mathematics

MDPI

Article

Sharper Sub-Weibull Concentrations

Huiming Zhang [1,2,†] and Haoyu Wei [3,*,†]

[1] Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau 999078, China; huimingzhang@um.edu.mo
[2] Zhuhai UM Science & Technology Research Institute, Zhuhai 519031, China
[3] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
* Correspondence: hwei4@ncsu.edu
† These authors contributed equally to this work.

Abstract: Constant-specified and exponential concentration inequalities play an essential role in the finite-sample theory of machine learning and high-dimensional statistics area. We obtain sharper and constants-specified concentration inequalities for the sum of independent sub-Weibull random variables, which leads to a mixture of two tails: sub-Gaussian for small deviations and sub-Weibull for large deviations from the mean. These bounds are new and improve existing bounds with sharper constants. In addition, a new *sub-Weibull parameter* is also proposed, which enables recovering the tight concentration inequality for a random variable (vector). For statistical applications, we give an ℓ_2-error of estimated coefficients in negative binomial regressions when the heavy-tailed covariates are sub-Weibull distributed with sparse structures, which is a new result for negative binomial regressions. In applying random matrices, we derive non-asymptotic versions of Bai-Yin's theorem for sub-Weibull entries with exponential tail bounds. Finally, by demonstrating a sub-Weibull confidence region for a log-truncated Z-estimator without the second-moment condition, we discuss and define the *sub-Weibull type robust estimator* for independent observations $\{X_i\}_{i=1}^n$ without exponential-moment conditions.

Keywords: constants-specified concentration inequalities; exponential tail bounds; heavy-tailed random variables; sub-Weibull parameter; lower bounds on the least singular value

MSC: 60E15; 62F25; 62F99

Citation: Zhang, H.; Wei, H. Sharper Sub-Weibull Concentrations. *Mathematics* **2022**, *10*, 2252. https://doi.org/10.3390/math10132252

Academic Editor: Christophe Chesneau

Received: 21 April 2022
Accepted: 21 June 2022
Published: 27 June 2022

1. Introduction

In the last two decades, with the development of modern data collection methods in science and techniques, scientists and engineers can access and load a huge number of variables in their experiments. Over hundreds of years, probability theory lays the mathematical foundation of statistics. Arising from data-driving problems, various recent statistics research advances also contribute to new and challenging probability problems for further study. For example, in recent years, the rapid development of high-dimensional statistics and machine learning have promoted the development of the probability theory and even pure mathematics, such as random matrices, large deviation inequalities, and geometric functional analysis, etc.; see [1]. More importantly, the concentration inequality (CI) quantifies the concentration of measures that are at the heart of statistical machine learning. Usually, CI quantifies how a random variable (r.v.) X deviates around its mean $EX =: \mu$ by presenting as one-side or two-sided bounds for the tail probability of $X - \mu$

$$\mathrm{P}(X - \mu > t) \text{ or } \mathrm{P}(|X - \mu| > t) \leq \text{some small } \delta, \ \forall\, t \geq 0.$$

The classical statistical models are faced with fixed-dimensional variables only. However, contemporary data science motivates statisticians to pay more attention to studying $p \times p$ random Hessian matrices (or sample covariance matrices, [2]) with $p \to \infty$, arising

from the likelihood functions of high-dimensional regressions with covariates in $\mathbb{R}^p$. When the model dimension increases with sample size, obtaining asymptotic results for the estimator is potentially more challenging than the fixed dimensional case. In statistical machine learning, concentration inequalities (large derivation inequalities) are essential in deriving non-asymptotic error bounds for the proposed estimator; see [3,4]. Over recent decades, researchers have developed remarkable results of matrix concentration inequalities, which focuses on non-asymptotic upper and lower bounds for the largest eigenvalue of a finite sum of random matrices. For a more fascinated introduction, please refer to the book [5].

Motivated from sample covariance matrices, a random matrix is a specific matrix $\mathbf{A}_{p\times p}$ with its entries A_{jk} drawn from some distributions. As $p \to \infty$, random matrix theory mainly focuses on studying the properties of the p eigenvalues of $\mathbf{A}_{p\times p}$, which turn out to have some limit law. Several famous limit laws in random matrix theory are different from the CLT for the summation of independent random variables since the p eigenvalues are dependent and interact with each other. For convergence in distribution, some pioneering works are the Wigner's semicircle law for some symmetric Gaussian matrices' eigenvalues, the Marchenko-Pastur law for Wishart distributed random matrices (sample covariance matrices), and the Tracy-Widom laws for the limit distribution for maximum eigenvalues in Wishart matrices. All these three laws can be regarded as the CLT of random matrix versions. Moreover, the limit law for the empirical spectral density is some circle distribution, which sheds light on the non-communicative behaviors of the random matrix, while the classic limit law in CLT is for normal distribution or infinite divisible distribution. For strong convergence, Bai-Yin's law complements the Marchenko-Pastur law, which asserts that almost surely convergence of the smallest and largest eigenvalue for a sample covariance matrix. The monograph [2] thoroughly introduces the limit law in random matrices.

This work aims to extend non-asymptotic results from sub-Gaussian to sub-Weibull in terms of exponential concentration inequalities with applications in count data regressions, random matrices, and robust estimators. The contributions are:

(i) We review and present some new results for sub-Weibull r.v.s, including sharp concentration inequalities for weighted summations of independent sub-Weibull r.v.s and negative binomial r.v.s, which are useful in many statistical applications.
(ii) Based on the generalized Bernstein-Orlicz norm, a sharper concentration for sub-Weibull summations is obtained in Theorem 1. Here we circumvent Stirling's approximation and derive the inequalities more subtly. As a result, the confidence interval based on our result is sharper and more accurate than that in [6] (For example, see Remark 2) and [7] (see Proposition 1 with unknown constants) gave.
(iii) By sharper sub-Weibull concentrations, we give two applications. First, from the proposed negative binomial concentration inequalities, we obtain the $O_P(\sqrt{p/n})$ (up to some log factors) estimation error for the estimated coefficients in negative binomial regressions under the increasing-dimensional framework $p = p_n$ and heavy-tailed covariates. Second, we provide a non-asymptotic Bai-Yin's theorem for sub-Weibull random matrices with exponential-decay high probability.
(iv) We propose a new *sub-Weibull parameters*, which is enabled of recovering the tight concentration inequality for a single non-zero mean random vector. The simulation studies for estimating sub-Gaussian and sub-exponential parameters show these parameters could be estimated well.
(v) We establish a unified non-asymptotic confidence region and the convergence rate for general log-truncated Z-estimator in Theorem 5. Moreover, we define a sub-Weibull type estimator for a sequence of independent observations $\{X_i\}_{i=1}^n$ without the second-moment condition, beyond the definition of the sub-Gaussian estimator.

2. Sharper Concentrations for Sub-Weibull Summation

Concentration inequalities are powerful in high-dimensional statistical inference, and it can derive explicit non-asymptotic error bounds as a function of sample size, sparsity

level, and dimension [3]. In this section, we present preparation results of concentration inequalities for sub-Weibull random variables.

2.1. Properties of Sub-Weibull Norm and Orlicz-Type Norm

In empirical process theory, sub-Weibull norm (or other Orlicz-type norms) is crucial to derive the tail probability for both single sub-Weibull random variable and summation of random variables (by using the Chernoff's inequality). A benefit of Orlicz-type norms is that the concentration does not need the zero mean assumption.

Definition 1 (Sub-Weibull norm)**.** *For $\theta > 0$, the sub-Weibull norm of X is defined as*

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

The $\|\cdot\|_{\psi_\theta}$ is also called the ψ_θ-norm. We define X as a *sub-Weibull random variable* with *index* θ if it has a bounded ψ_θ-norm (denoted as $X \sim \mathrm{subW}(\theta)$). Actually, the sub-Weibull norm is a special case of Orlicz norms below.

Definition 2 (Orlicz Norms)**.** Let $g : [0,\infty) \to [0,\infty)$ be a non-decreasing convex function with $g(0) = 1$. The "g-Orlicz norm" of a real-valued r.v. X is given by

$$\|X\|_g := \inf\{\eta > 0 : \mathrm{E}[g(|X|/\eta)] \le 2\}. \tag{1}$$

Using exponential Markov's inequality, we have

$$\mathrm{P}(|X| \ge t) = \mathrm{P}(g(|X|/\|X\|_g) \ge g(t/\|X\|_g)) \le g^{-1}(t/\|X\|_g)\mathrm{E}g(X/\|X\|_g) \le 2g^{-1}(t/\|X\|_g) \tag{2}$$

by Definition 2. For example, let $g(x) = e^{x^\theta}$, which leads to sub-Weibull norm for $\theta \ge 1$.

Example 1 (ψ_θ-norm of bounded r.v.)**.** *For a r.v. $|X| \le M < \infty$, we have*

$$\|X\|_{\psi_\theta} = \inf\{t > 0 : \mathrm{E}e^{|X|^\theta / t^\theta} \le 2\} \le \inf\{t > 0 : \mathrm{E}e^{M^\theta / t^\theta} \le 2\} = M(\log 2)^{-1/\theta}.$$

In general, we have following corollary to determine $\|X\|_{\psi_\theta}$ based on moment generating functions (MGF). It would be useful for doing statistical inference of ψ_θ-norm.

Corollary 1. *If $\|X\|_{\psi_\theta} < \infty$, then $\|X\|_{\psi_\theta} = \left(m^{-1}_{|X|^\theta}(2)\right)^{-1/\theta}$ for the MGF $\phi_Z(t) := \mathrm{E}e^{tZ}$.*

Remark 1. *If we observe i.i.d. data $\{X_i\}_{i=1}^n$ from a sub-Weibull distribution, one can use the empirical moment generating function (EMGF, [8]) to estimate the sub-Weibull norm of X. Since the EMGF $\hat{m}_{|X|^\theta}(t) = \frac{1}{n}\sum_{i=1}^n \exp\{t|X_i|^\theta\}$ converge to MGF $m_{|X|^\theta}(t)$ in probability for t in a neighbourhood of zero, the value of the inverse function of EMGF at 2. Then, under some regularity conditions, $\left(\hat{m}_{|X|^\theta}\right)^{-1}(2)$, is a consistent estimate for $\|X\|_{\psi_\theta}$.*

In particular, if we take $\theta = 1$, we get the sub-exponential norm of X, which is defined as $\|X\|_{\psi_1} = \inf\{t > 0 : \mathrm{E}\exp(|X|/t) \le 2\}$. For independent r.v.s $\{X_i\}_{i=1}^n$, if $\mathrm{E}X_i = 0$ and $\|X_i\|_{\psi_1} < \infty$, by Proposition 4.2 in [4], we know $\forall\, t \ge 0$

$$\mathrm{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left\{-\frac{1}{4}\left(\frac{t^2}{\sum_{i=1}^n 2\|X_i\|^2_{\psi_1}} \wedge \frac{t}{\max\limits_{1\le i\le n}\|X_i\|_{\psi_1}}\right)\right\}. \tag{3}$$

Example 2. *An explicitly calculation of the sub-exponential norm is given in [9], they show that Poisson r.v.* $X \sim \text{Poisson}(\lambda)$ *has sub-exponential norm* $\|X\|_{\psi_1} \leq [\log(\log(2)\lambda^{-1}+1)]^{-1}$. *And Example 1 with triangle inequality implies*

$$\|X - \mathrm{E}X\|_{\psi_1} \leq \|X\|_{\psi_1} + \|\mathrm{E}X\|_{\psi_1} = \|X\|_{\psi_1} + \frac{\lambda}{\log 2} \leq [\log(\log(2)\lambda^{-1}+1)]^{-1} + \frac{\lambda}{\log 2}$$

based on following useful results.

Proposition 1 (Lemma A.3 in [9]). *For any* $\alpha > 0$ *and any r.v.s* X, Y *we have* $\|X+Y\|_{\psi_\theta} \leq K_\alpha\big(\|X\|_{\psi_\theta} + \|Y\|_{\psi_\theta}\big)$ *and*

$$\|\mathrm{E}X\|_{\psi_\theta} \leq \frac{1}{d_\alpha(\log 2)^{1/\alpha}}\|X\|_{\psi_\theta}, \quad \|X - \mathrm{E}X\|_{\psi_\theta} \leq K_\alpha\Big(1 + (d_\alpha \log 2)^{-1/\alpha}\Big)\|X\|_{\psi_\theta},$$

where $d_\theta := (\theta e)^{1/\theta}/2$, $K_\theta := 2^{1/\theta}$ *if* $\theta \in (0,1)$ *and* $K_\theta = 1$ *if* $\theta \geq 1$.

To extend Poisson variables, one can also consider concentration for sums of independent heterogeneous negative binomial variables $\{Y_i\}_{i=1}^n$ with probability mass functions:

$$\mathrm{P}(Y_i = y) = \frac{\Gamma(y+k_i)}{\Gamma(k_i)y!}(1-q_i)^{k_i}q_i^y \quad \big(q_i \in (0,1), y \in \mathbb{N}\big), \tag{4}$$

where $\{k_i\}_{i=1}^n \in (0,\infty)$ are variance-dependence parameters. Here, the mean and variance of $\{Y_i\}_{i=1}^n$ are $\mathrm{E}Y_i = \frac{k_i q_i}{1-q_i}$, $\mathrm{Var}\,Y_i = \frac{k_i q_i}{(1-q_i)^2}$ respectively. The MGF of $\{Y_i\}_{i=1}^n$ are $\mathrm{E}e^{sY_i} = \left(\frac{1-q_i}{1-q_i e^s}\right)^{k_i}$ for $i = 1, \cdots, n$. Based on (3), we obtain following results.

Corollary 2. *For any independent r.v.s* $\{Y_i\}_{i=1}^n$ *satisfying* $\|Y_i\|_{\psi_1} < \infty$, $t \geq 0$, *and non-random weight* $\mathbf{w} = (w_1, \cdots, w_n)^\top$, *we have*

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)| \geq t\Big) \leq 2e^{-\frac{1}{4}\left(\frac{t^2}{2\sum_{i=1}^n w_i^2(\|Y_i\|_{\psi_1}+|\mathrm{E}Y_i/\log 2|)^2} \wedge \frac{t}{\max_{1\leq i\leq n}|w_i|(\|Y_i\|_{\psi_1}+|\mathrm{E}Y_i/\log 2|)}\right)}.$$

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)| > 2\Big(2t\sum_{i=1}^n w_i^2\|Y_i - \mathrm{E}Y_i\|_{\psi_1}^2\Big)^{1/2} + 2t\max_{1\leq i\leq n}(|w_i|\|Y_i - \mathrm{E}Y_i\|_{\psi_1})\Big) \leq 2e^{-t}.$$

In particular, if Y_i *is independently distributed as* $\mathrm{NB}(\mu_i, k_i)$, *we have*

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)| \geq t\Big) \leq 2e^{-\frac{1}{4}\left(\frac{t^2}{2\sum_{i=1}^n w_i^2 a^2(\mu_i,k_i)} \wedge \frac{t}{\max_{1\leq i\leq n}|w_i|a(\mu_i,k_i)}\right)}, \tag{5}$$

where $a(\mu_i, k_i) := \left[\log\frac{1-(1-q_i)/\sqrt[k_i]{2}}{q_i}\right]^{-1} + \frac{\mu_i}{\log 2}$ *with* $q_i := \frac{\mu_i}{k_i+\mu_i}$.

Corollary 2 can play an important role in many non-asymptotic analyses of various estimators. For instance, recently [10] uses the above inequality as an essential role for deriving the non-asymptotic behavior of the penalty estimator in the counting data model.

Next, we study moment properties for sub-Weibull random variables. Lemma 1.4 in [11] showed that if $X \sim \mathrm{subG}(\sigma^2)$, then we have: (a). the tail satisfies $\mathrm{P}(|X| > t) \leq 2e^{-t^2/2\sigma^2}$ for any $t > 0$; (b). The (a) implies that moments $\mathrm{E}|X|^k \leq (2\sigma^2)^{k/2}k\Gamma(\frac{k}{2})$ and $[k^{-1/2}(\mathrm{E}(|X|^k))^{1/k}]^2 \leq \sigma^2 e^{2/e}$, $k \geq 2$. We extend Lemma 1.4 in [11] to sub-Weibull r.v. X satisfying following properties.

Corollary 3 (Moment properties of sub-Weibull norm). *(a). If* $\|X\|_{\psi_\theta} < \infty$, *then* $\mathrm{P}\{|X| > t\} \le 2e^{-(t/\|X\|_{\psi_\theta})^\theta}$ *for all* $t \ge 0$; *and then* $\mathrm{E}|X|^k \le 2\|X\|^k_{\psi_\theta}\Gamma(\frac{k}{\theta}+1)$ *for all* $k \ge 1$. *(2). Let* $C_\theta := \max_{k\ge 1}\left(\frac{2\sqrt{2\pi}}{\theta}\right)^{1/k}\left(\frac{k}{\theta}\right)^{1/(2k)}$, *for all* $k \ge 1$ *we have* $(\mathrm{E}|X|^k)^{1/k} \le C_\theta(\theta e^{11/12})^{-1/\theta}\|X\|_{\psi_\theta}k^{1/\theta}$.

Particularly, sub-Weibull r.v.s reduce to sub-exponential or sub-Gaussian r.v.s when θ = 1 or 2. It is obvious that the smaller θ is, the heavier tail the r.v. has. A r.v. is called heavy-tailed if its distribution function fails to be bounded by a decreasing exponential function, i.e.,

$$\int e^{\lambda x}dF(x) = \infty, \forall \lambda > 0 \text{ (the tail decays slower than some exponential r.v.s);}$$

see [12]. Hence for sub-Weibull r.v.s, we usually focus on the the sub-Weibull index $\theta \in (0,1)$. A simple example that the heavy-tailed distributions arises when we work more production on sub-Gaussian r.v.s. Via a power transform of $|X|$, the next corollary explains the relation of sub-Weibull norm with parameter θ and $r\theta$, which is similar to Lemmas 2.7.6 of [1] for sub-exponential norm.

