1}^n X_i$

$$P\left(|\bar{X}-\mu| \le \sqrt{\frac{2c^2\log(2/\delta)}{n}}\right) \ge 1-\delta \qquad \text{Hoeffding,}$$
$$P\left(|\bar{X}-\mu| \le \frac{c}{3n}\log\left(\frac{2}{\delta}\right) + \sqrt{\frac{2(\mathrm{Var}X)\log(2/\delta)}{n}}\right) \ge 1-\delta \qquad \text{Bernstein.}$$

For large n, the Bernstein's confidence interval is substantially shorter if X_i has relatively small variance, i.e. $\mathrm{Var}X \ll c^2$ (the factor $\sqrt{\frac{\log(2/\delta)}{n}}$ is a dominated term).

The Hoeffding's confidence is shorter as $\mathrm{Var}X = c^2$ (this extreme case attains the upper bound $\mathrm{Var}X \le c^2$ in (5.1) due to $b-a=2c$). But, for the case $\mathrm{Var}X < c^2$, if n is sufficient small s.t.

$$\frac{c}{3n}\log\left(\frac{2}{\delta}\right) \ge (c-\sqrt{\mathrm{VarX}})\sqrt{\frac{2\log(2/\delta)}{n}},$$

i.e. we need restrictions

$$n \le \frac{1}{18}\left(\frac{c}{c-\sqrt{\mathrm{Var}X}}\right)^2 \log\left(\frac{2}{\delta}\right) \ge 1$$

to ensure Hoeffding's confidence interval is more accurate when

$$\delta \le 2\exp\left\{-\frac{1}{18}\left(\frac{c-\sqrt{\mathrm{Var}X}}{c}\right)^2\right\}.$$

To prove Corollary 5.1, we need get the sharp bounds of the MGF of the single variable and then do aggregation for the summation. By the Taylor expansion, we have

$$\begin{aligned} \mathrm{E}e^{sX_i} &= 1+\sum_{k=2}^{\infty} s^k \frac{\mathrm{E}X_i^k}{k!} \le 1+\sum_{k=2}^{\infty} s^k \frac{M^{k-2}\mathrm{Var}X_i}{k!} \\ &\le 1+s^2\mathrm{Var}X_i \sum_{k=2}^{\infty} \frac{(|s|M)^{k-2}}{k!}, \quad 1\le i\le n. \end{aligned}$$

Applying the inequality $\frac{k!}{2} \ge 3^{k-2}$ for any $k \ge 2$, it implies

$$\begin{aligned} \mathrm{E}e^{sX_i} &\le 1+\frac{s^2\mathrm{Var}X_i}{2}\sum_{k=2}^{\infty}\left(\frac{|s|M}{3}\right)^{k-2} \\ &= 1+\frac{s^2\mathrm{Var}X_i/2}{1-|s|M/3} \le \exp\left(\frac{s^2\mathrm{Var}X_i/2}{1-|s|M/3}\right). \end{aligned} \tag{5.2}$$

The upper bounds of MGF essentially have the same form in comparison with Gamma distribution below whose MGF is bounded by (5.3) in following example.

Example 5.2 (Gamma RVs)**.** The Gamma distribution with density

$$f(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, \quad x \ge 0$$

is denoted by $\Gamma(a,b)$. We have $\mathrm{E}X=ab$ and $\mathrm{Var}X=ab^2$ for $X\sim\Gamma(a,b)$. The [14, p. 28] shows that the log-MGF of a centered $\Gamma(a,b)$ is bounded by

$$\log\left(\mathrm{E}e^{s(X-\mathrm{E}X)}\right)=a\big(-\log(1-sb)-sb\big)\le\frac{s^2ab^2}{[2(1-bs)]},\quad\forall\,0<s<b^{-1}.\tag{5.3}$$

Motivated by the MGF bounds in (5.3), [14] defines the sub-Gamma RV based on the right tail and left tail with variance factor v and scale factor b.

Definition 5.1 (Sub-Gamma RV)**.** *A centralized RV X is sub-Gamma with the variance factor $v>0$ and the scale parameter $c>0$ (denoted by $X\sim\mathrm{sub}\Gamma(v,c)$) if*

$$\log\left(\mathrm{E}e^{sX}\right)\le\frac{s^2v}{2(1-c|s|)},\quad\forall\,0<|s|<c^{-1}.\tag{5.4}$$

If the restriction $0<|s|<b^{-1}$ is replaced by one side conditions $0<s<b^{-1}$ (or $0<-s<b^{-1}$), we call it sub-Gamma on the right tail (or sub-Gamma on the left tail), denoted as $\mathrm{sub}\Gamma_+(v,c)$ (or $\mathrm{sub}\Gamma_-(v,c)$). In Example 5.2, the Gamma RV $X\sim\mathrm{sub}\Gamma_+(ab^2,b)$. The (5.4) is called two-sided Bernstein's condition.

Example 5.3 (Sub-exponential RVs)**.** The sub-exponential distribution with positive support implies the sub-Gamma condition

$$\log\left(\mathrm{E}e^{sX}\right)\le\frac{s^2\lambda^2}{2}\le\frac{s^2\lambda^2}{2(1-\lambda|s|)},\quad\forall|s|<\frac{1}{\lambda}.$$

This shows that $X\sim\mathrm{subE}(\lambda)$ implies $X\sim\mathrm{sub}\Gamma(\lambda^2,\lambda)$.

The sub-Gamma condition (5.4) leads to the useful tail bounds and moment bounds.

Lemma 5.1 (Sub-gamma properties, [14])**.** *If $X\sim\mathrm{sub}\Gamma(v,c)$, then*

$$P\big(|X|>t\big)\le 2e^{-\frac{v}{c^2}h\left(\frac{ct}{v}\right)}\le 2e^{-\frac{t^2/2}{v+ct}},\tag{5.5}$$

where $h(u)=1+u-\sqrt{1+2|u|}$. Moreover, we have

$$P\left\{|X|>\sqrt{2vt}+ct\right\}\le e^{-t}.$$

The tail bound in Lemma 5.1 verifies that, the sub-Gamma variable has sub-Gaussian tail behavior with parameter v for suitably small t, and it has exponential tail behavior for larger t. The proof is originated from [9].

Proof. By Chernoff's inequality,

$$P\big(X-\mathrm{E}X\ge t\big)\le \inf_{s>0} e^{-st}\mathrm{E}e^{s(X-\mathrm{E}X)}.$$

It remains to bound $\log(e^{-st}\mathrm{E}e^{s(X-\mathrm{E}X)})$. By definition of sub-Gamma variable for all $0<|s|<c^{-1}$

$$\begin{aligned}&\inf_{c^{-1}\ge s>0}\log\left(e^{-st}\mathrm{E}e^{s(X-\mathrm{E}X)}\right)\\ \le&\inf_{c^{-1}\ge u>0}\left(\frac{u^2}{2}\frac{v}{1-cu}-ut\right)=-\frac{v}{c^2}h\left(\frac{ct}{v}\right)\le-\frac{t^2/2}{v+ct},\end{aligned}$$

where the last inequality is from

$$h(u)=1+u-\sqrt{1+2|u|}\ge\frac{u^2/2}{1+u}.$$

So we conclude (5.5). □

Proposition 5.1 (Concentration for sub-Gamma sum). *Let $\{X_i\}_{i=1}^n$ be independent $\{\mathrm{sub}\Gamma(v_i,c_i)\}_{i=1}^n$ distributed with zero mean. Define $c=\max_{1\le i\le n}c_i$, we have*

(a) Closed under addition: $S_n:=\sum_{i=1}^n X_i\sim\mathrm{sub}\Gamma(\sum_{i=1}^n v_i,c)$.

(b) For every $t\ge 0$:

$$P\big(|S_n|\ge t\big)\le 2\exp\left(-\frac{t^2/2}{\sum\limits_{i=1}^n v_i+ct}\right),$$

$$P\left\{|S_n|>\left(2t\sum_{i=1}^n v_i\right)^{\frac{1}{2}}+ct\right\}\le 2e^{-t}.$$

(c) If $X\sim\mathrm{sub}\Gamma(v,c)$, the moments bounds satisfy for any integer $k\ge1$:

$$\mathrm{E}X^k\le k2^{k-2}\left[2(\sqrt{2v})^k\Gamma\left(\frac{k}{2}\right)+c(\sqrt{2v})^{k-1}\Gamma\left(\frac{k+1}{2}\right)+3c^k\Gamma(k)\right].$$

(d) If $X\sim\mathrm{sub}\Gamma(v,c)$, the even moments bounds satisfy

$$\mathrm{E}X^{2k}\le k!(8v)^k+(2k)!(4c)^{2k},\quad k\ge 1.$$

(e) If

$$P\left\{|X|>(2tv)^{\frac{1}{2}}+ct\right\}\le 2e^{-t},$$

then $X\sim\mathrm{sub}\Gamma(32(v+2c^2),8c)$.

Proof. (a) By definition of $\{\mathrm{sub}\Gamma(v_i,c_i)\}_{i=1}^n$, we have

$$\log\left(\mathrm{E}e^{sX_i}\right)\le\frac{s^2}{2}\frac{v_i}{1-c_i|s|},\quad\forall\, 0<|s|<c^{-1},$$

from which and the independence among $\{X_i\}_{i=1}^n$, thus

$$\log\left(\mathrm{E}e^{sS_n}\right)\le\frac{s^2}{2}\sum_{i=1}^n\frac{v_i}{1-c_i|s|}\le\frac{s^2}{2}\frac{\sum\limits_{i=1}^n v_i}{1-c|s|},\quad\text{for all}\quad 0<|s|<c^{-1}.$$

(b) Employing Proposition 5.1, we immediately obtain (b) due to (a).
(c) Applying the integration form of the expectation formula, it yields

$$\begin{aligned}
\mathrm{E}X^k\le\mathrm{E}|X|^k&=k\int_0^\infty x^{k-1}P\{|X|>x\}dx\\
&=k\int_0^\infty x^{k-1}P\left\{|X|>\sqrt{2vt}+ct\right\}\left(\frac{\sqrt{2v}}{2\sqrt{t}}+c\right)dt\\
&\le 2k\int_0^\infty\left(\sqrt{2vt}+ct\right)^{k-1}\left(\frac{\sqrt{2vt}+2ct}{2t}\right)e^{-t}dt\\
&=k\int_0^\infty\left[\left(\sqrt{2vt}+ct\right)^k+ct\left(\sqrt{2vt}+ct\right)^{k-1}\right]\frac{e^{-t}}{t}dt.
\end{aligned}$$

From (b) and inequality (3.7),

$$\begin{aligned}
\mathrm{E}X^k&\le k\int_0^\infty\left\{2^{k-1}\left[\left(\sqrt{2vt}\right)^k+(ct)^k\right]+ct2^{k-2}\left[\left(\sqrt{2vt}\right)^{k-1}+(ct)^{k-1}\right]\right\}\frac{e^{-t}}{t}dt\\
&=k2^{k-2}\int_0^\infty\left[2\left(\sqrt{2v}\right)^k t^{(\frac{k}{2})-1}+c\left(\sqrt{2v}\right)^{k-1}t^{\frac{(k+1)}{2}-1}+3c^kt^{k-1}\right]e^{-t}dt\\
&=k2^{k-2}\left[2(\sqrt{2v})^k\Gamma\left(\frac{k}{2}\right)+c(\sqrt{2v})^{k-1}\Gamma\left(\frac{k-1}{2}\right)+3c^k(k-1)!\right].
\end{aligned}$$

(d,e) The proofs are in [14, Theorem 2.3]. □

Having obtained Proposition 5.1(b), from the upper bound in (5.2), we finish the proof of Proposition 5.1 by treating $X_i\sim\mathrm{sub}\Gamma(\mathrm{Var}X_i/2,M/3)$ for $i=1,\dots,n$.

5.2 Bernstein's growth of moments condition

In some settings, one can not assume the RVs being bounded. Bernstein's inequality for the sum of independent RVs allows us to estimate the tail probability by a weaker version of an exponential condition on the growth of the k-moment without the boundedness.

Corollary 5.2 (Bernstein's inequality with the growth of moment condition)**.** *If the centred independent RVs $X_1,\ldots,X_n$ satisfy the growth of moments condition*

$$\mathrm{E}|X_i|^k \le 2^{-1} v_i^2 \kappa_i^{k-2} k!, \quad i=1,\ldots,n \quad \textit{for all} \quad k\ge 2, \tag{5.6}$$

where $\{\kappa_i\}_{i=1}^n,\{v_i\}_{i=1}^n$ are constants independent of k. Let $\nu_n^2=\sum_{i=1}^n v_i^2$ (the fluctuation of sums) and $\kappa=\max_{1\le i\le n}\kappa_i$. Then, we have $X_i\sim \mathrm{sub}\Gamma(v_i,\kappa_i)$ and for $t>0$

$$\begin{aligned} &P\big(|S_n|\ge t\big)\le 2e^{-\frac{t^2}{2\nu_n^2+2\kappa t}},\\ &P\Big(|S_n|\ge \sqrt{2\nu_n^2 t}+\kappa t\Big)\le 2e^{-t}. \end{aligned} \tag{5.7}$$

Proof. Given that $\kappa_i|s|<1$ for all i, (5.6) implies that $X_i\sim\mathrm{sub}\Gamma(v_i,\kappa_i)$ for $1\le i\le n$

$$\mathrm{E}e^{sX_i}\le 1+\frac{v_i^2}{2}\sum_{k=2}^{\infty}|s|^k\kappa_i^{k-2}=1+\frac{s^2v_i^2}{2(1-|s|\kappa_i)}\le e^{\frac{s^2v_i^2}{(2-2\kappa_i|s|)}}.$$

The independence among $\{X_i\}_{i=1}^n$ and Proposition 5.1(a,b) implies (5.7). □

The (5.6) is also called Bernstein's moment condition. Corollary 5.2 slightly extends [82, Lemma 2.2.11] for the case $\kappa_i\equiv\kappa$ (a fixed number). It should be noted that (5.6) can be replaced by

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}|X_i|^k\le\frac{1}{2}v^2\kappa^{k-2}k!, \quad k=3,4,\ldots, \quad \forall i,$$

where the v^2 is a variance-depending constant such that $\frac{1}{n}\sum_{i=1}^n\mathrm{E}|X_i|^2\le v^2$. Then (5.7) still holds with $\nu_n^2=nv$, see [14, Theorem 2.10].

Example 5.4 (Normal RV)**.** Applying the relation between MGF and moment, the k-th moment of $X\sim N(\mu,\sigma^2)$ is

$$\begin{aligned} &\mathrm{E}X^{2k-1}=0,\\ &\mathrm{E}|X|^{2k}=\sigma^{2k}(2k-1)(2k-3)\cdots 3\cdot 1\le 2^{-1}(2\sigma^2)\sigma^{2k-2}(2k)!, \end{aligned}$$

which satisfies (5.6) with $v^2=2\sigma^2$, $\kappa=\sigma^2$.

5.3 Concentration of exponential family without compact space

Theory and statistical applications of natural exponential family (3.5) have attracted renewed attention in the past years [52]. In Lasso penalized generalized linear models (GLMs), the results of oracle inequalities lie on CIs of a quantity that can be represent as Karush-Kuhn-Tucker conditions (see (8.22)) related to the centralized exponential family empirical process: $\sum_{i=1}^{n} w_i(Y_i - \mathrm{E}Y_i)$ for no-random weights $\{w_i\}_{i=1}^{n}$ depending on the fixed design. [46] has studied the sub-exponential growth of the cumulants of an exponential family distribution and studied oracle inequalities of Lasso regularized GLMs, but the constant in their result is not specific.

In this part, we obtain cental moments bounds with a specific constant, which gives the Bernstein's inequality for the general exponential family, and the proof is based on the Cauchy formula of higher-order derivatives for complex functions [76, Corollary 4.3].

Lemma 5.2 (Cauchy's derivative inequalities)**.** *If f is analytic in an open set that contains the closure of a disk D centered at z_0 of radius $0<r<\infty$, then*

$$\left|f^{(n)}(z_0)\right| \le \frac{n!}{r^n} \sup_{z:|z-z_0|=r} |f(z)|.$$

[99] adopts a similar approach for recovering the probability mass function (p.m.f.) from the characteristic function.

It is well-known that exponential families on the natural parameter space,

$$\Theta := \left\{ \theta \in \mathbb{R}^k : e^{b(\theta)} := \int c(y) e^{y\theta} \mu(dy) < \infty \right\}$$

have finite analytic (standardized) moments and cumulants, see [46, Lemma 3.3]. The natural parameter space of an exponential family is convex, see [52]. A nice property in Lehmann's measure-theoretical statistical inference book is that:

Lemma 5.3 (Analytic property of MGF in the exponential family)**.** *The MGF $m_{\theta_i}(s) := \mathrm{E}_{\theta_i} e^{sX_i}$ on $s \in \mathbb{C}$ of exponential family RVs indexed by θ_i, is analytic on Θ (see* [52, Theorem 2.7.1] *or* [67, Theorem 2]*).*

First, let us check the following lemma which is deduced by Cauchy's inequalities for the Taylor's series coefficients of a complex analytic function.

Proposition 5.2. *The $s \mapsto \bar{m}_{\theta_i}(s) := \mathrm{E}_{\theta_i} e^{s|Y_i - \dot{b}(\theta_i)|}$ is analytic on the natural parameter space Θ with radius $\mathrm{r}(\Theta)$, and the k-th absolute central moment of $\{w_i Y_i\}_{i=1}^{n}$ is bounded by*

$$\mathrm{E}_{\theta_i}|w_i(Y_i-\mathrm{E}Y_i)|^k \le \frac{k!}{2}\Big(w\sqrt{2C_{\theta_i}}\Big)^2(wC_{\theta_i})^{k-2},\quad k=2,3,\ldots,$$

where $\{w_i\}_{i=1}^n$ *are non-random with* $w:=\max_{1\le i\le n}|w_i|>0$*, and*

$$C_{\theta_i}:=\inf_{0<r\le \mathrm{r}(\Theta)} r^{-1}\mathrm{E}_{\theta_i}e^{r|X_i-\dot{b}(\theta_i)|}.$$

Proof. Let $s\in\mathbb{R}\mathrm{i}:=\{b\mathrm{i}:b\in\mathbb{R}\}$ be a given complex number on imaginary axis.

$$\begin{aligned}\bar{m}_{\theta_i}(s)&=\mathrm{E}_{\theta_i}\Big(e^{s[Y_i-\dot{b}(\theta_i)]}1\{Y_i\ge\dot{b}(\theta_i)\}\Big)+\mathrm{E}_{\theta_i}\Big(e^{s[\dot{b}(\theta_i)-Y_i]}1\{Y_i<\dot{b}(\theta_i)\}\Big)\\&=\int_{x\ge\dot{b}(\theta_i)}c(x)e^{x(\theta_i+s)}e^{-\dot{b}(\theta_i)}\mu(dx)+\int_{x<\dot{b}(\theta_i)}c(y)e^{x(\theta_i-s)}e^{-\dot{b}(\theta_i)}\mu(dx)\\&=e^{-\dot{b}(\theta_i)}\left[\int_{x\ge\dot{b}(\theta_i)}c(x)e^{x(\theta_i+s)}\mu(dx)+\int_{x<\dot{b}(\theta_i)}c(x)e^{x(\theta_i-s)}\mu(dx)\right].\end{aligned}\tag{5.8}$$

The natural parameter space implies $\int c(x)e^{x\theta_i}\mu(dx)$ is finite and analytic for $\theta_i\in\Theta$, so

$$\begin{aligned}&\int 1\{x\ge\dot{b}(\theta_i)\}c(x)e^{x(\theta_i+s)}\mu(dx),\\&\int 1\{x<\dot{b}(\theta_i)\}c(y)e^{x(\theta_i-s)}\mu(dx)\end{aligned}$$

are finite and analytic for $s\in\mathrm{i},\theta_i\in\Theta$. By Lemma 5.3, $\bar{m}_{\theta_i}(s)$ in (5.8) is analytic on

$$D_{\theta_i}:=\big\{s\in\mathbb{C}:\mathrm{Re}(\theta_i+s)\in\mathrm{Int}(\Theta)\text{ and }\mathrm{Re}(\theta_i-s)\in\mathrm{Int}(\Theta)\big\}$$

by using analytic continuation (i.e. the $\bar{m}_{\theta_i}(s)$ has an analytic continuation from $\bar{m}_{\theta_i}(s)$ on $s\in D_{\theta_i}$ to $\bar{m}_{\theta_i}(s)$ on $s\in\mathbb{C}$, see [76, Corollary 4.9]).

