Bernoulli **27**(1), 2021, 192–217
https://doi.org/10.3150/20-BEJ1233

High-dimensional CLT: Improvements, non-uniform extensions and large deviations

ARUN KUMAR KUCHIBHOTLA[*], SOMABHA MUKHERJEE[†] and DEBAPRATIM BANERJEE[‡]

Department of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut Street, Jon M. Huntsman Hall, Philadelphia, PA 19104, USA. E-mail: [*]arunku@upenn.edu; [†]somabha@upenn.edu; [‡]dban@upenn.edu

Central limit theorems (CLTs) for high-dimensional random vectors with dimension possibly growing with the sample size have received a lot of attention in the recent times. Chernozhukov et al. (*Ann. Probab.* **45** (2017) 2309–2352) proved a Berry–Esseen type result for high-dimensional averages for the class of sparsely convex sets including hyperrectangles as a special case and they proved that the rate of convergence can be upper bounded by $n^{-1/6}$ up to a polynomial factor of $\log p$ (where n represents the sample size and p denotes the dimension). Convergence to zero of the bound requires $\log^7 p = o(n)$. We improve upon their result, for hyperrectangles, which only requires $\log^4 p = o(n)$ (in the best case). This improvement is made possible by a sharper dimension-free anti-concentration inequality for Gaussian process on a compact metric space. In addition, we prove two non-uniform variants of the high-dimensional CLT based on the large deviation and non-uniform CLT results for random variables in a Banach space by Bentkus, Račkauskas, and Paulauskas. We apply our results in the context of post-selection inference in linear regression and of empirical processes.

Keywords: anti-concentration; Cramér type large deviation; empirical processes; nonuniform CLT; Orlicz norms; post-selection inference

1. Introduction

In modern statistical applications like high dimensional estimation and multiple hypothesis testing problems [4], the dimension of the data is often much larger than the sample size. As can be expected from the classical asymptotic theory, the central limit theorem plays a pivotal role for inference. In this paper, we prove three variants of the high-dimensional central limit theorem. The setting we use is as follows. Consider independent mean zero random vectors $X_1, \ldots, X_n \in \mathbb{R}^p$ with covariance matrices $\Sigma_i := \mathbb{E}[X_i X_i^\top] \in \mathbb{R}^{p\times p}$. Here p is allowed to be larger than n. Define the scaled average

$$S_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}} \in \mathbb{R}^p.$$

Let Y_i (for $1 \le i \le n$) represent a multivariate Gaussian random vector with mean zero and variance-covariance matrix Σ_i. Define the corresponding scaled average as

$$U_{n,0} := \frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \in \mathbb{R}^p.$$

The problem of central limit theorem is the comparison of the probabilities $\mathbb{P}(S_n \in A)$ and $\mathbb{P}(U_{n,0} \in A)$ for $A \subseteq \mathbb{R}^p$. When p is fixed (or only grows at most sublinearly in n) which we refer to as multivariate setting, classical results show the rate $O(p^{7/4}/n^{1/2})$; see [35,36] and [6]. The case where p is allowed to grow faster than n, which we refer to as high-dimensional setting, has received significant interest in the recent times. The series of papers [12,14,16] have studied this problem extensively under general

conditions on the random vectors when the sets A are sparsely convex sets and in particular hyperrectangles. The main result of [16] bounds the difference $|\mathbb{P}[S_n \in A] - \mathbb{P}[U_{n,0} \in A]|$ uniformly over $A \in \mathcal{A}^{re}$ as a function of n and p. Here $\mathcal{A}^{re}$ is the class of all hyperrectangles. Proposition 2.1 of [16] implies that

$$\sup_{A \in \mathcal{A}^{re}} \left|\mathbb{P}(S_n \in A) - \mathbb{P}(U_{n,0} \in A)\right| \leq C\left(\frac{\log^7(pn)}{n}\right)^{1/6}, \tag{1}$$

under certain exponential tail assumption and a constant C depending on the distribution of the random vectors X_i, $1 \leq i \leq n$. In their earlier papers, [12,14] a special sub-class of sets $\mathcal{A}^m \subset \mathcal{A}^{re}$ was considered, where $\mathcal{A}^m$ is the class of all sets A of the form $A = \{x \in \mathbb{R}^p : x(j) \leq a \text{ for all } 1 \leq j \leq p\}$. Here and throughout, we use the notation $x(j)$ for a vector $x \in \mathbb{R}^p$ to represent the j-th coordinate of x.

From the bound (1) we require $\log^7(pn) = o(n)$ for the difference of probabilities to converge to zero uniformly. One of the lingering questions in high-dimensional CLT is the "correct" exponent of $\log(pn)$ that needs to be $o(n)$ for convergence to zero. By proving a dimension-free anti-concentration inequality and comparing the proof techniques from [7,16,30], we reduce the requirement to $\mu^6 \log^4(ep) = o(n)$ when the sets A are restricted to l_∞ balls in $\mathbb{R}^p$. Here μ represents the median of $\|U_{n,0}\|_\infty$. In the best case where $\mu \asymp 1$ the requirement becomes $\log^4(ep) = o(n)$ and in the worst case where $\mu \asymp \sqrt{\log(ep)}$ the requirement becomes $\log^7(ep) = o(n)$ coinciding with the requirement from (1).

Since the dependence, $n^{-1/6}$, on the sample size in bound (1) is larger than dependence $n^{-1/2}$ that appears in multivariate Berry–Esseen bounds, the result (1) does not provide useful information when the probability $\mathbb{P}(U_{n,0} \in A)$ is smaller. This leads naturally to the question of non-uniform version of (1). In particular, an interesting question is to find quantitative upper bound on

$$\left|\frac{\mathbb{P}(S_n \in A^c)}{\mathbb{P}(U_{n,0} \in A^c)} - 1\right|, \tag{2}$$

as a function of p and n. In this paper, we consider a special class of sets of the form

$$A = \left\{x \in \mathbb{R}^p : -a \leq x(j) \leq a \text{ for all } 1 \leq j \leq p\right\}, \tag{3}$$

and find an upper bound on (2). Note that sets of the form (3) are l_∞ balls. Another variant of non-uniform CLT is to consider how the difference $|\mathbb{P}(S_{n\infty} \leq r) - \mathbb{P}(U_{n,0_\infty} \leq r)|$ scales with r for $r \geq 0$. To this end we prove upper bounds on

$$\sup_{r \geq 0} r^m \left|\mathbb{P}(S_{n\infty} \leq r) - \mathbb{P}(Y_\infty \leq r)\right| \quad \text{for } m \geq 0. \tag{4}$$

The first problem (2) is well-studied in the classical multivariate setting under the name "Cramér-type large deviation". We refer to the encyclopedic work [34] for a review of the extensive literature on Cramér-type large deviation for sums of independent random variables along with extensions to multivariate random vectors. The classical result for univariate ($p = 1$) i.i.d. random variables is of the form

$$\frac{\mathbb{P}(S_n > r)}{\mathbb{P}(U_{n,0} > r)} = \exp\left(\frac{Cr^3}{6\Sigma_1^{3/2} n^{1/2}}\right)\left[1 + C\left(\frac{r+1}{\sqrt{n}}\right)\right] \tag{5}$$

for $0 \leq r = O(n^{1/6})$ and C a constant depending on the distribution of X_1. See Theorem 5.23 (and Section 5.8) of [33] for details. The second problem (4) is usually referred to as a "non-uniform CLT".

A result of this kind is also useful in proving convergence of moments. The classical result for univariate i.i.d. random variables of type (4) is given by

$$\left|\mathbb{P}(S_n \le r) - \mathbb{P}(U_{n,0} \le r)\right| \le C(m)\big(1+|r|\big)^{-m}\left(\frac{\mathbb{E}[|X_1|^3]}{\Sigma_1^{3/2} n^{1/2}} + \frac{\mathbb{E}[|X_1|^m]}{\Sigma_1^{m/2} n^{(m-2)/2}}\right) \tag{6}$$

for all $r \in \mathbb{R}, m \ge 3$ and for some constant $C(m)$ depending only on m. See Theorem 5.15 and (Section 5.5) of [33] for details and other results in this direction. Also, see [35] for multivariate setting. The classical results (5) and (6) provide rates that scale like $n^{-1/2}$ (as a function of n) in the non-uniform versions of CLT as does the classical Berry–Esseen bound. In lines with the Berry–Esseen type result (1) in high-dimensional setting, we derive rates in large deviation and non-uniform CLT with a scaling of order $n^{-1/6}$ as a function of the sample size n.[1]

It is well-known [10] that the rate $n^{-1/6}$ is optimal in the central limit theorem for Banach spaces and the space $(\mathbb{R}^p, \cdot_\infty)$ with p diverging behaves like an infinite-dimensional space. In this respect it is of particular interest to look back at the rich literature on the CLTs for Banach space valued random variables. These old and well-known large deviation and non-uniform CLTs for Banach space play a central role in the derivation of ours presented here. The basic setting for these results is as follows: Suppose $X_1, \ldots, X_n$ are i.i.d. random variables taking values in a Banach space B such that $\mathbb{E}[X_1] = 0$ and Y is a mean zero B-valued Gaussian random variable with the same covariance (operator) as X_1. The problem as before is the study of closeness of the distributions of Y and S_n where $\cdot$ is a Banach space norm. Several results on this problem are available in [7,32]. For a historical account of these results, see [32], p. 142. The papers [8,9] and [31] are of particular interest to us since they provide bounds on (2) and (4) for Banach space valued random variables. In [9] and [8] the problem of the convergence of ratio $\mathbb{P}(S_n > r)/\mathbb{P}(Y > r)$ to 1 was considered and it was proved that

$$\left|\mathbb{P}(S_n > r)/\mathbb{P}(Y > r) - 1\right| \le M_2(r+1)n^{-1/6}$$

for $0 \le r \le -1 + M_1 n^{1/6}$ where M_1 and M_2 are constants depending on the distribution of X_1. The non-uniform version of central limit theorem is also available for the Banach spaces from [31]. Understanding the implication of these results for the high-dimensional case is one contribution of our paper.

1.1. Our contributions

As described in the Introduction, we study uniform and non-uniform variants of high dimensional central limit theorem. In the process, we prove a sharper version of anti-concentration inequality for centered Gaussians. We assume that $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent random vectors with mean 0 and covariance matrices $\Sigma_i = \mathbb{E}[X_i X_i^\top] \in \mathbb{R}^{p\times p}$. Let $Y_1, \ldots, Y_n$ be centered Gaussian random vectors in $\mathbb{R}^p$ satisfying $\mathbb{E}[Y_i Y_i^\top] = \Sigma_i$ for $1 \le i \le n$. Our results are intended for the case where $\log(ep)$ grows sublinearly in the sample size n, although we do not explicitly make an assumption that $\log(ep)$ or p grows with n. Define for $m \ge 0$

$$\Delta_n^{(m)} := \sup_{r\ge 0} r^m \left|\mathbb{P}\big(\|S_n\|_\infty \le r\big) - \mathbb{P}\big(\|U_{n,0}\|_\infty \le r\big)\right|.$$

[1] Koike [22] and Chernozhukov et al. [17] recently proved an improvement in (1) with a scaling of order $n^{-1/4}$. The proof techniques are far more complicated and will not be compared with in this paper.

(i) One of the main ingredients of central limit theorems (both uniform and non-uniform versions) is an anti-concentration inequality which bounds $\mathbb{P}(r-\varepsilon \le \|U_{n,0}\|_\infty \le r+\varepsilon)$ over all $r \ge 0$ and $\varepsilon > 0$. We prove that for any $\varepsilon > 0$, $m \ge 0$

$$\sup_{r\ge 0} r^m \mathbb{P}\big(r-\varepsilon \le \|U_{n,0}\|_\infty \le r+\varepsilon\big) \le \Phi_{\mathrm{AC},m}\varepsilon,$$

where $\Phi_{\mathrm{AC},m} = O(\mu^{m+1})$ (assuming $\sigma_{\max}$ and $\sigma_{\min}^{-1}$ are of constant order) and μ denotes the median of $\|U_{n,0}\|_\infty$. Quantities $\sigma_{\max}^2$ and $\sigma_{\min}^2$ are given by the maximum and minimum variances of $U_{n,0}(j)$, $1 \le j \le p$. This provides a refinement of Lemma 3.1 of [8] with exact constants. Chernozhukov et al. [16] (based on the result of [28]) prove the above result for $m = 0$ case with $\Phi_{\mathrm{AC},0} = C\sqrt{\log(ep)}$ and since μ is at most of the order $\sqrt{\log(ep)}$, our result is sharper.

(ii) We compare the modern proof technique of [16] and the classical proofs from the Banach space CLT literature [7,32]; see Section 6 for details. Based on this we improve upon the proof of [16] to get better rates in high-dimensional CLT.

(iii) If X_i are sub-Weibull of order α, that is, $\|X_i(j)\|_{\psi_\alpha} \le K_p < \infty$ for all $1 \le i \le n$, $1 \le j \le p$, then

$$\Delta_n^{(0)} \le \Theta_\alpha \Phi_{\mathrm{AC},0} K_p \left(\frac{\log^4(ep)}{n}\right)^{1/6} + \Theta_\alpha \Phi_{\mathrm{AC},0} K_p \frac{(\log(ep))^{1+1/\alpha}(\log n)^{1/\alpha}}{n^{1/2}}$$

for some constant Θ_α depending only on α. Proposition 2.1 of [16] for $\alpha = 1$ proves that $\Delta_n^{(0)} \le C(\log^7(pn)/n)^{1/6}$ which requires $\log(pn) = o(n^{1/7})$. In contrast if $\Phi_{\mathrm{AC},0} = O(1)$, then our result only requires $\log(ep) = o(n^{1/4})$ which is the weakest requirement till date.