Corollary 4. *For any* $\theta, r \in (0,\infty)$, *if* $X \sim \mathrm{subW}(\theta)$, *then* $|X|^r \sim \mathrm{subW}(\theta/r)$. *Moreover,*

$$\||X|^r\|_{\psi_{\theta/r}} = \|X\|^r_{\psi_\theta}. \tag{6}$$

Conversely, if $X \sim \mathrm{subW}(r\theta)$, *then* $X^r \sim \mathrm{subW}(\theta)$ *with* $\|X^r\|_{\psi_\theta} = \|X\|^r_{\psi_{r\theta}}$.

By Corollary 4, we obtain that d-th root of the absolute value of sub-Gaussian is $\mathrm{subW}(2d)$ by letting $r = 1/d$. Corollary 4 can be extended to product of r.v.s, from Proposition D.2 in [6] with the equality replacing by inequality, we state it as the following proposition.

Proposition 2. *If* $\{W_i\}_{i=1}^d$ *are (possibly dependent) r.vs satisfying* $\|W_i\|_{\psi_{\alpha_i}} < \infty$ *for some* $\alpha_i > 0$, *then*

$$\|\prod_{i=1}^d W_i\|_{\psi_\beta} \le \prod_{i=1}^d \|W_i\|_{\psi_{\alpha_i}} \text{ where } \frac{1}{\beta} := \sum_{i=1}^d \frac{1}{\alpha_i}.$$

For multi-armed bandit problems in reinforcement learning, [7] move beyond sub-Gaussianity and consider the reward under sub-Weibull distribution which has a much weaker tail. The corresponding concentration inequality (Theorem 3.1 in [7]) for the sum of independent sub-Weibull r.v.s is illustrated as follows.

Proposition 3 (Concentration inequality for sub-Weibull distribution). *Suppose* $\{X_i\}_{i=1}^n$ *are independent sub-Weibull random variables with* $\|X_i - \mathrm{E}X_i\|_{\psi_\theta} \le v$. *Then there exists absolute constants* $C_{1\theta}$ *and* $C_{2\theta}$ *only depending on* θ *such that with probability at least* $1 - e^{-t}$:

$$\left|\frac{1}{n}\sum_{i=1}^n \frac{X_i - \mathrm{E}X_i}{v}\right| \le C_{1\theta}\left(\frac{t}{n}\right)^{1/2} + C_{2\theta}\left(\frac{t}{n}\right)^{1/\theta} = \begin{cases} O(n^{-1/\theta}), \theta > 2 \\ O(n^{-1/2}), 0 < \theta \le 2 \end{cases}.$$

The weakness in the Proposition 3 is that the upper bound of $S_n^a := \sum_{i=1}^n a_iY_i - \mathrm{E}(\sum_{i=1}^n a_iY_i)$ is up to a unknown constants $C_{1\theta}, C_{2\theta}$. In the next section, we will give a constants-specified and high probability upper bound for $|S_n^a|$, which improve Proposition 3 and is sharper than Theorem 3.1 in [6].

2.2. Main Results: Concentrations for Sub-Weibull Summation

Based on the exponential moment condition, the Chernoff's tricks implies the following sub-exponential concentrations from Proposition 4.2 in [4].

Proposition 4. *For any independent r.v.s* $\{Y_i\}_{i=1}^n$ *satisfying* $\|Y_i\|_{\psi_1} < \infty$, $t \geq 0$, *and non-random weight* $\mathbf{w} = (w_1, \cdots, w_n)^\top$, *we have*

$$\mathrm{P}(|\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)| > 2(2t\sum_{i=1}^n w_i^2\|Y_i - \mathrm{E}Y_i\|_{\psi_1}^2)^{1/2} + 2t\max_{1\leq i\leq n}(|w_i|\|Y_i - \mathrm{E}Y_i\|_{\psi_1})) \leq 2e^{-t}.$$

But it is not easy to extend to sub-Weibull distributions. From Corollary 4, $Y_i \sim \mathrm{subW}(\theta) \Rightarrow |Y_i|^{1/\theta} \sim \mathrm{subW}(1)$. The MGF of $|Y_i|^{1/\theta}$ satisfies $\mathrm{E}e^{\lambda^{1/\theta}|Y_i|^{1/\theta}} \leq e^{\lambda^{1/\theta}K^{1/\theta}}$, $|\lambda| \leq \frac{1}{K}$ for some constant $K > 0$. The bound of $\mathrm{E}e^{\lambda^{1/\theta}|Y_i|^{1/\theta}}$ with $\theta \neq 1$ or 2 is not directly applicable for deriving the concentration of $\sum_{i=1}^n w_i(Y_i - \mathrm{E}Y_i)$ by using the independence and Chernoff's tricks, since the MGF of Weibull r.v. do not has closed form as exponential function. Thanks to the tail probability derived by Orlicz-type norms, instead of using the upper bound for MGF, an alternative method is given by [6] who defines the so-called Generalized Bernstein-Orlicz (GBO) norm. And the GBO norm can help us to derive tail behaviours for sub-Weibull r.v.s.

Definition 3 (GBO norm)**.** *Fix* $\alpha > 0$ *and* $L \geq 0$. *Define the function* $\Psi_{\theta,L}(\cdot)$ *as the inverse function* $\Psi_{\theta,L}^{-1}(t) := \sqrt{\log(t+1)} + L(\log(t+1))^{1/\theta}$ *for all* $t \geq 0$. *The GBO norm of a r.v.* 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 r.v.s. Like the sub-Weibull norm properties Corollary 3, the following proposition in [6] allows us to get the concentration inequality for r.v. with finite GBO norm.

Proposition 5. *If* $\|X\|_{\Psi_{\theta,L}} < \infty$, *then* $\mathrm{P}(|X| \geq \|X\|_{\Psi_{\theta,L}}\{\sqrt{t} + Lt^{1/\theta}\}) \leq 2e^{-t}\ \forall\ t \geq 0$.

With an upper bound of GBO norm, we could easily derive the concentration inequality for a single sub-Weibull r.v. or even the sum of independent sub-Weibull r.v.s. The sharper upper bounds for the GBO norm is obtained for the sub-Weibull summation, which refines the constant in the sub-Weibull concentration inequality. Let $||X||_p := (\mathrm{E}|X|^p)^{1/p}$ for all integer $p \geq 1$. First, by truncating more precisely, we obtain a sharper upper bound for $||X||_p$, comparing to Proposition C.1 in [6].

Corollary 5. *If* $\|X\|_p \leq C_1\sqrt{p} + C_2 p^{1/\theta}$ *for* $p \geq 2$ *and constants* C_1, C_2, *then*

$$\|X\|_{\Psi_{\theta,K}} \leq \gamma e C_1$$

where $K = \gamma^{2/\theta}C_2/(\gamma C_1)$ *and* $\gamma \approx 1.78$ *is the minimal solution of*

$$\left\{k > 1 : e^{2k^{-2}} - 1 + \frac{e^{2(1-k^2)/k^2}}{k^2 - 1} \leq 1\right\}.$$

The proof can be seen in the Appendix A. In below, we need the moment estimation for sums of independent symmetric r.v.s.

Lemma 1 (Khinchin-Kahane Inequality, Theorem 1.3.1 of [13])**.** *Let* $\{a_i\}_{i=1}^n$ *be a finite non-random sequence,* $\{\varepsilon_i\}_{i=1}^n$ *be a sequence of independent Rademacher variables and* $1 < p < q < \infty$. *Then* $\|\sum_{i=1}^n \varepsilon_i a_i\|_q \leq \left(\frac{q-1}{p-1}\right)^{1/2}\|\sum_{i=1}^n \varepsilon_i a_i\|_p$.

Lemma 2 (Theorem 2 of [14])**.** *Let* $\{X_i\}_{i=1}^n$ *be a sequence of independent* symmetric *r.v.s, and* $p \geq 2$. *Then,* $\frac{e-1}{2e^2}\|(X_i)\|_p \leq \|X_1 + \cdots + X_n\|_p \leq e\|(X_i)\|_p$, *where* $\|(X_i)\|_p := \inf\{t > 0 : \sum_{i=1}^n \log\phi_p(X_i/t) \leq p\}$ *with* $\phi_p(X) := \mathrm{E}|1 + X|^p$.

Lemma 3 (Example 3.2 and 3.3 of [14]). *Assume X be a symmetric r.v. satisfying $\mathrm{P}(|X| \geq t) = e^{-N(t)}$. For any $t \geq 0$, we have*

(a) *If $N(t)$ is concave, then $\log \phi_p(e^{-2}tX) \leq pM_{p,X}(t) := (t^p \|X\|_p^p) \vee (pt^2\|X\|_2^2)$.*

(b) *For convex $N(t)$, denote the convex conjugate function $N^*(t) := \sup_{s>0}\{ts - N(s)\}$ and $M_{p,X}(t) = \begin{cases} p^{-1}N^*(p|t|), & \text{if } p|t| \geq 2 \\ pt^2, & \text{if } p|t| < 2. \end{cases}$ Then $\log \phi_p(tX/4) \leq pM_{p,X}(t)$.*

With the help of three lemmas above, we can obtain the main results concerning the shaper and constant-specified concentration inequality for the sum of independent sub-Weibull r.v.s.

Theorem 1 (Concentration for sub-Weibull summation). *Let γ be given in Corollary 5. If $\{X_i\}_{i=1}^n$ are independent centralized r.v.s such that $\|X_i\|_{\psi_\theta} < \infty$ for all $1 \leq i \leq n$ and some $\theta > 0$, then for any weight vector $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$, the following bounds holds true:*

(a) *The estimate for GBO norm of the summation:*

$$\|\textstyle\sum_{i=1}^n w_i X_i\|_{\Psi_{\theta, L_n(\theta, \boldsymbol{b}_X)}} \leq \gamma e C(\theta) \|\boldsymbol{b}_X\|_2,$$

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

$$C(\theta) := \begin{cases} 2\Big[\log^{1/\theta} 2 + e^3\Big(\Gamma^{1/2}\big(\frac{2}{\theta}+1\big) + 3^{\frac{2-\theta}{3\theta}} \sup_{p\geq 2} p^{-\frac{1}{\theta}}\Gamma^{1/p}\big(\frac{p}{\theta}+1\big)\Big)\Big], & \text{if } \theta \leq 1, \\ 2[4e + (\log 2)^{1/\theta}], & \text{if } \theta > 1; \end{cases}$$

and $L_n(\theta, \boldsymbol{b}) = \gamma^{2/\theta} A(\theta) \frac{\|\boldsymbol{b}\|_\infty}{\|\boldsymbol{b}\|_2} 1\{0 < \theta \leq 1\} + \gamma^{2/\theta} B(\theta) \frac{\|\boldsymbol{b}\|_\beta}{\|\boldsymbol{b}\|_2} 1\{\theta > 1\}$ where $B(\theta) =: \frac{2e\theta^{-1/\theta}\big(1-\theta^{-1}\big)^{1/\beta}}{4e+(\log 2)^{1/\theta}}$ and $A(\theta) =: \inf_{p\geq 2} \frac{e^3 3^{\frac{2-\theta}{3\theta}} p^{-1/\theta}\Gamma^{1/p}\big(\frac{p}{\theta}+1\big)}{2[\log^{1/\theta} 2 + e^3(\Gamma^{1/2}(\frac{2}{\theta}+1) + 3^{\frac{2-\theta}{3\theta}} \sup_{p\geq 2} p^{-1/\theta}\Gamma^{1/p}(\frac{p}{\theta}+1))]}$. For the case $\theta > 1$, β is the Hölder conjugate satisfying $1/\theta + 1/\beta = 1$.

(b) *Concentration for sub-Weibull summation:*

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i X_i| \geq 2eC(\theta)\|\boldsymbol{b}_X\|_2\{\sqrt{t} + L_n(\theta, \boldsymbol{b}_X)t^{1/\theta}\}\Big) \leq 2e^{-t}. \tag{7}$$

(c) *Another form of for $\theta \neq 2$:*

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i X_i| \geq s\Big) \leq 2\exp\left\{-\left(\frac{s^\theta}{\big[4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n(\theta, \boldsymbol{b}_X)\big]^\theta} \wedge \frac{s^2}{16e^2C^2(\theta)\|\boldsymbol{b}_X\|_2^2}\right)\right\}$$

$$(\theta < 2) = \begin{cases} 2e^{-s^2/16e^2C^2(\theta)\|b\|_2^2}, \text{ if } s \leq 4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n^{\theta/(\theta-2)}(\theta, \boldsymbol{b}_X) \\ 2e^{-s^\theta/[4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n(\theta, \boldsymbol{b}_X)]^\theta}, \text{ if } s > 4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n^{\theta/(\theta-2)}(\theta, \boldsymbol{b}_X); \end{cases}$$

$$(\theta > 2) = \begin{cases} 2e^{-s^\theta/[4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n(\theta, \boldsymbol{b}_X)]^\theta}, \text{ if } s < 4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n^{\theta/(2-\theta)}(\theta, \boldsymbol{b}_X) \\ 2e^{-s^2/16e^2C^2(\theta)\|\boldsymbol{b}_X\|_2^2}, \text{ if } s \geq 4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n^{\theta/(2-\theta)}(\theta, \boldsymbol{b}_X). \end{cases}$$

Remark 2. *The constant $C(\theta)$ in Theorem 1 can be improved as $C(\theta)/2$ under symmetric assumption of sub-Weibull r.v.s $\{X_i\}_{i=1}^n$. Moreover, by the improved symmetrization theorem (Theorem 3.4 in [15]), one can replace the constant $C(\theta)$ in Theorem 1 by a sharper constant $(1 + o(1))C(\theta)/2$. Theorem 1 (b) also implies a potential empirical upper bound for $\sum_{i=1}^n w_i X_i$ for independent sub-Weibull r.v.s $\{X_i\}_{i=1}^n$, because the only unknown variable in $2eC(\theta)\|\boldsymbol{b}_X\|_2\{\sqrt{t} + L_n(\theta)t^{1/\theta}\}$ is $\boldsymbol{b}_X$. From Remark 1, estimating $\boldsymbol{b}_X$ is possible for i.i.d. observation $\{X_i\}_{i=1}^n$.*

Remark 3. *Compared with the newest result in [6], our method do not use the crude String's approximation will give sharper concentration. For example, suppose $X_1, \ldots, X_{10}$ are i.i.d. r.v.s with mean μ and $\|X_1 - \mu\|_{\psi_\theta} = 1$. Here we set $\theta = 0.5$, X is heavy-tailed (for example set the*

density of X as $f(x) = \frac{1}{2\sqrt{x}} e^{-\sqrt{x}} \cdot 1(x \geq 0)$*). We find that* $C(\theta) \approx 2825.89$, $A(\theta) \approx 0.07$, *and* $L_{10}(\theta, \mathbf{1}_{10}^{\top}) = 0.23$. *Hence, 95% confidence interval in our method will be*

$$\mu \in \overline{X} \pm 2e \times 2118.80,$$

while the 95% confidence interval in Theorem 3.1 of [6] is evaluated as

$$\mu \in \overline{X} \pm 2e \times 3969.94.$$

In this example, it can be seen that our method does give a much better (tighter) confidence interval.

Remark 4. *Theorem 1 (b) generalizes the sub-Gaussian concentration inequalities, sub-exponential concentration inequalities, and Bernstein's concentration inequalities with Bernstein's moment condition. For* $\theta < 2$ *in Theorem 1 (c), the tail behaviour of the sum is akin to a sub-Gaussian tail for small t, and the tail resembles the exponential tail for large t; For* $\theta > 2$*, the tail behaves like a Weibull r.v. with tail parameter* θ *and the tail of sums match that of the sub-Gaussian tail for large t. The intuition is that the sum will concentrate around zero by the Law of Large Number. Theorem 1 shows that the convergence rate will be faster for small deviations from the mean and will be slower for large deviations from the mean.*

Remark 5. *Recently, similar result presented in [16] is that*

$$\mathrm{P}\left(\left| \sum_{i=1}^{n} X_i \right| > x \right) \leq \exp\left\{ - \left(\frac{x}{nK_\theta} \right)^{1/\theta} \right\}, \text{ for } x \geq nK_\theta$$

where K_θ *is some constants only depends on X and* θ (K_θ *can be obtained by Proposition 3). But it is obvious to see this large derivation result cannot guarantee a* $\sqrt{n}$*-convergence rate (as presented in Proposition 3) whereas our result always give a* $\sqrt{n}$*-convergence rate, as presented in Theorem 1 (c) and Proposition 3.*

2.3. Sub-Weibull Parameter

In this part, a new *sub-Weibull parameters* is proposed, which is enable of recovering the tight concentration inequality for single non-zero mean random vector. Similar to characterizations of sub-Gaussian r.vs. in Proposition 2.5.2 of [1], sub-Weibull r.vs. has the equivalent definitions.

Proposition 6 (Characterizations of sub-Weibull r.v., [17])**.** *Let X be a r.v., then the following properties are equivalent. (1). The tails of X satisfy* $\mathrm{P}(|X| \geq x) \leq e^{-(x/K_1)^\theta}$*, for all* $x \geq 0$*; (2). The moments of X satisfy* $\|X\|_k := (\mathrm{E}|X|^k)^{1/k} \leq K_2 k^{1/\theta}$ *for all* $k \geq 1 \wedge \theta$*; (3). The MGF of* $|X|^{1/\theta}$ *satisfies* $\mathrm{E}e^{\lambda^{1/\theta}|X|^{1/\theta}} \leq e^{\lambda^{1/\theta} K_3^{1/\theta}}$ *for* $|\lambda| \leq \frac{1}{K_3}$*; (4).* $\mathrm{E}e^{|X/K_4|^{1/\theta}} \leq 2$.

