Sub-exponential and -Gaussian Parameters Estimation for Tight Non-asymptotic Inference

Haoyu Wei
Email: cute@pku.edu.cn

Working Paper Slides for Discussion

Collaborators: Guang Cheng (Purdue University); Huiming Zhang (University of Macau)

October 10, 2021

Overview

- 1.1 CIs from Simple Statistics and Probability
- 1.2 CIs from Berry-Esseen corrected CLT and Hoeffding's inequality
- 1.3 Optimal variance proxy in sub-class distributions

1. 2. Estimation of sub-exponential and sub-Gaussian norms

2. 3. Tight concentration inequalities for function of vectors
   - 3.1 Extended McDiarmid's inequalities
   - 3.2 Concentration for sub-Gaussian and -exponential vectors

3. 4. Non-asymptotic inference
   - 4.1 The UCB in Bandit Problem
   - 4.2 Non-asymptotic bounds in DNN

CIs from Undergraduate Statistics and Probability

- Given that $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(\mu_0, \sigma)$, if $\sigma = 1$

$$\text{Additivity} \ \Rightarrow \mathrm{P}\left(\mu_0 \in \left[\bar{X} \pm 1.96/\sqrt{n}\right]\right) = 95\% \text{ for any } n.$$

- Without knowing the law of $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} F(\mu_0, \sigma)$, if $\sigma = 1$

$$\text{CLT} \ \Rightarrow \mathrm{P}\left(\mu_0 \in \left[\bar{X} \pm 1.96/\sqrt{n}\right]\right) \to 95\% \text{ as } n \to \infty.$$

However, **the price is the "asymptotic" validity.**

- **Q1**. For $n < \infty$, what if $F(\mu_0, \sigma)$ is non-Gaussian and unbounded,

to get $\mathrm{P}(\mu_0 \in [\widehat{L}_n, \widehat{U}_n]) \geq 1 - \delta$ based on concentration inequalities?

(No assumption for **densities**, but a few **moment conditions**)

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. AOS. Causal treatment effect estimation;
Yang, Y., Shang, Z., and Cheng, G. (2020). Non-asymptotic Theory for Nonparametric Testing. COLT. Hypothesis testing;
Arlot, S., Blanchard, G., & Roquain, E. (2010). Some nonasymptotic results on resampling in high dimension, I: confidence regions. AOS. Bootstrapped CIs for Gaussian data

CIs with the a small sample size

In experimental science (Rousseeuw&Verboven,2002), $n = 4$ to 8.

- **Q2**. What if n is extremely small, in order to get

$$\mathrm{P}(\mu_0 \in [L_n, U_n]) \geq 1 - \delta \text{ based on a few } \textbf{moment conditions}?$$

The normal approximated CIs cannot work for small n (**B-E bounds**).

- Let $\{X_i\}_{i=1}^n$ be i.i.d. with $\mathrm{E}X_1 = 0, \mathrm{E}X_1^2 = \sigma^2 > 0$, $\mathrm{E}\left|X_1\right|^3 = \rho < \infty$. Shevtsova(2013) gave a tighter B-E bounds :

$$\Delta_n := \sup_{x\in\mathbb{R}} \left|\mathrm{P}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_n \leq x\right) - \Phi(x)\right| \leq \frac{0.3328\left(\rho + 0.429\sigma^3\right)}{\sigma^3\sqrt{n}}, \ \forall\ n \geq 1.$$

- Consider Bernoulli samples $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}\,(1/2)$, with $\sigma = 1/2$ and $\rho = 1/8$, and Zolotukhin et al. (2018) shown $\Delta_n \leq 0.409954/\sqrt{n}$.

Rousseeuw, P. J., & Verboven, S. (2002). CSDA, 40(4), 741-758.; Shevtsova, I. G. (2013). Informatics and its Applications, 7(1):124–125.; Zolotukhin, A., Nagaev, S., and Chebotarev, V. (2018). Modern Stochastics, 5(3):385–410.

Put $\delta = 0.05, 0.075$. Hoeffding's inequality

$$\mathrm{P}(|\bar{X}_n - 1/2| \leq \frac{1}{2\sqrt{n}} \cdot \sqrt{2\log(\frac{2}{\delta})}) \geq 1 - \delta \text{ for } n \geq 1.$$

B-E bounds

$$\mathrm{P}(|\bar{X}_n - 1/2| \leq \frac{-1}{2\sqrt{n}} \cdot \Phi^{-1}(\frac{\delta}{2} - \frac{0.409954}{\sqrt{n}})) \geq 1 - \delta \text{ for } n \geq (0.8199/\delta)^2,$$

which requires $n \geq 269, 120$ for $\delta = 0.05, 0.075$.