(iv) Under the same assumptions in (iii), we have for $m \ge 0$

$$\Delta_n^{(m)} \le \Theta \Phi_{\mathrm{AC},m} \left(\frac{\log^4(epn)}{n}\right)^{1/6} + \begin{cases} 0 & \text{if } \alpha > 1, \\ \Theta_{\alpha,m}\big(\big(\log(epn)\big)^{5/4+3/\alpha}/n\big)^{(m+1)/3} & \text{if } \alpha \in (0,1]. \end{cases}$$

More generally, for any function $\phi(\cdot)$ satisfying $\phi(x+y) \le \phi(x)\phi(y)$ our methods can be used to obtain bounds for

$$\Delta_n^{(\phi)} := \sup_{r\ge 0} \phi(r)\big|\mathbb{P}\big(\|S_n\|_\infty \le r\big) - \mathbb{P}\big(\|U_{n,0}\|_\infty \le r\big)\big|.$$

These are analogues to the results of [31].

(v) Finally, we derive a Cramér-type large deviation in the high dimensional CLT setting. We assume that $X_1, \ldots, X_n$ are independent and identically distributed and that $\mathbb{E}[\exp\{HX_{1\infty}\}] < \infty$ for some $H > 0$. Under these assumptions, we prove that

$$\big|\mathbb{P}(S_{n\infty} > r)/\mathbb{P}(Y_\infty > r) - 1\big| \le M_1(r+1)n^{-1/6}$$

for any $r \le -1 + M_2 n^{1/6}$. Here M_1 and M_2 are constants depending on the distribution of X_1 and H which can be bounded by polynomials of $\log p$, under certain tail assumptions on the coordinates of X_1. The proof is motivated by the techniques of [9] and is modified for the high dimensional set up. The constants M_1 and M_2 are also made explicit in Theorem 5.1.

1.2. Organization of the paper

The paper is organized as follows. In Section 2, we define some useful notations. Section 3 is dedicated to our main results. In this section, we state the anti-concentration inequalities and prove refined uniform as well as non-uniform central limit theorems. In Section 4,

we present applications of our results to post-selection inference where the anti-concentration constant $\Phi_{\mathrm{AC},0}$ can be of order much smaller than $\sqrt{\log(ep)}$ and to bounding the expectation of suprema of empirical processes over a (possibly infinite) weak VC-major function class. In Section 5, we prove a Cramér-type large deviation based on the results of [9]. In Section 6, we present an outline of our proof of CLTs with detailed discussion on differences of proofs from other works. Finally, we conclude with a summary and future directions in Section 7. Proofs of all the results are given in the Supplementary Material [25].

2. Preliminaries

2.1. Notation and setting

As discussed earlier, we shall consider independent random vectors $X_1, \ldots, X_n \in \mathbb{R}^p$ with mean zero and covariance matrices $\Sigma_i, 1 \le i \le n$. Let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ denote Gaussian random vectors with mean and covariance matching that of $X_i, 1 \le i \le n$. The l_∞ norm on $\mathbb{R}^p$ is denoted by $\cdot$. By writing $a \lesssim b$, we mean that for some constant C, $a \le Cb$. We also use the following notation throughout the paper.

$$S_n := n^{-1/2}(X_1 + X_2 + \cdots + X_n),$$
$$U_{n,k} := n^{-1/2}(X_1 + \cdots + X_k + Y_{k+1} + \cdots + Y_n) \quad \text{for } 0 \le k \le n.$$

Note that $U_{n,n} = S_n$ and hence proving the closeness (in distribution) of $U_{n,k}$ and $U_{n,0}$ for all k ensures closeness of S_n and $U_{n,0}$. In this regard, define for $m \ge 0$

$$\delta_{n,m} := \sup_{r \ge 0} \max_{1 \le k \le n} r^m \left| \mathbb{P}\big(\|U_{n,k}\| \le r\big) - \mathbb{P}\big(\|U_{n,0}\| \le r\big) \right|. \tag{7}$$

Also define for $1 \le i \le n$ the signed measure ζ_i by

$$\zeta_i(A) := \mathbb{P}(X_i \in A) - \mathbb{P}(Y_i \in A) \quad \text{for } A \subseteq \mathbb{R}^p.$$

Based on this signed measure, $|\zeta_i|$ denotes the variation of measure ζ_i. It is clear that

$$\int d\zeta_i(x) = \int x(j)\, d\zeta_i(x) = \int x(j)x(k)\, d\zeta_i(x) = 0 \quad \text{for } 1 \le i \le n, 1 \le j, k \le p. \tag{8}$$

Define the "weak" third pseudo-moment as

$$L_n := \frac{1}{n} \sum_{i=1}^n \max_{1 \le j \le p} \int \big|x(j)\big|^3 |\zeta_i|(dx), \tag{9}$$

and the truncated "strong" second pseudo-moment as

$$M_n(\phi) := \frac{1}{n} \sum_{i=1}^n \int \|x\|^2 \mathbb{1}\big\{\|x\| \ge n^{1/2}\phi / \log(ep)\big\} |\zeta_i|(dx) \quad \text{for } \phi > 0. \tag{10}$$

These are called pseudo-moments since they are defined with respect to the variation measure and becomes zero if the distributions of X_i's and Y_i's are the same. Most of the results in classical multivariate setting (of [36]) and in Banach spaces (of [31]) are derived in terms of pseudo-moments. We will present our results also in terms of the pseudo-moments defined above.

Quantity $M_n(\phi)$ defined above is similar to the one defined in [16] except that they have $\|x\|^3$ instead of $\|x\|^2$. This subtlety allows us to derive better rates when random vectors only have $(2+\tau)$-moments.

Further, set $\mu_i := \text{median}(\|Y_i\|)$ and $\sigma_i^2 := \max_{1\le j\le p} \text{Var}(Y_i(j))$ for $1 \le i \le n$. Define the weighted "weak" third moment as

$$\bar{L}_{n,0} := \frac{1}{n}\sum_{i=1}^{n}(\mu_i+\sigma_i)\max_{1\le j\le p}\int \big|x(j)\big|^3|\zeta_i|(dx), \tag{11}$$

and for $m > 0$,

$$\bar{L}_{n,m} := \frac{1}{n}\sum_{i=1}^{n}\big[\mu_i^{m+1} + \sigma_i^{m+1}\big((m+1)/e\big)^{(m+1)/2}\big]\max_{1\le j\le p}\int \big|x(j)\big|^3|\zeta_i|(dx).$$

It is clear that $\bar{L}_{n,m}$ is at most of order $(\sqrt{\log(ep)})^{m+1}$ for $m \ge 0$ (assuming σ_i's are all of order 1). More precisely, we have

$$\bar{L}_{n,m} \le \Theta_m L_n\Big(\max_{1\le i\le n}\sigma_i\Big)^{m+1}\big(\log(ep)\big)^{(m+1)/2} \quad \text{for all } m \ge 0. \tag{12}$$

Here Θ_m is a constant depending only on $m \ge 0$ scaling like $m^{m/2}$.

2.2. Tail of Gaussian processes and anti-concentration inequality

In this section, we prove various inequalities regarding the distribution function of the maximum of a Gaussian process on a compact metric space. These inequalities will lead to the sharper versions of anti-concentration inequalities and are crucial for the large deviation result derived later. The following result is a refinement of Lemma 3.1 in [8] and makes use of a result from [18].

Theorem 2.1. *Let Y be a sample continuous centered Gaussian process on a compact metric space S such that $\sigma_{\min}^2 \le \mathbb{E}[Y^2(s)] \le \sigma_{\max}^2$ for all $s \in S$. Let μ denote the median of $Y := \sup_{s\in S}|Y(s)|$. Then the following are true*:

1. *For all $r \ge 0$,*
$$\mathbb{P}(Y > r) \ge \frac{1}{6}\exp\Big(-\frac{r^2}{\sigma_{\max}^2}\Big).$$
2. *For all $r, \varepsilon \ge 0$,*
$$\mathbb{P}(Y > r-\varepsilon) \le 20\exp\big(\Phi_4(r+1)\varepsilon\big)\mathbb{P}(Y > r),$$
where Φ_4 is given by
$$\Phi_4 := 1 + \frac{56(\mu+1.5\sigma_{\max})(\mu+4.1\sigma_{\max})}{\sigma_{\max}^2\sigma_{\min}^2} + \frac{32\pi(2.6\sigma_{\min}+\mu)^2(\sigma_{\min}^2+32\sigma_{\min}\mu+12\mu^2)}{\sigma_{\min}^6}.$$

3. *For all $r \geq 0$ and $\varepsilon > 0$,*

$$\mathbb{P}(r - \varepsilon \leq Y \leq r + \varepsilon) \leq \Phi_2 \varepsilon (r+1) \mathbb{P}(Y > r - \varepsilon), \tag{13}$$

where Φ_2 is given by

$$\Phi_2 := \max\left\{ \frac{51(\mu + 4.1\sigma_{\max})}{\sigma_{\min}^2}, \frac{32\pi(\mu + 2.6\sigma_{\min})^2}{\sigma_{\min}^4} \right\}.$$

Observe that the set $\{1, \ldots, p\}$ is compact and discrete. Hence, the result can be used in the high dimension case. Theorem 2.1 is dimension-free and the dependence on the "complexity" of S appears only through the median of Y. The following anti-concentration inequalities for Y can be derived as immediate corollaries of Theorem 2.1.

Theorem 2.2. *Fix $m \geq 0$. Under the assumptions of Theorem 2.1, we have for all $\varepsilon \geq 0$*

$$\sup_{r \geq 0} r^m \mathbb{P}\big(r - \varepsilon \leq \|Y\| \leq r + \varepsilon\big) \leq \Phi_{\mathrm{AC},m} \varepsilon,$$

where $\Phi_{\mathrm{AC},m} := \Theta_m \sigma_{\min}^{-4}[(\mu + \sigma_{\max})^{m+1}\sigma_{\min}^2 + (1 + \sigma_{\max})^2 \sigma_{\max}^{m+2}]$, with Θ_m representing a constant that depends only on m.

The constant $\Phi_{\mathrm{AC},m}$ obtained in Theorem 2.2, for any $m \geq 0$, is expected to be optimal since if the coordinates of Y are independent then the discussion following Corollary 2.7 of [18] and Example 2 of [12] imply that the density of $\|Y\|$ at points of order $\sqrt{\log p}$ is lower bounded in rate by $\sqrt{\log p}$. Hence, in this case for any $m \geq 0$ as $\varepsilon \to 0$ the rate is lower bounded by μ^{m+1}. Lemma B.2 of [27] provides optimal lower bounds of order $\sqrt{\log p}$ on the median of $\|Y\|$ when the coordinates of Y are mostly negatively correlated with correlations bounded away from 1.

Remark 2.1 (Comparison with [14]). Theorem 3 of [14] implies the anti-concentration for Y and provides a dimension-free bound depending on the mean of Y only under the additional assumption of $\sigma_{\max} = \sigma_{\min}$; Lemma 3.1 of [26] implies that the mean and median of $\|Y\|$ are of the same order. For the general case, their bound has an additional term of $\log(1/\varepsilon)$ which makes their bound weaker than that in Theorem 2.2. In terms of the proof technique, we note that the techniques of both works are the same for $r \leq 3(\mu + \sigma_{\max})$. For the case $r \geq 3(\mu + \sigma_{\max})$, Chernozhukov et al. [14] apply the Gaussian concentration inequality that leads to an extra $\log(1/\varepsilon)$ factor while we use inequality (13) that leads to the sharper version. Theorem 2.2 is the first result on dimension-free anti-concentration inequality for Y and hence our result readily applies to Gaussian processes on (infinite dimensional) compact metric spaces. The results of [28] imply an anti-concentration inequality that explicitly depends on the dimension as $\sqrt{\log(ep)}$. Nazarov's result as proved in [15] cannot lead to a rate better than $\sqrt{\log(ep)}$ since the covariance structure of Y is completely ignored in the proof.

2.3. Smooth approximation of the maximum

The starting point of almost any Berry–Esseen type result is a smoothing inequality. In our scenario, we need a smooth approximation of $\mathbb{1}\{\|S_n\| \leq r\}$ where, recall, $\|\cdot\|$ denotes the l_∞-norm. Theorem 1 of [5] provides an infinitely differentiable (in S_n) approximation of this indicator with sharp bounds on

the derivatives. However to get better rates of convergence in the central limit theorem, a certain stability property of the derivatives is needed. This property is exactly the reason why Chernozhukov et al. [16] were able to get better rates than those implied by Banach space CLTs; see [2] for implications of Banach space CLTs. The smooth approximation (taken from [12]), along with the its properties, is summarized in the following result. Define the "softmax" function as $F_\beta(z) := \beta^{-1} \log \sum_{j=1}^{2p} \exp(\beta z(j))$ for $z \in \mathbb{R}^{2p}$. Also define $g_0(t) := 30\mathbb{1}\{0 \le t \le 1\} \int_t^1 s^2(1-s)^2\, ds$.

Lemma 2.1. *Fix $r \ge 0, \varepsilon > 0$ and set $\beta = 2\log(2p)/\varepsilon$. Define the function $\varphi : \mathbb{R}^p \to \mathbb{R}$ as*

$$\varphi(x) = \varphi_{r,\varepsilon}(x) = g_0\left(\frac{2(F_\beta(z_x - r\mathbf{1}_{2p}) - \varepsilon/2)}{\varepsilon}\right),$$

where $z_x = (x^\top : -x^\top)^\top$ and $\mathbf{1}_{2p}$ is the vector of 1's of dimension $2p$. This function $\varphi(\cdot)$ satisfies the following properties.