Since $0+\theta_i=\theta_i\in D_{\theta_i}\subset\mathrm{Int}(\Theta)$, $\bar{m}_{\theta_i}(s)$ is analytic at the point 0 and hence the function is also analytic in a neighborhood of 0. By the analyticity of the functions $\{\bar{m}_{\theta_i}(s)\}_{\theta_i\in\Theta}$ on $s\in\mathrm{Int}(\Theta)$, and Cauchy's derivative inequality with $z_0=0$, we have

$$\mathrm{E}_{\theta_i}|Y_i-\dot{b}(\theta_i)|^k=\bar{m}_{\theta_i}^{(k)}(0)\le k!r^{-k}\sup_{|s|=r}\big|\mathrm{E}_{\theta_i}e^{s|Y_i-\dot{b}(\theta_i)|}\big|,\quad 0<r\le\mathrm{r}(\Theta).\tag{5.9}$$

Let $s=r(\cos\omega+\mathrm{i}\sin\omega)$, $\omega\in[0,2\pi]$. Then, we get

$$\begin{aligned}\mathrm{E}_{\theta_i}e^{s|Y_i-\dot{b}(\theta_i)|}&=\mathrm{E}_{\theta_i}e^{r(\cos\omega+\mathrm{i}\sin\omega)|Y_i-\dot{b}(\theta_i)|}\\&=\mathrm{E}_{\theta_i}\Big[e^{r\cos\omega|Y_i-\dot{b}(\theta_i)|}e^{\mathrm{i}r\sin\omega|Y_i-\dot{b}(\theta_i)|}\Big].\end{aligned}$$

Hence, (5.9) gives

$$
\begin{aligned}
&k!\frac{1}{r^k}\sup_{|s|=r}\left|\mathrm{E}_{\theta_i}e^{s|Y_i-\dot{b}(\theta_i)|}\right| \\
&\le k!\frac{1}{r^k}\sup_{\omega\in[0,2\pi]}\mathrm{E}_{\theta_i}e^{r\cos\omega|Y_i-\dot{b}(\theta_i)|}\le k!\frac{1}{r^k}\mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|} \\
&=k!\left\{\frac{1}{r}\left[\mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|}\right]^{\frac{1}{k}}\right\}^k\le k!\left\{\frac{1}{r}\left[\mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|}\right]\right\}^k \quad \left(\text{Due to } \mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|}\ge 1\right).
\end{aligned}
$$

From (5.9), it shows that by take infimum over $0<r\le \mathrm{r}(\Theta)$,

$$
\begin{aligned}
\mathrm{E}_{\theta_i}|Y_i-\dot{b}(\theta_i)|^k &\le k!\left\{\inf_{0<r\le\mathrm{r}(\Theta)} r^{-1}\left[\mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|}\right]\right\}^k \\
&\le k!C_{\theta_i}^k=\frac{k!}{2}\left(\sqrt{2C_{\theta_i}}\right)^2 C_{\theta_i}^{k-2},
\end{aligned}
$$

where $C_{\theta_i}:=\inf_{0<r\le\mathrm{r}(\Theta)} r^{-1}[\mathrm{E}_{\theta_i}e^{r|Y_i-\dot{b}(\theta_i)|}]$. Then for $\{w_i(Y_i-\mathrm{E}Y_i)\}_{i=1}^n$, we have

$$
\begin{aligned}
\mathrm{E}_{\theta_i}|w_i(Y_i-\mathrm{E}Y_i)|^k &\le \frac{1}{2}k!\left(\sqrt{2C_{\theta_i}}\right)^2 C_{\theta_i}^{k-2}w^k \\
&=\frac{1}{2}k!\left(w\sqrt{2C_{\theta_i}}\right)^2 (wC_{\theta_i})^{k-2}, \quad k=2,3,\ldots.
\end{aligned}
$$

The proof is complete. □

Therefore, $w_iX_i\sim \mathrm{sub}\Gamma(w\sqrt{2C_{\theta_i}},wC_{\theta_i})$ by Proposition 5.2 and we can apply the Bernstein's inequality with the growth of moments condition to get the following concentration of exponential family on a natural parameter space.

Theorem 5.1 (Concentration of exponential family)**.** *Let $\{Y_i\}_{i=1}^n$ be a sequence of independent RVs with their densities $\{f(y_i;\theta_i)\}_{i=1}^n$ belong to canonical exponential family* (3.5) *on the natural parameter space $\theta_i\in\Theta$. Given non-random weights $\{w_i\}_{i=1}^n$ with $w=\max_{1\le i\le n}|w_i|>0$, then*

$$
P\left(\left|\sum_{i=1}^n w_i(Y_i-\mathrm{E}Y_i)\right|\ge t\right)\le 2\exp\left(-\frac{t^2}{4w^2\sum_{i=1}^n C_{\theta_i}+2w\max\limits_{1\le i\le n} C_{\theta_i}t}\right). \tag{5.10}
$$

Theorem 5.1 has no compact space assumption. If we impose the compact space assumption (E.1) in Proposition 3.2, it leads to the sub-Gaussian concentration as presented in Proposition 5.1. The constant C_{θ_i} in Theorem 5.1 is hard to determine in general exponential family with infinite support. However, if the exponential family is Poisson, the C_{θ_i} can be obtained as an explicit form.

Theorem 5.2 (Concentration for weighted Poisson summation)**.** *Let* $\{Y_i\}_{i=1}^n$ *be independent* $\{\text{Poisson}(\lambda_i)\}_{i=1}^n$ *distributed. For non-random weights* $\{w_i\}_{i=1}^n$ *with* $w=\max_{1\le i\le n}|w_i|>0$*, put* $S_n^w:=\sum_{i=1}^n w_i(Y_i-\mathrm{E}Y_i)$*, then for all* $t\ge 0$

$$
\begin{aligned}
&P\big(|S_n^w|\ge t\big)\le 2\exp\left(-\frac{t^2/2}{w^2\sum_{i=i}^n\lambda_i+wt/3}\right),\\
&P\left\{|S_n^w|>w\left[\left(2t\sum_{i=1}^n\lambda_i\right)^{\frac{1}{2}}+\frac{t}{3}\right]\right\}\le e^{-t}.
\end{aligned}
\tag{5.11}
$$

Proof. We evaluate the log-MGF of centered Poisson RVs $\{Y_i-\mathrm{E}Y_i\}_{i=1}^n$

$$
\begin{aligned}
&\log \mathrm{E}e^{sw_i(Y_i-\mathrm{E}Y_i)}=-sw_i\mathrm{E}Y_i+\log \mathrm{E}e^{sw_iY_i}\\
&=-\lambda_i sw_i+\log e^{\lambda_i(e^{sw_i}-1)}=\lambda_i(e^{sw_i}-sw_i-1).
\end{aligned}
$$

Note that, for s in a small neighbourhood of zero,

$$
\begin{aligned}
\lambda_i\big(e^{sw_i}-sw_i-1\big)&=\lambda_i\sum_{k=2}^{\infty}\frac{(sw_i)^k}{k!}\le\lambda_i\sum_{k=2}^{\infty}\frac{(|sw|)^k}{k!}\\
&=\frac{\lambda_i s^2w^2}{2}\sum_{k=2}^{\infty}\frac{w^{k-2}}{k(k-1)\cdots 3}\le\frac{\lambda_i s^2w^2}{2}\sum_{k=2}^{\infty}\left(\frac{|sw|}{3}\right)^{k-2}\\
&=\frac{s^2}{2}\frac{w^2\lambda_i}{1-w|s|/3}
\end{aligned}
\tag{5.12}
$$

for $|s|\le\frac{3}{w}$, which implies $w_i(Y_i-\mathrm{E}Y_i)\sim \mathrm{sub}\Gamma(w^2\lambda_i,\frac{w}{3})$. By Proposition 5.1(a), we have $S_n^w\sim\mathrm{sub}\Gamma(w^2\sum_{k=i}^n\lambda_i,\frac{w}{3})$. Then applying Proposition 5.1(b), we get (5.11). □

Before ending this section, we show a result for checking Bernstein's moment condition by the moment recurrence condition of log-concave distributions.

Definition 5.2 (Moment recurrence condition)**.** *A RV* Z *is called moment bounded with parameter* $L>0$ *if it has recurrence condition* $\mathrm{E}|Z|^p\le p\,L\cdot\mathrm{E}|Z|^{p-1}$ *for any integer* $p\ge 1$.

By the recursion relation, Definition 5.2 implies that any moment bounded RV Z satisfies $\mathrm{E}|Z|^p\le p!L^P$. Hence, the tails of its moment bounded RVs decay as the Bernstein's growth of moment condition. So the constant C_{θ_i} in Theorem 5.1 is relatively easy to find. [74, Lemmas 7.2, 7.3, 7.6, 7.7] showed that any log-concave continuous distribution (see Section 3.4) and log-concave discrete distribution X with density f is moment bounded with parameter $L\propto\mathrm{E}|X|$.

Example 5.5 (Log-concave continuous distributions, [4])**.** Many continuous distributions, such as normal distribution, exponential distribution, uniform distribution over any convex set, logistic distribution, extreme value distribution, chi-square distribution, chi distribution, hyperbolic secant distribution, Laplace distribution, Weibull distribution (the shape parameter $\theta \geq 1$), Gamma distribution (the shape parameter $a \geq 1$) and Beta distribution (both shape parameters are ≥ 1) have log-concave continuous densities.

Analogous to the log-concave continuous function in (3.12), we can define log-concave sequence for the p.m.f. of discrete RV, which also has Bernstein-type concentrations.

Definition 5.3 (Log-concave discrete distributions)**.** *A sequence $\{p_i\}_{i\in\mathbb{Z}}$ (or $\{p_i\}_{i\in\mathbb{N}}$) is said to be log-concave if $p_{i+1}^2 \geq p_i p_{i+2}$ for all $i\in\mathbb{Z}$ (or $i\in\mathbb{N}$). An integer-valued RV X is log-concave if its probability mass function (p.m.f.) $p_i := P(X=i)$ is log-concave sequence.*

Example 5.6 (Log-concave discrete distributions)**.** Bernoulli and binomial distributions, Poisson distribution, geometric distribution, and negative binomial distribution (with number of success >1) and hypergeometric distribution have log-concave integer-valued p.m.f., see [45].

6 Sub-Weibull distributions

6.1 Sub-Weibull RVs and ψ_θ-norm

A RV is heavy-tailed if its distribution function fails to be bounded by a decreasing exponential function [31]. We first give a simple example of the heavy-tailed distributions arisen by multiplying sub-Gaussian RVs. The proof is motivated by [84, Lemmas 2.7.7].

Lemma 6.1 (The product of sub-Gaussians)**.** *Suppose $\{X^{(m)}\}_{m=1}^d$ are sub-Gaussian (may be dependent). Then $\prod_{m=1}^d |X^{(m)}|^{\frac{2}{d}}$ is sub-exponential and*

$$\left\|\prod_{m=1}^d [X^{(m)}]^{\frac{2}{d}}\right\|_{\psi_1} \leq \prod_{m=1}^d \|X^{(m)}\|_{\psi_2}^{\frac{2}{d}}.$$

Proof. By the definition of sub-Gaussian norm,

$$\mathrm{E}e^{|X^{(m)}/\|X^{(m)}\|_{\psi_2}|^2} \leq 2, \quad m=1,\ldots,d.$$

Applying the elementary inequality $\prod_{m=1}^{d} a_m \le \frac{1}{d}\sum_{m=1}^{d} a_m^d$, we get by Jensen's inequality

$$\begin{aligned}
&\mathrm{E}e^{\prod\limits_{m=1}^{d}\left[|X^{(m)}|^{2/d}/\|X^{(m)}\|_{\psi_2}^{2/d}\right]} \\
&\le \mathrm{E}e^{\frac{1}{d}\sum\limits_{m=1}^{d}\left[X^{(m)}/\|X^{(m)}\|_{\psi_2}\right]^2} \le \frac{1}{d}\sum_{m=1}^{d}\mathrm{E}e^{\left[X^{(m)}/\|X^{(m)}\|_{\psi_2}\right]^2} \le 2. \qquad (6.1)
\end{aligned}$$

The proof is finished by the definition of the sub-exponential norm. □

In probability, Weibull RVs are generated from the power of the exponential RVs.

Example 6.1 (Weibull RVs)**.** The Weibull RV $X\in\mathbb{R}^+$ is defined by its survival function

$$P(X\ge x)=e^{-bx^\theta},\quad x\ge 0$$

for the scale parameter $b>0$ and the shape parameter $\theta>0$.

Sub-Weibull distribution is characterized by the right tail of the Weibull distribution and is a generalization of both sub-Gaussian and sub-exponential distributions.

Definition 6.1 (Sub-Weibull distributions)**.** *A RV X satisfying* $P(|X|\ge x)\le ae^{-bx^\theta}$ *for given* $a,b,\theta>0$, *is called a sub-Weibull RV with tail parameter* θ *(denoted by* $X\sim \mathrm{subW}(\theta)$*).*

A $\mathrm{subW}(\theta)$'s tail is no heavier than that of a Weibull RV with tail parameter θ. It is emphasized that $X\sim\mathrm{subW}(\theta)$ RVs with $\theta<1$ belongs to heavy-tailed RVs. Recently, the Weibull-like tail condition is also studied in high-dimensional statistics and random matrix theory (see [50, 79, 89]). [36] names $\mathrm{subW}(\theta)$ as θ-sub-exponential RV. There are 4 equivalent conditions to reveal the sub-Weibull tail condition which is useful in applications.

Corollary 6.1 (Characterizations of sub-Weibull condition)**.** *Let X be an RV. Then the following properties are equivalent.*

(1) The tails of X satisfy $P(|X|\ge x)\le e^{-(x/K_1)^\theta}$ *for all* $x\ge 0$.

(2) The moments of X satisfy $\|X\|_k := (\mathrm{E}|X|^k)^{1/k}\le K_2 k^{1/\theta}$ *for all* $k\ge 1$.

(3) The MGF of $|X|^{1/\theta}$ *satisfies* $\mathrm{E}e^{\lambda^{1/\theta}|X|^{1/\theta}}\le e^{\lambda^{1/\theta}K_3^{1/\theta}}$ *for* $|\lambda|\le\frac{1}{K_3}$.

(4) The MGF of $|X|^{1/\theta}$ *is bounded at some point:* $\mathrm{E}e^{|X/K_4|^{1/\theta}} \le 2$.

The proof can be found in [85, 89] by mimicking the proof of [84, Proposition 2.5.2]. It follows from Corollary 6.1(4) that X is sub-Weibull with tail parameter θ if and only if $|X|^{1/\theta}$ is sub-exponential.

Let θ_1 and θ_2 $(0<\theta_1\le\theta_2)$ be two sub-Weibull parameters. Corollary 6.1 implies $\mathrm{subW}(\theta_1)\subset\mathrm{subW}(\theta_2)$. The following Orlicz-type norms play crucial roles in deriving tail and maximal inequality for sub-Weibull RVs without the zero-mean assumption.

Definition 6.2 (Sub-Weibull norm or ψ_θ-norm)**.** *Let* $\psi_\theta(x)=e^{x^\theta}-1$*. The sub-Weibull norm of X for any* $\theta>0$ *is defined as*

$$\|X\|_{\psi_\theta} := \inf\left\{C\in(0,\infty): \mathrm{E}e^{|X|^\theta/C^\theta}\le 2\right\}.$$

From Corollary 6.1(4), a second useful definition of sub-Weibull RVs is the RVs with finite ψ_θ-norm. Sub-Weibull norm is a special case of the Orlicz norm [88].

Definition 6.3 (Orlicz norms)**.** *Let* $g:[0,\infty)\to[0,\infty)$ *be a non-decreasing convex function with* $g(0)=0$*. The "g-Orlicz norm" of a RV X is*

$$\|X\|_g := \inf\{\eta>0 : \mathrm{E}[g(|X|/\eta)]\le 1\}.$$

Let $g(x)=e^{x^\theta}-1$ and $\mathrm{E}[g(|X|/\eta)]\le 1$ implies $\mathrm{E}[\exp(|X|^\theta/\eta^\theta)]\le 2$, which is the definition of sub-Weibull norm. Similar to sub-exponential, [85, 89, 96] attained the following.

Corollary 6.2 (Properties of sub-Weibull norm)**.** *If* $\mathrm{E}e^{|X/\|X\|_{\psi_\theta}|^\theta}\le 2$*, then*

(a) $P(|X|>t)\le 2e^{-(t/\|X\|_{\psi_\theta})^\theta}$ *for all* $t\ge 0$.

(b) Moment bounds: $\mathrm{E}|X|^k \le 2\|X\|_{\psi_\theta}^k \Gamma(\frac{k}{\theta}+1)$.

6.2 Concentrations for sub-Weibull summation

The Chernoff inequality trick in the derivation of Corollary 4.2 for sub-exponential concentration is not valid for sub-Weibull distributions, since the exponential moment equivalent conditions of sub-Weibull are on the absolute value $|X|$. However, Bernstein's moment condition is the exponential moment of the absolute value. An alternative method is given by [50], who defines the so-called

Generalized Bernstein-Orlicz (GBO) norm. Fixed $\alpha>0$ and $L\geq 0$, define a function $\Psi_{\theta,L}(\cdot)$ with its inverse function

$$\Psi_{\theta,L}^{-1}(t):=\sqrt{\log(t+1)}+L[\log(t+1)]^{\frac{1}{\theta}},\quad \forall t\geq 0.$$

A promising development is that the following GBO norm helps us derive tail behaviors for sub-Weibull RVs.

Definition 6.4 (Generalized Bernstein-Orlicz Norm). *The generalized Bernstein-Orlicz (GBO) norm of a RV X is then given by*

$$\|X\|_{\Psi_{\theta,L}}:=\inf\{\eta>0:\mathrm{E}[\Psi_{\theta,L}(|X|/\eta)]\leq 1\}.$$

The monotone function $\Psi_{\theta,L}(\cdot)$ is motivated by the classical Bernstein's inequality for sub-exponential RVs. Like the sub-Weibull norm properties Corollary 6.2(a), the following proposition in [50] allows us to get the concentration inequality for RVs with finite GBO norms.

Corollary 6.3 (GBO norm concentration). *For any RV X with $\|X\|_{\Psi_{\theta,L}}<\infty$, we have*

$$P\Big(|X|\geq\|X\|_{\Psi_{\theta,L}}\Big\{\sqrt{t}+Lt^{\frac{1}{\theta}}\Big\}\Big)\leq 2e^{-t},\quad \textit{for all}\quad t\geq 0.$$

From Corollary 6.3, it is easy to derive the concentration inequality for a single sub-Weibull RV or even the sum of independent sub-Weibull RVs. [50, Theorem 3.1] obtains an upper bound for the GBO norm of the summation.

Corollary 6.4 (Concentration for sub-Weibull summation). *If $\{X_i\}_{i=1}^n$ are independent centralized RVs such that $\|X_i\|_{\psi_\theta}<\infty$ for all $1\leq i\leq n$ and some $\theta>0$, then for any weight vector $\boldsymbol{w}=(w_1,\ldots,w_n)\in\mathbb{R}^n$, we have*

$$\Big\|\sum_{i=1}^n w_iX_i\Big\|_{\Psi_{\theta,L_n(\theta)}}\leq 2eC(\theta)\|\boldsymbol{b}\|_2$$

and

$$P\left(\left|\sum_{i=1}^n w_iX_i\right|\geq 2eC(\theta)\|\boldsymbol{b}\|_2\Big\{\sqrt{t}+L_n(\theta)t^{\frac{1}{\theta}}\Big\}\right)\leq 2e^{-t}, \tag{6.2}$$

where $\boldsymbol{b}=(w_1\|X_1\|_{\psi_\theta},\ldots,w_n\|X_i\|_{\psi_\theta})^T\in\mathbb{R}^n$,

$$L_n(\theta):=\frac{4^{\frac{1}{\theta}}}{\sqrt{2}\|\boldsymbol{b}\|_2}\times\begin{cases}\|\boldsymbol{b}\|_\infty, & \textit{if}\quad \theta<1,\\ \frac{4e\|\boldsymbol{b}\|_{\theta/(1-\theta)}}{C(\theta)}, & \textit{if}\quad \theta\geq 1,\end{cases}$$

and

$$C(\theta):=\max\left\{\sqrt{2},2^{\frac{1}{\theta}}\right\}\times\begin{cases}\sqrt{8}e^3(2\pi)^{\frac{1}{4}}e^{\frac{1}{24}}(e^{\frac{2}{e}}/\theta)^{\frac{1}{\theta}}, & \text{if} \quad \theta<1,\\ 4e+2(\log 2)^{\frac{1}{\theta}}, & \text{if} \quad \theta\geq 1.\end{cases}$$

The upper bound of sub-Weibull norm for summation provided by Corollary 6.4 dependents on $\|X_i\|_{\psi_\theta}$ and the $\boldsymbol{w}$. [100] gives a sharper version of Corollary 6.4. The $\theta=1$ is the phrase transition point, and reflect the fact that Weibull RVs are log-convex for $\theta\leq 1$ and log-concave for $\theta\geq 1$. At last, we mention a generalized Hanson-Wright inequality for sub-Weibull RVs in [36, Proposition 1.5]. Let $\max_{i=1,\dots,n}\|(a_{ij})_j\|_2:=\|A\|_{2\to\infty}$, where

$$\|A\|_{p\to q}:=\sup\left\{\|Ax\|_q:\|x\|_p\leq 1\right\}.$$

Corollary 6.5 (Concentration for the quadratic form of sub-Weibull RVs.). *Let $q\in\mathbb{N}$, $A=(a_{ij})$ be a symmetric $n\times n$ matrix and let $\{X_i\}_{i=1}^n$ be independent and centered RVs with $\|X_i\|_{\Psi_{2/q}}\leq M$ and $\mathrm{E}X_i^2=\sigma_i^2$. We have*

$$P\left(\left|\sum_{i,j}a_{ij}X_iX_j-\sum_{i=1}^n\sigma_i^2a_{ii}\right|\geq t\right)\leq 2e^{-\eta(A,q,t/M^2)/C}, \quad \forall t\geq 0,$$

where

$$\eta(A,q,t):=\min\left(\frac{t^2}{\|A\|_{\mathrm{F}}^2},\frac{t}{\|A\|_{\mathrm{op}}},\left(\frac{t}{\max\limits_{i=1,\dots,n}\|(a_{ij})_j\|_2}\right)^{\frac{2}{4+T}},\left(\frac{t}{\|A\|_\infty}\right)^{\frac{1}{q}}\right)$$

and C is a constant.