From the upper bound of $(\mathrm{E}|X|^k)^{1/k}$ in Proposition 6(2), an alternative definition of the sub-Weibull norm $\|X\|_{\psi_\theta} := \sup_{k \geq 1} k^{-1/\theta} (\mathrm{E}|X|^k)^{1/k}$ is given by [17]. Let $\theta = 1$. An alternative definition of the sub-exponential norm is $\|X\|_{\psi_1} := \sup_{k \geq 1} k^{-1} (\mathrm{E}|X|^k)^{1/k}$ see Proposition 2.7.1 of [1]. The sub-exponential r.v. X satisfies equivalent properties in Proposition 6 (Characterizations of sub-exponential with $\theta = 1$). However, these definition is not enough to obtain the sharp parameter as presented in the sub-Gaussian case. Here, we redefine the sub-Weibull parameter by our Corollary 3(a).

Definition 4 (Sub-Weibull r.v.,$X \sim \mathrm{subW}(\theta, v)$)**.** *Define the sub-Weibull norm*

$$\|X\|_{\varphi_\theta} = \sup_{k \geq 1} \left(\mathrm{E}|X|^{\theta k} / k! \right)^{1/(\theta k)}.$$

We denote the sub-Weibull r.v. as $X \sim \text{subW}(\theta, v)$ if $v = \|X\|_{\varphi_\theta} < \infty$ for a given $\theta > 0$. For $\theta \geq 1$, the $\|\cdot\|_{\varphi_\theta}$ is a norm which satisfies triangle inequality by Minkowski's inequality: $\text{E}(|X+Y|^r)^{1/r} \leq [\text{E}(|X|^r)]^{1/r} + [\text{E}(|Y|^r)]^{1/r}, (r \geq 1)$ comparing to Proposition 1. Definition 4 is free of bounding MGF, and it avoids Stirling's approximation in the proof of the tail inequality. We obtain following main results for this moment-based norm.

Corollary 6. *If* $\|X\|_{\varphi_\theta} < \infty$, *then* $\text{P}\{|X| > t\} \leq 2\exp\{-\frac{t^\theta}{2\|X\|^\theta_{\varphi_\theta}}\}$ *for all* $t \geq 0$.

Theorem 2 (sub-Weibull concentration)**.** *Suppose that there are n independent sub-Weibull r.v.s* $X_i \sim \text{subW}(\theta, v_i)$ *for* $i = 1, 2, \cdots, n$. *We have*

$$\text{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq \exp\left\{-\frac{\theta e^{11/12} t^\theta}{2[e(\sum_{i=1}^n v_i)C_\theta]^\theta}\right\}, \text{ for } t \geq e(\sum_{i=1}^n v_i)C_\theta(2^{-1}\theta e^{11/12})^{-1/\theta},$$

and $\text{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \leq e\bar{v}2^{1/\theta}C_\theta\left(\frac{\log(\alpha^{-1})}{\theta e^{11/12}}\right)^{1/\theta}\right) \geq 1-\alpha \in (1-e^{-1}, 1]$. *Moreover, we have*

$$\text{P}\left(|\sum_{i=1}^n X_i| \geq e(\sum_{i=1}^n (\text{E}|X_i|)^t)^{1/t} + e(\sum_{i=1}^n v_i)2^{1/\theta}C_\theta(\frac{t}{\theta e^{11/12}})^{1/\theta}\right) \leq e^{-t}, \ \forall\, t \geq 0.$$

The proof of Theorem 2 can be seen in Appendix A.8. The concentration in this Theorem 2 will serve a critical role in many statistical and machine learning literature. For instance, the sub-Weibull concentrations in [7] contain unknown parameters, which makes the algorithm for general sub-Weibull random rewards is infeasible. However, when using our results, it will become feasible as we give explicit constants in these concentrations.

Importantly, the sub-exponential parameter is a special case of sub-Weibull norm by letting $\theta = 1$. Denote the **sub-exponential parameter** for r.v X as

$$\|X\|_{\varphi_1} := \sup_{k \geq 1}\left(\frac{\text{E}|X|^k}{k!}\right)^{1/k}.$$

We denote $X \sim \text{sE}_{\varphi_1}(v)$ if $v = \|X\|_{\varphi_2}$. For exponential r.v. $X \sim \text{Exp}(\mu)$, the moment is $\text{E}X^k = k!\lambda^k$ and $\|X\|_{\varphi_1} = \lambda$. Another case of sub-Weibull norm is $\theta = 2$, which defines **sub-Gaussian parameter**:

$$\|X\|_{\varphi_2} := \sup_{k \geq 1}\left(\frac{\text{E}|X|^{2k}}{k!}\right)^{1/2k} \geq (\text{Var}\, X)^{1/2}.$$

Like the generalized method of moments, we can give the higher-moment estimation procedure for the norm $\|X\|_{\varphi_2}$. Unfortunately, the method in Remark 1 for estimating MGF is not stable in the simulation since the exponential function has a massive variance in some cases.

- **Estimation procedure for $\|X\|_{\varphi_2}$ and $\|X\|_{\varphi_1}$.** *Consider*

$$\widehat{\|X\|}_{\varphi_2} = \sup_{k \geq 1}\left(\frac{1}{n \times k!}\sum_{i=1}^n |X_i|^{2k}\right)^{1/(2k)}, \widehat{\|X\|}_{\varphi_1} = \sup_{k \geq 2}\left(\frac{1}{k!} \cdot \frac{1}{n}\sum_{i=1}^n |X_i|^k\right)^{1/k} \quad (8)$$

as a discrete optimization problem. We can take $k_{\max}$ *big enough to minimize*

$$\left(\tfrac{1}{n\times k!}\textstyle\sum_{i=1}^n |X_i|^{2k}\right)^{1/(2k)}, \left(\tfrac{1}{k!} \cdot \tfrac{1}{n}\textstyle\sum_{i=1}^n |X_i|^k\right)^{1/k} \text{ on } k \in \{1, \ldots, p_{\max}\}.$$

At the first glimpse, the bigger p is, the larger n is required in this method. Nonetheless, often, most of common distributions only require a median-size of p to give a relatively

good result, then only the median-size of n in turn is required. For standard Gaussian random, centralized Bernoulli (successful probability $\mu = 0.3$), and uniform distributed (on $[-1, 1]$) variable X,

$$\|X\|_{\varphi_2} = \sqrt{2}\left[\frac{\Gamma\left((1+p)/2\right)}{\Gamma(1/2)\Gamma(1+p/2)}\right]^{1/p}, \quad \left[\frac{\mu(1-\mu)^p + (1-\mu)\mu^p}{\Gamma(p/2+1)}\right]^{1/p}, \quad \frac{\Gamma^{-1/p}(p/2+1)}{(p+1)^{1/p}}.$$

It can be shown that $\|X\|_{\varphi_2} \approx 1, 0.4582576, 0.5773503$. The Figures 1–3 show the estimated value from different n under estimate method (8) for the three distributions mentioned above. The estimate method (8) is a correct estimated method for sub-Gaussian parameter to our best knowledge.

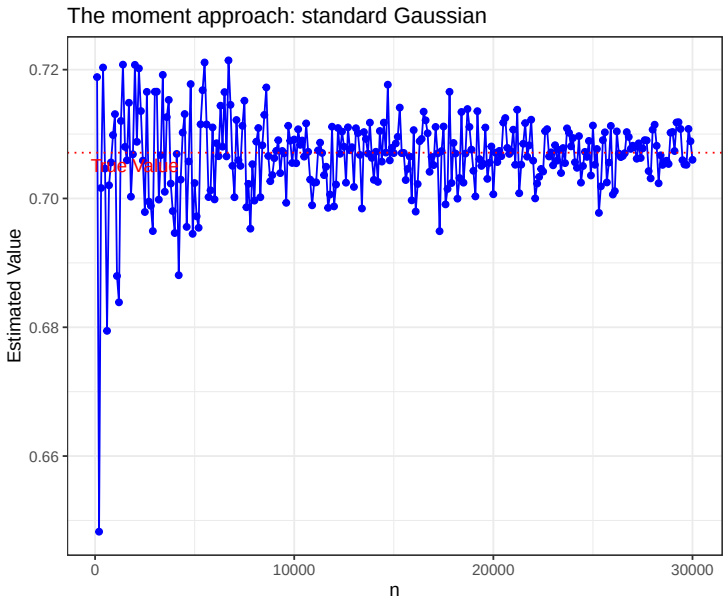

Figure 1. standard Gaussian.

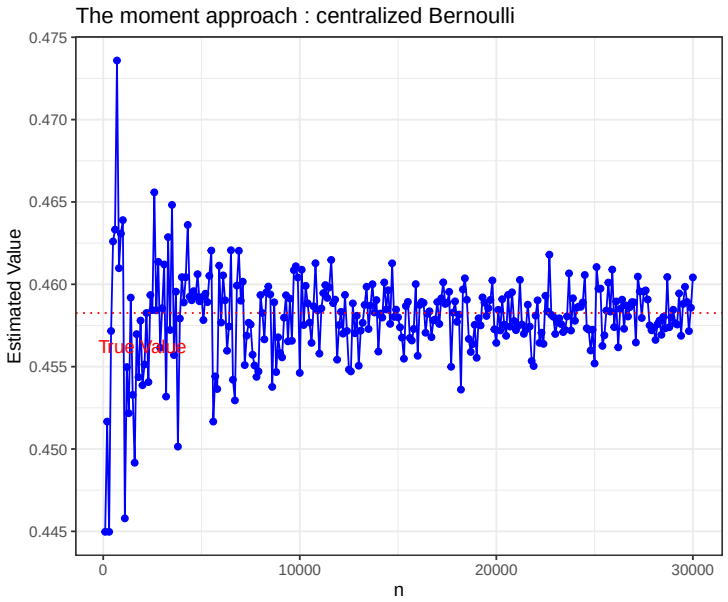

Figure 2. centralized Bernoulli.

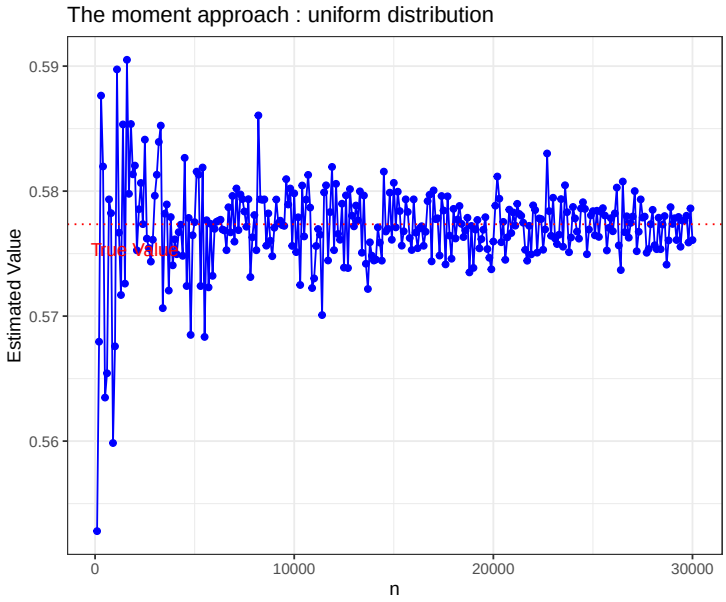

Figure 3. Uniform on $[-1, 1]$.

For centralized negative binomial, and centralized Poisson ($\lambda = 1$) variable X, $\|X\|_{\varphi_1} = 2.460938, 0.7357589$, respectively. The Figures 4 and 5 show the estimated value from different n under estimate method (8) for the four distributions mentioned above.

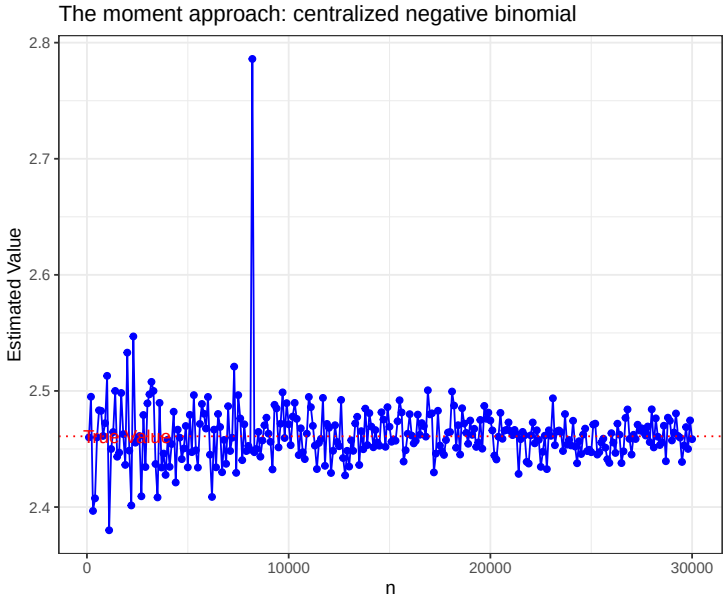

Figure 4. centralized negative binomial.

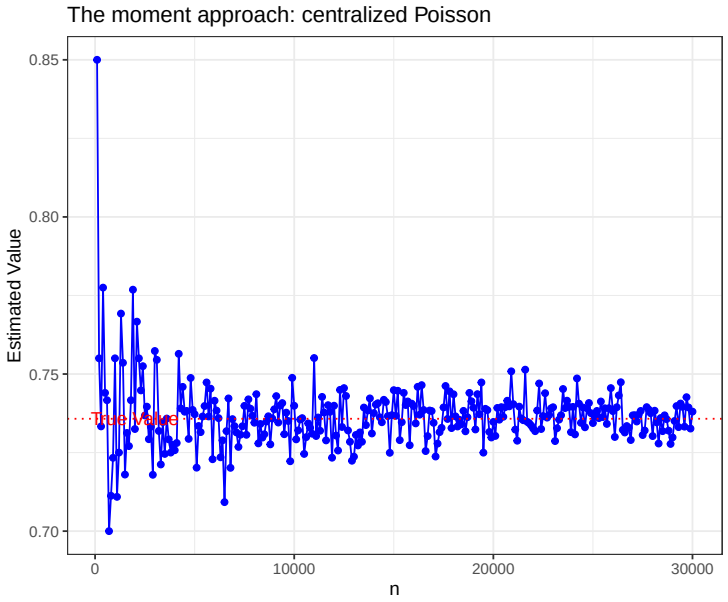

Figure 5. centralized Poisson.

The five figures mentioned above show litter bias between the estimated norm and true norm. It is worthy to note that the norm estimator for centralized negative binomial case has a peak point. This is caused by sub-exponential distributions having relatively heavy tails, and hence the norm estimation may not robust as that in sub-Gaussian under relatively small sample sizes.

Moreover, sub-Gaussian and sub-exponential parameter is extensible for random vectors with values in a normed space $(\mathcal{X}, \|\cdot\|)$, we define *norm-sub-Gaussian parameter* and *norm-sub-exponential parameter*: The norm-sub-Gaussian parameter:

$$\|\boldsymbol{X}\|_{\varphi_2} = \sup_{k\geq 1}(k!)^{-1/(2k)}\left(\mathrm{E}\|\boldsymbol{X}\|^{2k}\right)^{1/(2k)};$$

the norm-sub-exponential parameter:

$$\|\boldsymbol{X}\|_{\varphi_1} = \sup_{k\geq 1}(k!)^{-1/k}\left(\mathrm{E}\|\boldsymbol{X}\|^{k}\right)^{1/k}.$$

We denote $\boldsymbol{X} \sim \mathrm{nsubG}_{\varphi_1}(\sigma^2)$ and $\boldsymbol{X} \sim \mathrm{nsubG}_{\varphi_2}(\sigma^2)$ for $\sigma^2 = \|\boldsymbol{X}\|_{\varphi_2}$ and $\|\boldsymbol{X}\|_{\varphi_1}$, respectively.

3. Statistical Applications of Sub-Weibull Concentrations

3.1. Negative Binomial Regressions with Heavy-Tail Covariates

In statistical regression analysis, the responses $\{Y_i\}_{i=1}^n$ in linear regressions are assume to be continuous Gaussian variables. However, the category in classification or grouping may be infinite with index by the non-negative integers. The categorical variables is treated as countable responses for distinction categories or groups; sometimes it can be infinite. In practice, random count responses include the number of patients, the bacterium in the unit region, or stars in the sky and so on. The responses $\{Y_i\}_{i=1}^n$ with covariates $\{X_i\}_{i=1}^n$ belongs to generalized linear regressions. We consider i.i.d. random variables $\{(X_i, Y_i)\}_{i=1}^n \sim (X, Y) \in \mathbb{R}^p \times \mathbb{N}$. By the methods of the maximum likelihood or the M-estimation, the estimator $\hat{\beta}_n$ is given by

$$\hat{\beta}_n := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top \beta, Y_i), \tag{9}$$

where the loss function $\ell(\cdot,\cdot)$ is convex and twice differentiable in the first argument.

In high-dimensional regressions, the dimension $\boldsymbol{\beta}$ may be growing with sample size n. When $\{Y_i\}_{i=1}^n$ belongs to the exponential family, [18] studied the asymptotic behavior of $\hat{\beta}_n$ in the generalized linear models (GLMs) as $p_n := \dim(X)$ is increasing. In our study, we focus on the case that the covariates is subW(θ) heavy-tailed for $\theta < 1$.