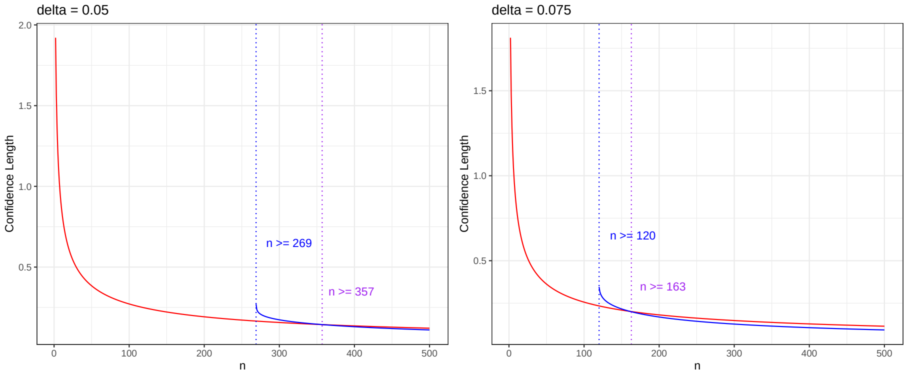


Figure: CIs via Hoeffding's inequality (red line) and B-E-corrected CLT (blue line).

Optimal variance proxy in sub-class distributions

Assume mean is zero for X in all cases.

Definition 0.1 ($X \sim \text{subG}(\sigma^2)$, Sub-Gaussian and parameter)

X is called sub-Gaussian with **variance proxy** σ^2 if MGF $\mathrm{E}e^{tX} \leq e^{\sigma^2 t^2/2}$, $\forall\, t \in \mathbb{R}$.
Optimal variance proxy is the minimal σ^2

$$\sigma^2_{opt}(X) := \inf\left\{\sigma^2 \geq 0 : \mathrm{E}e^{tX} \leq e^{\sigma^2 t^2/2}, \forall\, t \in \mathbb{R}\right\}.$$

By $\mathrm{E}e^{tX} \leq e^{\sigma^2_{opt} t^2/2}$, Chernoff's inequality implies

$$\mathrm{P}\left(X \geq t\right) \leq \inf_{s>0} e^{-st}\mathrm{E}e^{sX} \leq \inf_{s>0} e^{-st+\frac{\sigma^2_{opt}s^2}{2}} \xlongequal{s=t/\sigma^2_{opt}} e^{-\frac{t^2}{2\sigma^2_{opt}}}.$$

For $\{X_i\}_{i=1}^n \overset{\text{ind.}}{\sim} \text{subG}(\sigma^2_{opt}(X_i))$, we have **sub-G Hoeffding's inequality**

$$\mathrm{P}(|\textstyle\sum_{i=1}^n X_i| \geq t) \leq 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n \sigma^2_{opt}(X_i)}\right\},\ t \geq 0.$$

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2):235–256. subG(1) data

Motivation from estimating non-asymptotical CIs

Assume $\{X_i - \mu\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{subG}(\sigma^2_{opt}(X))$ with $\mu = \mathrm{E}X$. Hoeffding gives

$$\mu \in [\bar{X}_n - \sqrt{2\sigma^2_{opt}(X)n^{-1}\log(2/\alpha)}, \bar{X}_n + \sqrt{2\sigma^2_{opt}(X)n^{-1}\log(2/\alpha)}].$$

$\sigma^2_{opt}(X)$ is need to estimate!

- The sub-Gaussian MGF bound $\mathrm{E}e^{sX} \le e^{\frac{\sigma^2 s^2}{2}}$, $\forall s \in \mathbb{R}$ is too strong!

Definition 0.2 (Petrov(1975), $X \sim \text{subE}(\lambda, a)$)

X is sub-exponential with parameters (λ, α) if $\mathrm{E}e^{sX} \le e^{\frac{s^2\lambda^2}{2}}$ for all $|s| < 1/a$.

Cramer's condition:

X is sub-exponential if its MGF exists in a neighborhood of zero.

Suppose $\{X_i^2 - \text{Var}\, X\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \text{subE}(\lambda, a)$, $(1-\alpha)$-CI for $\text{Var}\, X$ is

$$\{\overline{X_n^2} - [\lambda\sqrt{2n^{-1}\log(2/\alpha)} + 2a\log(2/\alpha)/n], \overline{X_n^2} + [\lambda\sqrt{2n^{-1}\log(2/\alpha)} + 2a\log(2/\alpha)/n]\}.$$

To avoid estimating two parameters (λ, a), we resort to sub-Gamma MGF.

Cramér, H. (1938). Actual. Sci. Ind., 736:5–23.; Petrov, V. V. (1975). Springer.

Sub-R norm for sub-exponential and -Gaussian Parameters

- Bernstein's moment conditions:

$$\mathrm{E}|X_i|^z \leq v^2\kappa^{z-2}z!/2, \text{ for all } z \geq 2 \text{ where } \kappa > 0 \text{ and } v = \mathrm{Var}X_1,$$

- then it gives Bernstein's inequality for $t \geq 0$

$$\mathrm{P}(|\sum_{i=1}^n X_i| \geq t) \leq 2e^{-\frac{t^2}{2nv^2+2t\kappa}}, \quad \mathrm{P}(|\sum_{i=1}^n X_i| \geq \sqrt{2tn}v + t\kappa) \leq 2e^{-t}.$$

Now, we only focus on the estimation of κ as the sub-exponential parameter.

Definition 1.1 (Sub-R norm)

Given R with $\mathrm{E}\,R = 0$ and $\mathrm{Var}\,R = 1$ s.t. $\{\mathrm{E}R^k \propto r(k)\}_{k=2}^{\infty}$ (explicit sequence). Let X is data s.t. $\max_{k\geq 2} \mathrm{E}X^k/r(k) < \infty$ for each k. We define **sub-R norm** of X as

$$\|X\|_R = \max_{k\geq 2} [\frac{\mathrm{E}X^k}{r(k)}]^{1/k}.$$

- R is simple **comparison** r.v. with $N(0,1) =: G$ and $(\mathrm{Exp}(\lambda) - \lambda)/\lambda =: E$.