7 Concentration for extremes

The CIs presented so far only concern with linear combinations of independent RVs or Lipschitz function of random vectors. In many statistics applications, we have to control the maximum of the n RVs when deriving the error bounds, while these RVs may be arbitrarily dependent. This section is developed on advanced proof skills. So we present the proofs even for existing results, which are applications of CIs in a probability aspect.

7.1 Maximal inequalities

This section presents the maximal inequalities for RVs $\{X_i\}_{i=1}^n$ which may not be independent. In the theory of empirical process, it is of interest to bound $\mathrm{E}\max_{1\le i\le n}|X_i|$ [82, Section 2.2]. If $\{X_i\}_{i=1}^n$ is arbitrary sequence of real-valued RVs and has finite r-th moments ($r\ge 1$), [3] gives a crude upper bounds for $\mathrm{E}\max_{1\le i\le n}X_i$ by Jensen's inequality

$$\begin{aligned}\mathrm{E}\Big\{\max_{1\le i\le n}|X_i|^r\Big\}^{\frac{1}{r}}&\le\Big\{\mathrm{E}\max_{1\le i\le n}|X_i|^r\Big\}^{\frac{1}{r}}\le\left\{\sum_{i=1}^n\mathrm{E}|X_i|^r\right\}^{\frac{1}{r}}\\&\le n^{\frac{1}{r}}\max_{1\le i\le n}\left(\mathrm{E}|X_i|^r\right)^{\frac{1}{r}}.\end{aligned}\tag{7.1}$$

[81, p. 314] mentions a sharper version of (7.1) without the proof. Below, we introduce the proof by the truncation technique.

Corollary 7.1 (Sharper maximal inequality)**.** *Let $\{X_i\}_{i=1}^n$ be identically distributed but not necessarily independent and assume $\mathrm{E}(|X_1|^p)<\infty,(p\ge 1)$. Then,*

$$\mathrm{E}\max_{1\le i\le n}|X_i|=o\left(n^{\frac{1}{p}}\right).$$

Proof. Let $M_n:=\max_{1\le i\le n}|X_i|$. For any $\epsilon>0$, we truncate M_n by $\epsilon n^{\frac{1}{p}}$,

$$\begin{aligned}\mathrm{E}M_n&=\int_0^{\epsilon n^{1/p}}P(M_n>t)dt+\int_{\epsilon n^{1/p}}^\infty P(M_n>t)dt\\&\le\int_0^{\epsilon n^{1/p}}1dt+\int_{\epsilon n^{1/p}}^\infty nP(|X_1|>t)dt\\&=\epsilon n^{\frac{1}{p}}+n^{\frac{1}{p}}\int_{\epsilon n^{1/p}}^\infty n^{\frac{(p-1)}{p}}P(|X_1|>t)dt\\&\le\epsilon n^{\frac{1}{p}}+\frac{n^{\frac{1}{p}}}{\epsilon^{p-1}}\int_{\epsilon n^{1/p}}^\infty t^{p-1}P(|X_1|>t)dt.\end{aligned}$$

Thus, dividing by $n^{\frac{1}{p}}$ we have

$$\frac{\mathrm{E}M_n}{n^{1/p}}\le\epsilon+\frac{1}{\epsilon^{p-1}}\int_{\epsilon n^{1/p}}^\infty t^{p-1}P(|X_1|>t)dt=\epsilon+o(1),$$

where we adopt the fact

$$\int_{\epsilon n^{1/p}}^\infty t^{p-1}P(|X_1|>t)dt=o(1)$$

from moment condition: $\mathrm{E}|X_1|^p < \infty$. Finally, it implies that

$$\limsup_{n\to\infty} \frac{\mathrm{E}M_n}{n^{1/p}} \le \epsilon,$$

which gives $\mathrm{E}M_n = o(n^{\frac{1}{p}})$ by letting $\epsilon \to 0$. □

Corollary 7.1 reveals that $\max_{1\le i\le n}|X_i|$ diverges at rate slower than $n^{\frac{1}{r}}$ under the r-th moment condition. If we have arbitrary finite r-th moment conditions (such as Gaussian distribution), it means that the divergence rate of maxima is slower than any polynomial rate $n^{\frac{1}{r}}$. This suggests that the rate may be logarithmic. With the sub-Gaussian assumptions, the logarithmic divergence rate is possible and the proof is based on controlling the expectation of the supremum of variables, from the argument in [66].

Corollary 7.2 (Sub-Gaussian maximal inequality, [70]). *Let $\{X_i\}_{i=1}^n$ be RVs (without independence assumption) such that $X_i \sim \mathrm{subG}(\sigma^2)$. Then*

(a) $\mathrm{E}\left[\max_{1\le i\le n} X_i\right] \le \sigma\sqrt{2\log n}$ *and* $\mathrm{E}\left[\max_{1\le i\le n} |X_i|\right] \le \sigma\sqrt{2\log(2n)}$.

(b) $P\left(\max_{1\le i\le n} X_i > t\right) \le ne^{-\frac{t^2}{2\sigma^2}}$ *and* $P\left(\max_{1\le i\le n} |X_i| > t\right) \le 2ne^{-\frac{t^2}{2\sigma^2}}$.

Proof. (a) By the property of maximum, sub-Gaussian MGF and Jensen's inequality,

$$\begin{aligned}
\mathrm{E}\max_{1\le i\le n} X_i &= \inf_{s>0} s^{-1}\mathrm{E}\log e^{s\max_{1\le i\le n} X_i} \le \inf_{s>0} s^{-1}\log \mathrm{E} e^{s\max_{1\le i\le n} X_i} \\
&\le \inf_{s>0} s^{-1}\log\sum_{i=1}^n \mathrm{E}e^{sX_i} \le \inf_{s>0} s^{-1}\log\sum_{i=1}^n e^{\frac{\sigma^2 s^2}{2}} \\
&= \inf_{s>0}\left(\frac{\log n}{s} + \frac{\sigma^2 s}{2}\right) = \sigma\sqrt{2\log n},
\end{aligned}$$

where we set

$$s = \sqrt{\frac{2\log n}{\sigma^2}}$$

as the optimal bound.

Let $Y_{2i-1} = X_i$ and $Y_{2i} = -X_i (1 \le i \le n)$. It gives

$$\mathrm{E}\max_{1\le i\le n}|X_i| = \mathrm{E}\max_{1\le i\le n}\max\{X_i, -X_i\} = \mathrm{E}\max_{1\le i\le 2n} Y_i.$$

The previous result for sample size $2n$ finishes the proof of the second part.

(b) By Chernoff inequality and the sub-Gaussian MGF, we have

$$P\Big(\max_{1\le i\le n} X_i > t\Big) \le \inf_{s>0} e^{-st}\mathrm{E}e^{s\max_{1\le i\le n} X_i} \le \inf_{s>0} e^{-st}\sum_{i=1}^{n}\mathrm{E}e^{sX_i}$$
$$\le \inf_{s>0} ne^{-st+\frac{\sigma^2 s^2}{2}} \overset{s=t/\sigma^2}{=\!=\!=} ne^{-\frac{t^2}{2\sigma^2}}.$$

For the second part, note that

$$P\Big(\max_{1\le i\le n}|X_i|>t\Big) = P\Big(\max_{1\le i\le n} X_i > t, \max_{1\le i\le n} -X_i \ge t\Big) \le 2P\Big(\max_{1\le i\le n} X_i > t\Big).$$

The proof is complete. □

By a similar proof, Corollary 7.2 can be extended to other RVs, such as sub-Gamma RVs and RVs characterized by sub-Weibull norm (or Orlicz norm) as presented before.

Corollary 7.3 (Concentration for maximum of sub-Gamma RVs). *Let $\{X_i\}_{i=1}^n$ be independent zero-mean $\{\mathrm{sub}\Gamma(v_i,c_i)\}_{i=1}^n$. Then, for $\max_{i=1,\dots,n} v_i =: v$ and $\max_{i=1,\dots,n} c_i =: c$,*

$$\mathrm{E}\Big(\max_{i=1,\cdots,n}|X_i|\Big) \le \big[2v\log(2n)\big]^{\frac{1}{2}} + c\log(2n).$$

See [33, Theorem 3.1.10] for the proof of Corollary 7.3.

Bellow, based on the sub-Weibull norm condition, a fundamental result due to [66] is given for obtaining the divergence rate of the maxima of sub-Weibull RVs.

Corollary 7.4 (Maximal inequality for sub-Weibull RVs). *For $\theta>0$, consider the sub-Weibull norm $\|X\|_{\psi_\theta} := \inf_{C\in(0,\infty)}\{\mathrm{E}e^{|X|^\theta/C^\theta}\le 2\}$ for $\psi_\theta(x)=e^{x^\theta}-1$. For any RVs $\{X_i\}_{i=1}^n$,*

$$\mathrm{E}\Big(\max_{1\le i\le n}|X_i|\Big) \le \psi_\theta^{-1}(n)\max_{1\le i\le n}\|X_i\|_{\psi_\theta} = \big(\log(1+n)\big)^{\frac{1}{\theta}}\max_{1\le i\le n}\|X_i\|_{\psi_\theta}. \tag{7.2}$$

If the function $\psi_\theta(x)$ is replaced by any non-decreasing convex function $g(x)$ with $g(0)=0$ in the definition of Orlicz norm: $\|X\|_g := \inf\{\eta>0 : \mathrm{E}[g(|X|/\eta)]\le 1\}$, then

$$\mathrm{E}\Big(\max_{1\le i\le n}|X_i|\Big) \le g^{-1}(n)\max_{1\le i\le n}\|X_i\|_g \quad \textit{for finite} \quad \max_{1\le i\le n}\|X_i\|_g.$$

Proof. From Jensen's inequality, for $C\in(0,\infty)$ and $\psi_\theta(x)=e^{x^\theta}-1$ we get

$$\psi_\theta\left[\mathrm{E}\Big(\max_{1\le i\le n}|X_i|/C\Big)\right]\le \mathrm{E}\Big[\max_{1\le i\le n}\psi_\theta(|X_i|/C)\Big]\le\sum_{i=1}^n \mathrm{E}\psi_\theta(|X_i|/C)\le n, \tag{7.3}$$

where the last inequality is by the definition of sub-Weibull norm $\mathrm{E}\psi_\theta(|X_i|/C)$ ≤ 1.

Let $C=\max_{1\le i\le n}\|X_i\|_{\psi_\theta}$. Applying the non-decreasing property of $\psi_\theta(x)$ (so does its inverse $\psi_\theta^{-1}(x)$), the (7.3) implies $\mathrm{E}(\max_{1\le i\le n}|X_i|/C)\le\psi_\theta^{-1}(n)$ by operating the map ψ_θ^{-1}, and so we have (7.2). The derivation of Orlicz norm case is the same. □

By Hoeffding's lemma, the following results on the maxima of the sum of independent RVs, is useful for bounding empirical processes.

Corollary 7.5 (Maximal inequality for bounded RVs, [18, Lemma 14.14])**.** *Let $\{X_i\}_{i=1}^n$ be independent RVs on $\mathcal{X}$ and $\{f_i\}_{i=1}^n$ be real-valued functions on $\mathcal{X}$ which satisfy $\mathrm{E}f_j(X_i)=0$, $|f_j(X_i)|\le a_{ij}$ for all $j=1,\dots,p$ and all $i=1,\dots,n$. Then*

$$\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^n f_j(X_i)\right|\right)\le\left[2\log(2p)\right]^{\frac12}\max_{1\le j\le p}\left(\sum_{i=1}^n a_{ij}^2\right)^{\frac12}.$$

Proof. Let $V_j=\sum_{i=1}^n f_j(X_i)$. By Jensen's inequality and Hoeffding's lemma

$$\begin{aligned}\mathrm{E}\max_{1\le j\le p}|V_j|&=\frac{1}{\lambda}\mathrm{E}\log e^{\lambda\max\limits_{1\le j\le p}|V_j|}\le\frac{1}{\lambda}\log\mathrm{E}e^{\lambda\max\limits_{1\le j\le p}|V_j|}\le\frac{1}{\lambda}\log\sum_{i=1}^p\mathrm{E}e^{\lambda|V_j|}\\&\le\frac{1}{\lambda}\log\left[\sum_{j=1}^p 2e^{\frac12\lambda^2\sum\limits_{i=1}^n a_{ij}^2}\right]\le\frac{1}{\lambda}\log\left[2pe^{\frac12\lambda^2\max\limits_{1\le j\le p}\sum\limits_{i=1}^n a_{ij}^2}\right]\\&=\frac{1}{\lambda}\log(2p)+\frac12\lambda\max_{1\le j\le p}\sum_{i=1}^n a_{ij}^2 \qquad :\text{[Corollary 2.1]}.\end{aligned}$$

Then

$$\begin{aligned}\mathrm{E}\max_{1\le j\le p}|V_j|&\le\inf_{\lambda>0}\left\{\frac{1}{\lambda}\log(2p)+\frac12\lambda\max_{1\le j\le p}\sum_{i=1}^n a_{ij}^2\right\}\\&=\sqrt{2\log(2p)}\max_{1\le j\le p}\left(\sum_{i=1}^n a_{ij}^2\right)^{\frac12}.\end{aligned}$$

The proof is complete. □

If Hoeffding's lemma for moment is replaced by Bernstein's moment conditions, then the maximal inequality for the sum of independent bounded RVs in Corollary 7.5 can be extended to Bernstein's moment conditions. We give a modified version of [18, Corollary 14.1] based on truncated Jensen's inequality.

Proposition 7.1 (Maximal inequality with Bernstein's moment conditions). *If $\{X_{ij}\}$, $j=1,\ldots,p$ are read-valued independent variables across $i=1,\ldots,n$. Assume $\mathrm{E}X_{ij}=0$ and Bernstein's moment conditions*

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left|X_{ij}\right|^{k}\le\frac{1}{2}v^{2}\kappa^{k-2}k!,\quad k=2,3,\ldots,\quad\forall j.$$

Then for any $1\le m\le 1+\log p$ and $p\ge 2$, one has

$$\mathrm{E}\left(\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}\right|^{m}\right)\le\left[\frac{\kappa\log(2p)}{n}+(v^{2}+1)\sqrt{\frac{\log(2p)}{n}}\right]^{m}.$$

Proof. Let

$$M_{n,m}=\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}\right|^{m}.$$

First, we show for any RV X and all $m\ge 1$,

$$\mathrm{E}|X|^{m}\le\log^{m}\left(\mathrm{E}e^{|X|}-1+e^{m-1}\right). \tag{7.4}$$

The function $g(x)=\log^{m}(x+1)$, $x\ge 0$ is concave for all $x\ge e^{m-1}-1$. By the truncated Jensen's inequality in Lemma 2.5 with $Z:=e^{|X|}-1$, $c=e^{m-1}-1$, we have

$$\begin{aligned}\mathrm{E}|X|^{m}&=\mathrm{E}\log^{m}\left(e^{|X|}-1+1\right)\le\log^{m}\left[\mathrm{E}\left(e^{|X|}-1\right)+1+\left(e^{m-1}-1\right)\right]\\&=\log^{m}\left[\mathrm{E}\left(e^{|X|}-1\right)+e^{m-1}\right].\end{aligned}$$

Then for all $L,m>0$,

$$\begin{aligned}\left(\frac{L}{n}\right)^{-m}\mathrm{E}M_{n,m}&\le\log^{m}\left[\mathrm{E}e^{\max\limits_{1\le j\le p}|\sum\limits_{i=1}^{n}X_{ij}/L|}-1+e^{m-1}\right]\\&\le\log^{m}\left[\sum_{j=1}^{p}\mathrm{E}\left(e^{|\sum\limits_{i=1}^{n}X_{ij}/L|}-1\right)+e^{m-1}\right].\end{aligned} \tag{7.5}$$

Therefore, it is sufficient to bound $\mathrm{E}e^{|\sum_{i=1}^{n}X_{ij}|/L}$ uniformly in j.

Second. To bound the MGF in (7.5), then we show that for any real-valued RV X,

$$\mathrm{E}e^{X} \leq e^{\mathrm{E}e^{|X|}-1-\mathrm{E}|X|} \quad \text{with} \quad \mathrm{E}X=0. \tag{7.6}$$

Indeed, for any $c>0$, we have

$$e^{X-c}-1 \leq \frac{e^{|X|}}{1+c}-1=\frac{e^{X}-1-X+X-c}{1+c} \leq \frac{e^{|X|}-1-|X|+X-c}{1+c}.$$

Let $c=\mathrm{E}e^{|X|}-1-\mathrm{E}|X|$. Note that $\mathrm{E}X=0$, so

$$\mathrm{E}e^{X-c}-1 \leq \frac{\mathrm{E}e^{|X|}-1-\mathrm{E}|X|-c}{1+c}=0.$$

Hence, $\log(\mathrm{E}e^{X}) \leq c$.

Using Taylor's expansion, the (7.6) and $e^{|x|} \leq e^{x}+e^{-x}$ give

$$\begin{aligned}
&\mathrm{E}e^{|\sum_{i=1}^{n} X_{ij}|/L}-1 \leq \mathrm{E}e^{\sum_{i=1}^{n} X_{ij}/L}+\mathrm{E}e^{-\sum_{i=1}^{n} X_{ij}/L}-1 \qquad (7.7)\\
&\leq 2e^{\sum_{i=1}^{n}\mathrm{E}(e^{|X_{ij}|/L}-1-|X_{ij}|/L)}-1=2e^{\sum_{m=2}^{\infty}\sum_{i=1}^{n}\mathrm{E}|X_{ij}|^{m}/L^{m}m!}-1\\
&\leq 2e^{nv^{2}\sum_{m=2}^{\infty}(\kappa/L)^{m-2}/2L^{2}}-1=2e^{nv^{2}/(2L^{2}(1-\kappa/L))}-1 \quad \text{[By moment conditions]}.
\end{aligned}$$

Combining (7.7) and (7.5), we obtain for $L>\kappa=:L-\sqrt{n/2\log(p+e^{m-1})}$

$$\begin{aligned}
\mathrm{E}M_{n,m} &\leq \left(\frac{L}{n}\right)^{m}\log^{m}\left[p\left(2e^{\frac{nv^{2}}{2L^{2}(1-\kappa/L)}}-1\right)+e^{m-1}\right]\\
&\leq \left(\frac{L}{n}\right)^{m}\log^{m}\left[\left(p+e^{m-1}\right)e^{\frac{nv^{2}}{2(L^{2}-L\kappa)}}\right]\\
&=\left(\frac{1}{n}L\log\left(p+e^{m-1}\right)+\frac{v^{2}L}{2(L^{2}-L\kappa)}\right)^{m}\\
&=\left(\frac{\kappa}{n}\log\left(p+e^{m-1}\right)+\sqrt{\frac{1}{2n}\log\left(p+e^{m-1}\right)}+\frac{v^{2}}{2(L-\kappa)}\right)^{m}\\
&\leq \left[\frac{\kappa}{n}\log\left(p+e^{m-1}\right)+(v^{2}+1)\sqrt{\frac{1}{2n}\log\left(p+e^{m-1}\right)}\right]^{m}\\
&\leq \left[\frac{\kappa}{n}\log(2p)+(v^{2}+1)\sqrt{\frac{1}{2n}\log(2p)}\right]^{m},
\end{aligned}$$

where the second and last inequality is by

$$\frac{v^2}{L-\kappa} = \left[\sqrt{\frac{n}{2\log(p+e^{m-1})}}\right]^{-1} v^2$$

and $e^{m-1} \le p$. □

7.2 Concentration for suprema of empirical processes

Let $\{X_i\}_{i=1}^n$ be a random sample from a measure $\mathbb{P}$ on a measurable space $(\mathcal{X},\mathcal{A})$. The empirical distribution $\mathbb{P}_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability mass of 1 at x. Given a measurable function $f:\mathcal{X}\mapsto\mathbb{R}$, let $\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^n f(X_i)$ be the expectation of f under the empirical measure $\mathbb{P}_n$, and $Pf := \int f dP$ be the expectation under $\mathbb{P}$. The $\mathbb{P}_n f$ is called the empirical process index by n.