1. *It "approximates" the indicator of the l_∞-ball, that is,*

$$\varphi(x) = \begin{cases} 1 & \text{if } \|x\| \le r, \\ 0 & \text{if } \|x\| > r + \varepsilon, \end{cases} \quad \text{or equivalently} \quad \mathbb{1}\big\{\|x\| \le r\big\} \le \varphi(x) \le \mathbb{1}\big\{\|x\| \le r + \varepsilon\big\}.$$

2. *There exists functions $D_j(\cdot), D_{jk}(\cdot)$ and $D_{jkl}(\cdot)$ for $1 \le j, k, l \le p$ as well as constant $C_0 > 0$ such that $|\partial_j \varphi(x)| \le D_j(x), |\partial_{jk}\varphi(x)| \le D_{jk}(x), |\partial_{jkl}\varphi(x)| \le D_{jkl}(x)$, where $\partial_j, \partial_{jk}, \partial_{jkl}$ denote the partial derivatives of φ with respect to the indices in subscript and for all $x \in \mathbb{R}^p$,*

$$\sum_{j=1}^{p} D_j(x) \le C_0 \varepsilon^{-1}, \qquad \sum_{j,k=1}^{p} D_{jk}(x) \le C_0 \log(ep)\varepsilon^{-2},$$

$$\sum_{j,k,l=1}^{p} D_{jkl}(x) \le C_0 \log^2(ep)\varepsilon^{-3}.$$

3. *The functions D_j, D_{jk}, D_{jkl} also satisfy a ratio stability property: there exists universal constant $\mathfrak{C} > 0$ such that for all $x, w \in \mathbb{R}^p$,*

$$e^{-\mathfrak{C}\log(ep)\|w\|_\infty/\varepsilon} \le \frac{D_j(x+w)}{D_j(x)}, \frac{D_{jk}(x+w)}{D_{jk}(x)}, \frac{D_{jkl}(x+w)}{D_{jkl}(x)} \le e^{\mathfrak{C}\log(ep)\|w\|_\infty/\varepsilon}. \tag{14}$$

The smooth approximation result above is the bottleneck in attaining better dependence on $\log(ep)$ in CLTs. The $\log^4(ep)$ dependence in the uniform and non-uniform CLTs presented in Section 1.1 comes only from the $\log(ep)$ factors in the bounds of derivatives and stability property of smooth approximation.

The approximating functions in Lemma 2.1 are constructed to work for any high-dimensional distribution; the difference between $F_\beta(z)$ and $\max_j z(j)$ is at most $\log(2p)/\beta$ over all vectors z and this can be smaller if z comes from a specific distribution. This universality can be seen clearly in the construction of [5] who defines the ε approximation of $\|\cdot\|$ based on $f_\varepsilon(x) = \mathbb{E}[\|x + \varepsilon\eta\|]$, where $\eta \sim N_p(0, I_p)$. However one can replace η by other random vectors taking into account the dependence of $U_{n,0}$. A specific choice that we *conjecture* works is $f_\varepsilon(x) = \mathbb{E}[\|x + \varepsilon U_{n,0}\|]$. Since $\|\cdot\|$ is Lipschitz, we get that $|\|x\| - f_\varepsilon(x)| \le \varepsilon\mathbb{E}[\|U_{n,0}\|] = \varepsilon O(\mu)$. Since $U_{n,0}$ and S_n share the same dependence structure, we only need to bound the derivatives of f_ε at $U_{n,j}$ which would lead to better rates in CLT using the proofs here; see Section 6 for details.

3. Main results

We are now ready to state the main results of this paper. The proofs of all the results in this section are given in the Supplementary Material. The main steps of these proofs are presented, for readers' convenience, in Section 6. Recall the notation $\delta_{n,m}$ from (7) in Section 2.1. Also recall that $M_n(\cdot)$ and the quantity $\bar{L}_{n,0}$ are defined in (10) and (11), respectively. The quantity L_n (in (9)) denotes the "weak" third pseudo-moment and if $L_n = 0$ then $\delta_{n,0} = \delta_{n,m} = 0$ for any $m \geq 0$. For this reason, we assume $L_n > 0$. Let $\Phi_{\mathrm{AC},m}$ denote the anti-concentration constant in Theorem 2.2 for random vector $U_{n,0}$.

3.1. Uniform CLT

Theorem 3.1. *For independent random vectors $X_1, \ldots, X_n \in \mathbb{R}^p$, we have*

$$\delta_{n,0} \leq 4\Phi_{\mathrm{AC},0}\varepsilon_n + \frac{2C_0 \log(ep) M_n(\varepsilon_n)}{\varepsilon_n^2} + \frac{\log^{1/3}(ep)\bar{L}_{n,0}}{n^{1/3}L_n^{4/3}(2e^{5\mathfrak{C}}C_0)^{1/3}},$$

where $\varepsilon_n := (2e^{2\mathfrak{C}}C_0 \log^2(ep)L_n)^{1/3}/n^{1/6}$.

Theorem 3.1 is qualitatively the same as Theorem 2.1 of [16] for l_∞-balls. More importantly, note that if $\Phi_{\mathrm{AC},0} \asymp \sqrt{\log(ep)}$ then Theorem 3.1 has a dominating term of order $(\log^7(ep)/n)^{1/6}$ and hence the result above is as good as Theorem 2.1 of [16]. The last term in the bound of $\delta_{n,0}$ is of lower order compared to the first term. From inequality (12), we obtain that

$$\frac{\log^{1/3}(ep)\bar{L}_{n,0}}{n^{1/3}L_n^{4/3}(2e^{5\mathfrak{C}}C_0)^{1/3}} = O\left(\frac{(\log(ep))^{5/6}}{(nL_n)^{1/3}}\right) = O\left(\frac{(\log(ep))^{5/2}}{nL_n}\right)^{1/3},$$

which is dominated by the first term which is at least of order $(\log^4(ep)/n)^{1/6}$. Further the quantity $M_n(\varepsilon_n)$ (defined in (10)) is exactly what appears in the classic Lindeberg condition.

Rates under $(2+\tau)$-moments. Recall that $M_n(\phi)$ only involves truncated second moment instead of third moment used in [16]. This subtle difference allows for deriving better rates when $\|X_i\|$ have $(2+\tau)$-moments. In this case, the choice of ε_n in Theorem 3.1 is not large enough since $M_n(\varepsilon_n)/\varepsilon_n^2$ does not converge to zero. Even though our results involve third "moments" $L_n, \bar{L}_{n,0}$, from the proof (in particular (22) in the Supplementary Material) it can be seen that the integral in the definition of $L_n, \bar{L}_{n,0}$ can be changed to integral over a truncated set; see the Appendix and Step E in Section 6 for details. Very recently Sun [37] considered CLT under $(2+\tau)$-moments using Lindeberg method but no explicit rates were provided. In the following, we provide details for $\tau \geq 1$ and the calculations for $\tau < 1$ are provided in the Appendix. Set

$$\nu_{2+\tau}^{2+\tau} := \frac{1}{n}\sum_{i=1}^n \int \|x\|^{2+\tau}|\zeta_i|(dx) \quad \Rightarrow \quad M_n(\varepsilon) \leq \frac{\nu_{2+\tau}^{2+\tau}}{(n^{1/2}\varepsilon/\log(ep))^\tau} = \nu_{2+\tau}^{2+\tau}\left(\frac{\log(ep)}{n^{1/2}\varepsilon}\right)^\tau.$$

The proof of Theorem 3.1 actually proves a bound that holds for all $\varepsilon > 0$ and we now choose $\varepsilon = \varepsilon_n := \max\{(2e^{2\mathfrak{C}}C_0 \log^2(ep)L_n)^{1/3}/n^{1/6}, r_n\nu_{2+\tau}((\log(ep))^{1+\tau}/n^{\tau/2})^{1/(2+\tau)}\}$ for some $r_n \geq 1$. This choice implies that

$$\frac{e^{2\mathfrak{C}}C_0 \log^2(ep)L_n}{n^{1/2}\varepsilon_n^3} \leq \frac{1}{2} \quad \text{and} \quad \frac{C_0 \log(ep) M_n(\varepsilon_n)}{\varepsilon_n^2} \leq \frac{C_0}{r_n^{2+\tau}}.$$

Following the proof of Theorem 3.1, for any $r_n \geq 1$, we get

$$\delta_{n,0} \lesssim \frac{1}{r_n^{2+\tau}} + \Phi_{\mathrm{AC},0} \frac{(\log^2(ep)L_n)^{1/3}}{n^{1/6}} + \Phi_{\mathrm{AC},0} \frac{r_n \nu_{2+\tau} (\log(ep))^{(1+\tau)/(2+\tau)}}{n^{\tau/2(2+\tau)}}.$$

Here $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Minimizing over $r_n \geq 1$, we get

$$\delta_{n,0} \lesssim \begin{cases} \big[\Phi_{\mathrm{AC},0} L_{n,\tau}^{1/(2+\tau)} + (\Phi_{\mathrm{AC},0}\nu_{2+\tau})^{(2+\tau)/(3+\tau)}\big] \dfrac{(\log(ep))^{(\tau+1)/(\tau+2)}}{n^{\tau/(6+2\tau)}} & \text{if } \tau < 1, \\ \Phi_{\mathrm{AC},0} \dfrac{(\log^2(ep)L_n)^{1/3}}{n^{1/6}} + (\Phi_{\mathrm{AC},0}\nu_{2+\tau})^{(2+\tau)/(3+\tau)} \dfrac{(\log(ep))^{(\tau+1)/(\tau+3)}}{n^{\tau/(6+2\tau)}} & \text{if } \tau \geq 1, \end{cases} \tag{15}$$

where $L_{n,\tau} = n^{-1} \sum_{i=1}^n \max_{1 \leq j \leq p} \int |x(j)|^{2+\tau} |\zeta_i|(dx)$. As mentioned before, the proof for $\tau < 1$ is given in the Appendix. The dependence on sample size obtained above is better than the one obtained in Proposition 2.1 of [16] for all $\tau > 0$. In particular for $\tau = 1$, we got the dependence $n^{-1/8}$ whereas [16] obtained $n^{-1/9}$ on the sample size. We note here that following the proof of Theorem 2.1 of [7] it is possible to get $n^{-1/6}$ dependence on n whenever $\tau \geq 1$ as shown in the following proposition. This is similar to the case of multivariate CLT where the rate of convergence scales as $n^{-1/2}$ whenever the random vectors have more than three moments; i.e., after some number of moments the rate stabilizes. The following result is proved based on Theorem 2.1 of [7] and does not use the stability property (14).

Proposition 3.1. *For i.i.d. random vectors $X_1, \ldots, X_n$,*

$$\delta_{n,0} \leq 2^{1-n} + 8\Phi_{\mathrm{AC},0}\nu_3 \big(C_0 \log^2(ep)\big)^{1/3} n^{-1/6}.$$

It is clear that the rate obtained in Proposition 3.1 is sharper than the one obtained in (15) for $1 \leq \tau \leq 1.5$. Note that for $\tau = 1.5$, $\tau/(6+2\tau) = 1/6$ and (15) leads to $n^{-1/6}$ rate. A notable difference between the rates is that while Theorem 3.1 decouples (sub-optimally) the main rate $\Phi_{\mathrm{AC},0}(\log^2(ep))^{1/3}/n^{1/6}$ and the term depending on moments $\nu_{2+\tau}$, Proposition 3.1 puts these two terms together which implies that Proposition 3.1 does not lead to the "right" requirement between $\log(p)$ and n in case random vectors have exponential tails. This discussion raises an important point: What is the "right way" of proving rates in the high-dimensional CLT? The following result combines the proof techniques of Theorem 3.1 and Proposition 3.1 in the "right way" so that the rate scales like $n^{-1/6}$ for all $\tau \geq 1$ (as Proposition 3.1) and becomes the same as (15) for $\tau \to \infty$.

Theorem 3.2. *For i.i.d. random vectors $X_1, \ldots, X_n$, and for any $\tau \geq 1$,*

$$\delta_{n,0} \lesssim \frac{1}{2^n} + \Phi_{\mathrm{AC},0} \left(\frac{L_n^2 \log^4(ep)}{n} \right)^{1/6} + \Phi_{\mathrm{AC},0} \nu_{2+\tau} \frac{(\log(ep))^{(\tau+1)/(\tau+2)}}{n^{\tau/(4+2\tau)}},$$

where $\lesssim$ hides only universal constants.

The assumption of i.i.d. random vectors in Proposition 3.1 and Theorem 3.2 can be relaxed to observations $X_1, \ldots, X_n$ having the same covariance matrix. It is clear that Theorem 3.2 leads to better rates than (15) for all $\tau \geq 1$ since $\tau/(6+2\tau) < \tau/(4+2\tau)$.

3.2. Non-uniform CLT

As an extension of Theorem 3.1, we have the following non-uniform version of CLT. Since the statement is exact with explicit constants, it is cumbersome.

Theorem 3.3. *For independent $\mathbb{R}^p$ random vectors $X_1, \ldots, X_n$, for any $m, r_{n,m} \in (0, \infty)$,*

$$\delta_{n,m} \le 2^{2m^2/3+8m/3+1}\varepsilon_n^{m+1}\Phi_{\mathrm{AC},0} + \left(2^{2m/3+1}+2\right)\Phi_{\mathrm{AC},m}\varepsilon_n + 2C_0\log(ep)r_{n,m}^m\varepsilon_n^{-2}M_n\left(2^{2m/3}\varepsilon_n\right) + \frac{r_{n,m}^m\bar{L}_{n,m}}{L_n 2^m e^{\mathfrak{C}}}\left(\frac{\log(ep)}{n(2^{2m+1}e^{2\mathfrak{C}}C_0L_n)}\right)^{(m+1)/3} + \sup_{r\ge r_{n,m}}\max_{0\le k\le n} r^m\mathbb{P}\left(\|U_{n,k}\|\ge r\right).$$

Here ε_n is same as the one defined in Theorem 3.1.