- For $X \sim \mathrm{Exp}(\lambda)$ with $f_X(x) = \lambda^{-1} e^{-x/\lambda},\ x, \lambda > 0$ and $E := (X - \lambda)/\lambda$ with $\mathrm{E}E^k = k! \sum_{i=0}^{k} \frac{(-1)^i}{i!} :=!k$ (the sub-factorial of k).

Put $R = G$, we obtain sub-exponential norm with $r(k) =!k$ and $\|X - \lambda\|_E = \lambda$.

Definition 1.2 (Sub-exponential norm)

$$\|X\|_E = \max_{k \geq 2} \left(\tfrac{1}{!k}\mathrm{E}X^k\right)^{1/k} \text{ if } \mathrm{E}X = 0.$$

Theorem 1.3 (Tight Bernstein-type concentration)

For ind. $\{X_i\}_{i=1}^n$ *with* $\max_{i \in [n]} \|X_i\|_E < \infty$, *then*

$$\mathrm{P}\{|\textstyle\sum_{i=1}^n X_i| > (2t \sum_{i=1}^n \|X_i\|_E^2)^{1/2} + \max_{i \in [n]} \|X_i\|_E\, t\} \leq 2e^{-t}\ \forall t \geq 0.$$

Sub-Gaussian norm

$$\mathrm{E}[N(0,1)]^p = \begin{cases} 0 & \text{odd } p \\ (p-1)!! & \text{even } p \end{cases}, \ (2k-1)!! := \prod_{i=1}^{k}(2i-1) \text{ (double factorial).}$$

Put $R = G$, we obtain sub-Gaussian norm with $r(k) = !k$ and $\|N(0,v^2)\|_G = v$.

Definition 1.4 (Sub-Gaussian norm, Buldygin & Kozachenko(2000))

$$\|X\|_G = \sup_{k\geq 1}\left[\frac{\mathrm{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)} \text{ if } \mathrm{E}X = 0.$$

Theorem 1.5 (Tight Hoeffding-type concentration)

(a). *If $\{X_i\}_{i=1}^n$ are sym. about zero and ind. with finite sub-G norm,*

$$\mathrm{E}e^{tX_i} \leq e^{t^2\|X_i\|_G^2/2} \text{ and } \mathrm{P}(|\textstyle\sum_{i=1}^n X_i| \geq t) \leq 2\exp\{-t^2/[2\sum_{i=1}^n \|X_i\|_G^2]\}.$$

(b). *If $\{X_i\}_{i=1}^n$ are not sym. about zero, then $\mathrm{E}e^{tX_i} \leq e^{t^2(\sqrt{2}\|X_i\|_G)^2/2}$ and*

$$\mathrm{P}(|\textstyle\sum_{i=1}^n X_i| \geq t) \leq 2e^{-t^2/[4\sum_{i=1}^n \|X_i\|_G^2]}.$$

Summary of sub-Gaussian and-exponential norms

Norms	References
$\sigma_{opt}^2(X) = \inf\left\{\sigma^2 > 0 : \mathrm{E}e^{tX} \le e^{\sigma^2 t^2/2}, \forall\, t \in \mathbb{R}\right\}$; $\tau_a^2(X) = \inf\left\{\tau^2 > 0 : \mathrm{E}e^{tX} \le e^{\tau^2 t^2/2}, \forall\, \lvert t\rvert < 1/a\right\}$.	Buldygin & Kozachenko(2000)
$\lVert X\rVert_{w_2} = \inf\{c > 0 : \mathrm{E}e^{\lvert X\rvert^2/c^2} \le 2\}$; $\lVert X\rVert_{w_1} = \inf\{c > 0 : \mathrm{E}e^{\lvert X\rvert/c} \le 2\}$.	Orlicz norms in van der Vaart & Wellner(1996)
$\lVert X\rVert_{\psi_2} = \sup_{k\ge 2} k^{-1/2}\left(\mathrm{E}\lvert X\rvert^k\right)^{1/k}$; $\lVert X\rVert_{\psi_1} = \sup_{k\ge 2} k^{-1}\left(\mathrm{E}\lvert X\rvert^k\right)^{1/k}$.	Vershynin (2010)
$\lVert X\rVert_G = \sup_{k\ge 1} [\frac{2^k k!}{(2k)!}\mathrm{E}X^{2k}]^{1/(2k)}$ $\lVert X\rVert_E = \sup_{k\ge 2} ([k! \sum_{i=0}^k (-1)^i/i!]^{-1}\mathrm{E}X^k)^{1/k}$	Buldygin & Kozachenko(2000) Sub-exponential norm (Our)

Empirical MGF estimation,

$$\widehat{\sigma}_{opt}^2(X;\lambda) = \arg\min_{\sigma>0} \int \left|\frac{1}{n}\sum_{i=1}^n e^{tX_i} - \lambda e^{\sigma^2 t^2/2}\right|\omega(t)\, dt + P(\lambda).$$

Buldygin, V. V. and Kozachenko, I. V. (2000). Metric characterization of random variables and random processes, AMS.
Vershynin, R. (2010).Introduction to the non-asymptotic analysis of random matrices.arXiv preprint arXiv:1011.3027.