The study of the empirical processes begins with the uniform limit law of EDF in Example 2.1. The Glivenko-Cantelli theorem extends the LLN for EDF and gives uniform convergence: $\|\mathbb{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \stackrel{\text{as}}{\to} 0$. Moreover, a stronger result than Example 2.1 is the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [25]

$$P\big(\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| > \varepsilon\big) \le 2e^{-2n\varepsilon^2}, \quad \forall \varepsilon > 0. \tag{7.8}$$

The DKW inequality is a uniform version of Hoeffding's inequality, which also strengthens the Glivenko-Cantelli theorem since (7.8) implies Glivenko-Cantelli $\|\mathbb{F}_n - F\|_\infty \xrightarrow{a.s.} 0$ by Borel-Cantelli lemma

$$X_n \xrightarrow{a.s.} 0 \;\Leftrightarrow\; \sum_{n=1}^{\infty} P\big(|X_n| \ge \epsilon\big) < \infty \quad \text{for any} \quad \varepsilon > 0.$$

[25] proves $P(\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| > \varepsilon) \le Ce^{-2n\varepsilon^2}$ with an unspecified constant C. [59] attains the sharper constant $C = 2$. In some statistical applications, given an estimator $\hat{\theta}$, and $f_{\hat{\theta}}(X_i)$ is a function of X_i and $\hat{\theta}$. We want to study its asymptotic properties for sums of $f_{\hat{\theta}}(X_i)$ that changes with both n and $\hat{\theta}$,

$$\frac{1}{n}\sum_{i=1}^n \left[f_{\hat{\theta}}(X_i) - \mathrm{E} f_{\hat{\theta}}(X_i)\right] \qquad \text{(a dependent sum)}.$$

A possible route to attain results is via the suprema of the empirical process for all possible the "true" parameter θ_0 on a set K

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[f_{\hat{\theta}}(X_i)-\mathrm{E}f_{\hat{\theta}}(X_i)\right] &\le \sup_{\theta_0\in K}\left|\frac{1}{n}\sum_{i=1}^{n}\left[f_{\theta_0}(X_i)-\mathrm{E}f_{\theta_0}(X_i)\right]\right| \\ &=: \sup_{\theta_0\in K}\left|(\mathbb{P}_n-P)f_{\theta_0}\right|.\end{aligned}$$

Fortunately, the summation in the sup enjoys independence. So, the study of convergence rate suprema of empirical processes is important if we consider a functional class $\mathcal{F}$ instead of the set K such that

$$\sup_{f\in\mathcal{F}}\left|(\mathbb{P}_n-P)f\right| = \sup_{\theta_0\in K}\left|(\mathbb{P}_n-P)f_{\theta_0}\right|.$$

Let $(\mathcal{F},\|\cdot\|)$ be a normed space of real functions $f:\mathcal{X}\to\mathbb{R}$. For a probability measure Q, define the $L_r(Q)$-space with $L_r(Q)$-norm by

$$\|f\|_{L_r(Q)} = \left(\int |f|^r dQ\right)^{\frac{1}{r}}.$$

Given two functions $l(\cdot)$ and $u(\cdot)$, the bracket $[l,u]$ is the set of all functions $f\in\mathcal{F}$ with $l(x)\le f(x)\le u(x)$, for all $x\in\mathcal{X}$. An ε-bracket is a bracket $[l,u]$ with $\|l-u\|_{L_r(Q)}<\varepsilon$. The bracketing number $N_{[\,]}(\varepsilon,\mathcal{F},L_r(Q))$ is minimum number of ε-brackets needed to cover $\mathcal{F}$, i.e.

$$\begin{aligned}N_{[\,]}\big(\varepsilon,\mathcal{F},L_r(Q)\big) = \inf\Big\{n: \exists l_1,u_1,\ldots,l_n,u_n \text{ s.t. } \cup_{i=1}^{n}[l_i,u_i] \\ = \mathcal{F} \text{ and } \|l_n-u_n\|_{L_r(Q)}<\varepsilon\Big\}.\end{aligned}$$

The covering number $N(\varepsilon,\mathcal{F},L_r(Q))$ is the minimal number of $L_r(Q)$-balls of radius ε needed to cover the set $\mathcal{F}$. The uniform covering numbers is

$$\sup_{Q} N\big(\varepsilon\|F\|_{L_r(Q)},\mathcal{F},L_r(Q)\big),$$

where the supremum is taken over all probability measures Q for which $\|F\|_{L_r(Q)} > 0$. Two conditions to get the convergence of $\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-P)f_\theta|$ are the finite bracketing number condition with $L_1(P)$-norm in [81, Theorem 19.4] (or finite uniform covering numbers in [81, Theorem 19.13]).

Lemma 7.1 (Glivenko-Cantell class). *For every class $\mathcal{F}$ of measurable functions, if $N_{[\,]}(\varepsilon,\mathcal{F},L_1(P))<\infty$ (or $\sup_Q N(\varepsilon\|F\|_{L_1(Q)},\mathcal{F},L_1(Q))<\infty$ with $P^*F<\infty$) for every $\varepsilon>0$, then $\mathcal{F}$ is P-Glivenko-Cantelli, i.e. $\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-P)f|\stackrel{as}{\to}0$.*

Example 7.1 (Empirical process with indicator functions)**.** Let $\mathcal{F}$ be the collection of all indicator functions of the form $f_t(x) = 1_{(-\infty,t]}(x)$ with t ranging over $\mathbb{R}$. Then $\mathcal{F}$ is P-Glivenko-Cantelli, see [81, Example 19.4].

Example 7.2 (Weighted empirical process with dependent weights)**.** Suppose we observe a sequence of IID observations $\{(X_i,Y_i)\}_{i=1}^n$ drawn from a random pair (X,Y). Given some weighted functions $W(\cdot)$ and a bounded estimator $\hat{t}\in(0,\tau]$, we want to study the stochastic convergence of dependent weighted empirical processes

$$\left|\frac{1}{n}\sum_{i=1}^n 1\left(Y_i\geq\hat{t}\right)W(X_i)-\mu\left(\hat{t};W\right)\right|\left(\leq\sup_{0\leq t\leq\tau}\left|\frac{1}{n}\sum_{i=1}^n 1(Y_i\geq t)W(X_i)-\mu(t;W)\right|\right),$$

where

$$\mu(t;W)=\mathrm{E}_{X,Y}\{1(Y\geq t)W(X_i)\}<\infty,\quad W(X_i)\leq U_f<\infty$$

and $\tau<\infty$.

Consider the class of functions indexed by t,

$$\mathcal{F}=\{1(y\geq t)W(x)/U_f : t\in[0,\tau], y\in\mathbb{R}, W(x)\leq U_f\}.$$

It is crucial to evaluate $N_{[\,]}(\varepsilon,\mathcal{F},L_1(Q))$. Given $\epsilon\in(0,1)$, let t_i be the i-th $\lceil\frac{1}{\varepsilon}\rceil$ quantile of Y, thus

$$P(Y\leq t_i)=i\varepsilon,\quad i=1,\ldots,\left\lceil\frac{1}{\varepsilon}\right\rceil-1.$$

Furthermore, take $t_0=0$ and $t_{\lceil\frac{1}{\varepsilon}\rceil}=+\infty$. For $i=1,\ldots,\lceil\frac{1}{\varepsilon}\rceil$, define brackets $[L_i,U_i]$ with

$$L_i(x,y)=1(y\geq t_i)\frac{W(x)}{U_f},\quad U_i(x,y)=1(y>t_{i-1})\frac{W(x)}{U_f}$$

such that $L_i(x,y)\leq 1(y\geq t)e^{f_\theta(x)}/U_f\leq U_i(x,y)$ as $t_{i-1}<t\leq t_i$. The Jensen's inequality gives

$$\begin{aligned}\mathrm{E}|U_i-L_i|&\leq\left|E\left[\frac{W(X_i)}{U_f}\{1(Y\geq t_i)-1(Y>t_{i-1})\}\right]\right|\\&\leq\left|P(t_{i-1}<Y\leq t_i)\right|=\varepsilon.\end{aligned}$$

Therefore, $N_{[\,]}(\varepsilon,\mathcal{F},L_1(P))\leq\lceil\frac{1}{\varepsilon}\rceil<\infty$ for every $\varepsilon>0$. So the class $\mathcal{F}$ is P-Glivenko-Cantelli.

If the upper bounds of $N_{[\,]}(\varepsilon,\mathcal{F},L_2(P))$ and $\sup_Q N(\varepsilon,\mathcal{F},L_2(Q))$ have polynomial rates with respect to $\mathcal{O}(\frac{1}{\varepsilon})$, the following tail bound estimate gives the convergence rate of suprema of empirical processes in Lemma 7.1 obtained by [77]. It extends DKW inequality to general empirical processes with the bounded function classes.

Lemma 7.2 (Sharper bounds for suprema of empirical processes)**.** *Consider a probability space (Ω,Σ,P), and consider n IID RVs $\{X_i\}_{i=1}^n$ valued in Ω, of law P. Let $\mathcal{F}$ be a class of measurable functions $f:\mathcal{X}\mapsto[0,1]$ that satisfies bracketing number conditions $N_{[\,]}(\varepsilon,\mathcal{F},L_2(P))\le(\frac{K}{\varepsilon})^V$ (or $\sup_Q N(\varepsilon,\mathcal{F},L_2(Q))\le(\frac{K}{\varepsilon})^V$) for every $0<\varepsilon<K$. Then for every $t>0$*

$$P\left(\sqrt{n}\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-\mathbb{P})f|\ge t\right)\le\left(\frac{D(K)t}{\sqrt{V}}\right)^V e^{-2t^2}$$

with a constant $D(K)$ depending on K only.

The explicit constant $D(K)$ can be found in [97], who studied the tail bounds for the suprema of the unbounded and non-IID empirical process. [48] derived the rate of convergence for the Lasso regularized Cox models by using sharper concentration inequality for the suprema of empirical processes in Example 7.2 related to the negative log-partial likelihood function. In Example 7.2, we have

$$\begin{aligned}\left\{\mathrm{E}|U_i-L_i|^2\right\}^{\frac{1}{2}}&\le\left\{E\left[\frac{W(X_i)}{U_m}\{1(Y\ge t_i)-1(Y>t_{i-1})\}\right]^2\right\}^{\frac{1}{2}}\\&\le\left|P(t_{i-1}<Y\le t_i)\right|^{\frac{1}{2}}=\sqrt{\varepsilon},\end{aligned}$$

which implies $N_{[\,]}(\sqrt{\varepsilon},\mathcal{F},L_2(P))\le\lceil\frac{1}{\varepsilon}\rceil\le\frac{2}{\varepsilon}$ for every $\varepsilon>0$. Hence, $N_{[\,]}(\varepsilon,\mathcal{F},L_2(P))\le\frac{2}{\varepsilon^2}$. By applying Lemma 7.2 with $V=2$, $K=\sqrt{2}$, we have

$$P\left(\sup_{0\le t\le\tau}\left|\frac{1}{U_f\sqrt{n}}\sum_{i=1}^n\left[1(Y_i\ge t)W(X_i)-\mu(t;W)\right]\right|\ge t\right)\le\frac{D^2(\sqrt{2})}{2}t^2e^{-2t^2}.$$

The next two results are the symmetrization theorem and the contraction theorem, which are fundamental tools to get sharper bounds for suprema of empirical processes.

Lemma 7.3 (Symmetrization theorem)**.** *Let $\{\mathbf{X}_i\}_{i=1}^n$ be independent RVs with values in some space $\mathcal{X}$ and $\mathcal{F}$ be a class of measurable real-valued functions on $\mathcal{X}$. Let $\{\epsilon_i\}_{i=1}^n$*

be a Rademacher sequence with uniform distribution on $\{-1,1\}$*, independent of* $\{\boldsymbol{X}_i\}_{i=1}^n$ *and* $f\in\mathcal{F}$*. If* $\mathrm{E}|f(\boldsymbol{X}_i)|<\infty\ \forall\ i$*, then*

$$\mathrm{E}\left\{\sup\nolimits_{f\in\mathcal{F}}\Phi\left(\sum_{i=1}^n\left[f(\boldsymbol{X}_i)-\mathrm{E}f(\boldsymbol{X}_i)\right]\right)\right\}\leq\mathrm{E}\left\{\sup\nolimits_{f\in\mathcal{F}}\Phi\left[2\sum_{i=1}^n\epsilon_i f(\boldsymbol{X}_i)\right]\right\}$$

for every nondecreasing, convex $\Phi(\cdot):\mathbb{R}\mapsto\mathbb{R}$ *and class of measurable functions* $\mathcal{F}$*.*

Lemma 7.4 (Contraction theorem)**.** *Let* $x_1,\ldots,x_n$ *be the non-random elements of* $\mathcal{X}$ *and* $\varepsilon_1,\ldots,\varepsilon_n$ *be Rademacher sequence. Consider c-Lipschitz functions* g_i*, i.e.* $|g_i(s)-g_i(t)|\leq c|s-t|,\forall s,t\in R$*. Then for any function* f *and* h *in* $\mathcal{F}$*, we have*

$$\begin{aligned}&\mathrm{E}_\epsilon\left[\sup\nolimits_{f\in\mathcal{F}}\left|\sum_{i=1}^n\varepsilon_i\left[g_i\{f(x_i)\}-g_i\{h(x_i)\}\right]\right|\right]\\&\leq 2c\mathrm{E}_\epsilon\left[\sup\nolimits_{f\in\mathcal{F}}\left|\sum_{i=1}^n\varepsilon_i\{f(x_i)-h(x_i)\}\right|\right].\end{aligned}$$

A gentle introduction to suprema of empirical processes and its statistical applications are nicely presented in [75]. To further bound $\mathrm{E}\{\sup_{f\in\mathcal{F}}\sqrt{n}|(\mathbb{P}_n-\mathbb{P})f|\}$ in Lemma 7.3 with $\Phi(t)=|t|$, [33, Theorem 3.5.4] gave a constants-specified upper bound for the expectation of suprema of unbounded empirical processes.

Lemma 7.5 (Moment bound for suprema of unbounded empirical processes)**.** *Let* $\mathcal{F}$ *be a countable class of measurable functions with* $0\in\mathcal{F}$*, and let* F *be a strictly positive envelope for* $\mathcal{F}$*. Assume that*

$$J(\mathcal{F},F,t):=\int_0^t\sup_Q\sqrt{\log\left[2N(\mathcal{F},L_2(Q),\tau\|F\|_{L_2(Q)})\right]}d\tau<\infty$$

for some (for all) $t>0$*. Given* $\mathcal{X}$*-valued* IID *RVs* $\{X_i\}_{i=1}^n$ *with law* P *s.t.* $PF^2<\infty$*. Set* $U=\max_{1\leq i\leq n}F(X_i)$ *and* $\delta=\sup_{f\in\mathcal{F}}\sqrt{Pf^2}/\|F\|_{L^2(P)}$*. Then, for* $A_1=8\sqrt{6}$ *and* $A_2=2^{15}3^{\frac{5}{2}}$*,*

$$\begin{aligned}&\mathrm{E}\left\{\sup\nolimits_{f\in\mathcal{F}}\sqrt{n}\left|(\mathbb{P}_n-P)f\right|\right\}\\&\leq A_1\|F\|_{L^2(P)}J(\mathcal{F},F,\delta)\vee\left[A_2\|U\|_{L^2(P)}J^2(\mathcal{F},F,\delta)/(\sqrt{n}\delta^2)\right].\end{aligned}\tag{7.9}$$

The bound in (7.9) matches the non-asymptotically sub-exponential CLT in (4.6), and it reveals that $\sup_{f\in\mathcal{F}}\sqrt{n}|(\mathbb{P}_n-P)f|$ has the sub-exponential behaviour, although with a huge parameter (the constant $A_2=2^{15}3^{\frac{5}{2}}$ is so large). Recently, [6, Theorem 2] sharpened bound (7.9) when $\mathcal{F}$ takes values in $[-1,1]$. Applying [1] tail inequalities for $Z_n:=\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-P)f|$ with unbounded $\mathcal{F}$, they obtained following result.

Lemma 7.6 (Tail estimates for suprema of empirical processes under sub-Weibull norms)**.** *Let $\{X_i\}_{i=1}^n$ be independent $\mathcal{X}$-valued RVs and let $\mathcal{F}$ be a countable class of measurable functions $f:\mathcal{X}\to\mathbb{R}$. For some $\alpha\in(0,1]$, assume $\|\sup_{f\in\mathcal{F}}|f(X_i)-\mathrm{E}f(X_i)|\|_{\psi_\alpha}<\infty$ for every i. Define $\sigma_n^2=\sup_{f\in\mathcal{F}}\sum_{i=1}^n\mathrm{Var}f(X_i)$. For all $\eta\in(0,1)$ and $\delta>0$, then there exists a constant $C_{\alpha,\eta,\delta}>0$ s.t. both $P(Z_n\ge(1+\eta)\mathrm{E}Z_n+t)$ and $P(Z_n\le(1-\eta)\mathrm{E}Z_n-t)$ are bounded by*

$$\delta_{n,t,\eta,\delta}(\sigma_n^2,\alpha):=\exp\left(-\frac{t^2}{2(1+\delta)\sigma_n^2}\right)+3\exp\left(-\left(\frac{t}{C_{\alpha,\eta,\delta}\|\max_i\sup_{f\in\mathcal{F}}|f(X_i)|\|_{\psi_\alpha}}\right)^{\alpha}\right)$$

for all $t\ge0$.

So, Lemmas 7.5 and 7.6 give

$$\begin{aligned}&P\Big(n^{-1}Z_n\ge(1+\eta)n^{-\frac{1}{2}}[\text{Right hand side of (7.9)}]+t\Big)\\ \le& P\Big(Z_n\ge(1+\eta)n^{\frac{1}{2}}\mathrm{E}[n^{-\frac{1}{2}}Z_n]+nt\Big)\le\delta_{n,nt,\eta,\delta}(\sigma_n^2,\alpha).\end{aligned}$$

We have

$$\begin{aligned}&P\Big(\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-P)f|\le(1+\eta)n^{-\frac{1}{2}}[\text{Right hand side of (7.9)}]+t\Big)\\ \ge& 1-\delta_{n,nt,\eta,\delta}(\sigma_n^2,\alpha).\end{aligned}$$

The constant-unspecific version of Lemma 7.5 ([81, Lemma 19.36-19.38] or other versions) has wide applications in deriving the rate of convergence for kernel density estimations, M-estimators in high-dimensional and increasingly-dimensional regressions, see [18, 33, 64] and references therein.

8 Concentration for high-dimensional statistics

With the emergence of high-dimensional (HD) data such as the gene expression data, there are renewed interests on the CIs. One aspect of the HD data is such that the number of variables p can be comparable to or even greater than the sample size n. This section provides results in three commonly encountered settings: increasing-dimensional ($p_n=o(n)<n$), large-dimensional ($p_n=\mathcal{O}(n)$) and high-dimensional setting ($p_n\gg n$, $p_n=e^{o(n)}$).

8.1 Linear models with diverging number of covariates

Suppose that we have an n-dimensional random vector $\boldsymbol{Y}$ which contains n responses $\{Y_i\}_{i=1}^n$ to p covariates $\boldsymbol{X}_i=(x_{i1},\cdots,x_{ip})^T$, respectively. The n copies of $\boldsymbol{X}_i$ as row vectors make a $n\times p$ design matrix $\mathbf{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)^T$. The conditional expectation $\mathrm{E}[Y_i|\boldsymbol{X}_i]$ is linearly related to a coefficient vector $\boldsymbol{\beta}^*=(\beta_1^*,\cdots,\beta_p^*)^T$ such that

$$\boldsymbol{Y}=\mathbf{X}\boldsymbol{\beta}^*+\boldsymbol{\varepsilon}, \tag{8.1}$$

where $\{\varepsilon_i\}_{i=1}^n$ in the error vector $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)^T$ are IID with zero mean and finite variance σ^2. The $\boldsymbol{\beta}^*$ needs to be estimated.