Theorem 3.3 does not get the correct dependence on the sample size n when random vectors $X_1, \ldots, X_n$ only have $(2+\tau)$-moments as the calculation for Theorem 3.1 show. We now present an improvement for i.i.d. random vectors which can be seen as a non-uniform extension of Theorem 3.2.

Theorem 3.4. *Fix $m \ge 1$ and $\tau \ge m$. If $X_1, \ldots, X_n$, are i.i.d. random vectors, then*

$$\delta_{n,m} \le 2^{m/2}\left(\nu_m 2^{-n/2}\right)^m + 2^{2+2m}\varepsilon_n^m + 2.4\Phi_{\mathrm{AC},m}\varepsilon_n,$$

where

$$\varepsilon_n := \max\left\{\frac{(2^{2+3m/2}C_0e^{\mathfrak{C}}L_n\log^2(ep))^{1/3}}{n^{1/6}}, \frac{\nu_{2+\tau}(2^{3+5m/2}C_0\log^{\tau+1}(ep))^{1/(2+\tau)}}{n^{\tau/(4+2\tau)}}\right\}.$$

Theorem 3.4 always leads to $n^{-1/6}$ dependence on the sample size unlike Theorem 3.3.

Comments on the proof technique. The proofs of uniform and non-uniform CLTs in Banach space literature are based on Lindeberg method and smooth approximation; see, for example, [31] and [7]. Motivated by the proof technique of [16] who introduced the stability property, we combine Lindeberg method with a minor twist from the proof of [16] to prove all the above results; see Section 6 for a detailed outline of how our proofs differ from the ones in the above mentioned references.

Recently, Koike [21], Proposition 2.3, proved for sub-Gaussian random variables a rate of $(\log^6(ep)/n)^{1/6}$ for $\delta_{n,0}$ which could also be weaker than our result depending on $\Phi_{\mathrm{AC},0}$. It maybe possible that the methods in that paper could sharpen our result. We refrained from using their proof techniques for simplicity.

3.3. Corollaries for sub-Weibull random vectors

In this section, we provide simplified results under the assumption that the coordinates of the random vectors $X_1, \ldots, X_n$ are sub-Weibull. For a simplification of uniform CLTs, we choose a suitable τ in (15) and Theorem 3.2. For a simplification of Theorem 3.3, we simplify $M_n(\varepsilon_n)$ and $\bar{L}_{n,m}$ under sub-Weibull tails. Similar simplification can be done from Theorem 3.1 but this leads to sub-optimal dependence on $\log(ep)$ in the second order term.

Recall that $M_n(\varepsilon_n)$ and $\bar{L}_{n,m}$ are defined in terms of the variation measures ζ_j which we bound using the sum of the measures for simplicity. We now define Orlicz norms.

Definition 3.1. Let X be a real-valued random variable and $\psi : [0, \infty) \mapsto [0, \infty)$ be a non-decreasing function with $\psi(0) = 0$. Then, we define

$$X_\psi = \inf\bigl\{c > 0 : \mathbb{E}\psi\bigl(|X|/c\bigr) \le 1\bigr\},$$

where the infimum over the empty set is taken to be ∞.

Usually the definition of Orlicz "norm", $\|\cdot\|_\psi$, includes convexity assumption of ψ which we did not include since we also work with non-convex ψ below. It follows from Jensen's inequality, that when ψ is a non-decreasing, convex function, $\cdot_\psi$ is a norm on the set of random variables X for which $X_\psi < \infty$. The commonly used Orlicz norms are derived from $\psi_\alpha(x) := \exp(x^\alpha) - 1$, for $\alpha \ge 1$, which are obviously increasing and convex. For $0 < \alpha < 1$, ψ_α is not convex, and X_{ψ_α} is not a norm, but a quasinorm. A random variable X is called sub-exponential if $X_{\psi_1} < \infty$, and a random variable X is called sub-Gaussian if $X_{\psi_2} < \infty$.

Recall that the j-coordinate of X is $X(j)$ and set $\sigma^2_{\max} := \max_{1\le j\le p} \operatorname{Var}(U_{n,0}(j))$, and $\sigma^2_{\min} := \min_{1\le j\le p} \operatorname{Var}(U_{n,0}(j))$. Throughout the following corollaries, Θ stands for a universal constant that does not depend on p, n or any of the other distributional properties. Θ with subscripts (such as Θ_α) represents constants that only depend on those subscripts.

Corollary 3.1 (Uniform CLT). *Suppose there exists a constant $1 \le K_p < \infty$ such that*

$$\max_{1\le i\le n} \max_{1\le j\le p} \bigl\|X_i(j)\bigr\|_{\psi_\alpha} \le K_p \tag{16}$$

for some $0 < \alpha \le 2$. If $\log(n^{3/2}/(K_p\Theta_\alpha\Phi_{\mathrm{AC},0}(\log(ep))^{2+1/\alpha})) \ge 4$, then

$$\delta_{n,0} \le \Theta\Phi_{\mathrm{AC},0}\frac{(L_n\log^2(ep))^{1/3}}{n^{1/6}} + \Theta_\alpha K_p\Phi_{\mathrm{AC},0}\frac{(\log(ep))^{1+1/\alpha}}{n^{1/2}}\log^{1/\alpha}\left(\frac{\Theta_\alpha\Phi^{-1}_{\mathrm{AC},0}n^{3/2}}{(\log(ep))^{2+1/\alpha}}\right). \tag{17}$$

Instead if $\alpha\log(n/\log(ep)) \ge 3$ and $X_1, \dots, X_n$ share the same covariance matrix, then

$$\delta_{n,0} \le \Theta\Phi_{\mathrm{AC},0}\frac{(L_n\log^2(ep))^{1/3}}{n^{1/6}} + \Theta_\alpha K_p\Phi_{\mathrm{AC},0}\frac{(\log(ep))^{1+1/\alpha}}{n^{1/2}}\log^{1/\alpha}\left(\frac{n}{\log(ep)}\right). \tag{18}$$

Proof. Note that assumption (16) implies that $\nu_{2+\tau} \le K_p\Theta_\alpha(2+\tau)^{1/\alpha}(\log(ep))^{1/\alpha}$ for all $\tau \ge 1$. Then (17) follows from (15) by taking $\tau + 3 = \log(n^{3/2}/(K_p\Theta_\alpha\Phi_{\mathrm{AC},0}(\log(ep))^{2+1/\alpha}))$ which is possible for $\tau \ge 1$ since the right-hand side is assumed to be larger than 4. Further, (18) follows from Theorem 3.2 by taking $\tau + 2 = \alpha\log(n/\log(ep))$ (which is possible for $\tau \ge 1$ since $\alpha\log(n/\log(ep)) \ge 3$). □

Simplifying a little further, the bounds on $\delta_{n,0}$ can be written as

$$\delta_{n,0} \lesssim \Phi_{\mathrm{AC},0}\frac{(L_n\log^2(ep))^{1/3}}{n^{1/6}} + K_p\Phi_{\mathrm{AC},0}\frac{(\log(ep))^{1+1/\alpha}(\log n)^{1/\alpha}}{n^{1/2}}.$$

The following corollary is obtained by controlling $M_n(\varepsilon_n)$ and $\bar{L}_{n,m}$ in Theorem 3.3 for sub-Weibull random vectors. By choosing an appropriate of τ in Theorem 3.4, we get a much simpler form under the assumption that $X_1, \dots, X_n$ share the same covariance structure.

Corollary 3.2 (Non-uniform CLT). *Fix $m \geq 0$. If $n \geq \Theta K_p^3 L_n^{-1} (2e \log(ep))^{1+3/\alpha}$, then we have* (i) *for $1 < \alpha \leq 2$,*

$$\delta_{n,m} \leq \Theta_m \Phi_{\mathrm{AC},m} \left(\frac{L_n^2 \log^4(ep)}{n} \right)^{1/6} + \Theta_{\alpha,m} \left(\frac{K_p^{(2m+1)\alpha} \log^4(epn)}{n L_n^{(m+1)\alpha/3}} \right)^{1/(\alpha-1)} + \Theta_m K_p^m \sigma_{\max}^{m+1} \left(\frac{\log^4(epn)}{n L_n} \right)^{(m+1)/3} + \Theta_m \frac{K_p^{m+2}}{\sigma_{\max}^2 n^{2/3}},$$

and (ii) *for $0 < \alpha \leq 1$,*

$$\delta_{n,m} \leq \Theta_m \Phi_{\mathrm{AC},m} \left(\frac{L_n^2 \log^4(ep)}{n} \right)^{1/6} + \Theta_{\alpha,m} \frac{K_p^{3+m}}{L_n} \left(\frac{K_p^3 (\log(epn))^{5/4+3/\alpha}}{n L_n} \right)^{12/\alpha+2m} + \Theta_m K_p^m \sigma_{\max}^{m+1} \left(\frac{(\log(epn))^{1+3/\alpha}}{n L_n} \right)^{(m+1)/3} + \Theta_m \frac{K_p^{m+2}}{\sigma_{\max}^2 n^{2/3}}.$$

Instead if $\alpha \log(2^{3+5m/2} n / \log(ep)) \geq m + 2$ and $X_1, \ldots, X_n$ share the same covariance matrix, then

$$\delta_{n,m} \leq 2^{m/2} \left(\frac{\nu_m}{2^{n/2}} \right)^m + 2^{2+2m} \varepsilon_n^m + 2.4 \Phi_{\mathrm{AC},m} \varepsilon_n, \tag{19}$$

where

$$\varepsilon_n := \max \left\{ \frac{(2^{2+3m/2} C_0 e^{\mathfrak{C}} L_n \log^2(ep))^{1/3}}{n^{1/6}}, \Theta_{\alpha,m} \frac{(\log(ep))^{1+1/\alpha}}{n^{1/2}} \log^{1/\alpha} \left(\frac{2C_0 n}{\log(ep)} \right) \right\}.$$

Inequality (19) follows by taking $2 + \tau = \alpha \log(2^{3+5m/2} C_0 n / \log(ep))$ in ε_n of Theorem 3.4.

Remark 3.1 (Even more simplified rates). The bounds in Corollary 3.2 are finite sample and show explicit dependence on $L_n, K_p, \sigma_{\max}$ and other distributionals constants. If $\max\{K_p, \sigma_{\max}, L_n^{-1}\} = O(1)$ then the bounds in Corollary 3.2 can simply be written as

$$\delta_{n,m} \lesssim \Phi_{\mathrm{AC},m} \frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}} + \begin{cases} 0 & \text{if } \alpha > 1, \\ \left((\log(ep))^{5/4+3/\alpha} / n \right)^{(m+1)/3} & \text{if } \alpha \in (0, 1]. \end{cases}$$

The implication from (19) can be simply written as

$$\delta_{n,m} \lesssim \left(\frac{\nu_m}{2^{n/2}} \right)^m + \Phi_{\mathrm{AC},m} \frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}} + \Phi_{\mathrm{AC},m} \frac{(\log(ep))^{1+1/\alpha} (\log n)^{1/\alpha}}{n^{1/2}}.$$

Remark 3.2 (Convergence of moments). Theorems 3.1 and 3.3 are useful in proving convergence of m-th moment of S_n to that of $U_{n,0}$ at an $n^{-1/6}$ rate (up to factors depending on $\log(ep)$). To prove an explicit bound, note that for $m \geq 1$:

$$\left| \mathbb{E}\left[{S_n}^m \right] - \mathbb{E}\left[{U_{n,0}}^m \right] \right| = \left| \int_0^\infty m r^{m-1} \left(\mathbb{P}(S_n \geq r) - \mathbb{P}(U_{n,0} \geq r) \right) dr \right| \leq m \int_0^1 \delta_{n,0} \, dr + \int_1^\infty \frac{m}{r^{1+\beta}} \delta_{n,m+\beta} \, dr \leq m \delta_{n,0}(r) + \frac{m}{\beta} \delta_{n,m+\beta}$$

for any $\beta \geq 0$. The second term in the right-hand side above can be bounded using Theorem 3.3 or 3.4. Taking the main terms in the bounds for $\delta_{n,0}$, $\delta_{n,m+\beta}$ and using $\Phi_{\mathrm{AC},m} \lesssim \mu^{m+1}$ from Theorem 2.2, we can write the bound on moment difference as

$$\left|\mathbb{E}\left[{S_n}^m\right] - \mathbb{E}\left[{U_{n,0}}^m\right]\right| \lesssim m\mu \frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}} + \frac{m}{\beta}\mu^{m+\beta+1}\frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}}.$$

Note that we assumed here $\sigma_{\min}^{-1} = O(1)$. Choosing $\beta = 1/\log(\mu)$, we get

$$\left|\mathbb{E}\left[{S_n}^m\right] - \mathbb{E}\left[{U_{n,0}}^m\right]\right| \lesssim m\mu \frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}} + m\log(\mu)\mu^{m+1}\frac{(L_n \log^2(ep))^{1/3}}{n^{1/6}}.$$

Since the median and $(\mathbb{E}[\|U_{n,0}\|^\ell])^{1/\ell}$, $\ell \geq 1$ are of the same order, we get that for $m \geq 1$

$$\mathbb{E}\left[\|S_n\|^m\right] = \left(1 + O\left(\frac{\mu\log(\mu)(L_n \log^2(ep))^{1/3}}{n^{1/6}}\right)\right)\mathbb{E}\left[\|U_{n,0}\|^m\right].$$

Hence our results imply that the moments of $\|S_n\|$ match the moments of $\|U_{n,0}\|$ up to a lower order term if $\mu\log(\mu)(L_n \log^2(ep))^{1/3} = o(n^{1/6})$.