Properties of optimal variance proxy

- **Variance upper bounds**: $\sigma^2_{opt}(X) \geq \operatorname{Var} X$. The $\sigma^2_{opt}(X)$ not only characterizes the speed of decay in the tail prob. but also is an upper bounds for the Var X as well:

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

- Bernoulli r.v. $X \in \{0,1\} \sim \operatorname{Ber}(\mu)$ with mean $\mu \in (0,1)$ is sub-Gaussian with

$$\sigma^2_{opt}(X-\mu) = \frac{(1-2\mu)}{2\log\frac{1-\mu}{\mu}} \leq \mu(1-\mu) = \mathrm{Var}(X) \leq 1/4 \text{ in Kearns \& Saul (1998)},$$

while H.'s ineq. shows $X \sim \operatorname{subG}(1/4)$, and $\sigma^2_{opt}(X-1/2) = 1/4 = \operatorname{Var} X$.

Definition 1.6 (Buldygin and Kozachenko(2000), $\operatorname{ssubG}(\sigma^2_{opt}(X))$)

$X \sim \operatorname{subG}(\sigma^2_{opt}(X))$ is called **strictly sub-Gaussian** if $\operatorname{Var} X = \sigma^2_{opt}(X)$, and we redenote it as $X \sim \operatorname{ssubG}(\sigma^2_{opt}(X))$.

- $\operatorname{ssubG}(\sigma^2_{opt}(X))$ gives sharpest sub-G H.'s ineq. Gaussian; $U[-c,c]$; symmetric Beta, Bernoulli and binomial ($\operatorname{Bin}(n,\mu)$), triangular; see Arbel et al.(2020).

Kearns, M. & Saul, L. (1998). Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 311–319.

Arbel, J., Marchal, O., & Nguyen, H. D. (2020). On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. ESAIM: Probability and Statistics, 24, 39-55.

Comparison of optimal variance proxy and other norms

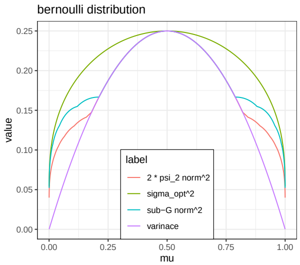


Figure: Bernoulli

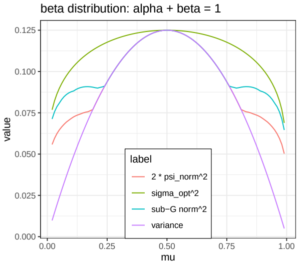


Figure: Beta(α_1, β_1)

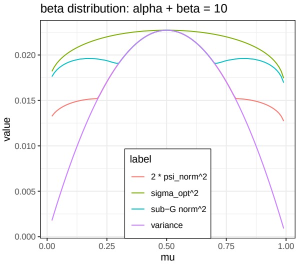


Figure: Beta(α_2, β_2)

Marchal & Arbel (2017) showed by a second order ODE (with a unique solution of the Cauchy problem)

$$\text{Beta}(\alpha, \beta) \text{ has } \sigma^2_{\text{opt}}(\alpha, \beta) = \frac{\alpha}{(\alpha+\beta)x_0}\left(\frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)} - 1\right) \geq \text{Var}[\text{Beta}(\alpha, \beta)],$$

where x_0 is a unique solution of $\log\left({}_1F_1\left(\alpha; \alpha+\beta; x_0\right)\right) = \frac{\alpha x_0}{2(\alpha+\beta)}\left(1 + \frac{{}_1F_1(\alpha+1;\alpha+\beta+1;x_0)}{{}_1F_1(\alpha;\alpha+\beta;x_0)}\right)$.

Marchal, O., & Arbel, J. (2017). On the sub-Gaussianity of the Beta and Dirichlet distributions. ECP, 22, 1-14.

Comparison of sub-Gaussian norms

1. **Variance upper bounds**: $\mathrm{norm}^2(X) \geq c \operatorname{Var} X$ for $\mathrm{E}X = 0$;
2. **Recover sharp tail inequality**: Derive exponential concentrations for single r.v.;
3. **Recover sharp MGF bounds**: Derive tight H- and B-type concentrations for sum of r.vs.;
4. **Easy estimations**: Plugging estimators are available.

Norms	Var upper bound	sharp Tail ineq.	sharp MGF bound	Easy estimation
$\sigma_{opt}^2(X)$	$\sigma_{opt}^2(X) \geq \operatorname{Var} X$	Yes	Yes	NO(exp-moments)
$\|X\|_{w_2}$	NO	Yes	NO	NO(exp-moments)
$\|X\|_{\psi_2}$	$2\|X\|_{\psi_2}^2 \geq \operatorname{Var} X$	NO	NO	Yes(High-moments)
$\|X\|_G$	$\|X\|_G^2 \geq \operatorname{Var} X$	Yes	Yes	Yes(High-moments)

- $\sigma_{opt}^2(X)$: $\mathrm{E}e^{tX} \leq e^{\sigma_{opt}^2(X)t^2/2}$. Empirical MGF (unstable).
- $\|X\|_{w_2}$: If $\|X\|_{w_2} = \sigma$ then $\mathrm{E}e^{tX} \leq e^{4\sigma^2 t^2/2}$. Empirical moments.
- $\|X\|_{\psi_2}$: $\mathrm{P}(|X| \geq t) \leq 2e^{-t^2/(2e \cdot 2\|X\|_{\psi_2}^2)}$. For $X = N(0,1)$, $\|X\|_{\psi_2} = \sqrt{2}$ and $\mathrm{P}(|X| \geq t) \leq 2e^{-t^2/(8e)}$. Empirical moments.
- $\|X\|_G$: $\mathrm{E}e^{tX} \leq e^{t^2\|X\|_G^2/2}$ if X is *sym. about 0*. Empirical moments.