This subsection only considers the case that p is increasing but $p<n$. The ordinary least square (OLS) estimator is

$$\hat{\boldsymbol{\beta}}_{LS}=\mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{8.2}$$

Assume $\mathrm{rank}(\mathbf{X})=p$, which is not hard to meet since $p<n$, $\hat{\boldsymbol{\beta}}_{LS}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ is the unique solution of the (8.2). The following result for the OLS estimator is well known.

Lemma 8.1. *Under the assumptions on the linear models and the rank of $\boldsymbol{X}$ is p, then*

(i) Let $\mathbf{A}$ be a $p\times n$ matrix, then $\mathrm{E}\|\mathbf{A}\boldsymbol{\varepsilon}\|_2^2=\mathrm{E}(\boldsymbol{\varepsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\varepsilon})=\sigma^2\mathrm{tr}(\mathbf{A}^T\mathbf{A})$.

(ii) (The curse of dimensionality.) *The mean square error and the average in-sample ℓ_2 risk of the OLS estimator are*

$$\mathrm{E}\|\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}^*\|_2^2=\mathrm{tr}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)\sigma^2,$$
$$\frac{1}{n}\mathrm{E}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}^*)\|_2^2=\frac{p\sigma^2}{n}.$$

Remark 8.1. As $p,n\to\infty$ with $p<n$, part (ii) implies that the OLS estimator may had poor performance if $\frac{p}{n}\to c>0$. The average in-sample ℓ_2-risk tends to zero if $p_n=o(n)$.

Put $\hat{\boldsymbol{\beta}}:=\hat{\boldsymbol{\beta}}_{LS}$. Let $\{\lambda_i(\mathbf{X}^T\mathbf{X})\}_{i=1}^k$ be the eigenvalue values of $\mathbf{X}^T\mathbf{X}$. Markov's inequality and Lemma 8.1 with $\mathbf{A}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ implies

$$\begin{aligned}P\left\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2>t\right\}&\le\frac{\sigma^2}{t^2}\mathrm{tr}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]=\frac{\sigma^2}{t^2}\sum_{i=1}^p\frac{1}{\lambda_i(\mathbf{X}^T\mathbf{X})}\\&\le\frac{\sigma^2}{t^2}\frac{p}{\lambda_{\min}(\mathbf{X}^T\mathbf{X})}=:\delta_n,\end{aligned}$$

which implies that, with probability greater than $1-\delta_n$,

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 \le \sigma\sqrt{\frac{p}{n}}\cdot\left[\delta_n\lambda_{\min}\left(\frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)\right]^{-\frac{1}{2}}. \tag{8.3}$$

Assume that $p:=p_n=o(n^r)$ as $n\to\infty$, $p_n<n$. We specific two groups of regularity conditions and the value of r such that the l_2 consistency ($\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 \xrightarrow{p} 0$) is true.

(1) By Lemma 8.1, if $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is uniformly positive definite ($\exists\, c>0$ s.t. $\frac{1}{n}\mathbf{X}^T\mathbf{X}\succ c\mathbf{I}_p$) then

$$\frac{p\sigma^2}{n}=\frac{1}{n}\mathrm{E}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2=\mathrm{E}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\frac{1}{n}\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]\ge c\mathrm{E}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2.$$

If $p=o(n)$, then $\mathrm{E}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2=o(1)$ which implies $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2=o_p(1)$.

(2) From (8.3), if $p=o(\lambda_{\min}(\mathbf{X}^T\mathbf{X}))$, we have $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2=o_p(1)$. In this case, if we consider: "$\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is positive definite" in (1), it also leads to $p=o(\lambda_{\min}(\mathbf{X}^T\mathbf{X}))=o(n)$.

In (8.1) with fixed design, suppose that the $\varepsilon_1,\dots,\varepsilon_n$ are sub-Gaussian zero-mean noise for which there exists a $\sigma>0$ such that

$$\mathrm{E}e^{\sum_{i=1}^n\alpha_i\varepsilon_i}\le e^{\sigma^2\sum_{i=1}^n\alpha_i^2},\quad \forall\alpha_1,\dots,\alpha_n\in\mathbb{R}.$$

Suppose that the Gram matrix $\boldsymbol{S}_n:=\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is invertible. The excess in-sample prediction error $R(\hat{\boldsymbol{\beta}})$ is the difference between the expected squared error for $\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$ and for $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$

$$\begin{aligned}
R(\hat{\boldsymbol{\beta}}):&=\frac{1}{n}\left\{\mathrm{E}\left[\sum_{i=1}^n\left(\boldsymbol{X}_i^T\boldsymbol{\beta}-Y_i\right)^2\right]-\mathrm{E}\left[\sum_{i=1}^n\left(X_i^T\boldsymbol{\beta}^*-Y_i\right)^2\right]\right\}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\
&=\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2+\frac{1}{n}\mathrm{E}\left[\sum_{i=1}^n\left(\boldsymbol{X}_i^T\boldsymbol{\beta}-\boldsymbol{X}_i^T\boldsymbol{\beta}^*\right)\cdot\varepsilon_i\right]\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\
&=\frac{1}{n}\left\|\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2^2,
\end{aligned}\tag{8.4}$$

which is a quadratic form of sub-Gaussian vector.

By Corollary 4.7 with $\mathbf{A}:=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T/\sqrt{n}$, $\boldsymbol{\xi}:=\boldsymbol{\varepsilon}$, $\boldsymbol{\mu}=\mathbf{0}$ and $\boldsymbol{\Sigma}:=\mathbf{A}^T\mathbf{A}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T/n$,

$$\mathrm{tr}(\boldsymbol{\Sigma})=\frac{1}{n}\mathrm{tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\right)=\frac{p}{n},\quad \mathrm{tr}(\boldsymbol{\Sigma}^2)=\frac{p}{n^2},\quad \|\boldsymbol{\Sigma}\|_2=\frac{1}{n},$$

where last identity is due to $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ being a projection matrix. Thus

$$P\left[R(\hat{\boldsymbol{\beta}})>\frac{\sigma^2}{n}\left(p+2\sqrt{pt}+2t\right)\right]\le e^{-t},$$

i.e. with probability $1-e^{-t}$

$$R(\hat{\boldsymbol{\beta}})\le\frac{\sigma^2}{n}\left(p+2\sqrt{pt}+2t\right).$$

For Gaussian noise, $\mathrm{E}R(\hat{\boldsymbol{\beta}})=\frac{\sigma^2 p}{n}$ in Lemma 8.1, so

$$P\left\{R(\hat{\boldsymbol{\beta}})-\mathrm{E}R(\hat{\boldsymbol{\beta}})\le\frac{\sigma^2}{n}(2\sqrt{pt}+2t)\right\}\ge 1-e^{-t}.$$

8.2 Non-asymptotic Bai-Yin theorem for random matrix

Let $\mathbf{A}$ be a $p\times p$ Hermitian matrix with real eigenvalues

$$\lambda_{\max}:=\lambda_1\ge\cdots\ge\lambda_p=:\lambda_{\min}.$$

The empirical spectral distribution (ESD) of $\mathbf{A}$ is

$$F_{\mathbf{A}}(x)=\frac{1}{p}\sum_{j=1}^{p}1(\lambda_j\le x),$$

which resembles the EDF of IID samples. Let $\{\mathbf{A}_n\}_{n\ge1}$ be a sequence of $p\times p$ Hermitian random matrices indexed by the sample size n, and $F_{\mathbf{A}_n}$ be the ESD of $\mathbf{A}_n$.

A major interest in random matrix theory is to investigate the convergence of $F_{\mathbf{A}_n}$ as a sequence of distributions to a limit F. In multivariate statistics, it is of interest to study the sample covariance matrix $\mathbf{S}_n:=\frac{1}{n}\mathbf{X}\mathbf{X}^T$ where the double array $\mathbf{X}=\{X_{ij},i=1,\ldots,p;j=1,\ldots,n\}$ contains zero-mean IID RVs $\{X_{ij}\}$ with variance σ^2. Suppose that the dimensions n and p grow to infinity while $\frac{p}{n}$ converges to a constant in $[0,1]$. [58] gives the limit behavior of the ESD of $\mathbf{S}_n$. [5] obtained a strong version of the Marčenko-Pastur law.

Corollary 8.1 (Bai-Yin theorem)**.** *Let $\mathbf{X}$ be an $n\times p$ random matrix whose entries are independent copies of a RV with zero mean, unit variance, and finite fourth moment (*$\mathrm{E}|X_{11}|^4<\infty$*). As $n\to\infty,p\to\infty,\frac{p}{n}\to y\in(0,1)$, then*

$$\lim_{n\to\infty}\lambda_{\min}(\mathbf{S}_n)=\sigma^2(1-\sqrt{y})^2,\quad \lim_{n\to\infty}\lambda_{\max}(\mathbf{S}_n)=\sigma^2(1+\sqrt{y})^2.$$

Note that $\lambda_i(\mathbf{S}_n)=\lambda_i(\frac{\mathbf{X}}{\sqrt{n}})$ for all i, Bai-Yin's law asserts that if $\sigma^2=1$

$$\lambda_{\min}\left(\frac{\mathbf{X}}{\sqrt{n}}\right)=1-\sqrt{\frac{p}{n}}+o\left(\sqrt{\frac{p}{n}}\right),\quad \lambda_{\max}\left(\frac{\mathbf{X}}{\sqrt{n}}\right)=1+\sqrt{\frac{p}{n}}+o\left(\sqrt{\frac{p}{n}}\right)\quad \text{a.s..}$$

[84, Theorem 4.6.1] studies the non-asymptotic upper and lower bounds of the extreme eigenvalues of $\mathbf{S}_n$ with independent sub-exponential entries, but the bounds contained un-specific constants. We give a constant-specified version.

Proposition 8.1 (Constants-specified non-asymptotic Bai-Yin theorem)**.** *Let* $\mathbf{X}$ *be an* $n\times p$ *matrix whose rows* $\boldsymbol{X}_i$ *are independent sub-Gaussian random vectors in* $\mathbb{R}^p$ *with* $\mathrm{Var}(\boldsymbol{X}_i)=\mathbf{I}_p$. *Define* $Z_i:=|\langle \boldsymbol{X}_i,\boldsymbol{x}\rangle|$, $\forall\, \boldsymbol{x}\in S^{n-1}$. *Further assume that* $\{Z_i^2-1\}_{i=1}^n$ *are* $\mathrm{subE}(\theta)$, *then*

$$P\left\{\left\|n^{-1}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|\le 2c\theta\max\left(\delta,\delta^2\right)\right\}\ge 1-2e^{-ct^2},\quad t\ge 0, \tag{8.5}$$

where $\delta=2c(\sqrt{\frac{p}{n}}+\frac{t}{\sqrt{n}})$ *with* $t=c\theta\max(\delta,\delta^2)$ *and* $c\ge\frac{2n\log 9}{p}$. *Moreover,*

$$P\left\{1-t^2\le\lambda_{\min}(\mathbf{S}_n)\le\lambda_{\max}(\mathbf{S}_n)\le 1+t^2\right\}\ge 1-2e^{-ct^2}. \tag{8.6}$$

Proposition 8.1 does not require $\frac{p}{n}\to y\in(0,1)$ as in Corollary 8.1.

Proof. Step 1. We introduce a counting measure for measuring the complexity of a set in some space. The covering number $\mathcal{N}(K,\varepsilon)$ is the smallest number of closed balls centered at K with radii ε whose union covers K. For some $\varepsilon\in[0,1)$, a subset $\mathcal{N}_\varepsilon\subset\mathbb{R}$ is an ε -net for S^{n-1} if for all $\boldsymbol{x}\in S^{n-1}$, there is an $\boldsymbol{y}\in\mathcal{N}_\varepsilon$, such that $\|\boldsymbol{x}-\boldsymbol{y}\|<\varepsilon$. We use the following results in [83, Lemmas 5.2 and 5.4].

Lemma 8.2 (Covering numbers of the sphere)**.** $\mathcal{N}(S^{n-1},\varepsilon)\le(1+\frac{2}{\varepsilon})^n$ *for every* $\varepsilon>0$.

Lemma 8.3 (Computing the spectral norm on a net)**.** *Let* $\mathbf{B}$ *be an* $p\times p$ *matrix. Then*

$$\left\|\mathbf{B}\right\|:=\max_{\|\boldsymbol{x}\|_2=1}\left\|\mathbf{B}\boldsymbol{x}\right\|_2=\sup_{\boldsymbol{x}\in S^{p-1}}|\langle\mathbf{B}\boldsymbol{x},\boldsymbol{x}\rangle|\le(1-2\varepsilon)^{-1}\sup_{\boldsymbol{x}\in\mathcal{N}_\varepsilon}|\langle\mathbf{B}\boldsymbol{x},\boldsymbol{x}\rangle|.$$

Lemma 8.3 shows that

$$\left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|\le 2\max_{\boldsymbol{x}\in\mathcal{N}_{1/4}}\left|\frac{1}{n}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|.$$

Indeed, note that

$$\left\langle \frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{x}-\boldsymbol{x},\boldsymbol{x}\right\rangle=\left\langle \frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{x},\boldsymbol{x}\right\rangle-1=\frac{1}{n}\|\mathbf{X}\boldsymbol{x}\|_2^2-1.$$

By setting $\varepsilon=\frac{1}{4}$ in Lemma 8.3, we get

$$\begin{aligned}\left\|n^{-1}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\| &\le (1-2\varepsilon)^{-1}\sup_{\boldsymbol{x}\in\mathcal{N}_\varepsilon}\left|\left\langle n^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{x}-\boldsymbol{x},\boldsymbol{x}\right\rangle\right| \\ &=2\sup_{x\in\mathcal{N}_{\frac{1}{4}}}\left|n^{-1}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|. \end{aligned} \tag{8.7}$$

By (8.7), we have

$$\begin{aligned} P\left\{\left\|n^{-1}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|\ge 2t\right\} &\le P\left\{2\sup_{x\in\mathcal{N}_{\frac{1}{4}}}\left|n^{-1}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|\ge 2t\right\} \\ &\le \sum_{x\in\mathcal{N}_{\frac{1}{4}}} P\left\{\left|n^{-1}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|\ge t\right\} \\ &\le \mathcal{N}\left(S^{n-1},\frac{1}{4}\right)P\left\{\left|n^{-1}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|\ge t\right\} \\ &\le 9^n P\left\{\left|n^{-1}\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|\ge t\right\}, \quad \forall x\in\mathcal{N}_{\frac{1}{4}}, \end{aligned} \tag{8.8}$$

where the last inequality follows Lemma 8.2 with $\varepsilon=\frac{1}{4}$.

Step 2. It is sufficient to bound $P\{|\frac{1}{n}\|\mathbf{X}\boldsymbol{x}\|_2^2-1|\ge t\}$. Let $Z_i:=|\langle \boldsymbol{X}_i,\boldsymbol{x}\rangle|$, $\forall \boldsymbol{x}\in S^{n-1}$. Observe that

$$\|\mathbf{X}\boldsymbol{x}\|_2^2=\sum_{i=1}^n|\langle \boldsymbol{X}_i,\boldsymbol{x}\rangle|^2=:\sum_{i=1}^n Z_i^2.$$

Apply the sub-exponential concentration inequality in Corollary 4.2

$$P\left(n^{-1}\left|\|\mathbf{X}\boldsymbol{x}\|_2^2-1\right|\ge t\right)=P\left(n^{-1}\left|\sum_{i=1}^n(Z_i^2-1)\right|\ge t\right)\le 2e^{-\frac{n}{2}\left(\frac{t^2}{\theta^2}\wedge\frac{t}{\theta}\right)}.$$

Specially, let

$$t=c\theta\max(\delta,\delta^2)=c\theta\left[\delta \mathrm{I}_{\{\delta\le 1\}}+\delta^2\mathrm{I}_{\{\delta>1\}}\right]$$

with $\delta := 2c(\frac{p}{n}+\frac{t}{\sqrt{n}})$. From (8.8),

$$
\begin{aligned}
P\left\{\left\|n^{-1}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|\geq 2t\right\} &\leq 9^n \mathrm{P}\left\{\left|n^{-1}\|\mathbf{X}x\|_2^2-1\right|\geq c\theta\max(\delta,\delta^2)\right\}\\
&\leq 2\cdot 9^n e^{[-\frac{cn}{2}\min\{\delta^2 \mathrm{I}_{\{\delta\leq 1\}}+\delta^4\mathrm{I}_{\{\delta>1\}},\delta\mathrm{I}_{\{\delta\leq 1\}}+\delta^2\mathrm{I}_{\{\delta>1\}}\}]}\\
&= 2\cdot 9^n e^{-\frac{cn}{2}\delta^2} = e^{-\frac{c}{2}(\sqrt{p}+t)^2} \leq 2\cdot 9^n e^{-\frac{c}{2}(p+t^2)},
\end{aligned}
$$

where the last inequality is obtained by using the inequality $(a+b)^2\geq a^2+b^2$ for $a,b\geq 0$. For $c\geq n\log\frac{9}{p}$, $2\cdot 9^n e^{-c(p+t^2)}\leq 2e^{-ct^2}$, which proves (8.5) .

Step 3. To show (8.6), the

$$
\max_{\|x\|_2=1}\left|\left\|\frac{1}{\sqrt{n}}\mathbf{X}x\right\|_2^2-1\right| = \max_{\|x\|_2=1}\left\|\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right)x\right\|_2^2 = \left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|^2\leq t^2
$$

implies that $1-t^2\leq\lambda_{\max}(\mathbf{S}_n)\leq 1+t^2$. Similarly, for $\lambda_{\min}(\mathbf{S}_n)$,

$$
\begin{aligned}
\min_{\|x\|_2=1}\left|\left\|\frac{1}{\sqrt{n}}\mathbf{X}x\right\|_2^2-1\right| &= \min_{\|x\|_2=1}\left\|\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right)x\right\|_2^2\\
&\leq \max_{\|x\|_2=1}\left\|\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right)x\right\|_2^2\leq t^2.
\end{aligned}
$$

So $\lambda_{\min}(\mathbf{S}_n)\in[1-t^2,1+t^2]$ and

$$
\left\{\left\|\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|^2\leq t^2\right\} \subset \left\{1-t^2\leq\lambda_{\min}(\mathbf{S}_n)\leq\lambda_{\max}(\mathbf{S}_n)\leq 1+t^2\right\}.
$$

Then

$$
\begin{aligned}
&P\left\{1-t^2\leq\lambda_{\min}(\mathbf{S}_n)\leq\lambda_{\max}(\mathbf{S}_n)\leq 1+t^2\right\}\\
\geq\, &P\left\{\left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}-\mathbf{I}_p\right\|^2\leq t^2\right\}\geq 1-2e^{-ct^2}.
\end{aligned}
$$

The proof is complete. □

8.3 Oracle inequalities for penalized linear models

This section introduces the proofs of the error bounds from the perspective of Lasso penalized linear models with the ℓ_2-loss function. When $p>n$, the OLS estimator is no longer available as $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T$ is of invertible. A common way for obtaining a plausible estimator for the true parameter $\boldsymbol{\beta}^*$ is by adding penalized function to the square loss function. For $0<q\le\infty$, we write $\|\boldsymbol{\beta}\|_q:=(\sum_{i=1}^{p}|\beta_i|^q)^{1/q}$ as the ℓ_q-norm for $\boldsymbol{\beta}\in\mathbb{R}^p$. If $q=\infty$, $\|\boldsymbol{\beta}\|_\infty:=\max_{i=1,\ldots,p}|\beta_i|$; if $q=0$, $\|\boldsymbol{\beta}\|_0:=\sum_{i=1}^{p}1(\beta_i\neq 0)$. There are two types statistical guarantees of $\hat{\boldsymbol{\beta}}$ as mentioned in [7].

1. **Persistence**: $\hat{\boldsymbol{\beta}}$ performs well on a new sample $\boldsymbol{X}^*\overset{d}{=}\boldsymbol{X}$ (equal in distribution), i.e. $\mathrm{E}\{[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2|\boldsymbol{X}^*\}\to 0$.

2. ℓ_q**-consistency** ($q\ge 1$): $\hat{\beta}$ approximates β^*, i.e. with high probability $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_q\to 0$.