4. Applications

In this section, we present two applications. The first is to post-selection inference that shows the impact of dimension-free anti-concentration constant. The second is to maximal inequalities for empirical processes that shows the importance of non-uniform CLT.

4.1. Many approximate means and post-selection inference

In this section, we use the "many approximate means" (MAM) framework of [4] for post-selection inference (PoSI). In this PoSI problem, we show scenarios where $\Phi_{\mathrm{AC},0}$ (or μ) grows almost *like* a constant.

The MAM framework is as follows. Suppose we have a parameter $\theta_0 = (\theta_0(1), \ldots, \theta_0(p))^\top$ and an estimator $\hat{\theta} = (\hat{\theta}(1), \ldots, \hat{\theta}(p))^\top$ of parameter θ_0 that has an approximate linear form:

$$n^{1/2}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(Z_i) + R_n, \tag{20}$$

where $\psi(\cdot) = (\psi_1(\cdot), \ldots, \psi_p(\cdot))^\top \in \mathbb{R}^p$ and $R_n = (R_n(1), \ldots, R_n(p))^\top \in \mathbb{R}^p$. Here $Z_1, \ldots, Z_n$ are independent random variables based on which $\hat{\theta}$ is constructed. The function $\psi_j(\cdot)$, $1 \leq j \leq p$ represents the influence function for estimator $\hat{\theta}(j)$. An estimator $\hat{\theta}$ satisfying (20) is called asymptotically linear and most (of the commonly used) M-estimators satisfy this assumption. Based on the linear approximation (20) and the anti-concentration result Theorem 2.2, we have the following result. Define

$$\bar{\Delta}_n(r) := \left|\mathbb{P}\left(\sqrt{n}\|\hat{\theta} - \theta_0\| \leq r\right) - \mathbb{P}\left(\left\|Y^\psi\right\| \leq r\right)\right|,$$

where $Y^\psi \sim N_p(0, n^{-1}\sum_{i=1}^n \mathbb{E}[\psi(Z_i)\psi^\top(Z_i)])$. Let $\Phi_{\mathrm{AC},m}^\psi$ denote the anti-concentration constant from Theorem 2.2 for Y^ψ. Set $S_n^\psi := n^{-1/2}\sum_{i=1}^n \psi(Z_i)$.

Proposition 4.1. *For any $\delta > 0$ and $m \geq 0$, we have*

$$r^m \bar{\Delta}_n(r) \leq 2\big[(5/4)^m \Delta_{n,m}^{\psi} + 5^m \Phi_{\mathrm{AC},0}^{\psi} \delta^{m+1}\big] + \Phi_{\mathrm{AC},m}^{\psi} \delta + r^m \mathbb{P}\big(\|R_n\| > \delta\big),$$

where $\Delta_{n,m}^{\psi} := \sup_{r \geq 0} r^m |\mathbb{P}(\|S_n^{\psi}\| \leq r) - \mathbb{P}(\|Y^{\psi}\| \leq r)|$.

This result is a slight generalization of Theorem 2.1 of [4] for l_∞-balls. To derive a proper non-uniform CLT result from Proposition 4.1, one needs to choose δ depending on r and apriori bound moments of $\|R_n\|$. For the case $m = 0$ (which we focus on from now), the *pf* of Proposition 4.1 implies $\bar{\Delta}_n(r) \leq \Delta_{n,m}^{\psi} + \Phi_{\mathrm{AC},0}^{\psi} \delta + \mathbb{P}(\|R_n\| > \delta)$. Hence $\bar{\Delta}_n(r)$ converges to zero if $\|R_n\| = o_p(r_n)$ such that $r_n \Phi_{\mathrm{AC},0} = o(1)$ as $n, p \to \infty$.

We now describe the framework of post-selection inference. Suppose $Z_i := (X_i^\top, Y_i)^\top \in \mathbb{R}^d \times \mathbb{R}$ for $1 \leq i \leq n$ represent regression data with d-dimensional covariates X_i. For any subset $M \subseteq \{1, 2, \ldots, d\}$, let $X_{i,M}$ denote a subvector of X_i with indices in M. Based on a loss function $\ell(\cdot, \cdot)$ define the regression "slope" estimator $\hat{\beta}_M$ as

$$\hat{\beta}_M := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(X_{i,M}^\top \theta, Y_i\big).$$

The target for the estimator $\hat{\beta}_M$ is given by $\beta_M := \arg\min_{\theta \in \mathbb{R}^{|M|}} n^{-1} \sum_{i=1}^n \mathbb{E}[\ell(X_{i,M}^\top \theta, Y_i)]$. Some examples of loss functions are related to linear regression $\ell(u, v) = (u - v)^2/2$, logistic regression $\ell(u, v) = uv - \log(1 + e^u)$, Poisson regression $\ell(u, v) = uv - \exp(u)$.

As is often done in practical data analysis, suppose we choose a model $\hat{M}$ based on the data $\{Z_i : 1 \leq i \leq n\}$. The problem of post-selection inference refers to the statistical inference for the (random) target $\beta_{\hat{M}}$. In particular, PoSI problem refers to construction of confidence regions $\{\hat{\mathcal{R}}_M : M \subseteq \{1, 2, \ldots, d\}\}$ (depending on $\alpha \in [0, 1]$) such that

$$\liminf_{n \to \infty} \mathbb{P}(\beta_{\hat{M}} \in \hat{\mathcal{R}}_{\hat{M}}) \geq 1 - \alpha, \tag{21}$$

holds for *any* randomly selected model $\hat{M}$. It was proved in Theorem 3.1 of [24] that the PoSI problem (21) is equivalent to the simultaneous inference problem

$$\liminf_{n \to \infty} \mathbb{P}\bigg(\bigcap_M \{\beta_M \in \hat{\mathcal{R}}_M\}\bigg) \geq 1 - \alpha,$$

where the intersection is taken over all $M \subseteq \{1, 2, \ldots, d\}$. A straightforward construction of such simultaneous confidence regions can be based on finding quantiles of the statistic

$$\max_M \max_{1 \leq j \leq |M|} \bigg| \frac{\sqrt{n}(\hat{\beta}_M(j) - \beta_M(j))}{\hat{\sigma}_M(j)} \bigg|, \tag{22}$$

where $v(j)$, for a vector v, represents the j-th coordinate of v and $\hat{\sigma}_M(j)$ represents an estimator of the standard deviation of $\sqrt{n}(\hat{\beta}_M(j) - \beta_M(j))$. In order to apply Proposition 4.1 for finding quantiles of the statistic (22), we need a linear approximation result such as (20). In case of linear regression ($\ell(u, v) = (u - v)^2/2$), it was proved in [23] that for any $1 \leq k \leq d$,

$$\max_{|M| \leq k} \bigg\| \hat{\beta}_M - \beta_M + \frac{1}{n} \sum_{i=1}^{n} \Omega_M^{-1} X_{i,M} \big(Y_i - X_{i,M}^\top \beta_M\big) \bigg\|_2 = O_p\bigg(\frac{k \log(ed/k)}{n}\bigg),$$

where $\Omega_M := n^{-1}\sum_{i=1}^n \mathbb{E}[X_{i,M}X_{i,M}^\top]$. Extensions for general ℓ are available in that paper. This result is proved under certain tail assumptions on the observations and holds both for random and fixed covariates. For linear regression with fixed covariates, we have

$$\hat{\beta}_M - \beta_M = \left(\frac{1}{n}\sum_{i=1}^n x_{i,M}x_{i,M}^\top\right)^{-1}\frac{1}{n}\sum_{i=1}^n x_{i,M}\big(Y_i - \mathbb{E}[Y_i]\big) = \frac{1}{n}\sum_{i=1}^n \Omega_M^{-1}x_{i,M}\big(Y_i - \mathbb{E}[Y_i]\big),$$

where we write x_i to note fixed covariates and the variance of $n^{1/2}(\hat{\beta}_M - \beta_M)$ is given by

$$\mathrm{Var}\big(\sqrt{n}(\hat{\beta}_M - \beta_M)\big) := \frac{1}{n}\sum_{i=1}^n \big(\Omega_M^{-1}x_{i,M}\big)\big(\Omega_M^{-1}x_{i,M}\big)^\top \mathrm{Var}(Y_i). \tag{23}$$

Set $\psi_{j,M}(x_i, Y_i) = (Y_i - \mathbb{E}[Y_i])(\Omega_M^{-1}x_{i,M})(j)/\sigma_M(j)$, with $\sigma_M(j)$ representing the j-th diagonal element of the variance matrix (23). Define

$$\Delta_{\mathtt{PoSI}} := \sup_{r\geq 0}\left|\mathbb{P}\left(\max_{\substack{|M|\leq k,\\ 1\leq j\leq |M|}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{j,M}(x_i, Y_i)\right| \leq r\right) - \mathbb{P}\Big(\max_{\substack{|M|\leq k,\\ 1\leq j\leq |M|}} |G_{j,M}| \leq r\Big)\right|,$$

where $(G_{j,M})_{j,M}$ has a multivariate normal distribution such that

$$\mathrm{Cov}(G_{j,M}, G_{j',M'}) = \mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{j,M}(x_i, Y_i), \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{j',M'}(x_i, Y_i)\right).$$

By taking the maximum over $|M| \leq k$ in $\Delta_{\mathtt{PoSI}}$, we are restricting the final selected model to have cardinality at most k. Without loss of generality, we can take $G_{j,M} = n^{-1/2}\sum_{i=1}^n g_i(\Omega_M^{-1}x_{i,M})(j)/\sigma_M(j)$, for $g_i \overset{ind}{\sim} N(0, \mathrm{Var}(Y_i))$. To bound $\Delta_{\mathtt{PoSI}}$ from our results, we need the median $\mu_{\mathtt{PoSI}}$:

$$\mu_{\mathtt{PoSI}} := \mathrm{median}\Big(\max_{|M|\leq k}\max_{1\leq j\leq |M|}|G_{j,M}|\Big).$$

Note that the Gaussian vector $(G_{j,M})_{j,M}$ has dimension $p \asymp (ed/k)^k$ and hence $\mu_{\mathtt{PoSI}} = O(\sqrt{k\log(ed/k)})$. Theorem 3.1 implies that $\Delta_{\mathtt{PoSI}} = O(\mu_{\mathtt{PoSI}}(k^4\log^4(ed/k)/n)^{1/6})$. Using the pessimistic bound on $\mu_{\mathtt{PoSI}}$, we get that

$$\Delta_{\mathtt{PoSI}} = O(1)\big(k^7\log^7(ed/k)/n\big)^{1/6}, \tag{24}$$

and this requires $k\log(ed/k) = o(n^{1/7})$. In words, this means that the selected model cardinality is at most $o(n^{1/7})$. Using Theorem 3.1, we can get a better requirement on k if we have better bounds for $\mu_{\mathtt{PoSI}}$. It was proved in Proposition 5.5 of [11] that if the covariates are orthogonal, that is, $n^{-1}\sum_{i=1}^n x_i x_i^\top = I_d$, then $\mu_{\mathtt{PoSI}} = O(\sqrt{\log d})$ for any $1 \leq k \leq d$. This result was further improved by Theorem 3.3 of [1] in that if there exists a $\kappa \in [0, 1)$ such that for all $|M| \leq k$

$$(1-\kappa)\|\theta\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \big(x_{i,M}^\top\theta\big)^2 \leq (1+\kappa)\|\theta\|_2^2 \quad \text{for all } \theta \in \mathbb{R}^{|M|}, \tag{25}$$

holds then $\mu_{\mathtt{PoSI}} \leq \sqrt{2\log(2d)} + C(\kappa)\kappa\sqrt{2k\log(6d/k)}$, for a function $C(\cdot)$ satisfying $C(\delta) \to 1$ as $\delta \to 0$. Condition (25) is called the restricted isometry property (RIP) and is (trivially) satisfied

with $\kappa = 0$ if the covariates are orthogonal. Hence, if $\kappa\sqrt{k}$ converges to zero then also $\mu_{\texttt{PoSI}} = O(\sqrt{\log d})$. Hence, under the RIP condition (25) with $\kappa\sqrt{k} \to 0$, we get from Theorem 3.1 that $\Delta_{\texttt{PoSI}} = O(\sqrt{\log(ed)}(k^4 \log^4(ed/k)/n)^{1/6})$. This result only requires $k(\log(ed))^{7/4} = o(n^{1/4})$ which is much weaker than that implied by (24). To get a better perspective take $k = d$ and for this case, we obtain

$$\Delta_{\texttt{PoSI}} = O(1)\big(d^4 \log^3(d)/n\big)^{1/6}, \tag{26}$$

which converges to zero if $d^4 = o(n/\log^3 n)$. On the other hand, the Berry–Esseen bound ([6]) for the linear regression estimator on the full model $M_{\texttt{full}} = \{1, 2, \ldots, d\}$ (no simultaneity involved) requires $d^{3.5} = o(n)$ which is close to the requirement from (26).