The target vector $\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \mathrm{E}\ell(X^T\beta, Y)$ is assumed to be the loss under the population expectation, comparing to (9). Let $\dot{\ell}(u,y) := \frac{\partial}{\partial t}\ell(t,y)\Big|_{t=u}$, $\ddot{\ell}(u,y) := \frac{\partial}{\partial t}\dot{\ell}(t,y)\Big|_{t=u}$ and $C(u,y) := \sup_{|s-t|\le u} \frac{\ddot{\ell}(s,y)}{\ddot{\ell}(t,y)}$. Finally, define the score function and Hessian matrix of the empirical loss function are $\hat{\mathcal{Z}}_n(\beta) := \frac{1}{n}\sum_{i=1}^n \dot{\ell}(X_i^T\beta, Y_i)X_i$ and $\hat{\mathcal{Q}}_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ddot{\ell}(X_i^T\beta, Y_i)X_iX_i^T$, respectively. The population version of Hessian matrix is $\mathcal{Q}(\beta) := \mathrm{E}[\ddot{\ell}(X^T\beta, Y)XX^T]$. The following so-called determining inequalities guarantee the ℓ_2-error for the estimator obtained from the smooth M-estimator defined as (9).

Lemma 4 (Corollary 3.1 in [19]). *Let $\delta_n(\beta) := \frac{3}{2}\|[\hat{\mathcal{Q}}_n(\beta)]^{-1}\hat{\mathcal{Z}}_n(\beta)\|_2$ for $\beta \in \mathbb{R}^p$. If $\ell(\cdot,\cdot)$ is a twice differentiable function that is convex in the first argument and for some $\beta^* \in \mathbb{R}^p$: $\max_{1\le i\le n} C(\|X_i\|_2\delta_n(\beta^*), Y_i) \le \frac{4}{3}$. Then there exists a vector $\hat{\beta}_n \in \mathbb{R}^p$ satisfying $\hat{\mathcal{Z}}_n(\hat{\beta}_n) = 0$ as the estimating equation of (9),*

$$\frac{1}{2}\delta_n(\beta^*) \le \|\hat{\beta}_n - \beta^*\|_2 \le \delta_n(\beta^*).$$

Applications of Lemma 4 in regression analysis is of special interest when X is heavy tailed, i.e., the sub-Weibull index $\theta < 1$. For the negative binomial regression (NBR) with the known dispersion parameter $k > 0$, the loss function is

$$\ell(u,y) = -yu + (y+k)\log(k+e^u). \tag{10}$$

Thus we have $\dot{\ell}(u,y) = -\frac{k(y-e^u)}{k+e^u}$, $\ddot{\ell}(u,y) = \frac{k(y+k)e^u}{(k+e^u)^2}$, see [20] for details.

Further computation gives $C(u,y) = \sup_{|s-t|\le u} \frac{e^s(k+e^t)^2}{(k+e^s)^2e^t}$ and it implies that $C(u,y) \le e^{3u}$. Therefore, condition $\max_{1\le i\le n} C(\|X_i\|_2\delta_n(\beta^*), Y_i) \le \frac{4}{3}$ in Lemma 4 leads to

$$\max_{1\le i\le n}\|X_i\|_2\delta_n(\beta^*) \le \frac{\log(4/3)}{3}.$$

This condition need the assumption of the design space for $\max_{1\le i\le n}\|X_i\|_2$.

In NBR with loss (10), one has

$$\widehat{\mathcal{Q}}_n(\beta^*) := \tfrac{1}{n}\sum_{i=1}^n \frac{(Y_i+k)ke^{X_i^\top\beta^*}X_iX_i^\top}{(k+e^{X_i^\top\beta^*})^2} \text{ and } \widehat{\mathcal{Z}}_n(\beta^*) := \tfrac{-1}{n}\sum_{i=1}^n \frac{k(Y_i-e^{X_i^\top\beta^*})X_i}{k+e^{X_i^\top\beta^*}}.$$

To guarantee that $\hat{\beta}_n$ approximates β^* well, some regularity conditions are required.

- (C.1): For $M_Y, M_X > 0$, assume $\max_{1\le i\le n}\|Y_i\|_{\psi_1} \le M_Y$ and the heavy-tailed covariates $\{X_{ik}\}$ are uniformly sub-Weibull with $\max_{1\le i\le n,1\le k\le p}\|X_{ik}\|_{\psi_\theta} \le M_X$ for $0 < \theta < 1$.
- (C.2): The vector X_i is sparse or bounded. Let $\mathcal{F}_Y := \{\max_{1\le i\le n}\mathrm{E}Y_i = \max_{1\le i\le n} e^{X_i^\top\beta^*} \le B, \max_{1\le i\le n}\|\boldsymbol{X}_i\|_2 \le I_n\}$ with a slowly increasing function I_n, we have $\mathrm{P}\{\mathcal{F}_Y^c\} = \varepsilon_n \to 0$.

In addition, to bound $\max_{1\le i\le n,1\le i\le k}|X_{ik}|$, the sub-Weibull concentration determines:

$$\mathrm{P}\big(\max_{1\le i\le n,1\le i\le k}|X_{ik}| > t\big) \le np\mathrm{P}(|X_{11}| > t) \le 2npe^{-(t/\|X_{11}\|_{\psi_\theta})^\theta} \le \delta \Rightarrow t = M_X\log^{1/\theta}(\frac{2np}{\delta}),$$

by using Corollary 3. Hence, we define the event for the maximum designs:

$$\mathcal{F}_{\max} = \Big\{\max_{1\le i\le n,1\le k\le p}|X_{ik}| \le M_X\log^{1/\theta}(\frac{2np}{\delta})\Big\} \cap \mathcal{F}_Y.$$

To make sure that the optimization in (9) has a unique solution, we also require the minimal eigenvalue condition.

- (C.3): Suppose that $b^\top \mathrm{E}(\hat{\mathcal{Q}}_n(\beta))b \geq \mathrm{C}_{\min}$ is satisfied for all $b \in S^{p-1}$.

In the proof, to ensure that the random Hessian function has a non-singular eigenvalue, we define the event

$$\mathcal{F}_1 = \left\{ \max_{k,j} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i k e^{X_i^\top \beta^*} X_{ik} X_{ij}}{(k + e^{X_i^\top \beta^*})^2} - \mathrm{E}\left(\frac{Y_i k e^{X_i^\top \beta^*} X_{ik} X_{ij}}{(k + e^{X_i^\top \beta^*})^2} \right) \right] \right| \leq \frac{\mathrm{C}_{\min}}{4} \right\}$$

$$\mathcal{F}_2 = \left\{ \max_{k,j} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{k e^{X_i^\top \beta^*} X_{ik} X_{ij}}{(k + e^{X_i^\top \beta^*})^2} - \mathrm{E}\left(\frac{k e^{X_i^\top \beta^*} X_{ik} X_{ij}}{(k + e^{X_i^\top \beta^*})^2} \right) \right] \right| \leq \frac{\mathrm{C}_{\min}}{4} \right\}.$$

Theorem 3 (Upper bound for ℓ_2-error)**.** *In the NBR with loss* (10) *and* $(C.1 - C.3)$*, let*

$$M_{BX} = M_X + \frac{B}{\log 2}, \quad R_n := \frac{6 M_{BX} M_X}{\mathrm{C}_{\min}} \left[\sqrt{\frac{2p}{n} \log\left(\frac{2p}{\delta}\right)} + \frac{1}{n} \sqrt{p \log\left(\frac{2p}{\delta}\right)} \right] \log^{1/\theta}\left(\frac{2np}{\delta}\right),$$

and $\boldsymbol{b} := (k/n) M_X^2 (1, \ldots, 1)^\top \in \mathbb{R}^n$*. Under the event* $\mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_{\max}$*, for any* $0 < \delta < 1$*, if the sample size n satisfies*

$$R_n I_n \leq \frac{\log(4/3)}{3}, \tag{11}$$

Let $c_n := e^{-\frac{1}{4}\left(\frac{nt^2}{2M_X^4 \log^{4/\theta}(\frac{2np}{\delta}) M_{BX}^2} \wedge \frac{nt}{M_X^2 \log^{2/\theta}(\frac{2np}{\delta}) M_{BX}}\right)} + e^{-\left(\frac{t^{\theta/2}}{[4eC(\theta/2)\|\boldsymbol{b}\|_2 L_n(\theta/2, \boldsymbol{b})]^{\theta/2}} \wedge \frac{t^2}{16e^2 C^2(\theta/2)\|\boldsymbol{b}\|_2^2}\right)}$ *with* $t = \mathrm{C}_{\min}/4$*, then*

$$\mathrm{P}(\|\hat{\beta}_n - \beta^*\|_2 \leq R_n) \geq 1 - 2p^2 c_n - \delta - \varepsilon_n.$$

A few comment is made on this theorem. First, in order to get $\|\hat{\beta}_n - \beta^*\|_2 \xrightarrow{p} 0$, we need $p = o(n)$ under sample size restriction (11) with $I_n = o(\log^{-1/\theta}(np) \cdot [n^{-1} p \log p]^{-1/2})$. Second, note that the ε_n in provability $1 - 2p^2 c_n - \delta - \varepsilon_n$ depends on the models size and the fluctuation of the design by the event $\mathcal{F}_{\max}$.

3.2. Non-Asymptotic Bai-Yin's Theorem

In statistical machine learning, exponential decay tail probability is crucial to evaluate the finite-sample performance. Unlike Bai-Yin's law with the fourth-moment condition that leads to polynomial decay tail probability, under sub-Weibull conditions of data, we provide a exponential decay tail probability on the extreme eigenvalues of a $n \times p$ random matrix.

Let $\mathbf{A} = \mathbf{A}_{n,p}$ be an $n \times p$ random matrix whose entries are independent copies of a r.v. with zero mean, unit variance, and finite fourth moment. Suppose that the dimensions n and p both grow to infinity while the aspect ratio p/n converges to a constant in $[0, 1]$. Then Bai-Yin's law [21] asserted that the standardized extreme eigenvalues satisfying

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

Next we introduce a special *counting measure* for measuring the complexity of a certain set in some space. The $\mathcal{N}_\varepsilon$ is called an *ε-net* of K in $\mathbb{R}^n$ if K can be covered by balls with centers in K and radii ε (under Euclidean distance). The *covering number* $\mathcal{N}(K, \varepsilon)$ is defined by the smallest number of closed balls with centers in K and radii ε whose union covers K.

For purposes of studying random matrices, we need to extend the definition of sub-Weibull r.v. to sub-Weibull random vectors. The n-dimensional unit Euclidean sphere S^{n-1}, is denoted by $S^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$. We say that a random vector $\boldsymbol{X}$ in $\mathbb{R}^n$ is sub-Weibull if the one-dimensional marginals $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ are sub-Weibull r.v.s for all $\boldsymbol{a} \in \mathbb{R}^n$. The sub-Weibull norm of a random vector $\boldsymbol{X}$ is defined as $\|\boldsymbol{X}\|_{\psi_\theta} := \sup_{\boldsymbol{a} \in S^{n-1}} \|\langle \boldsymbol{X}, \boldsymbol{a} \rangle\|_{\psi_\theta}$.

Similarly, define the spectral norm for any $p \times p$ matrix $\mathbf{B}$ as $\|\mathbf{B}\| = \max_{||\mathbf{x}||_2=1} \|\mathbf{B}\boldsymbol{x}\|_2 = \sup_{\boldsymbol{x}\in S^{p-1}} |\langle \mathbf{B}\boldsymbol{x}, \boldsymbol{x}\rangle|$. Spectral norm has many good properties, see [1] for details.

Furthermore, for simplicity, we assume that the rows in random matrices are isotropic random vectors. A random vector $\boldsymbol{Y}$ in $\mathbb{R}^n$ is called isotropic if $\mathrm{Var}(\boldsymbol{Y}) = \mathbf{I}_p$. Equivalently, $\boldsymbol{Y}$ is isotropic if $\mathrm{E}[\langle \boldsymbol{Y}, \boldsymbol{a}\rangle^2] = \|\boldsymbol{a}\|_2^2$ for all $\boldsymbol{a} \in \mathbb{R}^n$. In the non-asymptotic regime, Theorem 4.6.1 in [1] study the upper and lower bounds of maximum (minimum) eigenvalues of random matrices with independent sub-Gaussian entries which are sampled from high-dimensional distributions. As an extension of Theorem 4.6.1 in [1], the following result is a non-asymptotic versions of Bai-Yin's law for sub-Weibull entries, which is useful to estimate covariance matrices from heavy-tailed data [subW(θ), $\theta < 1$].

Theorem 4 (Non-asymptotic Bai-Yin's law)**.** *Let* $\mathbf{A}$ *be an* $n \times p$ *matrix whose rows* $\boldsymbol{A}_i$ *are independent isotropic sub-Weibull random vectors in* $\mathbb{R}^p$ *with covariance matrix* $\mathbf{I}_p$ *and* $\max_{1\le i\le n}\|\boldsymbol{A}_i\|_{\psi_\theta} \le K$*. Then for every* $s \ge 0$*, we have*

$$\mathrm{P}\left\{ \|\frac{1}{n}\mathbf{A}^\top\mathbf{A} - \mathbf{I}_p\| \le H(cp + s^2, n; \theta) \right\} \ge 1 - 2e^{-s^2},$$

where

$$H(t, n; \theta) := 2eKC(\theta/2)K_{\theta/2}[1 + ([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2})\left[\sqrt{\tfrac{t}{n}} + \left\{ \begin{array}{ll} A(\theta/2)\frac{(\gamma^2 t)^{2/\theta}}{n}, & \theta \le 2 \\ B(\theta/2)\frac{(\gamma^2 t)^{2/\theta}}{n^{1/\theta}}, & \theta > 2 \end{array} \right.\right],$$

where $K_\alpha := 2^{1/\alpha}$ *if* $\alpha \in (0, 1)$ *and* $K_\alpha = 1$ *if* $\alpha \ge 1$*;* $A(\theta/2)$*,* $B(\theta/2)$ *and* $C(\theta/2)$ *defined in Theorem 1a.*

Moreover, the concentration inequality for extreme eigenvalues hold for $c \ge n\log 9/p$

$$\mathrm{P}\left\{ \sqrt{1 - H^2(cp + s^2, n; \theta)} \le \tfrac{\lambda_{min}(\mathbf{A})}{\sqrt{n}} \le \tfrac{\lambda_{max}(\mathbf{A})}{\sqrt{n}} \le \sqrt{1 + H^2(cp + s^2, n; \theta)} \right\} \ge 1 - 2e^{-s^2}. \tag{12}$$

3.3. General Log-Truncated Z-Estimators and sub-Weibull Type Robust Estimators

Motivated from log-truncated loss in [22,23], we study the almost surely continuous and non-decreasing function $\varphi^c : \mathbb{R} \to \mathbb{R}$ for truncating the original score function

$$-\log[1 - x + c(|x|)] \le \varphi^c(x) \le \log[1 + x + c(|x|)], \quad \forall x \in \mathbb{R} \tag{13}$$

where $c(|x|) > 0$ is a *high-order function* [23] of $|x|$ which is to be specified. For example, a plausible choose for $\varphi^c(x)$ in (13) should have following form

$$\begin{aligned} \varphi^c(x) &= \log[1 + x + c(|x|)]1(x \ge 0) - \log[1 - x + c(|x|)]1(x \le 0) \\ &= \mathrm{sign}(x)\log(1 + |x| + c(|x|)). \end{aligned} \tag{14}$$

For (14), we get $\varphi^c(x) \approx x$ for sufficiently smaller x and $\varphi^c(x) \ll x$ for larger x. Under (13), now we show that $c(|x|)$ must obey a key inequality. For all $x \in \mathbb{R}$, it suffices to verify $-\log[1 - x + \mathrm{c}(|x|)] \le \log[1 + x + \mathrm{c}(|x|)]$, which is equivalent to check $\log[(1 + \mathrm{c}(|x|) + x)(1 + \mathrm{c}(|x|) - x)] \ge 0$, namely $(1 + \mathrm{c}(|x|))^2 - x^2 \ge 1 \Leftrightarrow \mathrm{c}(|x|) \ge \sqrt{1 + x^2} - 1$.

For independent r.v.s $\{X_i\}_{i=1}^n$, using the score function (14), we define the score function of data

$$\hat{Z}_{\alpha_n}(\theta) = \tfrac{1}{n\alpha_n}\textstyle\sum_{i=1}^n \varphi^c[\alpha_n(X_i - \theta)] \text{ for any } \theta \in \mathbb{R}.$$

Then the influence of the heavy-tailed outliers is weaken by $\varphi^c[\alpha_n(X_i - \theta)]$ by choosing an optimal α_n. We aim to estimate the *average mean*: $\mu_n := \frac{1}{n}\sum_{i=1}^n \mathrm{E}X_i$ for non-i.i.d. samples $\{X_i\}_{i=1}^n$. Define the Z-estimator $\hat{\theta}_{\alpha_n}$ as

$$\hat{\theta}_{\alpha_n} \in \{\theta \in \mathbb{R} : \hat{Z}_{\alpha_n}(\theta) = 0\}, \tag{15}$$

where α_n is the tuning parameter (will be determined later).

To guarantee consistency for log-truncated Z-estimators (15), we require following assumptions of $c(\cdot)$.

- (C.1): For a constant $c_2 > 1$, the $c(x)$ satisfies *weak triangle inequality* and *scaling property*,

$$(\text{C.1.1}) : c(x+y) \le c_2[c(x)+c(y)], \ \ (\text{C.1.2}) : c(tx) \le f(t)c(x)$$

for $f(t)$ satisfies
(C.1.3): $f(t)$ and $f(t)/|t|$ are non-constant increasing functions and $\lim_{t\to 0} f(t)/|t| = 0$.

Remark 6. *Note that* $|x| \ge \sqrt{1+x^2} - 1$ *and we could put* $c(|x|) = |x|$. *However,* $c(|x|) = |x|$ *does not satisfy (C.1.3) since* $f(t) = |t|$ *and* $f(t)/|t|$ *are constant functions of* t.