The persistence and ℓ_1-consistency are respectively obtained by error bounds:

$$\big\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1\le O_p(s\lambda_n),\quad \mathrm{E}\Big\{\big[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big]^2|\boldsymbol{X}^*\Big\}\le O_p(s\lambda_n^2),$$

(says oracle inequalities),

where $\lambda_n\to 0$ is a tuning parameter and $s:=\|\boldsymbol{\beta}^*\|_0$. In the following, we focus on the ℓ_1 estimation and prediction consistencies for the penalized linear models. Let $\lambda>0$ be a tuning parameter, the Lasso estimator [80] for Model (8.1) is

$$\hat{\boldsymbol{\beta}}_L=\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2}{n}+\lambda\|\boldsymbol{\beta}\|_1\right\}. \tag{8.9}$$

By sub-derivative techniques in convex optimizations, the Karush-Kuhn-Tucker (KKT) condition of Lasso optimization function is

$$\begin{cases}\dfrac{2}{n}\Big[\boldsymbol{X}^T(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_L)\Big]_j=-\lambda\operatorname{sign}(\hat{\beta}_{Lj}), & \text{if}\quad \hat{\beta}_j\neq 0,\\ \dfrac{2}{n}\left|\Big[\mathbf{X}^T(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})_L\Big]_j\right|\le\lambda, & \text{if}\quad \hat{\beta}_{Lj}=0,\end{cases}$$

which implies $\|\frac{1}{n}\mathbf{X}^T(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\|_\infty\le\frac{\lambda}{2}$. Another approach to get the Lasso-like sparse estimator is attained by Dantzig selector (DS)

$$\hat{\boldsymbol{\beta}}_{DS}=\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\|\boldsymbol{\beta}\|_1:\frac{\|\mathbf{X}^T(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})\|_\infty}{n}\le\frac{\lambda}{2}\right\}, \tag{8.10}$$

see [21]. Lasso and DS are capable of producing sparse estimates with only a few (hence sparse) nonzero coefficients among the p coefficients of the covariates. The idea of Lasso and DS was presented in a geophysics literature [55]. By (8.10), we get $\|\hat{\beta}_{DS}\|_1 \leq \|\hat{\beta}_L\|_1$, which signifies that the DS may be more sparse than the Lasso.

It is well-known that $\mathbf{\Sigma} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ is singular when $p>n$. To obtain oracle inequalities for the Lasso estimator with the minimax optimal rate [94], the restricted eigenvalues proposed in [11] is usually needed. Let $S(\boldsymbol{\beta}^*) := \{j : \beta_j^* \neq 0,\ \boldsymbol{\beta}^* = (\beta_1^*,\cdots,\beta_p^*)^T\}$ and $s := |S(\boldsymbol{\beta}^*)|$. For any vector $\boldsymbol{b} \in \mathbb{R}^p$ and any index set $H \subset \{1,\ldots,p\}$, define the sub-vector indexed by H as $\boldsymbol{b}_H = (\cdots,\tilde{b}_j,\cdots)^T \in \mathbb{R}^p$ with $\tilde{b}_j = b_j$ if $j \in H$ and $\tilde{b}_j = 0$ if $j \notin H$. Define the conic set for a sparse $\boldsymbol{\beta}^*$ with support $S(\boldsymbol{\beta}^*)$

$$\mathrm{C}(\eta,S(\boldsymbol{\beta}^*)) = \Big\{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}_{S(\boldsymbol{\beta}^*)^c}\|_1 \leq \eta \|\boldsymbol{b}_{S(\boldsymbol{\beta}^*)}\|_1\Big\}, \quad \eta > 0. \tag{8.11}$$

Denote the restricted eigenvalue condition (RE) as

$$RE(\eta,S(\boldsymbol{\beta}^*),\mathbf{\Sigma}) = \inf_{0\neq \boldsymbol{b} \in \mathrm{C}(\eta,S(\boldsymbol{\beta}^*))} \frac{(\boldsymbol{b}^T\mathbf{\Sigma}\boldsymbol{b})^{\frac{1}{2}}}{\|\boldsymbol{b}\|_2} > 0$$

for any $p\times p$ matrix $\mathbf{\Sigma}$. In the following, we present a modified version of [11, Theorem 7.2] from [56, Lemma 2.5] beyond Gaussian noise.

Proposition 8.2 (The rate of convergence of the Lasso). *Suppose that* $\mathbf{X}$ *is the fixed design matrix and the error sequence* $\{\varepsilon_i\}_{i=1}^n \overset{\text{IID}}{\sim} N(0,\sigma^2)$ *or* $\{\varepsilon_i/\sigma\}_{i=1}^n \overset{\text{IID}}{\sim} 2-$*strongly log-concave distribution satisfying Lemma* 3.2. *Let* $\{\boldsymbol{X}_{(j)}\}_{j=1}^p \in \mathbb{R}^n$ *be column vectors of* $\mathbf{X}$. *We assume that* $\frac{1}{n}\boldsymbol{X}_{(j)}^T\boldsymbol{X}_{(j)} = 1$. *If* $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$ *satisfies the KKT condition for* $\boldsymbol{\beta}^*$,

$$\Big\|\frac{1}{n}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}^*)\Big\|_\infty \leq \frac{\lambda}{2}. \tag{8.12}$$

(1) Then the estimated error $\boldsymbol{u} := \hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*$ *satisfies* $\|\boldsymbol{u}_{S(\boldsymbol{\beta}^*)^c}\|_1 \leq 3\|\boldsymbol{u}_{S(\boldsymbol{\beta}^*)}\|_1$, *i.e.* $\boldsymbol{u} \in \mathrm{C}(3,S(\boldsymbol{\beta}^*))$.

(2) Suppose that $\mathbf{X}$ *satisfies the RE condition* $\gamma := \mathrm{RE}(3,S(\boldsymbol{\beta}^*),\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T) > 0$. *We have non-asymptotic oracle inequalities with probability greater than* $1-2p^{1-\frac{A^2}{8}}$:

$$\begin{aligned}
&(a) \quad \|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*\|_1 \leq \frac{3A\sigma}{\gamma^2} s\sqrt{\frac{\log p}{n}}, \\
&(b) \quad \|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*\|_2^2 \leq \frac{9A\sigma^2}{\gamma^2}\frac{s\log p}{n},
\end{aligned} \tag{8.13}$$

$$(c) \quad \frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*)\|_2^2 \leq \frac{9A\sigma}{\gamma}\frac{s\log p}{n}, \quad A > 2\sqrt{2}. \tag{8.14}$$

Proof. The proof consists of 3 steps.

Step 1: By the Lasso optimization (8.9),

$$(2n)^{-1}\|\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_L\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}_L\|_1 \le (2n)^{-1}\|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2+\lambda\|\boldsymbol{\beta}^*\|_1. \tag{8.15}$$

From

$$\begin{aligned}
&\frac{\|\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_L\|_2^2}{2n}=\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta}^*+\boldsymbol{\varepsilon}-\mathbf{X}\hat{\boldsymbol{\beta}}_L\|_2^2\\
=&\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta}^*-\mathbf{X}\hat{\boldsymbol{\beta}}_L\|_2^2+\frac{\|\boldsymbol{\varepsilon}\|_2^2}{2n}-\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}^*),\\
&\frac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}\beta^*\|_2^2=\frac{\|\boldsymbol{\varepsilon}\|_2^2}{2n},
\end{aligned}$$

thus

$$\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta}^*-\mathbf{X}\hat{\boldsymbol{\beta}}_L\|_2^2+\frac{\|\boldsymbol{\varepsilon}\|_2^2}{2n}-\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+\lambda\|\hat{\boldsymbol{\beta}}_L\|_1\le\frac{\|\boldsymbol{\varepsilon}\|_2^2}{2n}+\lambda\|\boldsymbol{\beta}^*\|_1.$$

Then,

$$(2n)^{-1}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}^*)\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}_L\|_1\le n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}^*)+\lambda\|\boldsymbol{\beta}^*\|_1. \tag{8.16}$$

The (8.16) is usually called the basic inequality in the proof of Lasso oracle inequalities. The first term in the left side of inequality (8.16) is the empirical prediction error, while on the right side, $\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$ is random and $\lambda\|\boldsymbol{\beta}^*\|_1$ is still fixed and unknown. For $\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$, if we can get a sharper upper bound and it approaching 0 as $n\to\infty$, then we can achieve a sharper oracle inequality in below. By (8.12)

$$\begin{aligned}
\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2}{2n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1&\le\left\|\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}\right\|_\infty\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\|\boldsymbol{\beta}^*\|_1\\
&\le\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\|\boldsymbol{\beta}^*\|_1.
\end{aligned} \tag{8.17}$$

Let $S:=S(\boldsymbol{\beta}^*)$ and note that

$$\|\hat{\boldsymbol{\beta}}_S\|_1=\|\boldsymbol{\beta}_S^*+(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*)\|_1\ge\|\boldsymbol{\beta}_S^*\|_1-\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1,$$

then

$$\|\hat{\boldsymbol{\beta}}\|_1=\|\hat{\boldsymbol{\beta}}_{S^c}\|_1+\|\hat{\boldsymbol{\beta}}_S\|_1\ge\|\boldsymbol{\beta}_S^*\|_1-\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+\|\hat{\boldsymbol{\beta}}_{S^c}\|_1. \tag{8.18}$$

From (8.17), we get $\|\boldsymbol{u}_{S^c}\|_1 \leq 3\|\boldsymbol{u}_S\|_1$ by checking

$$\begin{aligned}
0 &\leq (2n)^{-1}\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2 \leq \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\|\boldsymbol{\beta}^*\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\leq \frac{\lambda}{2}\left\{\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+\|\hat{\boldsymbol{\beta}}_{S^c}\|_1\right\}+\lambda\|\boldsymbol{\beta}_S^*\|_1 \\
&\quad -\lambda\left\{\|\boldsymbol{\beta}_S^*\|_1-\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1+\|\hat{\boldsymbol{\beta}}_{S^c}\|_1\right\} \quad \text{[By (8.18)]} \\
&= \frac{3\lambda}{2}\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_S^*\|_1-\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 =: \frac{3\lambda}{2}\|\boldsymbol{u}_S\|_1-\frac{\lambda}{2}\|\boldsymbol{u}_{S^c}\|_1.
\end{aligned} \tag{8.19}$$

Step 2: The Gaussian error vector $\boldsymbol{\varepsilon}$ enables us to get the Gaussian concentration around its mean, we can shows that (8.12) occurs with a high probability. So next we need to check the Lipschitz condition in Lemma 3.1. Use Lemma 3.1, it implies that

$$P\left(n^{-1}\left|\boldsymbol{X}_{(j)}^T(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}^*)\right| \geq t\right) \leq 2pe^{-\frac{nt^2}{2\sigma^2}}, \quad \forall j \tag{8.20}$$

under the presupposition $\|\boldsymbol{X}_{(j)}\|_2^2=\boldsymbol{X}_{(j)}^T\boldsymbol{X}_{(j)}=n$. The Lipschitz condition depends on the design matrix $\mathbf{X}$. The different types of CIs require different assumptions on the design matrix (the random design is allowed if we adopt empirical process theory). In Lemma 3.1, put $f(\boldsymbol{a}):=\frac{1}{n}|\boldsymbol{X}_{(j)}^T(\sigma\boldsymbol{a}-\mathbf{X}\boldsymbol{\beta}^*)|$. Then, Cauchy's inequality implies

$$\begin{aligned}
f(\boldsymbol{a})-f(\boldsymbol{b}) &\leq \frac{\sigma}{n}\left|\boldsymbol{X}_{(j)}^T(\boldsymbol{b}-\boldsymbol{a})\right| \leq \frac{\sigma}{n}\|\boldsymbol{X}_{(j)}\|_2\cdot\|\boldsymbol{b}-\boldsymbol{a}\|_2 \\
&= \frac{\sigma}{\sqrt{n}}\|\boldsymbol{b}-\boldsymbol{a}\|_2, \quad \forall j.
\end{aligned}$$

Hence, $f(\boldsymbol{a})$ is $\frac{\sigma}{\sqrt{n}}$-Lipschitz. Recall $\lambda=A\sigma\sqrt{\frac{\log p}{n}}$. So (8.20) implies

$$\begin{aligned}
&P\left(\left\|\frac{1}{n}\mathbf{X}^T(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}^*)\right\|_\infty \geq \frac{\lambda}{2}\right) \\
&\leq \sum_{j=1}^p P\left(\frac{1}{n}\left|\boldsymbol{X}_{(j)}^T(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}^*)\right| \geq \frac{1}{2}A\sigma\sqrt{\frac{\log p}{n}}\right) \leq 2p^{1-\frac{A^2}{8}}.
\end{aligned}$$

By Lemma 3.2, (8.20) is also held for $\{\varepsilon_i/\sigma\}_{i=1}^n \sim$ 2-strongly log-concave distribution.

Step 3: Next we can start on the proof based on cone set condition (8.11). Since the $\mathbf{X}$ satisfies RE condition $\gamma := \mathrm{RE}(3,S,n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T) > 0$, by (8.11) we have

$$\begin{aligned}\gamma\|\boldsymbol{u}\|_2^2 \le \frac{1}{n}\|\mathbf{X}\boldsymbol{u}\|_2^2 &\overset{(8.19)}{\le} \lambda\big(3\|\boldsymbol{u}_S\|_1 - \|\boldsymbol{u}_{S^c}\|_1\big)\\ &\le 3\lambda\|\boldsymbol{u}_S\|_1 \le 3\lambda\sqrt{s}\|\boldsymbol{u}_S\|_2 \le 3\lambda\sqrt{s}\|\boldsymbol{u}\|_2,\end{aligned}$$

where the second last inequality is by Cauchy's inequality. Therefore,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*\|_2^2 =: \|\boldsymbol{u}\|_2^2 \le \frac{9\lambda^2 s}{\gamma^2} = \frac{9A^2\sigma^2}{\gamma^2}\frac{s\log p}{n},\\ \frac{\big\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*)\big\|_2^2}{n} =: \frac{\|\mathbf{X}\boldsymbol{u}\|_2^2}{n} \le \frac{9\lambda^2 s}{\gamma} = \frac{9A^2\sigma^2}{\gamma}\frac{s\log p}{n}.\end{aligned}$$

So

$$\big\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*\big\|_1 =: \|\boldsymbol{u}\|_1 \le \sqrt{s}\|\boldsymbol{u}\|_2 \le \frac{3\lambda s}{\gamma} = \frac{3A\sigma}{\gamma} s\sqrt{\frac{\log p}{n}}$$

by Cauchy's inequality. □

According to (8.3), the OLS with diverging number of covariates has the convergence rate $\mathcal{O}(\sqrt{\frac{p}{n}})$ under the minimal eigenvalue condition $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = \mathcal{O}(n)$. In contrast, due to the sparse restriction and the RE condition in Proposition 8.2, the factor $\sqrt{\log p}$ is much more small that the factor $\sqrt{p}$ in the convergence rate (8.3). Under the RE condition, Proposition 8.2 reveals that Lasso is ℓ_2-consistent if $\frac{s\log p}{n} \to 0$, and $s\sqrt{\frac{\log p}{n}} \to 0$ guarantees ℓ_1-consistency. [11, Theorem 7.1] also gives oracle inequalities (8.13) and (8.14) for the DS estimator (8.10).

8.4 High-dimensional Poisson regressions with random design

The Poisson regression [60] is a model for nonnegative integers response variables, i.e. $Y_i \overset{\mathrm{IID}}{\sim} \mathrm{Poisson}(\lambda_i)$, where $\log(\lambda_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ for $i = 1,\ldots,n$. We presume that the $\{\boldsymbol{X}_i\}_{i=1}^n$ are IID RVs on some space $\mathcal{X}$, and we observe n copies of $\{(Y_i,\boldsymbol{X}_i)\}_{i=1}^n \sim (Y,\boldsymbol{X}) \in \mathbb{R}\times\mathbb{R}^p$.

The average negative log-likelihood of Poisson regressions is

$$\ell_n(\boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^n \Big[Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\Big]$$

and the Lasso penalized estimator is

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\ell_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\}$$
$$\text{with a turning parameter } \lambda > 0. \tag{8.21}$$

[18, Lemma 4.2] shows the first-order conditions for the optimization in (8.21).

Lemma 8.4 (Necessary and sufficient condition)**.** *Let $j \in \{1,\dots,p\}$ and $\lambda > 0$. Then, a necessary and sufficient condition for the Lasso estimates* (8.21) *is*

$$\begin{cases} n^{-1}\sum_{i=1}^n X_{ij}\big(Y_i - e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}\big) = -\lambda \operatorname{sign}(\hat{\beta}_j), & \textit{if} \quad \hat{\beta}_j \neq 0, \\ \Big|n^{-1}\sum_{i=1}^n X_{ij}\big(Y_i - e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}\big)\Big| \leq \lambda, & \textit{if} \quad \hat{\beta}_j = 0. \end{cases} \tag{8.22}$$

Let $l(Y,\boldsymbol{X},\boldsymbol{\beta}) = -Y\boldsymbol{X}^T\boldsymbol{\beta} + e^{\boldsymbol{X}^T\boldsymbol{\beta}}$ be the Poisson loss function. The true coefficient $\boldsymbol{\beta}^*$ is the minimizer of the expected Poisson loss, i.e.

$$\boldsymbol{\beta}^* = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathrm{E}l(Y,\boldsymbol{X},\boldsymbol{\beta}). \tag{8.23}$$

The KKT condition of the ℓ_1-penalized likelihood is for the estimated parameter. But, here we use the true parameter version of the KKT conditions

$$\left|\frac{1}{n}\sum_{i=1}^n X_{ij}(Y_i - \mathrm{E}Y_i)\right| \leq \lambda, \quad j = 1,\dots,p$$

by replacing $e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}$ by $\mathrm{E}Y_i = e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}$ to approximate the estimated version (8.22). To motivate the next two propositions concerning high-probability events, let us consider the following notations and the decomposition of empirical process.

The Poisson loss $l(\boldsymbol{\beta},\boldsymbol{X},Y) = l_1(\boldsymbol{\beta},\boldsymbol{X},Y) + l_2(\boldsymbol{\beta},\boldsymbol{X})$ is decomposed into two parts where $l_1(\boldsymbol{\beta}) := l_1(\boldsymbol{\beta},\boldsymbol{X},Y) := -Y\boldsymbol{X}^T\boldsymbol{\beta}$ and $l_2(\boldsymbol{\beta}) := l_2(\boldsymbol{\beta},\boldsymbol{X}) := e^{\boldsymbol{X}^T\boldsymbol{\beta}}$ is free of response. Let $\mathbb{P}l(\boldsymbol{\beta}) := \mathrm{E}l(\boldsymbol{\beta},\boldsymbol{X},Y)$ be the expected loss. We are interested in the centralized empirical loss $(\mathbb{P}_n - \mathbb{P})l(\boldsymbol{\beta})$ representing fluctuations between the expected and empirical losses. Note that

$$(\mathbb{P}_n - \mathbb{P})l(\boldsymbol{\beta}) = (\mathbb{P}_n - \mathbb{P})l_1(\boldsymbol{\beta}) + (\mathbb{P}_n - \mathbb{P})l_2(\boldsymbol{\beta}), \tag{8.24}$$

which is crucial in attaining the convergence rate of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$. Motivated by the rate of convergence theorem ([82, Theorem 3.2.5]) for M-estimation with functional parameter in some metric space, we study the upper bounds (or the rate) for the first and second part of the difference of the centralized empirical process between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$: $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$, for $m = 1,2$.

Proposition 8.3 (Convergence rate of $(\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}}))$). *Suppose that*

$$\sup_{1\le i\le\infty}\|\boldsymbol{X}_i\|_\infty\le L<\infty \ \text{ a.s.},\quad \|\boldsymbol{\beta}^*\|_1\le B. \tag{8.25}$$

In the event of

$$\mathcal{A}:=\bigcap_{j=1}^{p}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}(Y_i-\mathrm{E}Y_i)\right|\le\frac{\lambda}{4}\right\},$$

we have

$$(\mathbb{P}_n-\mathbb{P})(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}}))\le\frac{\lambda}{4}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1. \tag{8.26}$$

If

$$\lambda\ge\max\left\{\frac{16A^2L\log(2p)}{3n},8ALe^{\frac{LB}{2}}\sqrt{\frac{\log(2p)}{n}}\right\}$$

with $A>1$, we have $P(\mathcal{A})\ge 1-(2p)^{1-A^2}$.