4.2. Maximal inequalities for empirical processes

In this section, we consider application of our non-uniform CLT result for the case of suprema of empirical processes. Suppose (ξ_i, X_i), $1 \le i \le n$ are independent random variables in a measurable space $\mathbb{R} \times \mathcal{X}$ such that $\mathbb{E}[\xi_i | X_i] = 0$ for $1 \le i \le n$ and let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. We consider the problem of bounding the moments of

$$Z_n(\mathcal{F}) := \sup_{f \in \mathcal{F}} \left| n^{-1/2} \sum_{i=1}^n \xi_i f(X_i) \right|.$$

Maximal inequalities that bound $\mathbb{E}[Z_n(\mathcal{F})]$ are of great importance in the study of M-estimators and empirical risk minimizers; see [3]. In the study of least squares estimator obtained by minimizing the quadratic loss over $f \in \mathcal{F}$ the rate is essentially determined by the behavior of $\mathbb{E}[Z_n(\mathcal{F}')]$ over subsets $\mathcal{F}'$ of $\mathcal{F}$; see [39], Lemma 3.1. Rates of convergence when ξ_i only has finite number of moments has received some interest recently. Even though $\mathcal{F}$ is usually countable and infinite, our results can be used in case of (weak) VC-major classes when ξ_i only has 3 moments. The calculation in this section could be used to prove rates for multiple isotonic regression when the errors are dependent on covariates and only have three moments; see [19], p. 24.

A class $\mathcal{F}$ is *weak VC-major* with dimension $d \ge 1$ if d is the smallest integer $k \ge 1$ such that for all $u \in \mathbb{R}$, the class $\mathcal{C}_u(\mathcal{F}) := \{\{x \in \mathcal{X} : f(x) > u\} : f \in \mathcal{F}\}$ is a VC-class of subsets of $\mathcal{X}$ with dimension not larger than k; see Section 2.6 of [40] for details on VC classes. It can be proved ([3], Eq. (2.5)) that for a weak VC-major class of dimension d, $\log|\mathcal{E}_u(x_1, \ldots, x_n)| \le d \log(2en/d)$ where $\mathcal{E}_u(x_1, \ldots, x_n) := \{\{i = 1, \ldots, n : x_i \in C\} : C \in \mathcal{C}_u(\mathcal{F})\}$. If $f(x) \in [0, 1]$ for all $x \in \mathcal{X}$, $f \in \mathcal{F}$ and $\mathcal{F}$ is a weak VC-major class of dimension d, then Baraud [3], Section 3.3, shows

$$Z_n(\mathcal{F}) \le \int_0^1 \sup_{C \in \mathcal{C}_u(\mathcal{F})} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \mathbb{1}\{X_i \in C\} \right| du = \int_0^1 \sup_{a \in \mathcal{A}_u} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i a_i \right| du,$$

where $\mathcal{A}_u = \{a = (\mathbb{1}\{X_1 \in C\}, \ldots, \mathbb{1}\{X_n \in C\}) \in \{0, 1\}^n : C \in \mathcal{C}_u(\mathcal{F})\}$. From the assumption of weak VC-major class in $\mathcal{F}$, the supremum over $C \in \mathcal{C}_u(\mathcal{F})$ is actually a finite maximum over $a \in \mathcal{A}_u$ which allows us to apply our uniform and non-uniform CLTs (by conditioning on $X_1, \ldots, X_n$ so that $\mathcal{A}_u$ can be treated non-random). To state the final bound, note that after conditioning on $X_1, \ldots, X_n$ the "limiting" Gaussian vector can be taken to as $(G_a)_{a \in \mathcal{A}_u}$ where $G_a := n^{-1/2} \sum_{i=1}^n g_i \sigma(X_i) a_i$, for $\sigma^2(X_i) := \mathbb{E}[\xi_i^2 | X_i]$ and $g_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. Set $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. If $\sup_x \mathbb{E}[\xi_i^3 | X_i = x] < \infty$, then Re-

mark 3.2 implies that

$$\mathbb{E}_{|\mathcal{X}_n}\left[\sup_{a\in\mathcal{A}_u}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i a_i\right|\right]=\mu_u+O(1)\mu_u^2\log(\mu_u)\frac{(L_{n,u}d^2\log^2(en/d))^{1/3}}{n^{1/6}},$$

where $\mu_u:=\mathbb{E}_{|\mathcal{X}_n}[\sup_{a\in\mathcal{A}_u}|n^{-1/2}\sum_{i=1}^{n}g_i\sigma(X_i)a_i|]$. Here $\mathbb{E}_{|\mathcal{X}_n}$ represents the expectation conditional on $\mathcal{X}_n$. Taking expectations with respect to $X_1,\dots,X_n$, we get if $d=O(1)$,

$$\mathbb{E}\big[Z_n(\mathcal{F})\big]\le\big(1+o(1)\big)\int_0^1\mathbb{E}\left[\sup_{C\in\mathcal{C}_u(\mathcal{F})}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i\sigma(X_i)\mathbb{1}\{X_i\in C\}\right|\right]du.$$

Note that the expectation on the right-hand side is the Gaussian mean width of sets $\mathcal{A}_u$. The right-hand side can be bounded using several existing results; see, for example, [38] Chapters 2, 10, [20], Theorem 1.

5. Cramér-type large deviation

In this section, we prove a Cramér-type large deviation result for S_n. A version of this result appeared in [9], Theorem 1, for the case of Banach space valued random variables. In the following result, we make the dependences on distributional constants precise. For this result, we assume that the observations $X_1,\dots,X_n$ are i.i.d. and write the variation measure as ζ instead of ζ_i. Set $C_2=C_0\log(ep)$, $C_3=C_0\log^2(ep)$. We assume the following condition for the result: there exists $H\in(0,\infty)$ such that $\int\exp(H\|x\|)|\zeta|(dx)\le 4$. We also need the following quantities. Set $B:=2(1+\sigma_{\max}^{-2})H^{-1}$, and

$$\tilde{\Pi}_n:=4+12\Phi_{\mathrm{AC},0}\big(8C_3L_ne^{\mathfrak{C}}\big)^{1/3}+\frac{20\Phi_{\mathrm{AC},0}\log(ep)\log(8C_0n)}{Hn^{1/3}}+\frac{5.1\log(ep)}{C_0n^{5/6}},$$

$$\Pi:=\max\left\{\tilde{\Pi},(132\Phi_2)^{4/3},\frac{19C_3L_ne^{\mathfrak{C}}}{\Phi_4},\big(37C_3L_ne^{\mathfrak{C}}\big)^{4/7},\left(\frac{24C_2}{\Phi_4^5H^2}\right)^{4/11}\right\},$$

$$M:=\max\left\{2\Pi,(112\Phi_2+83C_3L_n)^{\frac{4}{3}},\frac{(48C_2)^{10/23}}{(\Phi_4^{32}H^{20})^{1/23}},36(C_3L_n\Phi_2)^{\frac{2}{3}},\frac{(124C_3L_n)^2}{(\mu+1)^{17/8}n^{5/16}}\right\}.$$

The quantity M determines how big the ratio (2) is relative to $(r+1)n^{-1/6}$.

Theorem 5.1 (Cramér-type large deviation). *Under the setting above, we have for* $n\ge 4$,

$$\left|\frac{\mathbb{P}(S_n>r)}{\mathbb{P}(Y>r)}-1\right|\le 1.02M(r+1)n^{-1/6}$$

for all r *such that* $(r+1)n^{-1/6}\le\mathfrak{B}_0\exp(-3M^{1/4}(\mu+1)^{-17/16}\log(en)n^{-5/32})$, *where*

$$\mathfrak{B}_0:=\min\left\{\frac{1}{3(\Phi_4^4\Pi)^{1/3}},\frac{1}{4(\Phi_4^4M)^{4/15}},\right.$$
$$\left.\frac{(\log(en))^{-1/3}}{2(\Phi_4B)^{1/3}},\frac{\Pi^{1/9}(\log(en))^{-4/9}}{2(B\log(ep))^{4/9}},\frac{M^{1/8}(\log(en))^{-1/2}}{(6\mathfrak{C}B\log(ep))^{1/2}}\right\}.$$

The following corollary is obtained for sub-Weibull random vectors using $\Phi_2 \le \Theta\mu^2 \le \Theta\log(ep)$ and $\Phi_4 \le \Theta\mu^4 \le \Theta\log^2(ep)$. For examples where μ is of smaller order than $\sqrt{\log(ep)}$ (as in post-selection inference example; Section 4.1) better rates follow easily. It is noteworthy that the rates in the corollary does not depend on $\alpha \in [1, 2]$.

Corollary 5.1. *Suppose that assumption* (16) *holds for some* $1 \le \alpha \le 2$. *Then, there exist positive constants* $\Theta_1, \Theta_2 \in (0, \infty)$ *depending on* $K_p, \sigma_{\min}$ *and* $\sigma_{\max}$, *such that*

$$\left|\mathbb{P}(S_n > r)/\mathbb{P}(Y > r) - 1\right| \le \Theta_1\big(\log(ep)\big)^{8/3}(r+1)n^{-1/6}$$

for all $n \ge (\log(ep))^{64/15}(\log(en))^{32/5}(\mu+1)^{-34/5}$ *and* $(r+1)n^{-1/6} \le \Theta_2(\log(ep+en))^{-28/9}$.

The corollary requires n to be slightly larger than $\log^4(ep)$ which becomes relaxed if $\mu \ge \Theta\sqrt{\log(ep)}$. Note that $n \ge \log^4(ep)$ is required for the bound on $\delta_{n,0}$ to be less than 1.

6. Outline of the proofs

The following represents the main steps in our proofs of uniform and non-uniform CLTs. The outline for the large deviation result is given at the end of this section. We also detail the differences between the proofs of [16] and [7,31]. We define and control for $r \ge 0$,

$$\Delta_{n,k}(r) := \left|\mathbb{P}\big(\|U_{n,k}\| \le r\big) - \mathbb{P}\big(\|U_{n,0}\| \le r\big)\right|.$$

Recall that $\delta_{n,0} = \sup_{r\ge 0}\max_{1\le k\le n}\Delta_{n,k}(r)$ and $\delta_{n,m} = \sup_{r\ge 0}\max_{1\le k\le n} r^m\Delta_{n,k}(r)$.

Step A: Smoothing inequality. The first step in all Berry–Esseen type results is a smoothing inequality which replaces the probabilities by expectations of smooth approximations of indicators. By Lemma 5.1.1 of [32], we have

$$\Delta_{n,k}(r) \le \max_{\ell=1,2}\left|\mathbb{E}\big[\varphi_\ell(U_{n,k}) - \varphi_\ell(U_{n,0})\big]\right| + \mathbb{P}\big(r-\varepsilon \le \|U_{n,0}\| \le r+\varepsilon\big), \tag{27}$$

where $\varphi_1(x) := \varphi_{r,\varepsilon}(x)$ and $\varphi_2(x) := \varphi_{r-\varepsilon,\varepsilon}(x)$ are given by Lemma 2.1. The second term in (27) is controlled by Theorem 2.2 for anti-concentration.

Step B: Lindeberg replacement. Write $\varphi(\cdot)$ for both $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ and fix any $k \ge 1$. To bound the first term in (27), we use Lindeberg method.

$$\begin{aligned}\left|\mathbb{E}\big[\varphi(U_{n,k}) - \varphi(U_{n,0})\big]\right| &\le \sum_{j=1}^{k}\left|\mathbb{E}\big[\varphi(U_{n,j-1}) - \varphi(U_{n,j})\big]\right| \\ &= \sum_{j=1}^{k}\left|\int \mathbb{E}\big[\varphi\big(W_{n,j} + n^{-1/2}x\big)\big]\zeta_j(dx)\right|,\end{aligned}$$

where $W_{n,j} = n^{-1/2}(X_1 + \cdots + X_{j-1} + Y_{j+1} + \cdots + Y_n)$. The last equality follows from $U_{n,j} = W_{n,j} + n^{-1/2}X_j$ and $U_{n,j-1} = W_{n,j} + n^{-1/2}Y_j$. Using the Taylor expansion of φ (possible because of

smoothness) and $\mathbb{E}[X_j] = \mathbb{E}[Y_j] = 0$, $\mathbb{E}[X_j X_j^\top] = \mathbb{E}[Y_j Y_j^\top]$, we get

$$\left| \int \mathbb{E}\big[\varphi\big(W_{n,j} + n^{-1/2}x\big)\big]\zeta_j(dx) \right| = \left| \int \mathbb{E}\big[\mathrm{Rem}_n(W_{n,j}, x)\big]\zeta_j(dx) \right| \\ \leq \int \mathbb{E}\big[\big|\mathrm{Rem}_n(W_{n,j}, x)\big|\big]|\zeta_j|(dx) =: I_j,$$

where

$$\mathrm{Rem}_n(y, x) = \varphi\big(y + xn^{-1/2}\big) - \varphi(y) - \frac{1}{\sqrt{n}} \sum_{j=1}^{p} x(j)\partial_j \varphi(y) - \frac{1}{2n} \sum_{j,k=1}^{p} x(j)x(k)\partial_{jk}\varphi(y).$$

Step C: Splitting the integral. Based on Step B, it remains to bound the integral of $\mathrm{Rem}_n(W_{n,j}, x)$ with respect to ζ_j. Using the mean value theorem and the bound on derivatives in Lemma 2.1, we have

$$\big|\mathrm{Rem}_n(W_{n,j}, x)\big| \leq \min\left\{ \frac{C_0 \log^2(ep)\varepsilon^{-3}\|x\|^3}{6n^{3/2}}, \frac{C_0 \log(ep)\varepsilon^{-2}\|x\|^2}{n} \right\}. \tag{28}$$

Since the support of $\varphi_1(\cdot), \varphi_2(\cdot)$ is contained in $\{x : \|x\| \in [r - \varepsilon, r + \varepsilon]\}$, the bound above on the remainder can be multiplied by $\mathbb{1}\{|\|W_{n,j}\| - r| \leq \varepsilon + n^{-1/2}\|x\|\}$. Noting that the change point (where the minimum changes from first to second term) in the minimum is at order $n^{1/2}\varepsilon/\log(ep)$, we split the integral of remainder into two parts. Set $\mathcal{E} := \{x \in \mathbb{R}^p : \|x\| \leq n^{1/2}\varepsilon/\log(ep)\}$ and this yields

$$I_j = \int_{\mathcal{E}} \mathbb{E}\big[\big|\mathrm{Rem}_n(W_{n,j}, x)\big|\big]|\zeta_j|(dx) + \int_{\mathcal{E}^c} \mathbb{E}\big[\big|\mathrm{Rem}_n(W_{n,j}, x)\big|\big]|\zeta_j|(dx) =: I_j^{(1)} + I_j^{(2)}.$$

It is intuitively clear that $I_j^{(1)}$ contributes to the main rate term in the bound and $I_j^{(2)}$ would be a second order term since it is an integral over a tail end of the distribution.