Estimation of sub-exponential and sub-Gaussian norms

1. Related to GAN estimation for comparison two distributions.

Definition 1.7 (Sub-level parameter(Hitting time))

Let $k_{X,R} := \min_k \arg\max_{k\geq 2} [\frac{\mathrm{E}X^k}{r(k)}]^{1/k} - 1$ be **sub-level parameter** s.t.

$\|X\|_R \leq \max_{k>k_{X,R}+1} [\frac{\mathrm{E}X^k}{r(k)}]^{1/k}$, to evaluate the gap between X and R.

- For example, $k_{R,R} = 1$ by definition. We assume that $k_{X,R} < \infty$.
- Let $\bar{\mathrm{E}}_n f(X) := \frac{1}{n}\sum_{i=1}^n \mathrm{E}f(X_i)$ for non-i.i.d. data $\{X_i\}_{i=1}^n$. Motivated by GANs (Liang,2018), define **sub-R GAN problem and estimator** :

$$\|X\|_{n,R} = \arg\inf_{\sigma \geq \sqrt{\bar{\mathrm{E}}_n X^2/r(2)}} \max_{2\leq k\leq k_{X,R}} |(\frac{\bar{\mathrm{E}}_n X^k}{r(k)})^{\frac{1}{k}} - \sigma|$$

$$\widetilde{\|X\|}_R = \arg\inf_{\sigma \geq \sqrt{\bar{\mathrm{E}}_n X^2/r(2)}} \max_{2\leq k\leq k_{X,R}} |(\frac{\sum_{i=1}^n X_i^k}{nr(k)})^{\frac{1}{k}} - \sigma|.$$

2. Related to GMM for estimations from ℓ_2-norm loss to ℓ_∞-norm loss.

Liang, T. (2018).How well generative adversarial networks learn distributions. arXiv preprint arXiv:1811.03179.

$X \sim \text{subW}(\eta)$ is *sub-Weibull* if $\|X\|_{w_\eta} := \inf\{C > 0 : \mathrm{E}e^{|X|^\eta / C^\eta} \le 2\} < \infty$.

Theorem 1.8 (Oracle inequality for the sub-R GAN estimator)

Let $\{X_i\}_{i=1}^n$ *be ind.* $\text{subW}(\eta)$*-distributed with* $\max\limits_{2 \le k \le k_{X,R}} \max\limits_{i \in [n]} \|X_i\|_{\psi_{\eta/k}} < \infty$ *and* $k_{X,R} < \infty$, $\arg\max\limits_{2 \le k \le k_{X,R}} |(\frac{1}{nr(k)} \sum_{i=1}^n X_i^k)^{1/k} - \sigma| = \arg\max\limits_{2 \le k \le k_{X,R}} |\frac{1}{nr(k)} \sum_{i=1}^n X_i^k - \sigma^k|$.

Then with probability as least $1 - 2k_{X,R}e^{-t}$,

$$\max_{2 \le k \le k_{X,R}} \left| \widetilde{\|X\|}_R^k - \frac{\bar{\mathrm{E}}_n X^k}{r(k)} \right| \le \inf_{\sigma \in \Theta} \max_{2 \le k \le k_{X,R}} \left| \sigma^k - \frac{\bar{\mathrm{E}}_n X^k}{r(k)} \right|$$

$$+ (\frac{t}{n})^{\frac{1}{2}} \max_{2 \le k \le k_{X,R}} \frac{2eC(\eta/k)}{r(k)} (\frac{2}{n} \sum_{i=1}^n \|X_i\|_{w_{\eta/k}}^2)^{\frac{1}{2}} + (\frac{t}{n})^{\frac{k_{X,R}}{\eta}} \max_{2 \le k \le k_{X,R}} D_n(X, k, \eta),$$

where $D_n(X, k, \eta)$ *is function of* (k, η) *and data.*

- If bias $= 0$, $\max\limits_{2 \le k \le k_{X,R}} |\widetilde{\|X\|}_R^k - \frac{\bar{\mathrm{E}}_n X^k}{r(k)}| = \begin{cases} O((\frac{1}{n})^{\frac{k_{X,R}}{\eta}}), 0 < \frac{k_{X,R}}{\eta} \le 1/2 \\ O((\frac{1}{n})^{\frac{1}{2}}), \quad \frac{k_{X,R}}{\eta} > 1/2 \end{cases}.$