Proof. Note that, on the event $\mathcal{A}$

$$\begin{aligned}(\mathbb{P}_n-\mathbb{P})\big(l_1(\boldsymbol{\beta}^*)-l_1(\hat{\boldsymbol{\beta}})\big)&=\frac{-1}{n}\sum_{i=1}^{n}(Y_i-\mathrm{E}Y_i)\boldsymbol{X}_i^T(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}})\\&=\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j^*)\frac{1}{n}\sum_{i=1}^{n}X_{ij}(Y_i-\mathrm{E}Y_i)\\&\le\sum_{j=1}^{p}|\hat{\beta}_j-\beta_j^*|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}(Y_i-\mathrm{E}Y_i)\right|\overset{\mathcal{A}}{\le}\frac{\lambda}{4}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1.\end{aligned}$$

Next, we show that $\mathcal{A}$ is a high probability event if λ is well chosen. For $j=1,\ldots,p$ and $i=1,\ldots,n$,

$$P(\mathcal{A}^c)\le\sum_{j=1}^{p}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}(Y_i-\mathrm{E}Y_i)\right|>\frac{\lambda}{4}\right\}.$$

Given $\mathbf{X}$, $\{S_{nj}(Y,X):=\frac{1}{n}X_{ij}(Y_i-\mathrm{E}Y_i)\}_{i=1}^{n}$ are conditional independent for each $j=1,\ldots,p$. Thus Corollary 5.2 with $w_i=\frac{X_{ij}}{n}$ gives

$$\begin{aligned}P\big(|S_{nj}(Y,X)|\ge t|\mathbf{X}\big)&\le 2\exp\left\{-\frac{nt^2/2}{\Big(\sum_{i=1}^{n}e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}\max_{1\le i\le n}X_{ij}^2+\max_{1\le i\le n}(|X_{ij}|t/3)\Big)/n}\right\}\\&\le 2\Big(e^{\frac{-nt^2}{4L^2e^{LB}}}\vee e^{\frac{-3nt}{4L}}\Big),\end{aligned} \tag{8.27}$$

where the last inequality is from $\mathrm{e}^{-\frac{a}{b+c}} \le \mathrm{e}^{\frac{-a}{2b}} \vee \mathrm{e}^{\frac{-a}{2c}}$ for any positive numbers a,b and c.

Let $t = \frac{\lambda}{4}$. Assumptions (8.25) and (8.27) give for $j = 1,\ldots,p$

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}(Y_i - \mathrm{E}Y_i)\right| \ge \frac{\lambda}{4}\right) &= \mathrm{E}P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}(Y_i - \mathrm{E}Y_i)\right| \ge \frac{\lambda}{4}\middle|\mathbf{X}\right) \\ &\le 2\max\left\{e^{\frac{-n\lambda^2}{64L^2e^{LB}}}, e^{\frac{-3n\lambda}{16L}}\right\}, \end{aligned}$$

which implies that $P(\mathcal{A}^c) \le 2p\max\{e^{\frac{-n\lambda^2}{64L^2e^{LB}}}, e^{\frac{-3n\lambda}{16L}}\}$.

Finally, if

$$\lambda \ge \max\left\{\frac{16A^2L\log(2p)}{3n}, 8ALe^{\frac{LB}{2}}\sqrt{\frac{\log(2p)}{n}}\right\}, \quad A > 1,$$

so $P(\mathcal{A}^c) \le (2p)^{1-A^2}$. □

Next, we provide a crucial lemma to bound $(\mathbb{P}_n - \mathbb{P})(l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta}))$. Let

$$\nu_n(\boldsymbol{\beta},\boldsymbol{\beta}^*) := \frac{(\mathbb{P}_n - \mathbb{P})\big(l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta})\big)}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1}$$

the normalized empirical process indexed by $\boldsymbol{\beta}$. Denote the ℓ_1-ball by $\mathcal{S}_M(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le M < \infty\}$, we define the local stochastic Lipschitz constant

$$Z_M(\boldsymbol{\beta}^*) := \sup_{\boldsymbol{\beta} \in \mathcal{S}_M(\boldsymbol{\beta}^*)} \big|\nu_n(\boldsymbol{\beta},\boldsymbol{\beta}^*)\big|$$

and a random event

$$\mathcal{B} := \left\{Z_M(\boldsymbol{\beta}^*) \le \frac{\lambda_1}{4}\right\}.$$

It is easy to see $|\nu_n(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*)| \le \sup_{\mathcal{S}_M(\boldsymbol{\beta}^*)} |\nu_n(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*)| \le \frac{\lambda_1}{4}$, which gives

$$\big|(\mathbb{P}_n - \mathbb{P})(l_2(\hat{\boldsymbol{\beta}}) - l_2(\boldsymbol{\beta}^*))\big| \le \frac{\lambda_1}{4}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$$

provided that $\hat{\boldsymbol{\beta}} \in \mathcal{S}_M(\boldsymbol{\beta}^*)$. Then we have the following result.

Proposition 8.4 (Convergence rate of $(\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\hat{\boldsymbol{\beta}}))$)**.** *Assume that there exists a large constant M such that $\hat{\boldsymbol{\beta}}$ is in the ℓ_1-ball $\mathcal{S}_M(\boldsymbol{\beta}^*)$. Under assumption* (8.25), *we have*

$$P\left(Z_M(\boldsymbol{\beta}^*)\geq 5ALe^{LB}\sqrt{\frac{\log 2p}{n}}\right)\leq (2p)^{-A^2}. \tag{8.28}$$

If $\lambda\geq 20ALe^{LB}\sqrt{\frac{2\log 2p}{n}}$, we get

$$P\left\{\left|(\mathbb{P}_n-\mathbb{P})(l_2(\hat{\boldsymbol{\beta}})-l_2(\boldsymbol{\beta}^*))\right|\leq\frac{\lambda}{4}(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1)\right\}\geq 1-(2p)^{-A^2}.$$

Proof. In the first step, we apply following McDiarmid's inequality to $Z_M(\boldsymbol{\beta}^*)$ by showing that $Z_M(\boldsymbol{\beta}^*)$ is fluctuated of no more than $\frac{2e^{LB}}{n}$. Let us check it. Put $\mathbb{P}_n:=\frac{1}{n}\sum_{j=1}^n 1_{\boldsymbol{X}_j,Y_j}$ and $\mathbb{P}_n^{'}:=\frac{1}{n}\sum_{j=1,j\neq i}^n 1_{\boldsymbol{X}_j,Y_j}+1_{\boldsymbol{X}_i^{'},Y_i^{'}}$, where $(\boldsymbol{X}_i^{'},Y_i^{'})$ is the independent copy of $(\boldsymbol{X}_j,Y_j)$.

Let $\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}_i$ ($\boldsymbol{X'}_i^T\tilde{\boldsymbol{\beta}}_i$) be an intermediate point between $\boldsymbol{X}_i^T\boldsymbol{\beta}$ ($\boldsymbol{X'}_i^T\boldsymbol{\beta}$) and $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$ ($\boldsymbol{X'}_i^T\boldsymbol{\beta}^*$) from the Taylor's expansion of function $F(x):=e^x$. It deduces

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in S_M}\frac{|(\mathbb{P}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta}))|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}-\sup_{\boldsymbol{\beta}\in S_M}\frac{|(\mathbb{P'}_n-\mathbb{P})(l_2(\boldsymbol{\beta}^*)-l_2(\hat{\boldsymbol{\beta}}))|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\\
\leq&\sup_{\boldsymbol{\beta}\in S_M}\frac{|l_2(\boldsymbol{\beta}^*,\boldsymbol{X}_i)-l_2(\boldsymbol{\beta},\boldsymbol{X}_i)-l_2(\boldsymbol{\beta}^*,\boldsymbol{X'}_i)+l_2(\boldsymbol{\beta},\boldsymbol{X'}_i)|}{n\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\\
\leq&\sup_{\boldsymbol{\beta}\in S_M}\frac{1}{n}e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}\cdot\frac{|\boldsymbol{X}_i^T\boldsymbol{\beta}^*-\boldsymbol{X}_i^T\boldsymbol{\beta}|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}+\sup_{\boldsymbol{\beta}\in S_M}\frac{1}{n}e^{\boldsymbol{X}_i^T\tilde{\boldsymbol{\beta}}}\cdot\frac{|\boldsymbol{X'}_i^T\boldsymbol{\beta}^*-\boldsymbol{X'}_i^T\boldsymbol{\beta}|}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\\
\leq&\sup_{\boldsymbol{\beta}\in S_M}\frac{2Le^{LB}}{n}\frac{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}=\frac{2Le^{LB}}{n}.
\end{aligned}$$

Apply McDiarmid's inequality to $Z_M(\boldsymbol{\beta}^*)$, we have

$$P\big(Z_M(\boldsymbol{\beta}^*)-\mathrm{E}Z_M(\boldsymbol{\beta}^*)\geq\lambda\big)\leq e^{-\frac{n\lambda^2}{2L^2e^{2LB}}}.$$

Let $(2p)^{-A^2}=\exp\{-\frac{n\lambda^2}{2L^2e^{2LB}}\}$, we get $\lambda\geq ALe^{LB}\sqrt{\frac{2\log(2p)}{n}}$ for $A>0$, therefore

$$P\big(Z_M(\boldsymbol{\beta}^*)-\mathrm{E}Z_M(\boldsymbol{\beta}^*)\geq\lambda\big)\leq(2p)^{-A^2}. \tag{8.29}$$

The next step is to estimate the sharper upper bounds of $\mathrm{E}Z_M(\boldsymbol{\beta}^*)$ by Lemma 7.3 with $\Phi(t)=|t|$ and Lemma 7.4. Note that

$$(\mathbb{P}_n-\mathbb{P})\big\{l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta})\big\}=\mathbb{P}_n\big\{l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta})\big\}-\mathrm{E}\big\{l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta})\big\}$$

by symmetrization theorem, the expected terms is canceled. To see contraction theorem, for

$$Z_M(\boldsymbol{\beta}^*)=\sup_{\boldsymbol{\beta}\in\mathcal{S}_M}\left\{\frac{1}{n\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\left|\sum_{i=1}^n\left(e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}-e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)-n\mathrm{E}\big[l_2(\boldsymbol{\beta}^*)-l_2(\boldsymbol{\beta})\big]\right|\right\}$$

it is required to check the Lipschitz property of g_i in Lemma 7.4 with $\mathcal{F}=\mathbb{R}^p$. Let

$$f(x_i)=\frac{x_i^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1},\quad h(x_i)=\frac{x_i^T\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1},$$
$$g_i(t)=\frac{e^{t\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}}{n\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\quad\left(|t|\le\frac{LB}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\right).$$

Then the function $g_i(t)$ here is $\frac{e^{LB}}{n}$-Lipschitz. In fact

$$|g_i(s)-g_i(t)|=\frac{e^{\tilde{t}}}{n}\cdot|s-t|\le\frac{e^{LB}}{n}|s-t|,\quad t,s\in\left[-\frac{LB}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1},\frac{LB}{\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1}\right],$$

where $\tilde{t}\in[-LB/\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1,LB/\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_1]$ is an intermediate point between t and s given by applying Lagrange mean value theorem.

The symmetrization theorem and the contraction theorem imply

$$\begin{aligned}\mathrm{E}Z_M(\boldsymbol{\beta}^*)&\le\frac{4e^{LB}}{n}\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\mathcal{S}_M}\left|\sum_{i=1}^n\frac{\epsilon_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^*-\boldsymbol{\beta})}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1}\right|\right)\\&\le\frac{4e^{LB}}{n}\mathrm{E}\left(\sup_{\boldsymbol{\beta}\in\mathcal{S}_M}\max_{1\le j\le p}\left|\sum_{i=1}^n\epsilon_iX_{ij}\right|\cdot\frac{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1}{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1}\right)\\&\le\frac{4e^{LB}}{n}\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^n\epsilon_iX_{ij}\right|\right)=\frac{4e^{LB}}{n}\mathrm{E}\left(\mathrm{E}\left[\max_{1\le j\le p}\left|\sum_{i=1}^n\epsilon_iX_{ij}\right|\bigg|\mathbf{X}\right]\right).\end{aligned}$$

From Corollary 7.5, with $\mathrm{E}_\epsilon[\epsilon_iX_{ij}|\mathbf{X}]=0$ we get

$$\begin{aligned}&\frac{4e^{LB}}{n}\mathrm{E}\left(\mathrm{E}\left[\max_{1\le j\le p}\left|\sum_{i=1}^n\epsilon_iX_{ij}\right|\bigg|\mathbf{X}\right]\right)\\&\le\frac{4e^{LB}}{n}\sqrt{2\log2p}\cdot\sqrt{nL^2}=4e^{LB}L\sqrt{\frac{2\log2p}{n}}.\end{aligned}$$

Thus, for $A \geq 1$,

$$\mathrm{E}Z_M(\boldsymbol{\beta}^*) \leq 4e^{LB} L\sqrt{\frac{2\log 2p}{n}} \leq 4ALe^{LB}\sqrt{\frac{2\log 2p}{n}}. \tag{8.30}$$

With $\lambda \geq ALe^{LB}\sqrt{\frac{2\log(2p)}{n}}$ and (8.30), we conclude from (8.29) that

$$P\left(Z_M(\boldsymbol{\beta}^*) \geq 5ALe^{LB}\sqrt{\frac{\log 2p}{n}}\right) \leq P\big(Z_M(\boldsymbol{\beta}^*) \geq \lambda + \mathrm{E}Z_M(\boldsymbol{\beta}^*)\big) \leq (2p)^{-A^2}.$$

Finally, we complete the proof of Proposition 8.4 by letting $\frac{\lambda}{4} \geq 5ALe^{LB}\sqrt{\frac{2\log 2p}{n}}$ and setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} \in Z_M(\boldsymbol{\beta}^*)$. $\square$

Let $S := S(\boldsymbol{\beta}^*)$ for $\boldsymbol{\beta}^*$ defined in (8.23) and $s := |S|$. To obtain sharp oracle inequalities for Lasso penalized Poisson regression, we consider the following regularity conditions:

- (H.1): The covariate $\boldsymbol{X}$ is almost surely bounded $\|\boldsymbol{X}\|_\infty \leq L$ a.s. for $L > 0$;
- (H.2): There exists a constant $B > 0$ such that $\|\boldsymbol{\beta}^*\|_1 \leq B$;
- (H.3): (Stabil Condition) For $\Sigma := \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)$, there exist a $k \in (0,1)$ such that

$$\delta^T \Sigma \delta \geqslant k \sum_{j\in S} \delta_j^2 \quad \text{for any} \quad \delta \in C(c_0, S) := \left\{\delta \in \mathbb{R}^p : \sum_{j\in S^c} |\delta_j| \leq c_0 \sum_{j\in S} |\delta_j|\right\}.$$

The Stabil Condition (H.3) is denoted as $S(c_0, S, k, \Sigma)$ which is a similar version of the RE condition in the Lasso linear models proposed in [19]. Due to the random variance, Poisson regression is more complex than the linear model with the constant variance assumption. Thus, (H.1) and (H.2) are stronger than those assumed for the linear models. Based on the high-probability event $\mathcal{A}$ and $\mathcal{B}$, we have the oracle inequalities for estimation and prediction for Lasso estimator $\hat{\boldsymbol{\beta}}$ in (8.21) for the Poisson regressions.

Theorem 8.1. *Assume conditions $(H.1)-(H.3)$ hold. Let λ be chosen such that*

$$\lambda \geq \max\left\{\frac{16A^2 L\log(2p)}{3n}, 8ALe^{\frac{LB}{2}}\sqrt{\frac{\log(2p)}{n}}, 20ALe^{LB}\sqrt{\frac{2\log 2p}{n}}\right\} \quad \textit{for} \quad A > \sqrt{2}. \tag{8.31}$$

Suppose that we have a new covariate vector $\boldsymbol{X}^*$ *(as the test data) which is an independent copy of* X *(as the training data), and* E^* *represents the expectation with respect to* $\boldsymbol{X}^*$ *only, then*

$$P\left(\mathrm{E}^*[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2\le\frac{12e^{10LB}}{k}s\lambda^2\right),$$
$$P\left(\|\hat{\beta}-\beta^*\|_1\le\frac{4e^{5LB}}{k}s\lambda\right)\ge 1-(2p)^{1-A^2}-(2p)^{-\frac{A^2}{2}}.$$

The Theorem 8.1 leads to the persistence and ℓ_1-consistency if $\max\{s\lambda,s\lambda^2\}\to 0$.

Proof. The proof consists of three steps. The techniques are adapted from [41,98] and references therein.

Step 1: Check $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\in C(3,S)$. From the definition of the Lasso estimates $\hat{\boldsymbol{\beta}}$ (see (8.21)),

$$\mathbb{P}_n l(\hat{\boldsymbol{\beta}})+\lambda\|\hat{\boldsymbol{\beta}}\|_1\le\mathbb{P}_n l(\boldsymbol{\beta}^*)+\lambda\|\boldsymbol{\beta}^*\|_1. \tag{8.32}$$

By adding $\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))+\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1$ to both sides of (8.32), we have

$$\begin{aligned}&\mathbb{P}\big(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)\big)+\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\\ \le&(\mathbb{P}_n-\mathbb{P})\big(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}})\big)+\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big),\end{aligned}$$

which leads

$$\begin{aligned}&\mathbb{P}\big(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)\big)+\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\\ \le&(\mathbb{P}_n-\mathbb{P})\big(l(\boldsymbol{\beta}^*)-l(\hat{\boldsymbol{\beta}})\big)+\frac{\lambda}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big)\\ \le&\lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big).\end{aligned} \tag{8.33}$$

By the definition of β^*, $\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))\ge 0$. The above inequality and the fact $|\hat{\beta}_j-\beta_j^*|+|\beta_j^*|-|\hat{\beta}_j|=0$ for $j\notin S$ and $|\hat{\beta}_j|-|\beta_j^*|\le|\hat{\beta}_j-\beta_j^*|$ for $j\in S$ lead to

$$\begin{aligned}\frac{\lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1}{2}&\le\lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big)\\ &\le 2\lambda\big\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\big\|_1.\end{aligned} \tag{8.34}$$

Thus, $\frac{\lambda}{2}\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_{S^c}\|_1\le 1.5\lambda\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\|_1$ and then $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\in C(3,S)$.

Step 2: Choosing λ. Since $\mathbb{P}(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*))\geq 0$, (8.33) implies

$$\begin{aligned}\frac{\lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1}{2}&\leq \lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big)\\ &\leq \lambda\|\hat{\boldsymbol{\beta}}\|_1+\lambda\|\boldsymbol{\beta}^*\|_1+\lambda\big(\|\boldsymbol{\beta}^*\|_1-\|\hat{\boldsymbol{\beta}}\|_1\big)=2\lambda\|\boldsymbol{\beta}^*\|_1. \end{aligned}\tag{8.35}$$

Thus (H.2) implies $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\leq 4B$. After having shown Propositions 8.3 and 8.4, we need the result on the high probability of the event $\mathcal{A}\bigcap\mathcal{B}$, whose proof is skipped.

Proposition 8.5. *Under the event* $\mathcal{A}\bigcap\mathcal{B}$ *with* (H.1)-(H.3), *we have* $\hat{\boldsymbol{\beta}}\in\mathcal{S}_{4B}(\boldsymbol{\beta}^*)$. *And if* λ *are chosen as* (8.31), *then* $P(\mathcal{A}\cap\mathcal{B})\geq 1-(2p)^{1-A^2}-(2p)^{-\frac{A^2}{2}}$.

Step 3: Error bounds from Stabil Condition. As $\boldsymbol{X}^*$ is an independent copy of $\boldsymbol{X}$,

$$\begin{aligned}\mathbb{P}\big\{l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)\big\}&=\mathrm{E}^*\Big[\mathrm{E}\big\{l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)|\boldsymbol{X}^*\big\}\Big]\\ &:=\mathrm{E}^*\Big\{\mathrm{E}\big[-Y\boldsymbol{X}^{*T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}}-e^{\boldsymbol{X}^T\boldsymbol{\beta}^*}\big]|\boldsymbol{X}^*\Big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &=\mathrm{E}^*\Big\{\mathrm{E}\big[-Y|\boldsymbol{X}^*\big]\boldsymbol{X}^{*T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+(e^{\boldsymbol{X}^{*T}\hat{\boldsymbol{\beta}}}-e^{\boldsymbol{X}^T\boldsymbol{\beta}^*})]|\boldsymbol{X}^*\Big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &\quad\Big(\mathrm{E}^*[Y|\boldsymbol{X}^*]=e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}\Big)\\ &=\mathrm{E}^*\Big\{-e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}+e^{\boldsymbol{X}^{*T}\boldsymbol{\beta}^*}+2^{-1}e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}\big[\boldsymbol{X}^{*T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big]^2\Big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &=2^{-1}\mathrm{E}^*\Big\{e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}\big[\boldsymbol{X}^{*T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big]^2\Big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},\end{aligned}$$

where $\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}=(1-t)\boldsymbol{X}^{*T}\boldsymbol{\beta}^*+t\boldsymbol{X}^{*T}\hat{\beta}$ is an intermediate point of $\boldsymbol{X}^{*T}$ and $\boldsymbol{X}^{*T}\hat{\beta}$ with $t\in[0,1]$.