In the following theorem, we establish the finite sample confidence interval and the convergence rate of the estimator $\hat{\theta}_{\alpha_n}$.

Theorem 5. *Let* $\{X_i\}_{i=1}^n$ *be independent samples drawn from an unknown probability distribution* $\{P_i\}_{i=1}^n$ *on* $\mathbb{R}$. *Consider the estimator* $\hat{\theta}_{\alpha_n}$ *defined as (15) with (C.1),* $\alpha_n \to 0$ *and* $\frac{1}{n}\sum_{i=1}^n \mathrm{E}[c(X_i-\theta)] = O(1)$. *Let* $B_n^+(\theta) = \mu_n - \theta + \frac{1}{n\alpha_n}\sum_{i=1}^n \mathrm{E}[c(\alpha_n(X_i-\theta))] + \frac{\log(\delta^{-1})}{n\alpha_n}$ *and* $B_n^-(\theta) = \mu_n - \theta - \frac{1}{n\alpha_n}\sum_{i=1}^n \mathrm{E}[c(\alpha_n(X_i-\theta))] - \frac{\log(\delta^{-1})}{n\alpha_n}$. *Let* θ_+ *be the smallest solution of the equation* $B_n^+(\theta) = 0$ *and* θ_- *be the largest solution of* $B_n^-(\theta) = 0$.
(a). *We have with the* $(1-2\delta)$*-confidence intervals*

$$\mathrm{P}(B_n^-(\theta) < \hat{Z}_{\alpha_n}(\theta) < B_n^+(\theta)) \ge 1-2\delta, \qquad \mathrm{P}(\theta_- \le \hat{\theta}_{\alpha_n} \le \theta_+) \ge 1-2\delta,$$

for any $\delta \in (0, 1/2)$ *satisfies the sample condition:*

$$\frac{1}{n\alpha_n}\sum_{i=1}^n \mathrm{E}[c(\alpha_n X_i - \alpha_n[\mu_n \pm d_n(c)])] + \frac{\log(\delta^{-1})}{n\alpha_n} < d_n(c), \tag{16}$$

where $d_n(c)$ *is a constant such that* $B_n^{\pm}(\mu_n \pm d_n(c)) < 0$.
(b). *Moreover, picking* $\alpha_n \ge f^{-1}\left(\frac{\log(\delta^{-1})}{c_2\sum_{i=1}^n \mathrm{E}[c(X_i-\mu_n)]}\right)$, *one has*

$$\mathrm{P}\left(|\hat{\theta}_{\alpha_n} - \mu_n| \le \left| g_{\alpha_n}^{-1}\left\{-\frac{2\log(\delta^{-1})}{n\alpha_n}\right\}\right|\right) \ge 1-2\delta, \text{ with } g_{\alpha_n}(t) := t + \frac{c_2}{\alpha_n}c(\alpha_n t). \tag{17}$$

The (17) in Theorem 5 is a fundamental extension of Lemma 2.1 (see Theorem 16 in [24]) with $c(x) = x^2/2$ from i.i.d. sample to independent sample. Let $c(x) = |x|^\beta/\beta$, for i.i.d. sample, Theorem 5 implies Lemmas 2.3, 2.4 and Theorem 2.1 in [22]. The $\alpha_n \ge f^{-1}\left(\frac{\log(\delta^{-1})}{c_2\sum_{i=1}^n \mathrm{E}[c(X_i-\mu_n)]}\right)$ in Theorem 5(b) gives a theoretical guarantee for choosing the tuning parameter α_n.

Proposition 7 (Theorem 2.1 in [22]). *Let* $\{X_i\}_{i=1}^n$ *be a sequence of i.i.d. samples drawn from an unknown probability distribution on* $\mathbb{R}$. *We assume* $\mathrm{E}|X_1|^\beta < \infty$ *for a certain* $\beta \in (1,2]$ *and denote* $\mu = \mathrm{E}[X_1]$, $v_\beta = \mathrm{E}|X_1-\mu|^\beta$. *Given any* $\epsilon \in (0,1/2)$ *and positive integer* $n \ge \left(\frac{2v_\beta+1}{\beta}\right)^{\frac{\beta}{\beta-1}}\frac{2\beta\log(\epsilon^{-1})}{v_\beta}$, *let* $\alpha_n = \frac{1}{2}\left(\frac{2\beta\log(\epsilon^{-1})}{nv_\beta}\right)^{\frac{1}{\beta}}$. *Then, with probability at least* $1-2\epsilon$,

$$|\hat{\theta}_{\alpha_n} - \mu| \le 2\left(\frac{2\beta\log(\epsilon^{-1})}{n}\right)^{\frac{\beta-1}{\beta}} v_\beta^{\frac{1}{\beta}}\left[\beta - \left(\frac{2\beta\log(\epsilon^{-1})}{nv_\beta}\right)^{\frac{\beta-1}{\beta}}\right]^{-1} = O\left(n^{-\frac{\beta-1}{\beta}}\right). \tag{18}$$

Comparing to the convergence rate in (18), put $O(n^{-\frac{\beta-1}{\beta}}) = O(n^{-1/\theta})$ for $\theta > 2$. It implies

$$\beta^{-1} + \theta^{-1} = 1, \ (\theta \geq 2 \text{ or } 0 < \beta \leq 2).$$

For example, let us deal with the Pareto distribution Pareto(α, k) with shape parameter $\alpha > 0$ and scale parameter $k > 0$, and the density function is $f(x) = \frac{\alpha k^{\alpha}}{x^{\alpha+1}} \cdot 1_{\{x \in [k,\infty)\}}$. For $\alpha \leq 2$, Pareto(α, k) has infinite variance, and it does not belong to the sub-Weibull distribution, so do the sample mean of i.i.d. Pareto distributed data. Proposition 7 shows that the estimator error for robust mean estimator enjoys sub-Weibull concentration as presented in Proposition 3, without finite sub-Weibull norm assumption of data. With the Weibull-tailed behavior, it motivates us to define general sub-Weibull estimators having the non-parametric convergence rate $O(n^{-1/\theta})$ in Proposition 3 for $\theta > 2$, even if the data do not have finite sub-Weibull norm.

Definition 5 (Sub-Weibull estimators). *An estimator $\hat{\mu} := \hat{\mu}(X_1, \cdots, X_n)$ based on i.i.d. samples $\{X_i\}_{i=1}^n$ from an unknown probability distribution P with mean μ_P, is called* (A, B, C)-subW(θ) *if*

$$\forall t \in (0, A), \quad \mathrm{P}(|\hat{\mu} - \mu_P| \leq B(t/n)^{1/\theta}) \geq 1 - Ce^{-t}.$$

For example, in Proposition 7, $\hat{\theta}_{\alpha_n}$ is $(\infty, B, 1)$-subW$(\frac{\beta}{\beta-1})$ with $B \sim 2\big(2\beta \log(\epsilon^{-1})\big)^{\frac{\beta-1}{\beta}} v_{\beta}^{\frac{1}{\beta}}$ in Definition 5. When $\theta = 2$, [25] defined sub-Gaussian estimators (includes Median of means and Catoni's estimators) for certain heavy-tailed distributions and discussed the nonexistence of sub-Gaussian mean estimators under β-moment condition for the data ($\beta \in (1, 2)$).

4. Conclusions

Concentration inequalities are far-reaching useful in high-dimensional statistical inferences and machine learnings. They can facilitate various explicit non-asymptotic confidence intervals as a function of the sample size and model dimension.

Future research includes sharper version of Theorem 2 that is crucial to construct non-asymptotic and data-driven confidence intervals for the sub-Weibull sample mean. Although we have obtained sharper upper bounds for sub-Weibull concentrations, the lower bounds on tail probabilities are also important in some statistical applications [26]. Developing non-asymptotic and sharp lower tail bounds of Weibull r.v.s is left for further study. For negative binomial concentration inequalities in Corollary 2, it is of interesting to study concentration inequalities of COM-negative binomial distributions (see [27]).

Author Contributions: Conceptualization, H.Z. and H.W.; Formal analysis, H.Z. and H.W.; Funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by National Natural Science Foundation of China Grant (12101630) and the University of Macau under UM Macao Talent Programme (UMMTP-2020-01). This work is also supported in part by the Key Project of Natural Science Foundation of Anhui Province Colleges and Universities (KJ2021A1034), Key Scientific Research Project of Chaohu University (XLZ-202105).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank Guang Cheng for the discussion about doing statistical inference in the non-asymptotic way and Arun Kumar Kuchibhotla for his help about the proof of Theorem 1. The authors also thank Xiaowei Yang for his helpful comments on Theorem 5.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1

Proof of Corollary 1. $\phi_{|X|^\theta}(t)$ is continuous for t a neighborhood of zero, by the definition, $2 \geq \mathrm{E}e^{(|X|/\|X\|_{\psi_\theta})^\theta} = m_{|X|^\theta}(\|X\|_{\psi_\theta}^{-\theta})$. Since $|X|^\theta > 0$, the MGF $m_{|X|^\theta}(t)$ is monotonic increasing. Hence, inverse function $m^{-1}_{|X|^\theta}(t)$ exists and $\|X\|_{\psi_\theta}^{-\theta} = m^{-1}_{|X|^\theta}(2)$. So $\|X\|_{\psi_\theta} = \left(m^{-1}_{|X|^\theta}(2)\right)^{-1/\theta}$. □

Appendix A.2

Proof of Corollary 2. The first inequality is the direct application of (3) by observing that for any constant $a \in \mathbb{R}$, and r.v. Y with $\|Y\|_{\psi_1} < \infty$, $\|aY\|_{\psi_1} = |a|\|Y\|_{\psi_1}$, $\|Y + a\|_{\psi_1} \leq \|Y\|_{\psi_1} + \|a\|_{\psi_1} = \|Y\|_{\psi_1} + |a|/\log 2$ and $\|X + a\|^2_{\psi_1} \leq (\|X\|_{\psi_1} + |a|/\log 2)^2$. The second inequality is obtained from (3) by considering two rate in $\left(\frac{t^2}{\sum_{i=1}^n 2\|Y_i\|^2_{\psi_1}} \wedge \frac{t}{\max_{1\leq i\leq n}\|Y_i\|_{\psi_1}}\right)$ separately. For (5), we only need to note that

$$\|Y_i\|_{\psi_1} = \inf\{t > 0 : \mathrm{E}e^{Y_i/t} \leq 2\} = \inf\{t > 0 : \left(\tfrac{1-q_i}{1-q_i e^{1/t}}\right)^{k_i} \leq 2\} = \left[\log \tfrac{1-(1-q_i)/\sqrt[k_i]{2}}{q_i}\right]^{-1}.$$

Then the third inequality is obtained by the first inequality and the definition of $a(\mu_i, k_i)$. □

Appendix A.3

Proof of Corollary 3. The first and second part of this proposition were shown in Lemma 2.1 of [28]. For the third result, using the bounds of Gamma function [see [29]]:

$$\sqrt{2\pi}x^{x-(1/2)}e^{-x} \leq \Gamma(x) \leq [\sqrt{2\pi}x^{x-(1/2)}e^{-x}] \cdot e^{1/(12x)}, (x > 0),$$

it gives

$$\begin{aligned}(\mathrm{E}|X|^k)^{1/k} &\leq \left\{2\|X\|^k_{\varphi_\theta}\left(\tfrac{k}{\theta}\right)[\sqrt{2\pi}(k/\theta)^{\frac{k}{\theta}-\frac{1}{2}}e^{-\frac{11k}{12\theta}}]\right\}^{1/k} = (\tfrac{2\sqrt{2\pi}}{\theta})^{1/k}\{\left(\tfrac{k}{\theta}\right)^{\frac{k}{\theta}+\frac{1}{2}}e^{-\frac{11k}{12\theta}}\}^{1/k}\|X\|_{\varphi_\theta} \\ &= (\tfrac{2\sqrt{2\pi}}{\theta})^{1/k}(k/\theta)^{\frac{1}{\theta}+\frac{1}{2k}}e^{-\frac{11}{12\theta}}\|X\|_{\varphi_\theta} \leq C_\theta(\theta e^{11/12})^{-1/\theta}\|X\|_{\varphi_\theta}k^{1/\theta}.\end{aligned}$$

□

Appendix A.4

Proof of Corollary 4. By the definition of ψ_θ-norm, $\mathrm{E}\exp\{|X/\|X\|_{\psi_\theta}|^\theta\} \leq 2$. Then $\mathrm{E}\exp\{||X|^r/\|X\|^r_{\psi_\theta}|^{\theta/r}\} \leq 2$. The result $|X|^r \sim \mathrm{subW}(\theta/r)$ follows by the definition of ψ_θ-norm again. Moreover,

$$\begin{aligned}\|X\|_{\psi_\theta} &:= \inf\{C \in (0,\infty) : \mathrm{E}[\exp(|X|^\theta/C^\theta)] \leq 2\} \\ &= [\inf\{C^r \in (0,\infty) : \mathrm{E}[\exp\{||X|^r/C^r|^{\theta/r}\}] \leq 2\}]^{1/r} = \||X|^r\|^{1/r}_{\psi_{\theta/r}},\end{aligned}$$

which verifies (6). If $X \sim \mathrm{subW}(r\theta)$, then $\mathrm{E}\exp\{|X^r/\|X\|^r_{\psi_{r\theta}}|^\theta\} = \mathrm{E}\exp\{|X/\|X\|_{\psi_{r\theta}}|^{r\theta}\} \leq 2$, which means that $X^r \sim \mathrm{subW}(\theta)$ with

$$\begin{aligned}\|X\|_{\psi_{r\theta}} &:= \inf\{C \in (0,\infty) : \mathrm{E}[\exp(|X|^{r\theta}/C^{r\theta})] \leq 2\} \\ &= [\inf\{C^r \in (0,\infty) : \mathrm{E}[\exp\{||X|^r/C^r|^\theta\}] \leq 2\}]^{1/r} = \||X|^r\|^{1/r}_{\psi_\theta}.\end{aligned}$$

□

Appendix A.5

Proof of Corollary 5. Set $\Delta := \sup_{p\geq 2} \frac{\|X\|_p}{\sqrt{p}+Lp^{1/\theta}}$ so that $\|X\|_p \leq \Delta\sqrt{p} + L\Delta p^{1/\theta}$ holds for all $p \geq 2$. By Markov's inequality for t-th moment ($t \geq 2$), we have

$$\mathrm{P}\Big(|X| \geq e\Delta\sqrt{t} + eL\Delta t^{1/\theta}\Big) \leq \left(\frac{||X||_t}{e\Delta[\sqrt{t}+Lt^{1/\theta}]}\right)^t \leq e^{-t}, \text{ [By the definition of } \Delta].$$

So, for any $t \geq 2$,

$$\mathrm{P}\Big(|X| \geq e\Delta\sqrt{t} + eL\Delta t^{1/\theta}\Big) \leq e^{-t}. \tag{A1}$$

Note the definition of Δ shows $\|X\|_t \leq \Delta\sqrt{t} + L\Delta t^{1/\theta}$ holds for all $t \geq 2$ and assumption $\|X\|_t \leq C_1\sqrt{t} + C_2 t^{1/\theta}$ for all $t \geq 2$. It gives $e\Delta\sqrt{t} + eL\Delta t^{1/\theta} \leq eC_1\sqrt{t} + eC_2 t^{1/\theta}$. This inequality with (A1) gives

$$\mathrm{P}\Big(|X| \geq eC_1\sqrt{t} + eC_2 t^{1/\theta}\Big) \leq 1\{0 < t < 2\} + e^{-t}\{t \geq 2\}, \ \ \forall t > 0. \tag{A2}$$

Take $K = k^{2/\theta}C_2/(kC_1)$, and define $\delta_k := keC_1$ for a certain constant $k > 1$,

$$\begin{aligned}
&\mathrm{E}\left[\Psi_{\theta,K}\left(\frac{|X|}{\delta_k}\right)\right] = \int_0^\infty \mathrm{P}\Big(|X| \geq \delta_k \Psi_{\theta,K}^{-1}(s)\Big)ds \\
&= \int_0^\infty \mathrm{P}(|X| \geq keC_1\sqrt{\log(1+s)} + keC_1K[\log(1+s)]^{1/\theta})ds \\
&= \int_0^\infty \mathrm{P}(|X| \geq eC_1\sqrt{\log(1+s)^{k^2}} + eC_2[\log(1+s)^{k^2}]^{1/\theta})ds \\
[\text{By (A2)}] \ &\leq \int_{0<k^2\log(1+s)<2} ds + \int_{k^2\log(1+s)\geq 2} \exp\Big\{-k^2\log(1+s)\Big\}ds \\
&\leq \int_0^{e^{2k^{-2}}-1} dt + \int_{e^{2k^{-2}}-1}^\infty \frac{dt}{(1+t)^{k^2}} \\
&= e^{2k^{-2}} - 1 + \frac{(1+t)^{1-k^2}}{1-k^2}\Big|_{e^{2k^{-2}}-1}^\infty = e^{2k^{-2}} - 1 + \frac{e^{2(1-k^2)/k^2}}{k^2-1} \leq 1.
\end{aligned}$$

Therefore, $\|X\|_{\Psi_{\theta,K}} \leq \gamma eC_1$ with γ defined as the smallest solution of the inequality $\{k > 1 : e^{2k^{-2}} - 1 + \frac{e^{2(1-k^2)/k^2}}{k^2-1} \leq 1\}$. An approximate solution is $\gamma \approx 1.78$. □

Appendix A.6

The main idea in the proof is by the sharper estimates of the GBO norm of the sum of *symmetric* r.v.s.

Proof of Theorem 1.