Median-of-mean estimators for $\|X\|_{n,E}$ and $\|X\|_{n,G}$

- Let m and b be positive integer s.t. $n = mb$ and let $B_1, \ldots, B_b$ be a partition of $[n]$ into subsets of equal cardinality m. For any $s \in [b]$, let $\mathbb{P}_m^{B_s} Y = m^{-1} \sum_{i \in B_s} Y_i$ for ind. $\{Y_i\}_{i=1}^n$.
- For ind. $\{X_i\}_{i=1}^n$, the MOM version sub-G and -E norms are

$$\widehat{\|X\|}_{n,G} : \max_{2 \le 2k \le 2k_{X,G}} \operatorname*{med}_{s \in [b]} \{[\tfrac{1}{(2k-1)!!} \cdot \mathbb{P}_m^{B_s} X^{2k}]^{1/(2k)}\} \text{ and}$$

$$\widehat{\|X\|}_{n,E} : \max_{2 \le k \le k_{X,E}} \operatorname*{med}_{s \in [b]} \{[\tfrac{1}{!k} \cdot \mathbb{P}_m^{B_s} X^{k}]^{1/k}\},\ k_{X,G} \ge 1, k_{X,E} \ge 2 \text{ (tuning)}$$

MOM estimators have **two merits**:

1. Finite moment-conditions, exponential concentration is still achieved;
2. It permits outliers and suitable non-i.i.d. data.

Theorem 1.9 (High-probability error bounds for estimated norms)

Let $o(1) \to 0$ *as* $m \to 0$. (a): *Under* $k_{X,G}/2$*th-moment conditions of data,*

$$\mathrm{P}\{\|X\|_{n,G} \le [1 - o(1)]^{-1} \widehat{\|X\|}_{n,G}\} > 1 - k_{X,G} e^{-b/8}/2;$$

and $\mathrm{P}\{\|X\|_{n,G} \ge [1 + o(1)]^{-1} \widehat{\|X\|}_{n,G}\} > 1 - k_{X,G} e^{-b/8}/2$.
(b): *Under* $k_{X,E}$*th-moment conditions of data,*

$$\mathrm{P}\{\|X\|_{n,E} \le [1 - o(1)]^{-1} \widehat{\|X\|}_{n,E}\} > 1 - k_{X,E} e^{-b/8}$$

and $\mathrm{P}\{\|X\|_{n,E} \ge [1 + o(1)]^{-1} \widehat{\|X\|}_{n,E}\} > 1 - k_{X,E} e^{-b/8}$.

Simulations:

- We adopt the empirical method (DE) using empirical moments $\frac{1}{n} \sum_{i=1}^{n} |X_i|^k$ to approximate $\mathrm{E}|X|^k$ directly, as well as the MOM method for comparison.
- $N(0, 1)$, centralized Bernoulli (successful $p = 0.3$), and $U(-1/2, 1/2)$ variable X; Centralized exponential ($\lambda = 1$), centralized chi-square ($\mathrm{df} = 2$), centralized negative binomial (NB,$\mu = 5$, $p = 1/2$) variables.

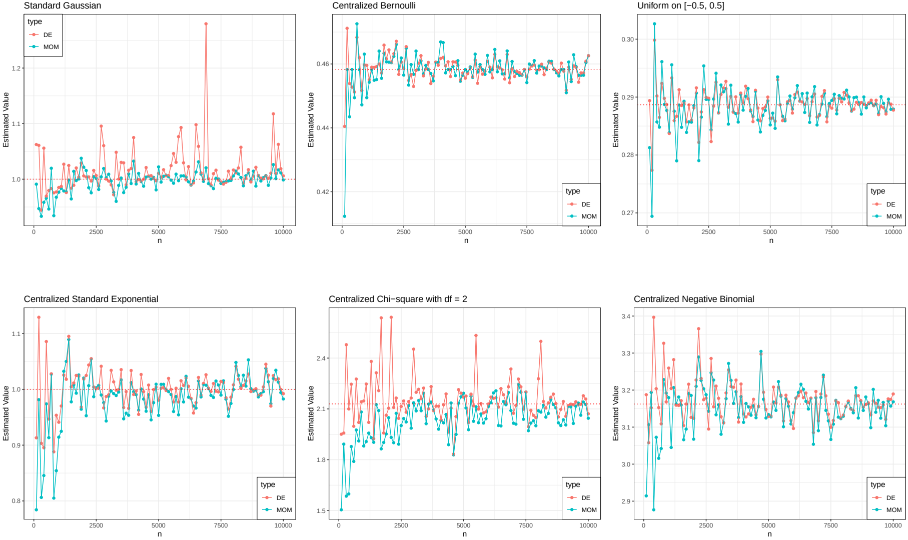


Figure: $\widehat{\|\cdot\|}_G$ and $\widehat{\|\cdot\|}_G$ variables by using MOM ($b = 20$) and DE. The red dot line in each figures represents the true value.