In the classical proof of [7], the splitting of integral is done at $\|x\| \leq n^{1/2}\varepsilon$ and in the modern proof of [16] the splitting is done as above.

Step D: Controlling $I_j^{(2)}$. Chernozhukov et al. [16] bound $I_j^{(2)}$ as

$$I_j^{(2)} \leq \int_{\mathcal{E}^c} \frac{C_0 \log^2(ep)\varepsilon^{-3}\|x\|^3}{6n^{3/2}} |\zeta_j|(dx).$$

This is sub-optimal since from (28) the second term in the minimum is a better upper bound. Hence in Theorem 3.1, we bound $I_j^{(2)}$ using

$$I_j^{(2)} \leq \int_{\mathcal{E}^c} \frac{C_0 \log(ep)\varepsilon^{-2}\|x\|^2}{n} |\zeta_j|(dx). \tag{29}$$

Summing over $1 \leq j \leq n$ leads to $M_n(\varepsilon)$ in Theorem 3.1 and leads to better rates under $(2 + \tau)$-moments of $\|X_j\|$.

In the classical proof of [7], $I_j^{(2)}$ is bounded using

$$I_j^{(2)} \leq \int_{\mathcal{E}^c} \frac{C_0 \log(ep)\varepsilon^{-2}\|x\|^2}{n} \mathbb{P}\big(r - \varepsilon - n^{-1/2}\|x\| \leq \|W_{n,j}\| \leq r + \varepsilon + n^{-1/2}\|x\|\big)|\zeta_j|(dx).$$

Note that this bound is better than (29) since the probability is bounded by 1. Since $W_{n,j}$ is related to $U_{n-1,j-1}\sqrt{1-1/n}$, we can use the definition of $\delta_{n-1,0}$ (or $\delta_{n-1,m}$ for non-uniform CLT) to obtain

$$\begin{aligned}
&\mathbb{P}\big(r-\varepsilon-n^{-1/2}\|x\| \le \|W_{n,j}\| \le r+\varepsilon+n^{-1/2}\|x\|\big) \\
&\quad\le \mathbb{P}\big(\sqrt{n/(n-1)}\big(r-\varepsilon-n^{-1/2}\|x\|\big) \le \|U_{n,0}\| \le \sqrt{n/(n-1)}\big(r+\varepsilon+n^{-1/2}\|x\|\big)\big) \\
&\qquad + 2\delta_{n-1,0} \\
&\quad\le \Phi_{\mathrm{AC},0}\big(\varepsilon+n^{-1/2}\|x\|\big)\sqrt{n/(n-1)} + 2\delta_{n-1,0}.
\end{aligned}$$

Here the last inequality follows from Theorem 2.2. This is used in the proof of Theorem 3.1 to get the "right" dependence on τ for the finite moment case. To make the recursion work, we need the anti-concentration inequality to not change with sample size and for this reason i.i.d. assumption is introduced in Proposition 3.1 and Theorem 3.2. In case of non-uniform CLT, we need to split $\mathcal{E}^c$ further in order to avoid dividing by zero when using $\delta_{n-1,m}$.

Step E: Controlling $I_j^{(1)}$. The simpler approach in bound $I_j^{(1)}$ is in the classical proof of [7] that uses

$$I_j^{(1)} \le \int_{\mathcal{E}} \frac{C_0 \log^2(ep)\varepsilon^{-3}\|x\|^3}{6n^{3/2}} \mathbb{P}\big(r-\varepsilon-n^{-1/2}\|x\| \le \|W_{n,j}\| \le r+\varepsilon+n^{-1/2}\|x\|\big)|\zeta_j|(dx).$$

Since $\|x\| \le n^{1/2}\varepsilon/\log(ep)$ for $x \in \mathcal{E}$, the probability can be bounded by $\mathbb{P}(r-2\varepsilon \le \|W_{n,j}\| \le r+2\varepsilon)$ (which does not involve x anymore). Now as in Step D, we can relate this probability to $\delta_{n-1,0}$ and $\Phi_{\mathrm{AC},0}\varepsilon$. This results in $\int \|x\|^3|\zeta_j|(dx) = \nu_3^3$ in the final bound in Proposition 3.1 and leads to a sub-optimal rate in case X_j have exponential tails.

A better way to control $I_j^{(1)}$ from [16] uses

$$\begin{aligned}
&\big|\mathrm{Rem}_n(W_{n,j},x)\big| \\
&\quad= \frac{1}{2}\left|\sum_{j_1,j_2,j_3=1}^{p} \frac{x(j_1)x(j_2)x(j_3)}{n^{3/2}} \int_0^1 (1-t)^2 \partial_{j_1j_2j_3}\big(W_{n,j}+tn^{-1/2}x\big)\,dt\right| \\
&\quad\le \sum_{j_1,j_2,j_3=1}^{p} \frac{|x(j_1)x(j_2)x(j_3)|}{2n^{3/2}} \\
&\qquad\times \int_0^1 (1-t)^2 D_{j_1j_2j_3}\left(W_{n,j}+\frac{tx}{\sqrt{n}}\right)\mathbb{1}\left\{\big|\|W_{n,j}\|-r\big| \le \varepsilon+\frac{\|x\|}{\sqrt{n}}\right\}dt.
\end{aligned}$$

By stability property (14) of $D_{j_1j_2j_3}$ along with $\|x\|/\sqrt{n} \le \varepsilon/\log(ep)$ for $x \in \mathcal{E}$, we get

$$e^{-\mathfrak{C}} D_{j_1j_2j_3}(W_{n,j}) \le D_{j_1j_2j_3}\big(W_{n,j}+txn^{-1/2}\big) \le e^{\mathfrak{C}} D_{j_1j_2j_3}(W_{n,j}).$$

This implies

$$I_j^{(1)} \le \frac{e^{\mathfrak{C}}}{2n^{3/2}} \sum_{j_1,j_2,j_3=1}^{p} \int_{\mathcal{E}} \big|x(j_1)x(j_2)x(j_3)\big||\zeta_j|(dx)\mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j})\mathbb{1}\big\{\big|\|W_{n,j}\|-r\big| \le 2\varepsilon\big\}\big].$$

By Hölder's inequality,

$$\int_{\mathcal{E}} \big|x(j_1)x(j_2)x(j_3)\big||\zeta_j|(dx) \le \max_{1\le j_1\le p} \int_{\mathcal{E}} \big|x(j_1)\big|^3|\zeta_j|(dx).$$

If X_j has $(2+\tau)$-moments for some $\tau \geq 1$, then the integral over $\mathcal{E}$ can be replaced by integral over $\mathbb{R}^p$ which leads to L_n (when summed over $1 \leq j \leq n$). In case X_j has only $(2+\tau)$-moments for some $\tau < 1$, then using $\mathcal{E} = \{\|x\| \leq n^{1/2}\varepsilon/\log(ep)\}$, we can write

$$\int_{\mathcal{E}} \big|x(j_1)\big|^3 |\zeta_j|(dx) \leq \left(\frac{n^{1/2}\varepsilon}{\log(ep)}\right)^{1-\tau} \int \big|x(j_1)\big|^{2+\tau} |\zeta_j|(dx),$$

which would lead to a version of L_n only involving "weak" $(2+\tau)$-moment rather than the "weak" third moment. Getting back to $I_j^{(1)}$, the above discussion leads to

$$I_j^{(1)} \leq \frac{e^{\mathfrak{C}}}{2n^{3/2}} \max_{1\leq j_1\leq p} \int_{\mathcal{E}} \big|x(j_1)\big|^3 |\zeta_j|(dx) \sum_{j_1,j_2,j_3=1}^{p} \mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j}) \mathbb{1}\big\{\big|\|W_{n,j}\| - r\big| \leq 2\varepsilon\big\}\big]. \tag{30}$$

There are two ways to bound the right-hand side. First, the way used in the proof of [16] is to split the indicator inside the expectation by adding $Y_j \in \mathcal{E}$ and $Y_j \in \mathcal{E}^c$:

$$\begin{aligned}
&\mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j}) \mathbb{1}\big\{\big|\|W_{n,j}\| - r\big| \leq 2\varepsilon\big\}\big] \\
&\quad = \mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j}) \mathbb{1}\big\{\big|\|W_{n,j}\| - r\big| \leq 2\varepsilon, Y_j \in \mathcal{E}\big\}\big] \\
&\qquad + \mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j}) \mathbb{1}\big\{\big|\|W_{n,j}\| - r\big| \leq 2\varepsilon, Y_j \in \mathcal{E}^c\big\}\big].
\end{aligned}$$

The first term on the right-hand side can be linked using the stability property (14) to $\mathbb{E}[D_{j_1j_2j_3}(U_{n,j-1}) \mathbb{1}\{|\|U_{n,j-1}\| - r| \leq 3\varepsilon\}]$. Further using the bound on derivative this can bounded by $C_3\varepsilon^{-3}\mathbb{P}(r - 3\varepsilon \leq \|U_{n,j-1}\| \leq r + 3\varepsilon)$ that can linked to $\delta_{n,0} + \Phi_{\mathrm{AC},0}\varepsilon$. The second term is, anyways, small since it involves $Y_j \in \mathcal{E}^c$ which is a small probability event since Y_j has Gaussian tails.

The calculations after (30) are used in the proof of [16] to get a bound in terms of $\delta_{n,0}$ and solve the inequality for $\delta_{n,0}$. This is what we followed in the proof of Theorem 3.1 but using the "right" bound on $I_j^{(2)}$.

For the proof of Theorem 3.2, we proceed from (30) by using

$$\sum_{j_1,j_2,j_3=1}^{p} \mathbb{E}\big[D_{j_1j_2j_3}(W_{n,j}) \mathbb{1}\big\{\big|\|W_{n,j}\| - r\big| \leq 2\varepsilon\big\}\big] \leq C_3\varepsilon^{-3}\mathbb{P}\big(r - 2\varepsilon \leq \|W_{n,j}\| \leq r + 2\varepsilon\big),$$

where the right-hand side can be bounded in terms of $\delta_{n-1,0}$ and anti-concentration term as shown above in the control of $I_j^{(2)}$.

Combining the bounds and solving recursions (when needed) completes the proofs of all three versions of uniform CLT. By appropriate (minor) modifications mentioned in the outline above, the two versions of non-uniform CLT also follow.

Sketch of the proof of Theorem 5.1. The proof of Theorem 5.1 relies upon Lindeberg method and a refined induction argument. The starting point for the proof is Step B of the outline. To control I_j we split it into two parts depending on whether $\{\|x\| \leq n^{1/2} f_n(r)\}$ or not for some function $f_n(\cdot)$ of r. The case when $x \geq n^{1/2} f_n(r)$ the integral is bounded using $\int \exp(H\|x\|)|\zeta|(dx) \leq 4$ and Markov's inequality by:

$$\frac{8\beta C_2\varepsilon^{-2}}{nH^2} \exp\big(-Hn^{1/2} f_n(r)/2\big) \leq \frac{8C_2\varepsilon^{-2}\beta}{n^{2+\Phi_1}H^2\Phi_0} \mathbb{P}\big(\|Y\| > r\big). \tag{31}$$

The integral corresponding to the case when x is smaller than $n^{1/2} f_n(r)$ is bounded by

$$\frac{C_3 \varepsilon^{-3} e^{\mathfrak{C} f_n(r) \log(ep)/\varepsilon} L_n}{n^{3/2}} \mathbb{P}\big(\bar{a}_n(r) \le \|W_{n,j}\| \le \bar{b}_n(r)\big) \tag{32}$$

for suitably chosen $\bar{a}_n(r)$ and $\bar{b}_n(r)$.