(a) Without loss of generality, we assume $\|X_i\|_{\psi_\theta} = 1$. Define $Y_i := \big(|X_i| - (\log 2)^{1/\theta}\big)_+$, then it is easy to check that $\mathrm{P}(|X_i| \geq t) \leq 2e^{-t^\theta}$ implies $\mathrm{P}(Y_i \geq t) \leq e^{-t^\theta}$. For independent Rademacher r.v. $\{\varepsilon_i\}_{i=1}^n$, the symmetrization inequality gives $\|\sum_{i=1}^n w_i X_i\|_p \leq 2\|\sum_{i=1}^n \varepsilon_i w_i X_i\|_p$. Note that $\varepsilon_i X_i$ is identically distributed as $\varepsilon_i|X_i|$,

$$
\begin{aligned}
\|\sum_{i=1}^n w_i X_i\|_p &\le 2\|\sum_{i=1}^n \varepsilon_i w_i |X_i|\|_p \le 2\|\sum_{i=1}^n \varepsilon_i w_i (Y_i + (\log 2)^{1/\theta})\|_p \\
&\le 2\|\sum_{i=1}^n \varepsilon_i w_i Y_i\|_p + 2(\log 2)^{1/\theta}\|\sum_{i=1}^n \varepsilon_i w_i\|_p \\
[\text{Khinchin-Kahane inequality}] &\le 2\|\sum_{i=1}^n \varepsilon_i w_i Y_i\|_p + 2(\log 2)^{1/\theta}\left(\frac{p-1}{2-1}\right)^{1/2}\|\sum_{i=1}^n \varepsilon_i w_i\|_2 \\
&< 2\|\sum_{i=1}^n \varepsilon_i w_i Y_i\|_p + 2(\log 2)^{1/\theta}\sqrt{p}(\mathrm{E}(\sum_{i=1}^n \varepsilon_i w_i)^2)^{1/2} \\
[\{\varepsilon_i\}_{i=1}^n \text{ are independent}] &= 2\|\sum_{i=1}^n \varepsilon_i w_i Y_i\|_p + 2(\log 2)^{1/\theta}\sqrt{p}\|\boldsymbol{w}\|_2. \qquad \text{(A3)}
\end{aligned}
$$

From Lemma 2, we are going to handle the first term in (A3) with the sum of symmetric r.v.s. Since $\mathrm{P}(Y_i \ge t) \le e^{-t^\theta}$, then

$$\|\textstyle\sum_{i=1}^n \varepsilon_i w_i Y_i\|_p = \|\sum_{i=1}^n w_i Z_i\|_p, \qquad Z_i := \varepsilon_i Y_i$$

for symmetric independent r.v.s $\{Z_i\}_{i=1}^n$ satisfying $|Z_i| \overset{d}{=} Y_i$ and $\mathrm{P}(Z_i \ge t) = e^{-t^\theta}$ for all $t \ge 0$.

Next, we proceed the proof by checking the moment conditions in Corollary 5.

Case $\theta \le 1$*:* $N(t) = t^\theta$ is concave for $\theta \le 1$. From Lemmas 2 and 3 (a), for $p \ge 2$,

$$
\begin{aligned}
&\Big\|\sum_{i=1}^n w_i Z_i\Big\|_p \le e \inf\left\{t > 0 : \sum_{i=1}^n \log \phi_p\Big(e^{-2}\Big(\frac{w_i e^2}{t}\Big) Z_i\Big) \le p\right\} \\
&\le e \inf\left\{t > 0 : \sum_{i=1}^n p M_{p,Z_i}\Big(\frac{w_i e^2}{t}\Big) \le p\right\} \\
&= e \inf\left\{t > 0 : \sum_{i=1}^n \left[\left\{\Big(\frac{w_i e^2}{t}\Big)^p \|Z_i\|_p^p\right\} \vee \left\{p\Big(\frac{w_i e^2}{t}\Big)^2 \|Z_i\|_2^2\right\}\right] \le p\right\} \\
&\le e \inf\left\{t > 0 : \Gamma\Big(\frac{p}{\theta}+1\Big)\frac{e^{2p}}{t^p}\|\boldsymbol{w}\|_p^p \le 1\right\} + e \inf\left\{t > 0 : p\Gamma\Big(\frac{2}{\theta}+1\Big)\frac{e^4}{t^2}\|\boldsymbol{w}\|_2^2 \le 1\right\},
\end{aligned}
$$

where the last inequality we use $\|Z_i\|_p^p = \int_0^\infty p t^{p-1} \mathrm{P}(|Z_i| \ge t)\, dt \le \int_0^\infty p t^{p-1} e^{-t^\theta}\, dt = p\Gamma\big(\frac{p}{\theta}+1\big)$. Hence

$$\|\textstyle\sum_{i=1}^n w_i Z_i\|_p \le e^3\Big[\Gamma^{1/p}\big(\frac{p}{\theta}+1\big)\|\boldsymbol{w}\|_p + \sqrt{p}\Gamma^{1/2}\big(\frac{2}{\theta}+1\big)\|\boldsymbol{w}\|_2\Big],$$

and

$$
\begin{aligned}
\|\sum_{i=1}^n w_i X_i\|_p &\le 2e^3\Big[\Gamma^{1/p}\Big(\frac{p}{\theta}+1\Big)\|\boldsymbol{w}\|_p + \sqrt{p}\Gamma^{1/2}\Big(\frac{2}{\theta}+1\Big)\|\boldsymbol{w}\|_2\Big] + 2(\log 2)^{1/\theta}\sqrt{p}\|\boldsymbol{w}\|_2 \\
&= 2e^3\Gamma^{1/p}\Big(\frac{p}{\theta}+1\Big)\|\boldsymbol{w}\|_p + 2\Big[(\log 2)^{1/\theta} + e^3\Gamma^{1/2}\Big(\frac{2}{\theta}+1\Big)\Big]\sqrt{p}\|\boldsymbol{w}\|_2.
\end{aligned}
$$

Using homogeneity, we can assume that $\sqrt{p}\|\boldsymbol{w}\|_2 + p^{1/\theta}\|\boldsymbol{w}\|_\infty = 1$. Then $\|\boldsymbol{w}\|_2 \le p^{-1/2}$ and $\|\boldsymbol{w}\|_\infty \le p^{-1/\theta}$. Therefore, for $p \ge 2$,

$$
\begin{aligned}
\|\boldsymbol{w}\|_p &\le \Big(\sum_{i=1}^n |w_i|^2 \|\boldsymbol{w}\|_\infty^{p-2}\Big)^{1/p} \le (p^{-1-(p-2)/\theta})^{1/p} = (p^{-p/\theta} p^{(2-\theta)/\theta})^{1/p} \\
&\le 3^{\frac{2-\theta}{3\theta}} p^{-1/\theta} = 3^{\frac{2-\theta}{3\theta}} p^{-1/\theta}\{\sqrt{p}\|\boldsymbol{w}\|_2 + p^{1/\theta}\|\boldsymbol{w}\|_\infty\},
\end{aligned}
$$

where the last inequality follows form the fact that $p^{1/p} \le 3^{1/3}$ for any $p \ge 2, p \in \mathbb{N}$. Hence

$$\Big\| \sum_{i=1}^n w_i X_i \Big\|_p \le 2e^{3+\frac{2-\theta}{e\theta}} \Gamma^{1/p}\Big(\frac{p}{\theta}+1\Big) \|\boldsymbol{w}\|_\infty + 2\Big[\log^{1/\theta} 2 + e^3 \Big(\Gamma^{1/2}\Big(\frac{2}{\theta}+1\Big) + 3^{\frac{2-\theta}{3\theta}} p^{-\frac{1}{\theta}} \Gamma^{1/p}\Big(\frac{p}{\theta}+1\Big)\Big)\Big] \sqrt{p} \|\boldsymbol{w}\|_2.$$

Following Corollary 5, we have

$$\Big\| \sum_{i=1}^n w_i X_i \Big\|_{\Psi_{\theta, L_n(\theta,p)}} \le \gamma e D_1(\theta),$$

where $L_n(\theta,p) = \frac{\gamma^{2/\theta} D_2(\theta,p)}{\gamma D_1(\theta)}$, $D_1(\theta) := 2[\log^{1/\theta} 2 + e^3(\Gamma^{1/2}(\frac{2}{\theta}+1) + \sup_{p\ge 2} 3^{\frac{2-\theta}{3\theta}} p^{-\frac{1}{\theta}} \Gamma^{1/p}(\frac{p}{\theta}+1))]\|\boldsymbol{w}\|_2 < \infty$, and $D_2(\theta,p) := 2e^3 3^{\frac{2-\theta}{3\theta}} p^{-1/\theta} \Gamma^{1/p}\big(\frac{p}{\theta}+1\big)\|\boldsymbol{w}\|_\infty$.

Finally, take $L_n(\theta) = \inf_{p\ge 1} L_n(\theta,p) > 0$. Indeed, the positive limit can be argued by (2.2) in [30]. Then by the monotonicity property of the GBO norm, it gives

$$\Big\| \sum_{i=1}^n w_i X_i \Big\|_{\Psi_{\theta, L_n(\theta)}} \le \Big\| \sum_{i=1}^n w_i X_i \Big\|_{\Psi_{\theta, L_n(\theta,p)}} \le \gamma e D_1(\theta).$$

Case $\theta > 1$: In this case $N(t) = t^\theta$ is convex with $N^*(t) = \theta^{-\frac{1}{\theta-1}}(1-\theta^{-1}) t^{\frac{\theta}{\theta-1}}$. By Lemmas 2 and 3(b), for $p \ge 2$, we have

$$\begin{aligned}\Big\| \textstyle\sum_{i=1}^n w_i Z_i \Big\|_p &\le e \inf\Big\{ t>0 : \textstyle\sum_{i=1}^n \log \phi_p\Big(\frac{4w_i}{t} Z_i/4\Big) \le p\Big\} + e \inf\Big\{ t>0 : \textstyle\sum_{i=1}^n p M_{p,Z_i}(\frac{4w_i}{t}) \le p\Big\} \\ &\le e \inf\Big\{ t>0 : \textstyle\sum_{i=1}^n p^{-1} N^*\Big(p\Big|\frac{4w_i}{t}\Big|\Big) \le 1\Big\} + e \inf\Big\{ t>0 : \textstyle\sum_{i=1}^n p(\frac{4w_i}{t})^2 \le 1\Big\} \\ &= 4e\big[\sqrt{p}\|\boldsymbol{w}\|_2 + (p/\theta)^{1/\theta}(1-\theta^{-1})^{1/\beta}\|\boldsymbol{w}\|_\beta\big]\end{aligned}$$

with β mentioned in the statement. Therefore, for $p \ge 2$, Equation (A3) implies

$$\|\textstyle\sum_{i=1}^n w_i X_i\|_p \le [8e + 2(\log 2)^{1/\theta}]\sqrt{p}\|\boldsymbol{w}\|_2 + 8e(p/\theta)^{1/\theta}(1-\theta^{-1})^{1/\beta}\|\boldsymbol{w}\|_\beta.$$

Then the following result follows by Corollary 5,

$$\|\textstyle\sum_{i=1}^n w_i X_i\|_{\Psi_{\theta, L'(\theta)}} \le \gamma e D_1'(\theta),$$

where $L_n(\theta) = \frac{\gamma^{2/\theta} D_2'(\theta)}{\gamma D_1'(\theta)}$, $D_1'(\theta) = \Big[8e + 2(\log 2)^{1/\theta}\Big]\|\boldsymbol{w}\|_2$, and $D_2'(\theta) = 8e\theta^{-1/\theta}(1-\theta^{-1})^{1/\beta}\|\boldsymbol{w}\|_\beta$.

Note that $w_i X_i = (w_i \|X_i\|_{\psi_\theta})(X_i/\|X_i\|_{\psi_\theta})$, we can conclude (a).

(b) It is followed from Proposition 5 and (a).

(c) For easy notation, put $L_n(\theta) = L_n(\theta, \boldsymbol{b}_X)$ in the proof. When $\theta < 2$, by the inequality $a+b \le 2(a \vee b)$ for $a, b > 0$, we have

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i X_i| \ge 4eC(\theta)\|\boldsymbol{b}\|_2 \sqrt{t} \Big) \le 2e^{-t}, \text{ if } \sqrt{t} \ge L_n(\theta) t^{1/\theta}.$$

Put $s := 4eC(\theta)\|\boldsymbol{b}\|_2\sqrt{t}$, we have

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i X_i| \ge s \Big) \le 2\exp\Big\{ -\frac{s^2}{16e^2 C^2(\theta)\|\boldsymbol{b}\|_2^2} \Big\}, \text{ if } s \le 4eC(\theta)\|\boldsymbol{b}\|_2 L_n^{\theta/(\theta-2)}(\theta).$$

For $\sqrt{t} \le L_n(\theta) t^{1/\theta}$, we obtain $\mathrm{P}(|\sum_{i=1}^n w_i X_i| \ge 4eC(\theta)\|\boldsymbol{b}_X\|_2 L_n(\theta) t^{1/\theta}) \le 2e^{-t}$. Let $s := 4eC(\theta)\|\boldsymbol{b}\|_2 L_n(\theta) t^{1/\theta}$, it gives

$$\mathrm{P}\Big(|\sum_{i=1}^n w_i X_i| \ge s \Big) \le 2\exp\Big\{ -\frac{s^\theta}{[4eC(\theta)\|\boldsymbol{b}\|_2 L_n(\theta)]^\theta} \Big\}, \text{ if } s > 4eC(\theta)\|\boldsymbol{b}\|_2 L_n^{\theta/(\theta-2)}(\theta).$$

Similarly, for $\theta > 2$, it implies

$$\mathrm{P}(|\textstyle\sum_{i=1}^{n} w_i X_i| \geq s) \leq 2e^{-\frac{s^\theta}{[4eC(\theta)\|\boldsymbol{b}\|_2 L_n(\theta)]^\theta}} \text{ if } s \leq 4eC(\theta)\|\boldsymbol{b}\|_2 L_n^{\theta/(2-\theta)}(\theta),$$

and $\mathrm{P}(|\sum_{i=1}^{n} w_i X_i| \geq s) \leq 2e^{-\frac{s^2}{16e^2C^2(\theta)\|\boldsymbol{b}\|_2^2}}$ if $s \geq 4eC(\theta)\|\boldsymbol{b}\|_2 L_n^{\theta/(2-\theta)}(\theta)$. □

Appendix A.7

Proof of Corollary 6. Using the definition of $\|X\|_{\varphi_\theta}$, it yields

$$\begin{aligned}
\mathrm{E}e^{(c^{-1}|X|)^\theta} &= 1 + \sum_{k=1}^{\infty} \frac{c^{-k}\mathrm{E}|X|^{k\theta}}{k!} \leq 1 + \sum_{k=1}^{\infty} \frac{c^{-k}k!\|X\|_{\varphi_\theta}^{k\theta}}{k!} \\
&= 1 + \sum_{k=1}^{\infty} (\frac{\|X\|_{\varphi_\theta}^{\theta}}{c^\theta})^k = 1 + \frac{\|X\|_{\varphi_\theta}^{\theta}}{c^\theta} \sum_{k=0}^{\infty} (\frac{\|X\|_{\varphi_\theta}^{\theta}}{c^\theta})^k \\
[\frac{\|X\|_{\varphi_2}^{\theta}}{c^\theta} < 1] &= 1 + (\frac{\|X\|_{\varphi_2}^{\theta}}{c^\theta}) \frac{1}{1 - \|X\|_{\varphi_2}^{\theta}/c^\theta} \leq 2
\end{aligned}$$

if $\frac{\|X\|_{\varphi_2}^{\theta}}{c^\theta} \leq \frac{1}{2}$ which implies that the minimal c is $2^{1/\theta}\|X\|_{\varphi_\theta}$. That is to say we have $\mathrm{E}e^{|X/[2^{1/\theta}\|X\|_{\varphi_\theta}]|^{1/\theta}} \leq 2$. Applying (2), we have

$$\mathrm{P}\{|X| > t\} \leq 2e^{-(t/[2^{1/\theta}\|X\|_{\varphi_\theta}])^\theta} = 2\exp\{-\frac{t^\theta}{2\|X\|_{\varphi_\theta}^{\theta}}\} \text{ for all } t \geq 0. \tag{A4}$$

□

Appendix A.8

Proof of Theorem 2. Minkowski's inequality for $p \geq 1$ and definition of $\|X\|_{\varphi_\theta}$ imply

$$\left\|\sum_{i=1}^{n} X_i\right\|_p \leq \sum_{i=1}^{n} \|X_i\|_p \leq \sum_{i=1}^{n} v_i \cdot 2^{1/\theta} C_\theta \left(\frac{p}{\theta e^{11/12}}\right)^{1/\theta},$$

where the last inequality by letting $C_\theta := \max_{k \geq 1} \left(\frac{2\sqrt{2\pi}}{\theta}\right)^{1/k} \left(\frac{k}{\theta}\right)^{1/(2k)}$ in Corollary 3b.