Extended McDiarmid's inequalities

Lemma 2.1 (McDiarmid's inequality)

Suppose $X_1, \cdots, X_n$ *are ind. r.vs in* $\mathcal{X}$, *and assume* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfies the* bounded difference condition *(Lip. property w.r.t. Hamming distance)*

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

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

- Applicable for **sup of bounded empirical processes** $f(X) := \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i)$.
- Given $f \in \mathcal{F}$, assume $\forall f \in \mathcal{F}, \mathrm{E}\left[f(X_i)\right] = 0$ and $f(X_i) \in [a_i, b_i]$. So $|\sup\limits_{f\in\mathcal{F}} \frac{1}{n}\sum\limits_{i=1}^n f(x_i) - \sup\limits_{f\in\mathcal{F}} \frac{1}{n}\sum\limits_{i=1}^n f(y_i)| \le \sum\limits_{i=1}^n \frac{b_i - a_i}{n} 1_{\{x_i \ne y_i\}}$. Then,

$$\mathrm{P}(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i) \le \mathrm{E}[\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i)] + \sqrt{n}u) \ge 1 - e^{-2(\sqrt{n}u)^2/[\frac{1}{n}\sum\limits_{i=1}^n (b_i-a_i)^2]}.$$

How about $f(X) \in$ supremum of unbounded EP?

Lemma 2.2 (Theorem 2.26, Wainwright(2019))

Let $X \sim N(0, \mathrm{I}_n)$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to (w.r.t.) the Euclidean norm: $|f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2$ for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$. Then,

$$\mathrm{P}\left(|f(X) - \mathrm{E}f(X)| \geq t\right) \leq 2e^{-t^2/(2L^2)}, \ \forall \ t > 0.$$

A function $\psi(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is γ-*strongly concave* if there is some $\gamma > 0$ s.t.

$$\lambda\psi(\boldsymbol{x}) + (1-\lambda)\psi(\boldsymbol{y}) - \psi(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \frac{\gamma}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \ \forall \ \lambda \in [0,1] \text{ and } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

A continuous density $p(\boldsymbol{x})$ is strongly log-concave if $\log f(\boldsymbol{x})$ is a strongly concave.

Lemma 2.3 (Theorem 3.16, Wainwright(2019))

Let $\mathbb{P}$ be any $\gamma-$strongly log-concave distribution on $\mathbb{R}^n$ with parameter $\gamma > 0$. Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz w.r.t. the Euclidean norm,

$$\mathrm{P}[f(X) - \mathrm{E}f(X) \geq t] \leq e^{-\gamma t^2/(4L^2)} \text{ for } X \sim \mathbb{P} \text{ and } t \geq 0.$$

The $\gamma-$strongly log-concave distribution is hard to check from data!

Concentration for the suprema of unbounded empirical processes

- Let $Z = (Z_1, \ldots, Z_n)$ be a vector of **independent r.vs** with values in a space $\mathcal{Z}$, define Z' as an independent copy of Z.
- For $w \in \mathcal{Z}$ and $k \in [n]$, define **substitution operator** $S_w^k : \mathcal{Z}^n \to \mathcal{Z}^n$ by

$$S_w^k(z) = (z_1, \ldots, z_{k-1}, w, z_{k+1}, \ldots, z_n)$$

and the centered conditional version of f as the r.v.

$$\begin{aligned} D_{f,Z_k}(z) &\equiv f\left(z_1, \ldots, z_{k-1}, Z_k, z_{k+1}, \ldots, z_n\right) - \mathbb{E}\left[f\left(z_1, \ldots, z_{k-1}, Z_k', z_{k+1}, \ldots, z_n\right)\right] \\ &= f\left(S_{Z_k}^k(z)\right) - \mathrm{E}[f(S_{Z_k'}^k(z))] = \mathbb{E}[f\left(S_{Z_k}^k(z)\right) - f(S_{Z_k'}^k(z)) | Z_k] \end{aligned} \tag{1}$$

where $D_{f,Z_k}(z)$ can be viewed as **random-valued functions** $z \in \mathcal{Z}^n \mapsto Y_{f,Z_k}(z)$.

- If $f(z) = \sum_{i=1}^n z_i$ then $D_{f,Z_k}(x) = Z_k - \mathrm{E}Z_k$ is independent of z.

Aim: **Concentrations for the suprema of unbounded EP.**

Proposition 2.4 (Theorems 3.1 and 3.2 in Maurer & Pontil(2021))

If $\{D_{f,Z_k}(z)\}_{i=1}^n$ have finite $\|\cdot\|_{\psi_2}$-norm [or finite $\|\cdot\|_{\psi_1}$-norm] for all $z \in \mathcal{Z}$ (called the stochastic bounded difference condition),then

$$\mathrm{P}\{f(Z) - \mathrm{E}f(Z) > t\} \leq \exp\{-(t^2/32e)/\sup_{z\in\mathcal{Z}} \sum_{k=1}^n \left\|D_{f,Z_k}(z)\right\|_{\psi_2}^2\}.$$

or $P\{f(Z) - \mathrm{E}f(Z) > (2e^2 \sup\limits_{z\in\mathcal{Z}} \sum_{i=1}^n \left\|D_{f,Z_i}(z)\right\|_{\psi_1}^2 t)^{1/2} + e\max\limits_{i\in[n]} \sup\limits_{z\in\mathcal{Z}} \left\|D_{f,Z_i}(z)\right\|_{\psi_1} t\} \leq 2e^{-t}.$

- Deep insights on thermodynamics and concentration, Maurer(2012). But constants are looser due to $\|Z\|_{\psi_1}$ and $\|Z\|_{\psi_2}$.