Since $W_{n,j} \stackrel{d}{=} ((n-1)/n)^{1/2} U_{n-1,j-1}$, we get

$$\begin{aligned}\Delta_n(r) \le \mathbb{P}\big(r - \varepsilon \le \|Y\| \le r + \varepsilon\big) + \frac{8 C_2 \varepsilon^{-2} \beta}{\Phi_0 H^2 n^{1+\Phi_1}} \mathbb{P}\big(\|Y\| > r\big) \\ + \frac{C_3 \varepsilon^{-3} L_n e^{\mathfrak{C} f_n(r) \log(ep)/\varepsilon}}{6 n^{1/2}} \max_{0 \le k \le n-1} \mathbb{P}\big(a_n(r) \le \|U_{n-1,k}\| \le b_n(r)\big)\end{aligned} \tag{33}$$

for some $a_n(r)$ and $b_n(r)$. Our next task is to bound the right-hand side of (33) in term of $\mathbb{P}[Y > r]$. We now inductively use a bound on $\mathbb{P}[\|U_{n-1,j}\| \ge a_n(r)]$ in terms of $\mathbb{P}[Y > r]$ to get a bound for (33). Finally, bounding $\mathbb{P}[r - \varepsilon \le Y \le r + \varepsilon]$ in terms of $\mathbb{P}[Y > r]$ using Theorem 2.1 and summing this with the bounds for (31), (32) we obtain the following result. For any $1 \le k \le n$, $|\mathbb{P}(U_{n,k} > r)/\mathbb{P}(Y > r) - 1| \le \Pi T_{n,r}^{1/4}$, for all $r \in \mathbb{R}$ satisfying

$$T_{n,r} \le \min\left\{ \frac{1}{16 \Phi_4^4 \Pi}, \frac{1}{2 \Phi_4 B \log(en)}, \frac{\Pi^{1/3}}{(B \log(en) \log(ep))^{4/3}} \right\},$$

where $T_{n,r} = (r+1)^3 n^{-1/2}$. The details are in Lemma S.6.3. Since $T_{n,r}^{1/4} = (r+1)^{3/4} n^{-1/8}$, Lemma S.6.3 gives a large deviation with rate $n^{-1/8}$ (for all $n \ge 1$). However this rate can be modified to rate $n^{-1/6}$ by a more refined induction argument. This is done in detail in the proof of Theorem 5.1.

7. Summary and future directions

In this paper, we proved non-uniform central limit theorems and large deviations for scaled averages of independent high-dimensional random vectors based on dimension-free anti-concentration inequalities. We further illustrated the usefulness of these results in the context of post-selection inference for linear regression and in bounding the expectation of suprema of empirical processes. All the proofs are based on the Lindeberg method which is an integral tool in Banach space CLTs. Using the stability property introduced in [16], we obtained refinements for uniform as well as non-uniform CLTs. It should be mentioned here that we credit Bentkus [9], Bentkus and Račkauskas [8] for the proof of Theorem 5.1. In comparison to [16], we mention that our setting is restrictive in the sense that we consider l_∞ balls while [16] consider general sparsely convex sets. Once an anti-concentration result such as Theorem 2.2 holds, it is fairly easy to follow our proofs to extend the results which we hope to pursue in the future along with the discussion of "best" anti-concentration inequalities.

In this work, we have presented the results in the simpler setting with independent random vectors. The Lindeberg method is well-known for its robustness to independence assumptions and hence extensions to the case of dependent random vectors (in particular martingales) form an interesting future direction. Apart from the application in PoSI, other areas of interest in terms of applications are bootstrap and high-dimensional vectors with a specified group structure. Our results can be used to obtain CLTs and large deviations to empirical processes which form an interesting direction. See [29] and [13] for some results.

Appendix. Uniform CLT for $(2+\tau)$-moments with $\tau < 1$

Following the outline in (Step E of) Section 6 for $\tau \le 1$, we get for any $\varepsilon > 0$:

$$\delta_{n,0} \le \frac{e^{2\mathfrak{C}} C_3 \varepsilon^{-3} L_{n,\tau}}{2n^{3/2}} \left(\frac{n^{1/2}\varepsilon}{\log(ep)} \right)^{1-\tau} [2\Phi_{\mathrm{AC},0}\varepsilon + \delta_{n,0}] + \frac{C_2 M_n(\varepsilon)}{\varepsilon^2} + \Phi_{\mathrm{AC},0\varepsilon}$$
$$+ \frac{e^{\mathfrak{C}} C_3 \varepsilon^{-3}}{2n^{3/2}} \left(\frac{n^{1/2}\varepsilon}{\log(ep)} \right)^{1-\tau} \sum_{j=1}^{n} \max_{1 \le j_1 \le p} \int \big|x(j_1)\big|^{2+\tau} |\zeta_j|(dx) \mathbb{P}\big(\|Y_j\| > n^{1/2}\varepsilon / \log(ep)\big),$$

where $L_{n,\tau} = n^{-1} \sum_{j=1}^{n} \max_{1 \le j_1 \le p} \int |x(j_1)|^{2+\tau} |\zeta_j|(dx)$. It is also clear that

$$M_n(\varepsilon) \le \nu_{2+\tau}^{2+\tau} \left(\frac{\log(ep)}{n^{1/2}\varepsilon} \right)^{2+\tau}.$$

Set $\varepsilon = \varepsilon_n$ so that

$$\frac{e^{2\mathfrak{C}} C_3 L_{n,\tau}}{2n^{3/2}} \left(\frac{n^{1/2}\varepsilon}{\log(ep)} \right)^{1-\tau} \le \frac{1}{2} \quad \text{and} \quad \frac{C_2 \nu_{2+\tau}^{2+\tau} \log^{\tau}(ep)}{n^{\tau/2}\varepsilon^{2+\tau}} \le \frac{1}{r_n^{2+\tau}}$$

for some $r_n \ge 1$. This implies that

$$\delta_{n,0} \lesssim \Phi_{\mathrm{AC},0}\varepsilon + \frac{1}{r_n^{2+\tau}}$$
$$\lesssim \Phi_{\mathrm{AC},0} \frac{(L_{n,\tau} \log^2(ep))^{1/(2+\tau)}}{n^{1/2} (\log(ep))^{(1-\tau)/(2+\tau)}} + r_n \Phi_{\mathrm{AC},0} \nu_{2+\tau} \frac{(\log(ep))^{(\tau+1)/(\tau+2)}}{n^{\tau/(4+2\tau)}} + \frac{1}{r_n^{2+\tau}}.$$

Choosing the "optimal" order of r_n by equating the last two terms, we get

$$\delta_{n,0} \lesssim \left[\frac{\Phi_{\mathrm{AC},0} L_{n,\tau}^{1/(2+\tau)}}{n^{1/2}} + \frac{(\Phi_{\mathrm{AC},0}\nu_{2+\tau})^{(\tau+2)/(\tau+3)}}{n^{\tau/(6+2\tau)}} \right] \big(\log(ep)\big)^{(\tau+1)/(\tau+2)}$$
$$\lesssim \big[\Phi_{\mathrm{AC},0} L_{n,\tau}^{1/(2+\tau)} + (\Phi_{\mathrm{AC},0}\nu_{2+\tau})^{(\tau+2)/(\tau+3)} \big] \frac{(\log(ep))^{(\tau+1)/(\tau+2)}}{n^{\tau/(6+2\tau)}}.$$

The last inequality follows since $n^{\tau/(6+2\tau)} \le n^{1/8} \le n^{1/2}$.

Acknowledgment

We would like to thank Prof. Jian Ding for comments that led to an improved presentation.

Supplementary Material

Supplement to "High-dimensional CLT: Improvements, non-uniform extensions and large deviations" (DOI: 10.3150/20-BEJ1233SUPP; .pdf). We provide the proofs of the all the results in the Supplementary Material.

References

[1] Bachoc, F., Blanchard, G. and Neuvial, P. (2018). On the post selection inference constant under restricted isometry properties. *Electron. J. Stat.* **12** 3736–3757. MR3878579 https://doi.org/10.1214/18-ejs1490

[2] Banerjee, D., Kuchibhotla, A.K. and Mukherjee, S. (2018). Cramér-type large deviation and non-uniform central limit theorems in high dimensions. Preprint. Available at arXiv:1806.06153v1.

[3] Baraud, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *Electron. J. Stat.* **10** 1709–1728. MR3522658 https://doi.org/10.1214/15-EJS1055

[4] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. and Kato, K. (2018). High-dimensional econometrics and regularized GMM. Preprint. Available at arXiv:1806.01888.

[5] Bentkus, V. (1990). Smooth approximations of the norm and differentiable functions with bounded support in Banach space l_∞^k. *Lith. Math. J.* **30** 223–230.

[6] Bentkus, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.* **49** 400–410. MR2144310 https://doi.org/10.1137/S0040585X97981123

[7] Bentkus, V., Götze, F., Paulauskas, V. and Račkauskas, A. (2000). The accuracy of Gaussian approximation in Banach spaces. In *Limit Theorems of Probability Theory* 25–111. Berlin: Springer.

[8] Bentkus, V. and Račkauskas, A. (1990). On probabilities of large deviations in Banach spaces. *Probab. Theory Related Fields* **86** 131–154. MR1065276 https://doi.org/10.1007/BF01474639

[9] Bentkus, V.Y. (1987). Large deviations in Banach spaces. *Theory Probab. Appl.* **31** 627–632.

[10] Bentkus, V.Yu. (1985). Lower bounds for the rate of convergence in the central limit theorem in Banach spaces. *Liet. Mat. Rink.* **25** 10–21. MR0823198

[11] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[12] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

[13] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230

[14] Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* **162** 47–70. MR3350040 https://doi.org/10.1007/s00440-014-0565-9

[15] Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Detailed proof of Nazarov's inequality. Preprint. Available at arXiv:1711.10696.

[16] Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963 https://doi.org/10.1214/16-AOP1113

[17] Chernozhukov, V., Chetverikov, D., Kato, K. and Koike, Y. (2019). Improved central limit theorem and bootstrap approximations in high dimensions. Preprint. Available at arXiv:1912.10529.

[18] Giné, E. (1976). Bounds for the speed of convergence in the central limit theorem in $C(S)$. *Z. Wahrsch. Verw. Gebiete* **36** 317–331. MR0418190 https://doi.org/10.1007/BF00532697

[19] Han, Q. (2019). Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions. Preprint. Available at arXiv:1905.12823.

[20] Han, Q. and Wellner, J.A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319. MR3953452 https://doi.org/10.1214/18-AOS1748

[21] Koike, Y. (2019). High-dimensional central limit theorems for homogeneous sums. Preprint. Available at arXiv:1902.03809.

[22] Koike, Y. (2019). Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. Preprint. Available at arXiv:1911.00160.

[23] Kuchibhotla, A.K., Brown, L.D., Buja, A., George, E.I. and Zhao, L. (2018). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. Preprint. Available at arXiv:1802.05801.

[24] Kuchibhotla, A.K., Brown, L.D., Buja, A., George, E.I. and Zhao, L. (2018). Valid post-selection inference in assumption-lean linear regression. Preprint. Available at arXiv:1806.04119.

[25] Kuchibhotla, A.K., Mukherjee, S. and Banerjee, D. (2020). Supplement to "High-dimensional CLT: Improvements, non-uniform extensions and large deviations." https://doi.org/10.3150/20-BEJ1233SUPP

[26] Ledoux, M. and Talagrand, M. (2011). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Classics in Mathematics*. Berlin: Springer. MR2814399

[27] Lopes, M.E., Lin, Z. and Müller, H.-G. (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *Ann. Statist.* **48** 1214–1229. MR4102694 https://doi.org/10.1214/19-AOS1844

[28] Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*. *Lecture Notes in Math.* **1807** 169–187. Berlin: Springer. MR2083397 https://doi.org/10.1007/978-3-540-36428-3_15

[29] Norvaiša, R. and Paulauskas, V. (1991). Rate of convergence in the central limit theorem for empirical processes. *J. Theoret. Probab.* **4** 511–534. MR1115160 https://doi.org/10.1007/BF01210322

[30] Paulauskas, V. and Račkauskas, A. (1989). *Approximation Theory in the Central Limit Theorem*: *Exact Results in Banach Spaces*. *Mathematics and Its Applications* (*Soviet Series*) **32**. Dordrecht: Kluwer Academic. MR1015294 https://doi.org/10.1007/978-94-011-7798-6

[31] Paulauskas, V. and Račkauskas, A. (1991). Nonuniform estimates in the central limit theorem in Banach spaces. *Liet. Mat. Rink.* **31** 483–496. MR1162240 https://doi.org/10.1007/BF00973059

[32] Paulauskas, V. and Račkauskas, A. (2012). *Approximation Theory in the Central Limit Theorem*: *Exact Results in Banach Spaces*. *Mathematics and Its Applications* **32**. Berlin: Springer.

[33] Petrov, V.V. (1995). *Limit Theorems of Probability Theory*: *Sequences of Independent Random Variables*. *Oxford Studies in Probability* **4**. New York: The Clarendon Press. MR1353441

[34] Saulis, L. and Statulevičius, V.A. (1991). *Limit Theorems for Large Deviations*. *Mathematics and Its Applications* (*Soviet Series*) **73**. Dordrecht: Kluwer Academic. MR1171883 https://doi.org/10.1007/978-94-011-3530-6

[35] Sazonov, V.V. (1981). *Normal Approximation – Some Recent Advances*. *Lecture Notes in Math.* **879**. Berlin: Springer. MR0643968

[36] Sazonov, V.V. and Ul'yanov, V.V. (1982). On the accuracy of normal approximation. *J. Multivariate Anal.* **12** 371–384. MR0666012 https://doi.org/10.1016/0047-259X(82)90072-0

[37] Sun, Q. (2020). Gaussian approximations for maxima of random vectors under $(2+\iota)$-th moments. *Statist. Probab. Lett.* **158** Art. ID 108523. MR4027763 https://doi.org/10.1016/j.spl.2019.05.022

[38] Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes*: *Modern Methods and Classical Problems*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge*. *A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series*. *A Series of Modern Surveys in Mathematics*] **60**. Heidelberg: Springer. MR3184689 https://doi.org/10.1007/978-3-642-54075-2

[39] van de Geer, S. and Wainwright, M.J. (2017). On concentration for (regularized) empirical risk minimization. *Sankhya A* **79** 159–200. MR3707417 https://doi.org/10.1007/s13171-017-0111-9

[40] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

Received June 2019 and revised February 2020