From Markov's inequality, it yields

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq t^{-p}\left\|\sum_{i=1}^{n} X_i\right\|_p^p \leq t^{-p}(\sum_{i=1}^{n} v_i)^p 2^{p/\theta} C_\theta \left(\frac{p}{\theta e^{11/12}}\right)^{p/\theta}.$$

Let $t^{-p}(\sum_{i=1}^{n} v_i)^p 2^{p/\theta} C_\theta \left(\frac{p}{\theta e^{11/12}}\right)^{p/\theta} = e^{-p}$, it gives

$$t = e(\sum_{i=1}^{n} v_i) 2^{1/\theta} C_\theta \left(\frac{p}{\theta e^{11/12}}\right)^{1/\theta} \text{ and } p = \frac{\theta e^{11/12} t^\theta}{[e(\sum_{i=1}^{n} v_i) 2^{1/\theta} C_\theta]^\theta}.$$

Therefore, for $p \geq 1$, we have

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq \mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq e(\sum_{i=1}^{n} v_i) C_\theta (2^{-1}\theta e^{11/12})^{-1/\theta}\right) \leq e^{-p} \in (0, e^{-1}]. \tag{A5}$$

So

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq \exp\left\{-\frac{\theta e^{11/12} t^{\theta}}{2[e(\sum_{i=1}^{n} v_i)C_{\theta}]^{\theta}}\right\}, \qquad t \geq e(\sum_{i=1}^{n} v_i)C_{\theta}(2^{-1}\theta e^{11/12})^{-1/\theta}.$$

Let $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ and $e^{-p} =: \alpha$. Then

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq e\bar{v}2^{1/\theta}C_{\theta}\left(\frac{\log(\alpha^{-1})}{\theta e^{11/12}}\right)^{1/\theta}\right) \geq 1-\alpha \in (1-e^{-1}, 1].$$

For $p < 1$, note that moment monotonicity show that $[\mathrm{E}(|X|^p)]^{1/p}$ is a non-decreasing function of p,, i.e.,

$$0 < p \leq 1 \Rightarrow [\mathrm{E}|X|^p]^{1/p} \leq \mathrm{E}|X|.$$

The c_r -inequality implies $\left\|\sum_{i=1}^{n} X_i\right\|_p^p \leq \sum_{i=1}^{n} \|X_i\|_p^p$. Using Markov's inequality again, we have

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq t^{-p}\left\|\sum_{i=1}^{n} X_i\right\|_p^p \leq t^{-p}\sum_{i=1}^{n}\|X_i\|_p^p \leq t^{-p}\sum_{i=1}^{n}(\mathrm{E}|X_i|)^p.$$

Put $t^{-p}\sum_{i=1}^{n}(\mathrm{E}|X_i|)^p = e^{-p}$ and $t = e(\sum_{i=1}^{n}(\mathrm{E}|X_i|)^p)^{1/p}$. Then, we obtain

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq e(\sum_{i=1}^{n}(\mathrm{E}|X_i|)^p)^{1/p}\right) \leq e^{-p} \in (e^{-1}, 1). \tag{A6}$$

Combine (A5) and (A6), we obtain for all $t \geq 0$,

$$\mathrm{P}\left(|\sum_{i=1}^{n} X_i| \geq e(\sum_{i=1}^{n}(\mathrm{E}|X_i|)^t)^{1/t} + e(\sum_{i=1}^{n} v_i)2^{1/\theta}C_{\theta}(\frac{t}{\theta e^{11/12}})^{1/\theta}\right) \leq e^{-t}.$$

This completes the proof. □

Appendix A.9

Proof of Theorem 3. Note that for $\forall b \in S^{p-1}$, it yields

$$\begin{aligned} b^{\top}\hat{\mathcal{Q}}_n(\beta^*)b - b^{\top}\mathrm{E}(\hat{\mathcal{Q}}_n(\beta^*))b &\geq -\|b\|\max_{k,j}|[\hat{\mathcal{Q}}_n(\beta^*) - \mathrm{E}\hat{\mathcal{Q}}_n(\beta^*)]_{kj}| \\ = -\max_{k,j}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i+k)ke^{X_i^{\top}\beta^*}X_iX_i^{\top}}{(k+e^{X_i^{\top}\beta^*})^2} - \mathrm{E}\left(\frac{(Y_i+k)ke^{X_i^{\top}\beta^*}X_iX_i^{\top}}{(k+e^{X_i^{\top}\beta^*})^2}\right)\right]_{kj}\right|. \end{aligned} \tag{A7}$$

Consider the decomposition

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(Y_i+k)ke^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2} - \mathrm{E}\left(\frac{(Y_i+k)ke^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_ike^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2} - \mathrm{E}\left(\frac{Y_ike^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2}\right)\right] + \frac{k}{n}\sum_{i=1}^{n}\left[\frac{ke^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2} - \mathrm{E}\left(\frac{ke^{X_i^{\top}\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^{\top}\beta^*})^2}\right)\right] \end{aligned}$$

For the first term, we have under the $\mathcal{F}_{\max}$ with $t = C_{\min}/4$

$$\begin{aligned}
&\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i k e^{X_i^\top\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^\top\beta^*})^2}-\mathrm{E}\left(\frac{Y_i k e^{X_i^\top\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^\top\beta^*})^2}\right)\right]\right|\geq t,\mathcal{F}_{\max}\right)\\
&\leq 2\exp\left\{-\frac{1}{4}\left(\frac{n^2t^2}{2\sum\limits_{i=1}^{n}(X_{ik}X_{ij})^2(\|Y_i\|_{\psi_1}+|\frac{\exp(X_i^\top\beta^*)}{\log 2}|)^2}\wedge\frac{nt}{\max\limits_{1\leq i\leq n}|X_{ik}X_{ij}|(\|Y_i\|_{\psi_1}+|\frac{\exp(X_i^\top\beta^*)}{\log 2}|)}\right)\right\}\\
&\leq 2\exp\left\{-\frac{1}{4}\left(\frac{nt^2}{2M_X^4\log^{4/\theta}(\frac{2np}{\delta})M_{BX}^2}\wedge\frac{nt}{M_X^2\log^{2/\theta}(\frac{2np}{\delta})M_{BX}}\right)\right\}
\end{aligned}$$

where we use $ke^{X_i^\top\beta^*}(k+e^{X_i^\top\beta^*})^{-2}\leq 1$ and the second last inequality is from Corollary 2.

For the second term, by Theorem 1 and $\|X_{ik}X_{ij}\|_{\psi_{\theta/2}}\leq\|X_{ik}\|_{\psi_\theta}\|X_{ij}\|_{\psi_\theta}\leq M_X^2$ we have

$$\begin{aligned}
&\mathrm{P}\left(\left|\frac{k}{n}\sum_{i=1}^{n}\left[\frac{ke^{X_i^\top\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^\top\beta^*})^2}-\mathrm{E}\left(\frac{ke^{X_i^\top\beta^*}X_{ik}X_{ij}}{(k+e^{X_i^\top\beta^*})^2}\right)\right]\right|\geq t,\mathcal{F}_{\max}\right)\\
&\leq 2\exp\left\{-\left(\frac{t^{\theta/2}}{[4eC(\theta/2)\|b\|_2L_n(\theta/2,b)]^{\theta/2}}\wedge\frac{t^2}{16e^2C^2(\theta/2)\|b\|_2^2}\right)\right\}
\end{aligned}$$

where $b=(k/n)M_X^2(1,\ldots,1)^\top\in\mathbb{R}^n$.

Assume that $b^\top\mathrm{E}(\hat{\mathcal{Q}}_n(\beta))b\geq C_{\min}$ for all $b\in S^{p-1}$. Under $\mathcal{F}_1$ and $\mathcal{F}_2$, it shows that by (A7): $b^\top\mathrm{E}\,(\hat{\mathcal{Q}}_n(\beta))b\geq C_{\min}-\frac{C_{\min}}{2}=\frac{C_{\min}}{2}$. Then

$$\begin{aligned}
&\mathrm{P}\{\lambda_{\min}(\hat{\mathcal{Q}}_n(\beta))\leq\frac{C_{\min}}{2}\}=\mathrm{P}\left\{b^\top\mathrm{E}(\hat{\mathcal{Q}}_n(\beta))b\leq\frac{C_{\min}}{2},\forall b\in S^{p-1}\right\} &\text{(A8)}\\
&\leq\mathrm{P}\left\{b^\top\mathrm{E}(\hat{\mathcal{Q}}_n(\beta))b\leq\frac{C_{\min}}{2},\forall b\in S^{p-1},\mathcal{F}_{\max}\right\}+\mathrm{P}(\mathcal{F}_{\max}^c)\\
&\leq\mathrm{P}\{\mathcal{F}_1,\mathcal{F}_{\max}\}+\mathrm{P}\{\mathcal{F}_2,\mathcal{F}_{\max}\}+\mathrm{P}(\mathcal{F}_R^c(n))\\
&\leq 2p^2\exp\left\{-\frac{1}{4}\left(\frac{nt^2}{M_X^4\log^{4/\theta}(\frac{2np}{\delta})M_{BX}^2}\wedge\frac{nt}{M_X^2\log^{2/\theta}(\frac{2np}{\delta})M_{BX}}\right)\right\}\\
&+2p^2\exp\left\{-\left(\frac{t^{\theta/2}}{[4eC(\theta/2)\|b\|_2L_n(\theta/2,b)]^{\theta/2}}\wedge\frac{t^2}{16e^2C^2(\theta/2)\|b\|_2^2}\right)\right\}+\mathrm{P}(\mathcal{F}_{\max}^c). &\text{(A9)}
\end{aligned}$$

Then we have by conditioning on $\mathcal{F}_1\cap\mathcal{F}_2$

$$\delta_n(\beta):=\tfrac{3}{2}\|[\hat{\mathcal{Q}}_n(\beta)]^{-1}\hat{\mathcal{Z}}_n(\beta)\|_2\leq\tfrac{3}{C_{\min}}\|\hat{\mathcal{Z}}_n(\beta)\|_2.$$

By $k/(k+e^{X_i^\top\beta^*})\leq 1$, Corollary 2 implies for any $1\leq k\leq p$,

$$\begin{aligned}
\mathrm{P}\Big[\Big|\sqrt{\frac{p}{n}}\sum_{i=1}^{n}\frac{k(Y_i-e^{X_i^\top\beta^*})X_{ik}}{k+e^{X_i^\top\beta^*}}\Big|>&2\Big(\frac{2tp}{n}\sum_{i=1}^{n}X_{ik}^2\|Y_i-\mathrm{E}Y_i\|_{\psi_1}^2\Big)^{1/2}\\
&+2t\sqrt{\frac{p}{n}}\max_{1\leq i\leq n}|X_{ik}|\|Y_i-\mathrm{E}Y_i\|_{\psi_1}\Big]\leq 2e^{-t}.
\end{aligned}\tag{A10}$$

Let

$$\lambda_{1n}(t,X):=2\Big(\tfrac{2tp}{n}\max_{1\leq k\leq n}\sum_{i=1}^{n}X_{ik}^2\|Y_i-\mathrm{E}Y_i\|_{\psi_1}^2\Big)^{1/2}+2t\sqrt{\tfrac{p}{n}}\max_{1\leq i\leq n,1\leq k\leq p}(|X_{ik}|\|Y_i-\mathrm{E}Y_i\|_{\psi_1}).$$

We bound $\max_{1\le i\le n,1\le k\le p}|X_{ik}| \le M_X \log^{1/\theta}(\frac{2np}{\delta})$ and $\max_{1\le k\le n}\frac{1}{n}\sum_{i=1}^{n}X_{ik}^2 \le M_X^2\log^{2/\theta}(\frac{2np}{\delta})$ under the event $\mathcal{F}_{\max}$. Note that $M_{BX} = M_X + \frac{B}{\log 2}$, then (C.1) and (C.2) gives

$$\begin{aligned}\lambda_{1n}(t,X) &\le 2\Big(2tpM_{BX}^2 \max_{1\le k\le p}\frac{1}{n}\sum_{i=1}^{n}X_{ik}^2\Big)^{1/2} + 2t\sqrt{\frac{p}{n}}\max_{1\le i\le n,1\le k\le p}|X_{ik}|M_{BX}\\ &\le 2M_{BX}M_X(\sqrt{2tp}+t\sqrt{p/n})\log^{1/\theta}(2np/\delta) =: \lambda_n(t).\end{aligned}$$

So, $\mathrm{P}\Big(|\sqrt{\frac{p}{n}}\sum_{i=1}^{n}\frac{k(Y_i-e^{X_i^\top\beta^*})X_{ik}}{k+e^{X_i^\top\beta^*}}| > \lambda_n(t)\Big) \le 2e^{-t}$, $k=1,2,\ldots,p$. Thus (A10) shows

$$\begin{aligned}&\mathrm{P}\{\sqrt{n}\|\widehat{\mathcal{Z}}_n(\beta^*)\|_2 > \lambda_{1n}(t)\} \le \mathrm{P}\{\sqrt{n}\|\widehat{\mathcal{Z}}_n(\beta^*)\|_2 > \lambda_{1n}(t),\mathcal{F}_{\max}\} + \mathrm{P}(\mathcal{F}_{\max}^c)\\ &\le \mathrm{P}(\bigcup_{k=1}^{p}\{\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{k(Y_i-e^{X_i^\top\beta^*})X_{ik}}{k+e^{X_i^\top\beta^*}}\| > \frac{\lambda_{1n}(t)}{\sqrt{p}}\}) + \mathrm{P}(\mathcal{F}_{\max}^c) \le 2pe^{-t} + \mathrm{P}(\mathcal{F}_{\max}^c) = \delta+\varepsilon_n,\end{aligned}$$

where $t := \log(\frac{2p}{\delta})$. Then $\|\hat{\beta}_n-\beta^*\|_2 \le \delta_n(\beta^*) \le \frac{3}{C_{\min}}\|\hat{\mathcal{Z}}_n(\beta^*)\|_2 \le \frac{3\lambda_{1n}(t)}{C_{\min}\sqrt{n}}$ via Lemma 4. Under $\mathcal{F}_1\cap\mathcal{F}_2\cap\mathcal{F}_{\max}$, we obtain

$$\|\hat{\beta}_n-\beta^*\|_2 \le \frac{6M_{BX}M_X}{C_{\min}}\left[\sqrt{\frac{2p}{n}\log\Big(\frac{2p}{\delta}\Big)} + \frac{1}{n}\sqrt{p\log\Big(\frac{2p}{\delta}\Big)}\right]\log^{1/\theta}\Big(\frac{2np}{\delta}\Big).$$

Furthermore, under $\mathcal{F}_1\cap\mathcal{F}_2\cap\mathcal{F}_{\max}$, it gives the condition of n: (11). □

Appendix A.10

Proof of Theorem 4. For convenience, the proof is divided into three steps.

Step 1. Adopting the lemma

Lemma A1 (Computing the spectral norm on a net, Lemma 5.4 in [1]). *Let* $\mathbf{B}$ *be an* $p\times p$ *matrix, and let* $\mathcal{N}_\varepsilon$ *be an* ε*-net of* S^{p-1} *for some* $\varepsilon\in[0,1)$. *Then*

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

Then show that $\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\| \le 2\max_{x\in\mathcal{N}_{1/4}}|\frac{1}{n}\|\mathbf{A}x\|_2^2-1|$. Indeed, note that $\langle\frac{1}{n}\mathbf{A}^\top\mathbf{A}x-x,x\rangle = \langle\frac{1}{n}\mathbf{A}^\top\mathbf{A}x,x\rangle - 1 = \frac{1}{n}\|Ax\|_2^2-1$. By setting $\varepsilon=1/4$ in Lemma 4, we can obtain:

$$\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\| \le (1-2\varepsilon)^{-1}\sup_{x\in\mathcal{N}_\varepsilon}|\langle\frac{1}{n}\mathbf{A}^\top\mathbf{A}x-x,x\rangle| = 2\max_{x\in\mathcal{N}_{1/4}}\Big|\frac{1}{n}\|\mathbf{A}x\|_2^2-1\Big|.$$

Step 2. Let $Z_i := |\langle A_i,x\rangle|$ fix any $x\in S^{n-1}$. Observe that $\|\mathbf{A}x\|_2^2 = \sum_{i=1}^{n}|\langle A_i,x\rangle|^2 = \sum_{i=1}^{n}Z_i^2$. The fact that $\{Z_i\}_{i=1}^n$ are subW(θ) with $\mathrm{E}Z_i^2=1, \max_{1\le i\le n}\|Z_i\|_{\psi_\theta}=K$. Then by Corollary 4, Z_i^2 are independent subW$(\theta/2)$ r.v.s with $\max_{1\le i\le n}\|Z_i^2\|_{\psi_{\theta/2}} = K^2$. The norm triangle inequality (Lemma A.3 in [9]) gives

$$\max_{1\le i\le n}\|Z_i^2-1\|_{\psi_{\theta/2}} \le K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K. \tag{A11}$$

where $K_\alpha := 2^{1/\alpha}$ if $\alpha\in(0,1)$ and $K_\alpha=1$ if $\alpha\ge1$.

Denote $\boldsymbol{b}_X := \frac{1}{n}(\|Z_1^2-1\|_{\psi_{\theta/2}},\ldots,\|Z_n^2-1\|_{\psi_{\theta/2}})^\top$ in Theorem 1. With (A11), we have

$$\|\boldsymbol{b}_X\|_2 = n^{-1}\sqrt{\textstyle\sum_{i=1}^{n}\|Z_i^2-1\|_{\psi_{\theta/2}}^2} \le \frac{K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K}{\sqrt{n}}$$

and $\|\boldsymbol{b}\|_\infty \le \frac{K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K}{n}$.

For $\beta := \frac{\theta}{\theta-1} > 1$, we obtain

$$\|\boldsymbol{b}_X\|_\beta = n^{-1}\{\sum_{i=1}^n \|Z_i^2-1\|_{\psi_{\theta/2}}^\beta\}^{1/\beta} \le n^{\beta^{-1}-1}[K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K] = n^{-\theta^{-1}}K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K.$$

Write $L_n(\theta/2, \boldsymbol{b}_X)$ as the constant defined in Theorem 1(a). Then,

$$\|\boldsymbol{b}_X\|_2 L_n(\theta/2,\boldsymbol{b}_X) = \gamma^{4/\theta}\begin{cases} A(\theta/2)\|\boldsymbol{b}\|_\infty, & \theta \le 2 \\ B(\theta/2)\|\boldsymbol{b}\|_\beta, & \theta > 2. \end{cases}$$
$$\le K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2}]K\gamma^{4/\theta}\begin{cases} A(\theta/2)/n, & \theta \le 2 \\ B(\theta/2)/n^{1/\theta}, & \theta > 2. \end{cases}$$

Hence

$$2eC(\theta/2)\{\|\boldsymbol{b}_X\|_2\sqrt{t}+\|\boldsymbol{b}\|_2 L_n(\theta/2,\boldsymbol{b}_X)t^{2/\theta}\}$$
$$\le 2eKC(\theta/2)K_{\theta/2}[1+([(e\theta/2)^{\theta/2}]\log 2)^{-\theta/2})\left[\sqrt{\frac{t}{n}}+\begin{cases} A(\theta/2)(\gamma^2 t)^{2/\theta}/n, & \theta \le 2 \\ B(\theta/2)(\gamma^2 t)^{2/\theta}/n^{1/\theta}, & \theta > 2 \end{cases}\right]$$
$$=: H(t,n;\theta).$$

Therefore, $\mathrm{P}(\frac{1}{n}|\sum_{i=1}^n (Z_i^2-1)| \ge H(t,n;\theta)) \le 2e^{-t}$. Let $t = cp+s^2$ for constant c, then

$$\mathrm{P}\left\{\left|\frac{1}{n}\|\mathbf{A}\boldsymbol{x}\|_2^2-1\right| \ge H(cp+s^2,n;\theta)\right\} \le 2e^{-(cp+s^2)}.$$

Step 3. Consider the follow lemma for covering numbers in [1].