Note that $\|\boldsymbol{\beta}^*\|_1\leq B$ by (H.1) and $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\leq 4B$, (H.2) yields

$$\begin{aligned}|\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}|&\leq t\big|\boldsymbol{X}^{*T}\hat{\beta}-\boldsymbol{X}^{*T}\boldsymbol{\beta}^*\big|+\big|\boldsymbol{X}^{*T}\boldsymbol{\beta}^*\big|\\ &\leq\|\boldsymbol{X}^*\|_\infty\cdot\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1+\big|\boldsymbol{X}^{*T}\boldsymbol{\beta}^*\big|\\ &\leq 4LB+LB=5LB,\end{aligned}$$

which implies for $c:=\frac{e^{-5LB}}{2}$

$$\begin{aligned}\mathbb{P}\big\{l(\hat{\beta})-l(\beta^*)\big\}&\geq \inf\nolimits_{|t|\leq 5LB}2^{-1}\mathrm{E}^*\Big\{e^{\boldsymbol{X}^{*T}\tilde{\boldsymbol{\beta}}}\big[\boldsymbol{X}^{*T}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big]^2\Big\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\\ &=:c\mathrm{E}^*\big[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big]^2.\end{aligned}\tag{8.36}$$

As $\mathrm{E}^*(\boldsymbol{X}^*\boldsymbol{X}^{*T})=\boldsymbol{\Sigma}$, $\mathrm{E}^*[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)]^2=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$.

Having checked the cone condition $C(3,S)$, we apply the Stabil Condition

$$c(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\geq ck\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\|_2^2. \tag{8.37}$$

From (8.33), (8.34) and (8.36), we get

$$\begin{aligned}&c\mathrm{E}^*\left[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2+\frac{\lambda}{2}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_1\\ \leq&\mathbb{P}\left(l(\hat{\boldsymbol{\beta}})-l(\boldsymbol{\beta}^*)\right)+\frac{\lambda}{2}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_1\leq 2\lambda\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_1,\end{aligned} \tag{8.38}$$

which gives

$$ck\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_2^2+\frac{\lambda}{2}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_1\leq 2\lambda\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_1$$

by plugging (8.37) into (8.38). Then, employing Cauchy's inequality, we have

$$\begin{aligned}&2ck\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_2^2+\lambda\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right\|_1\\ \leq&4\lambda\left(s\cdot\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_2^2\right)^{\frac{1}{2}}\leq 4t\lambda^2s+\frac{1}{t}\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\right\|_2^2,\end{aligned} \tag{8.39}$$

where the last inequality is from the elementary inequality $2xy\leq tx^2+\frac{y^2}{t}$ for all $t>0$. Let us set $t=(2ck)^{-1}$ in (8.39), thus

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\leq 4t\lambda s=\frac{2\lambda s}{ck}=\frac{4e^{5LB}}{k}s\lambda.$$

To derive the oracle inequality of prediction error, from (8.38), we obtain

$$c\mathrm{E}^*\left[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2\leq 1.5\lambda\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)_S\|_1\leq 1.5\lambda\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_1,$$

which implies

$$\mathrm{E}^*\left[\boldsymbol{X}^*(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right]^2\leq\frac{1.5\lambda}{c}\left\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_1\leq\frac{3s\lambda^2}{c^2k}=\frac{12e^{10LB}}{k}s\lambda^2,$$

where the last inequality is from $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1\leq\frac{4e^{5LB}}{k}s\lambda$. □

For general losses beyond linear models, the crucial techniques in the non-asymptotical analysis of increasing-dimensional and high-dimensional regressions, which are Bahadur representation's for the M-estimator [49, 64] and concentration for Lipschitz loss functions [18, 98], respectively. In large-dimensional regressions with $\frac{p}{n}\to c$, the theory of random matrix [93], leave-one-out analysis [27, 53] and approximate message passing [23, 26, 27] play important roles for obtaining asymptotical results.

9 Extensions

The review has been focused on the sum of independent RVs in the Euclidean space. However, independence structure may not be suitable for some applications, for instance, econometrics, survival analysis, and graphical models. At the same time, the Euclidean valued RVs may not be appropriate for functional data and image data. In the following we point out results in settings not covered to broaden this review.

By CIs for the martingales, oracle inequalities have been proposed for Lasso penalized Cox models, see [42]. Some statistical models, such as the Ising model involving Markov's chains. [62] applied Hoeffding's inequality for Markov's chains to deal with this difficulty, see [28] for a review. In time series analysis, [90] studies the square-root Lasso method for HD linear models with α,ρ, ϕ-mixing or m-dependent errors. The Hoeffding's and Bernstein's CIs for weakly dependent summations can be found in [13]. Via sub-Weibull concentrations under β-mixing, non-asymptotic inequalities for estimation errors, and the prediction errors are obtained by [89] for the Lasso-regularized sparse VAR model with sub-Weibull innovations. U-Statistic is another dependent sum, and Example 2.2 provides a concentration result by McDiarmid's inequality. [12] introduces the concentration for the Banach-valued U-statistics.

In non-parametric regressions, the corresponding score functions may be RVs in Banach (or Hilbert) space, see the monographs [51, 95] for introductions. Exponential tail bounds for Banach- or Hilbert-valued RVs are indispensable for deriving sharp oracle inequalities of the error bounds, see [54, 101]. Recently, Banach-valued CIs are applied to conceive non-asymptotic hypothesis testing for non-parametric regressions, see [92]. To extend the empirical covariance matrices from finite to infinite dimension, the sample covariance operator is treated as a random element in Banach spaces. The concentrations of empirical covariance operator also have been raised attention in kernel principal components analysis, and functional data analysis, see [20, 71].

Testing hypotheses on the regression coefficients are a necessity in measuring the effects of covariates on the certain response variables. Scientists are interested in testing the significance of a large number of covariates simultaneously. From this backgrounds, [102] proposed simultaneous tests for coefficients in HD linear models under the "large p, small n" situations by U-statistics motivated by [22]. However, their HD tests are asymptotical without a non-asymptotic guarantee. Motivated by [2, 104] invents a new methodology for testing the linearity hypothesis in HD linear models, and the test they proposed does not impose any restriction of model sparsity. Based on the concentration of Lipschitz functions of Gaussian distributions or strongly log-concave distribution, [103] developed a new concentration-based test in HD regressions. Recently, [87] studied non-asymptotical two-sample testing using Projected Wasserstein Distance, via McDiarmid's inequality.

In future, it would be essential and practical to study the estimator for the sub-exponential, sub-Gaussian, sub-Weibull and GBO norms as the unknown parameters when constructing non-asymptotical and data-driven confidence intervals.

Acknowledgments

The authors thank Haoyu Wei, Xiang Li, Xiaoyu Lei, Qiuping Wang, Yanpeng Li, Chang Cui, Shengming Zhong, and Han Run for comments on the early versions of this paper. The authors also thank two referees for helpful suggestions. This research is funded by National Natural Science Foundation of China (Grants 92046021, 12071013, 12026607, 71973005), and LMEQF at Peking University.

References

[1] R. Adamczak, *A tail inequality for suprema of unbounded empirical processes with applications to Markov chains*, Electronic Journal of Probability, 13(13) (2008) 1000–1034.

[2] S. Arlot, G. Blanchard, and E. Roquain, *Some nonasymptotic results on resampling in high dimension I: confidence regions*, The Annals of Statistics, 38(1) (2010) 51–82.

[3] T. Aven, *Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables*, Journal of Applied Probability, 22(3) (1985) 723–728.

[4] M. Bagnoli, and T. Bergstrom, *Log-concave probability and its applications*, Economic Theory, 26(2) (2005) 445–469.

[5] Z. D. Bai and Y. Q. Yin, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, The Annals of Probability, 21(3) (1993) 1275–1294.

[6] Y. Baraud and J. Chen, *Robust estimation of a regression function in exponential families*, arXiv:2011.01657, 2020.

[7] P. L. Bartlett, S. Mendelson and J. Neeman, *l1-regularized linear regression: persistence and oracle inequalities*, Probability Theory and Related Fields, 154(1) (2012) 193–224.

[8] P. C. Bellec, *Concentration of quadratic forms under a Bernstein moment assumption*, arXiv:1901.08736, 2019.

[9] G. Bennett, *Probability inequalities for the sum of independent random variables*, Journal of the American Statistical Association, 57(297) (1962) 33–45.

[10] S. Bernstein, *On a modification of Chebyshev's inequality and of the error formula of Laplace*, Uchenye Zapiski Nauch.-Issled. Kaf. Ukraine, Sect. Math, 1(4) (1924) 38–49.

[11] P. J. Bickel, Y. A. Ritov and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, The Annals of Statistics, (2009) 1705–1732.

[12] Y. V. Borovskikh, *U-Statistics in Banach Spaces*, VSB, 1996.

[13] D. Bosq, *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction, 2ed*, Springer, 1998.

[14] S. Boucheron, G. Lugosi and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[15] S. Boucheron and M. Thomas, *Concentration inequalities for order statistics*, Electronic Communications in Probability, 17 (2012) 1–12.

[16] J. Bourgain, *Random points in isotropic convex sets*, Convex Geometric Analysis, 34 (1996) 53–58.

[17] L. D. Brown, *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*, IMS, 1986.

[18] P. Bühlmann and S. A. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[19] F. Bunea, *Honest variable selection in linear and logistic regression models via l1 and l1+ l2 penalization*, Electronic Journal of Statistics, 2 (2008) 1153–1194.

[20] F. Bunea and L. Xiao, *On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA*, Bernoulli, 21(2) (2015) 1200–1230.

[21] E. Candes and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, The Annals of Statistics, 35(6) (2007) 2313–2351.

[22] S. X. Chen and Y. L. Qin, *A two-sample test for high-dimensional data with applications to gene-set testing*, The Annals of Statistics, 38(2) (2010) 808–835.

[23] D. Donoho and A. Montanari, *High dimensional robust m-estimation: Asymptotic variance via approximate message passing*, Probability Theory and Related Fields, 166(3-4) (2016) 935–969.

[24] R. Durrett, *Probability: Theory and Examples, 5ed*, Cambridge University Press, 2019.

[25] A. Dvoretzky, J. Kiefer and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, The Annals of Mathematical Statistics, 27(3) (1956) 642–669.

[26] N. El Karoui, *Random matrices and high-dimensional statistics: beyond covariance matrices*, In Proceedings of the International Congress of Mathematicians, 4 (2018) 2875–2894.

[27] N. El Karoui, D. Bean, P. J. Bickel, C. Lim and B. Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences, 110(36) (2013) 14557–14562.

[28] J. Fan, B. Jiang and Q. Sun, *Hoeffding's lemma for Markov Chains and its applications to statistical learning*, The Journal of Machine Learning Research (in press), 2021.

[29] J. Fan, R. Li, C. H. Zhang and H. Zou, *Statistical Foundations of Data Science*, CRC Press, 2020.

[30] Y. Fan, H. Zhang and T. Yan, *Asymptotic theory for differentially private generalized β-models with parameters increasing*, Statistics and Its Interface, 13(3) (2020) 385–398.

[31] S. Foss, D. Korshunov and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*, Springer, 2011.

[32] R. Fukuda, *Exponential integrability of sub-Gaussian vectors*, Probability Theory and Related Fields, 85(4) (1990) 505–521.

[33] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge University Press, 2016.

[34] C. Giraud, *Introduction to High-Dimensional Statistics*, (Vol. 138), CRC Press, 2014.

[35] R. D. Gordon, *Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument*, The Annals of Mathematical Statistics, 12(3) (1941) 364–366.

[36] F. Götze, H. Sambale and A. Sinulis, *Concentration inequalities for polynomials in α-sub-exponential random variables*, arXiv:1903.05964, 2019.

[37] D. L. Hanson and F. T. Wright, *A bound on tail probabilities for quadratic forms in independent random variables*, The Annals of Mathematical Statistics, 42(3) (1971) 1079–1083.

[38] C. Hillar and A. Wibisono, *Maximum entropy distributions on graphs*, arXiv:1301.3321, 2013.

[39] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, 58(301) (1963) 13–30.

[40] D. J. Hsu, S. M. Kakade and T. Zhang, *A tail inequality for quadratic forms of subgaussian random vectors*, Electronic Communications in Probability, 17 (2012) 1–6.

[41] H. Huang, Y. Gao, H. Zhang and B. Li, *Weighted Lasso Estimates for Sparse Logistic Regression: Non-asymptotic Properties with Measurement Error*, Acta Mathematica Scientia, 2021.

[42] J. Huang, T. Sun, Z. Ying, Y. Yu and C. H. Zhang, *Oracle inequalities for the lasso in the Cox model*, Annals of Statistics, 41(3) (2013) 1142–1165.

[43] K. Jamieson, M. Malloy, R. Nowak and S. Bubeck, *lil'UCB: An optimal exploration algorithm for multi-armed bandits*, Proceedings of The 27-th Conference on Learning Theory, PMLR 35 (2014) 423–439.

[44] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade and M. I. Jordan, *A short note on concentration inequalities for random vectors with subgaussian norm*, arXiv:1902.03736, 2019.

[45] N. L. Johnson, A. W. Kemp and S. Kotz, *Univariate Discrete Distributions, 3ed*, Wiley, 2005.

[46] S. Kakade, O. Shamir, K. Sindharan and A. Tewari, *Learning exponential families in high-dimensions: Strong convexity and sparsity*, In International Conference on Artificial Intelligence and Statistics, (2010) 381–388.

[47] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, Ecole dEté de Probabilités de Saint-Flour XXXVIII-2008 (Vol. 2033),

Springer, 2011.
[48] S. Kong and B. Nan, *Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso*, Statistica Sinica, 24(1) (2014) 25–42.
[49] A. K. Kuchibhotla, *Deterministic inequalities for smooth m-estimators*, arXiv:1809. 05172, 2018.
[50] A. K. Kuchibhotla and A. Chakrabortty, *Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression*, arXiv: 1804.02605, 2018.
[51] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, 1991.
[52] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses, 3rd*, Springer, 2006.
[53] L. Lei, P. J. Bickel and N. El Karoui, *Asymptotics for high dimensional regression M-estimates: fixed design results*, Probability Theory and Related Fields, 172(3-4) (2018) 983–1079.
[54] X. Lei and H. Zhang, *Non-asymptotic optimal prediction error for RKHS-based partially functional linear models*, arXiv:2009.04729, 2020.
[55] S. Levy and P. K. Fullagar, *Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution*, Geophysics, 46(9) (1981) 1235–1243.
[56] Y. Li and J. Jia, *L1 least squares for sparse high-dimensional LDA*, Electronic Journal of Statistics, 11(1) (2017) 2499–2518.
[57] Z. Lin and Z. Bai, *Probability Inequalities*, Springer, 2011.
[58] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices (in Russian)*, Mathematics of the USSR-Sbornik, 1(4) (1967) 457–483.
[59] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, The Annals of Probability, 18(3) (1996) 1269–1283.
[60] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, CRC, 1983.
[61] C. McDiarmid, *On the method of bounded differences*, Surveys in Combinatorics, 141(1) (1989) 148–188.
[62] B. Miasojedow and W. Rejchel, *Sparse estimation in ising model via penalized Monte Carlo methods*, The Journal of Machine Learning Research, 19(1) (2018) 2979–3004.
[63] J. P. Mills, *Table of the ratio: area to bounding ordinate, for any portion of normal curve*, Biometrika, 18(3-4) (1926) 395–400.
[64] X. Pan and W. X. Zhou, *Multiplier bootstrap for quantile regression: non-asymptotic theory under random design*, Information and Inference: A Journal of the IMA, 2020.
[65] V. V. Petrov, *Sums of Independent Random Variables*, (Vol. 82), Springer, 1975.
[66] G. Pisier, *Some applications of the metric entropy condition to harmonic analysis*, In Banach Spaces, Harmonic Analysis, and Probability Theory, Springer, (1983) 123–154.
[67] G. Pistone and H. P. Wynn, *Finitely generated cumulants*, Statistica Sinica, (1999) 1029–1052.

[68] D. Pollard, *Mini Empirical*, http://www.stat.yale.edu/pollard/Books/Mini, 2015.

[69] P. Rigollet, *Kullback-Leibler aggregation and misspecified generalized linear models*, The Annals of Statistics, 40(2) (2012) 639–665.

[70] P. Rigollet and J. C. Hütter, *High dimensional statistics*, http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf, 2019.

[71] L. Rosasco, M. Belkin and E. De Vito, *On learning with integral operators*, Journal of Machine Learning Research, 11 (2010) 905–934.

[72] M. Rudelson and R. Vershynin, *Hanson-Wright inequality and sub-Gaussian concentration*, Electronic Communications in Probability, 18 (2013) 1–9.

[73] A. Saumard and J. A. Wellner, *Log-concavity and strong log-concavity: a review*, Statistics Surveys, 8 (2014) 45–114.

[74] W. Schudy and M. Sviridenko, *Concentration and moment inequalities for polynomials of independent random variables*, In SODA 12 Proceedings of the twenty-third annual ACM-SIAM symposium on discrete algorithms, (2012) 437–446.

[75] B. Sen, *A Gentle Introduction to Empirical Process Theory and Applications*, http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf, 2018.

[76] E. M. Stein and R. Shakarchi, *Complex Analysis (Vol. 2)*, Princeton University Press, 2010.

[77] M. Talagrand, *Sharper bounds for Gaussian and empirical processes*, The Annals of Probability, 22(1) (1994) 28–76.

[78] T. Tao, *Topics in Random Matrix Theory (Vol. 132)*, Providence, RI: American Mathematical Society, 2012.

[79] T. Tao and V. Vu, *Random matrices: Sharp concentration of eigenvalues*, Random Matrices: Theory and Applications, 2(3) (2013) 1350007.

[80] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), 58(1) (1996) 267–288.

[81] A. W. van der Vaart, *Asymptotic Statistics (Vol. 3)*, Cambridge University Press, 1998.

[82] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: with Applications to Statistics*, Springer, 1996.

[83] R. Vershynin, *Introduction to the Non-Asymptotic Analysis of Random Matrices. in Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok*, Cambridge University Press, 2012.

[84] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science (Vol. 47)*, Cambridge University Press, 2018.

[85] M. Vladimirova, S. Girard, H. Nguyen and J. Arbel, *Sub-Weibull Distributions: Generalizing Sub-Gaussian and Sub-Exponential Properties to Heavier-Tailed Distributions*, Stat, 2020.

[86] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint (Vol. 48)*,

Cambridge University Press, 2019.

[87] J. Wang, R. Gao and Y. Xie, *Two-sample test using projected wasserstein distance: breaking the curse of dimensionality*, arXiv:2010.11970, 2020.

[88] J. A. Wellner, *The Bennett-Orlicz norm*, Sankhya A, 79(2) (2017) 355–383.

[89] K. C. Wong, Z. Li and A. Tewari, *Lasso guarantees for β-mixing heavy-tailed time series*, Annals of Statistics, 48(2) (2020) 1124–1142.

[90] F. Xie and Z. Xiao, *Square-root LASSO for high-dimensional sparse linear systems with weakly dependent errors*, Journal of Time Series Analysis, 39(2) (2018) 212–238.

[91] X. Yang, S. Song and H. Zhang, *Law of the iterated logarithm and model selection consistency for independent and dependent GLMs*, Frontiers of Mathematics in China, 2021. (To appear).

[92] Y. Yang, Z. Shang and G. Cheng, *Non-asymptotic analysis for nonparametric testing*, In Conference on Learning Theory, (2020) 3709–3755.

[93] J. Yao, S. Zheng and Z. D. Bai, *Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press, 2015.

[94] F. Ye and C. H. Zhang, *Rate minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls*, Journal of Machine Learning Research, 11(Dec) (2010) 3519–3540.

[95] V. V. Yurinsky, *Sums and Gaussian Vectors*, Springer, 1995.

[96] K. Zajkowski, *On norms in some class of exponential type Orlicz spaces of random variables*, Positivity, (2019) 1–10.

[97] D. X. Zhang, *Tail bounds for the suprema of empirical processes over unbounded classes of functions*, Acta Mathematica Sinica, 22 (2006) 339–345.

[98] H. Zhang and J. Jia, *Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signals detection*, Statistica Sinica, 32(1) 2022. (To appear). `https://doi.org/ 10.5705/ss.202019.0315`.

[99] H. Zhang, Y. Liu and B. Li, *Notes on discrete compound Poisson model with applications to risk theory*, Insurance: Mathematics and Economics, 59 (2014) 325–336.

[100] H. Zhang and H. Wei, *Sharper sub-Weibull concentrations: non-asymptotic Bai-Yin theorem*, arXiv:2102.02450, 2021.

[101] T. Zhang, *Learning bounds for Kernel regression using effective data dimensionality*, Neural Computation, 17(9) (2005) 2077–2098.

[102] P. S. Zhong and S. X. Chen, *Tests for high-dimensional regression coefficients with factorial designs*, Journal of the American Statistical Association, 106(493) (2011) 260–274.

[103] Y. Zhu, *Statistical inference and feasibility determination: a nonasymptotic approach*, arXiv:1808.07127, 2018.

[104] Y. Zhu and J. Bradic, *Linear hypothesis testing in dense high-dimensional linear models*, Journal of the American Statistical Association, 113(524) (2018) 1583–1600.