Lemma A2 (Covering numbers of the sphere). *For the unit Euclidean sphere S^{n-1}, the covering number $\mathcal{N}(S^{n-1},\varepsilon)$ satisfies $\mathcal{N}(S^{n-1},\varepsilon) \le (1+\frac{2}{\varepsilon})^n$ for every $\varepsilon > 0$.*

Then, we show the concentration for $\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\|$, and (12) follows by the definition of largest and least eigenvalues. The conclusion is drawn by Step 1 and 2:

$$\mathrm{P}\left\{\left\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\right\| \ge H(cp+s^2,n;\theta)\right\} \le \mathrm{P}\left\{2\max_{\boldsymbol{x}\in\mathcal{N}_{1/4}}\left|\frac{1}{n}\|\mathbf{A}\boldsymbol{x}\|_2^2-1\right| \ge H(cp+s^2,n;\theta)\right\}$$
$$\le \mathcal{N}(S^{n-1},1/4)\mathrm{P}\left\{\left|\frac{1}{n}\|\mathbf{A}\boldsymbol{x}\|_2^2-1\right| \ge H(cp+s^2,n;\theta)/2\right\} \le 2\cdot 9^n e^{-(cp+s^2)},$$

where the last inequality follows by Lemma A2 with $\varepsilon = 1/4$. When the $c \ge n\log 9/p$, then $2\cdot 9^n e^{-(cp+s^2)} \le 2e^{-s^2}$, and the (12) is proved.

Moreover, note that

$$\max_{||\boldsymbol{x}||_2=1}\left|\left\|\frac{1}{\sqrt{n}}\mathbf{A}\boldsymbol{x}\right\|_2^2-1\right| = \max_{||\boldsymbol{x}||_2=1}\left\|(\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p)\boldsymbol{x}\right\|_2^2 = \left\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\right\|^2 \le H^2(cp+s^2,n;\theta).$$

implies that

$$\sqrt{1-H^2(cp+s^2,n;\theta)} \le \tfrac{1}{\sqrt{n}}\lambda_{max}(\mathbf{A}) \le \sqrt{1+H^2(cp+s^2,n;\theta)}.$$

Similarly, for the minimal eigenvalue, we have

$$\min_{||\boldsymbol{x}||_2=1}\left|\left\|\frac{1}{\sqrt{n}}\mathbf{A}\boldsymbol{x}\right\|_2^2-1\right| = \min_{||\boldsymbol{x}||_2=1}\left\|(\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p)\boldsymbol{x}\right\|_2^2 = \left\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\right\|^2 \le H^2(cp+s^2,n;\theta).$$

This implies $\sqrt{1-H^2(cp+s^2,n;\theta)} \le \frac{1}{\sqrt{n}}\lambda_{min}(\mathbf{A}) \le \sqrt{1+H^2(cp+s^2,n;\theta)}$. So we obtain that the two events satisfy

$$\Big\{\|\frac{1}{n}\mathbf{A}^\top\mathbf{A}-\mathbf{I}_p\|^2 \le H^2(cp+s^2,n;\theta)\Big\}$$
$$\subset \Big\{\sqrt{1-H^2(cp+s^2,n;\theta)} \le \frac{1}{\sqrt{n}}\lambda_{min}(\mathbf{A}) \le \frac{1}{\sqrt{n}}\lambda_{max}(\mathbf{A}) \le \sqrt{1+H^2(cp+s^2,n;\theta)}\Big\}$$

Then we obtain the second conclusion in this theorem. □

Appendix A.11

Proof of Theorem 5. By independence and (13),

$$\mathrm{E}e^{\pm n\alpha_n\hat{Z}_{\alpha_n}(\theta)} = \prod_{i=1}^n \mathrm{E}\exp\{\pm\varphi^c[\alpha_n(X_i-\theta)]\} \le \prod_{i=1}^n \mathrm{E}[1\pm\alpha_n(X_i-\theta)+c(\alpha_n(X_i-\theta))]$$
$$\le \prod_{i=1}^n \exp\{\pm\alpha_n\mathrm{E}(X_i-\theta)+\mathrm{E}[c(\alpha_n(X_i-\theta))]\} = \exp\Big\{\pm\alpha_n\sum_{i=1}^n\mathrm{E}(X_i-\theta)+\sum_{i=1}^n\mathrm{E}[c(\alpha_n(X_i-\theta))]\Big\}. \tag{A12}$$

For convenience, let

$$B_n^+(\theta) = \mu_n-\theta+\frac{1}{n\alpha_n}\sum_{i=1}^n\mathrm{E}[c(\alpha_n(X_i-\theta))]+\frac{\log(\delta^{-1})}{n\alpha_n} \tag{A13}$$

and $B_n^-(\theta) = \mu_n-\theta-\frac{1}{n\alpha_n}\sum_{i=1}^n\mathrm{E}[c(\alpha_n(X_i-\theta))]-\frac{\log(\delta^{-1})}{n\alpha_n}$. Therefore, Equation (A12) and the Markov's inequality show

$$\mathrm{P}(\hat{Z}_{\alpha_n}(\theta)\ge B_n^+(\theta)) = \mathrm{P}(e^{n\alpha_n\hat{Z}_{\alpha_n}(\theta)}\ge e^{n\alpha_nB_n^+(\theta)}) \le \frac{\mathrm{E}e^{n\alpha_n\hat{Z}_{\alpha_n}(\theta)}}{e^{n\alpha_nB_n^+(\theta)}} \le \frac{e^{n\alpha_nB_n^+(\theta)-\log(\delta^{-1})}}{e^{n\alpha_nB_n^+(\theta)}} = \delta$$

and $\mathrm{P}(\hat{Z}_{\alpha_n}(\theta)\le B_n^-(\theta)) = \mathrm{P}(e^{-n\alpha_n\hat{Z}_{\alpha_n}(\theta)}\ge e^{-n\alpha_nB_n^-(\theta)}) \le \frac{\mathrm{E}e^{-n\alpha_n\hat{Z}_{\alpha_n}(\theta)}}{e^{-n\alpha_nB_n^-(\theta)}} \le \frac{e^{-n\alpha_nB_n^-(\theta)-\log(\delta^{-1}}}{e^{-n\alpha_nB_n^-(\theta)}} = \delta$. These two inequality yield $\mathrm{P}\big(B_n^-(\theta)<\hat{Z}_{\alpha_n}(\theta)\big) = 1-\mathrm{P}\big(\hat{Z}_{\alpha_n}(\theta)\le B_n^-(\theta)\big)-\mathrm{P}\big(\hat{Z}_{\alpha_n}(\theta)\ge B_n^+(\theta)\big)\ge 1-2\delta$.

The $\frac{\partial\hat{Z}_{\alpha_n}(\theta)}{\partial\theta} = -\frac{1}{n}\sum_{i=1}^n\dot{\varphi}^c[\alpha_n(X_i-\theta)] < 0$ implies the map $\theta\mapsto\hat{Z}_{\alpha_n}(\theta)$ is non-increasing. If $\theta=\mu_n$, we have $B_n^+(\mu_n)>0$ from (A13). As n is sufficient large and $\alpha_n\to 0$, in $B_n^+(\theta)$, from (C.1.2) the term $\frac{1}{n\alpha_n}\sum_{i=1}^n\mathrm{E}[c(\alpha_n(X_i-\theta))] \le \frac{f(\alpha_n)}{\alpha_n}\frac{1}{n}\sum_{i=1}^n\mathrm{E}[c(X_i-\theta)] = \frac{f(\alpha_n)}{\alpha_n}O(1)$ converges to 0 by (C.1.3). Then, there must be a constant $d_n(c)>0$ such that $B_n^+(\mu_n+d_n(c))<0$. So under (16), it implies that $B_n^+(\theta)=0$ has a solution and denote the smallest solution $\theta_+\in(\mu_n,\mu_n+d_n(c))$. Similarly, for $B_n^-(\theta)$, we have $B_n^-(\mu_n)<0$. The condition (16) implies $B_n^-(\mu_n-d_n(c))>0$, then $B_n^-(\theta)=0$ has a solution and denote the largest solution $\theta_-\in(\mu_n-d_n(c),\mu_n)$. Note that $\hat{Z}_{\alpha_n}(\theta)$ is a continuous and non-increasing function, the estimating equation $\hat{Z}_{\alpha_n}(\theta)=0$ has a solution $\hat{\theta}_{\alpha_n}\in[\theta_-,\theta_+]$ such that $\theta_-\le\hat{\theta}_{\alpha_n}\le\theta_+$ with a probability at least $1-2\delta$. Recall that

$$B_n^+(\theta_+) = \mu_n-\theta_++\frac{1}{n\alpha_n}\sum_{i=1}^n\mathrm{E}[c(\alpha_n(X_i-\theta_+))]+\frac{\log(\delta^{-1})}{n\alpha_n} = 0. \tag{A14}$$

has the smallest solution $\theta_+\in(\mu_n,\mu_n+d_n(c))$ under the condition (16). We have

$$\begin{aligned}\mu_n - \hat{\theta}_{\alpha_n} \geq \mu_n - \theta_+ &= -\frac{1}{n\alpha_n}\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n X_i - \alpha_n\theta_+)] - \frac{\log(\delta^{-1})}{n\alpha_n} \\ &= -\frac{1}{n\alpha_n}\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n(X_i-\mu_n) + \alpha_n(\mu_n-\theta_+))] - \frac{\log(\delta^{-1})}{n\alpha_n} \\ [\text{By (C.1.1)}] \ &\geq -\frac{c_2}{n\alpha_n}\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n X_i - \alpha_n\mu_n)] - \frac{c_2}{\alpha_n}\cdot c(\alpha_n(\mu_n-\theta_+)) - \frac{\log(\delta^{-1})}{n\alpha_n}\end{aligned} \tag{A15}$$

which implies

$$\mu_n - \theta_+ + \frac{c_2}{\alpha_n}\cdot c(\alpha_n(\mu_n-\theta_+)) \geq -\left(\frac{c_2}{n\alpha_n}\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n(X_i-\mu_n))] + \frac{\log(\delta^{-1})}{n\alpha_n}\right). \tag{A16}$$

Put $\frac{c_2}{n\alpha_n}\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n(X_i-\mu_n))] = \frac{\log(\delta^{-1})}{n\alpha_n}$, i.e., $\sum_{i=1}^{n} c_2\mathrm{E}[c(\alpha_n(X_i-\mu_n))] = \log(\delta^{-1})$. The scaling assumption $c(tx) \leq f(t)c(x)$ gives

$$f(\alpha_n)c_2\sum_{i=1}^{n}\mathrm{E}[c(X_i-\mu_n)] \geq c_2\sum_{i=1}^{n}\mathrm{E}[c(\alpha_n(X_i-\mu_n))] = \log(\delta^{-1})$$

and thus $\alpha_n \geq f^{-1}\left(\frac{\log(\delta^{-1})}{c_2\sum_{i=1}^{n}\mathrm{E}[c(X_i-\mu_n)]}\right)$. Let $g_{\alpha_n}(t) = t + \frac{c_2}{\alpha_n}c(\alpha_n t)$. Moreover, Equation (A16) and the value α_n yields

$$g_{\alpha_n}(\mu_n-\theta_+) = \mu_n - \theta_+ + \frac{c_2}{\alpha_n}c(\alpha_n(\mu_n-\theta_+)) \geq -\frac{2\log(\delta^{-1})}{n\alpha_n}.$$

Solve the above inequality in terms of $\mu_n - \theta_+$, we obtain

$$\mu_n - \hat{\theta}_{\alpha_n} \geq \mu_n - \theta_+ \geq g_{\alpha_n}^{-1}\left\{-\tfrac{2\log(\delta^{-1})}{n\alpha_n}\right\}.$$

Similarly, for θ_-, one has $\mu_n - \hat{\theta}_{\alpha_n} \leq \mu_n - \theta_- \leq -g_{\alpha_n}^{-1}\left\{-\tfrac{2\log(\delta^{-1})}{n\alpha_n}\right\}$. Then we obtain that (17) holds with probability at least $1-2\delta$. □

References

1. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge University Press: Cambridge, UK, 2018; Volume 47.
2. Bai, Z.; Silverstein, J.W. *Spectral Analysis of Large Dimensional Random Matrices*; Springer: New York, NY, USA, 2010; Volume 20.
3. Wainwright, M.J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*; Cambridge University Press: Cambridge, UK, 2019; Volume 48.
4. Zhang, H.; Chen, S.X. Concentration Inequalities for Statistical Inference. *Commun. Math. Res.* **2021**, *37*, 1–85.
5. Tropp, J.A. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **2015**, *8*, 1–230. [CrossRef]
6. Kuchibhotla, A.K.; Chakrabortty, A. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Inf. Inference J. Imag.* **2022**, *ahead of print*. [CrossRef]
7. Hao, B.; Abbasi-Yadkori, Y.; Wen, Z.; Cheng, G. Bootstrapping Upper Confidence Bound. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
8. Gbur, E.E.; Collins, R.A. Estimation of the Moment Generating Function. *Commun. Stat. Simul. Comput.* **1989**, *18*, 1113–1134. [CrossRef]
9. Götze, F.; Sambale, H.; Sinulis, A. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electron. J. Probab.* **2021**, *26*, 1–22. [CrossRef]
10. Li, S.; Wei, H.; Lei, X. Heterogeneous Overdispersed Count Data Regressions via Double-Penalized Estimations. *Mathematics* **2022**, *10*, 1700. [CrossRef]
11. Rigollet, P.; Hütter, J.C. High Dimensional Statistics. 2019. Available online: http://www-math.mit.edu/rigollet/PDFs/RigNotes17.pdf (accessed on 20 April 2022).
12. Foss, S.; Korshunov, D.; Zachary, S. *An Introduction to Heavy-Tailed and Subexponential Distributions*; Springer: New York, NY, USA, 2011.
13. De la Pena, V.; Gine, E. *Decoupling: From Dependence to Independence*; Springer: Berlin/Heidelberg, Germany, 2012.
14. Latala, R. Estimation of moments of sums of independent real random variables. *Ann. Probab.* **1997**, *25*, 1502–1513. [CrossRef]

15. Kashlak, A.B. Measuring distributional asymmetry with Wasserstein distance and Rademacher symmetrization. *Electron. J. Stat.* **2018**, *12*, 2091–2113. [CrossRef]
16. Vladimirova, M.; Girard, S.; Nguyen, H.; Arbel, J. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat* **2020**, *9*, e318. [CrossRef]
17. Wong, K.C.; Li, Z.; Tewari, A. Lasso guarantees for β-mixing heavy-tailed time series. *Ann. Stat.* **2020**, *48*, 1124–1142. [CrossRef]
18. Portnoy, S. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Stat.* **1988**, *16*, 356–366. [CrossRef]
19. Kuchibhotla, A.K. Deterministic inequalities for smooth m-estimators. *arXiv* **2018**, arXiv:1809.05172.
20. Zhang, H.; Jia, J. Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signals detection. *Stat. Sin.* **2022**, *32*,181–207. [CrossRef]
21. Bai, Z.D.; Yin, Y.Q. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*; World Scientific: Singapore, 1993; pp. 1275–1294.
22. Chen, P.; Jin, X.; Li, X.; Xu, L. A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1,2)$. *Electron. J. Stat.* **2021**, *15*, 5523–5544. [CrossRef]
23. Xu, L.; Yao, F.; Yao, Q.; Zhang, H. Non-Asymptotic Guarantees for Robust Statistical Learning under $(1+\varepsilon)$-th Moment Assumption. *arXiv* **2022**, arXiv:2201.03182.
24. Lerasle, M. Lecture notes: Selected topics on robust statistical learning theory. *arXiv* **2019**, arXiv:1908.10761.
25. Devroye, L.; Lerasle, M.; Lugosi, G.; Oliveira, R.I. Sub-gaussian mean estimators. *Ann. Stat.* **2016**, *44*, 2695–2725. [CrossRef]
26. Zhang, A.R.; Zhou, Y. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat* **2020**, *9*, e314. [CrossRef]
27. Zhang, H.; Tan, K.; Li, B. COM-negative binomial distribution: modeling overdispersion and ultrahigh zero-inflated count data. *Front. Math. China* **2018**, *13*, 967–998. [CrossRef]
28. Zajkowski, K. On norms in some class of exponential type Orlicz spaces of random variables. *Positivity* **2019**, *24*, 1231–1240. [CrossRef]
29. Jameson, G.J. A simple proof of Stirling's formula for the gamma function. *Math. Gaz.* **2015**, *99*, 68–74. [CrossRef]
30. Alzer, H. On some inequalities for the gamma and psi functions. *Math. Comput.* **1997**, *66*, 373–389. [CrossRef]