Corollary 2.5 (Shaper concentrations by sub-R norms)

If $\{D_{f,Z_i}(z)\}_{i=1}^n$ have finite $\|\cdot\|_G$-norm [or $\|\cdot\|_E$-norm] for all $z \in \mathcal{Z}$, we have $\forall t \geq 0$

$$\mathrm{P}\{f(Z) - \mathrm{E}f(Z) > t\} \leq \exp\{-(t^2/2)/\sup_{z\in\mathcal{Z}} \sum_{k=1}^n \left\|D_{f,Z_k}(z)\right\|_G^2\}.$$

or $\mathrm{P}\{f(Z) - \mathrm{E}f(Z) > (2\sup\limits_{z\in\mathcal{Z}} \sum_{i=1}^n \left\|D_{f,Z_i}(z)\right\|_E^2 t)^{1/2} + \max\limits_{i\in[n]} \sup\limits_{z\in\mathcal{Z}} \left\|D_{f,Z_i}(z)\right\|_E t\} \leq 2e^{-t}.$

Maurer, A.(2012). Bernoulli, 18(2):434–454.; Maurer, A., & Pontil, M. (2021). arXiv:2102.06304.; Lei, J. (2020). Bernoulli, 26(1):767–798. Convergence of empirical measures under Wasserstein distance in unbounded spaces.

Let $\{\boldsymbol{X}_i\}_{i=1}^n$ be ind. on $\mathbb{R}^d$. Jin et al.(2019) showed

$$\mathrm{P}(\|\sum_{i=1}^n \boldsymbol{X}_i\|_{\ell_2} \leq 2\sqrt{2}(\sum_{i=1}^n \|\|\boldsymbol{X}_i\|_{\ell_2}\|_{\psi_2}^2 \log(2d/\delta))^{1/2}) \geq 1-\delta,\ \delta \in (0,1). \tag{2}$$

By our sub-G norm, the following result sharper the factor $2\sqrt{2}$ in (2).

Theorem 2.6 (Hoeffding-type inequality for norm-subGaussian)

If $\{\boldsymbol{X}_i\}_{i=1}^n$ are zero-mean in $\mathbb{R}^d$ satisfying $\max_{i\in[n]}\|\|\boldsymbol{X}_i\|_{\ell_2}\|_G \leq \infty$, then

$$\mathrm{P}(\|\textstyle\sum_{i=1}^n \boldsymbol{X}_i\|_{\ell_2} \leq \sqrt{2}[\sum_{i=1}^n \|\|\boldsymbol{X}_i\|_{\ell_2}\|_G^2 \log(2d/\delta)]^{1/2}) \geq 1-\delta,\ \delta \in (0,1).$$

Example 2.7 (One-dimensional Gaussian variables)

For $X_i \sim N(0, v^2)$, we have $\sigma_i = v/\sqrt{2}$ in (2). Since $\|X_i\|_G = v$, the bound in Theorem 2.6 is

$$\sqrt{2}[\textstyle\sum_{i=1}^n \|X_i\|_G^2 \log(2d/\delta)]^{1/2} = \sqrt{2}nv(\log(2d/\delta))^{1/2}$$

with the sharper constant $\sqrt{2}$, comparing to the upper bound

$$2\sqrt{2}(\textstyle\sum_{i=1}^n \sigma_i^2 \log(2d/\delta))^{1/2} = 2nv(\log(2d/\delta))^{1/2} \text{ in (2) with looser constant 2.}$$

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). arXiv:1902.03736.

By sub-R norms and entropy method, the norm-concentration for $\mathcal{X}$-valued random vectors is obtained in bellow, without imposing a $\log d$ factor in Theorem 2.6.

Theorem 2.8

Suppose $\{\boldsymbol{X}_i\}_{i=1}^n$ are zero-mean ind. in a normed space $(\mathcal{X}, \|\cdot\|)$ s.t. $\max_{i\in[n]}\|\|\|\boldsymbol{X}_i\|\|\|_E < \infty$ or $\max_{i\in[n]}\|\|\|\boldsymbol{X}_i\|\|\|_G < \infty$. Let $\boldsymbol{X}_i'$ be an i.i.d. copy of $\boldsymbol{X}_i$. Then, with probability at least $1-\delta$

$$\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| - \mathrm{E}\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| \le \frac{2\sqrt{2}}{\sqrt{n}}[\frac{1}{n}\sum_{i=1}^n \|\|\|\boldsymbol{X}_i - \boldsymbol{X}_i'\|\|\|_G^2 \log(\tfrac{1}{\delta})]^{1/2}\} \text{ or}$$

$$\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| - \mathrm{E}\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| \le$$
$$\frac{1}{\sqrt{n}}[2\sum_{i=1}^n \frac{\|\|\|\boldsymbol{X}_i - \boldsymbol{X}_i'\|\|\|_E^2}{n}\log(\tfrac{1}{\delta})]^{1/2} + \frac{\max_{i\in[n]}\|\|\|\boldsymbol{X}_i - \boldsymbol{X}_i'\|\|\|_E}{n}\log(\tfrac{1}{\delta}).$$

Moreover, if $\mathcal{X}$ is a Hilbert space and $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d., we have with probability at least $1-\delta$,

$$\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| \le \frac{1}{\sqrt{n}}\{[\mathrm{E}\|\boldsymbol{X}_1 - \mathrm{E}\boldsymbol{X}_1\|^2]^{1/2} + 2\sqrt{2}[\|\|\|\boldsymbol{X}_1 - \boldsymbol{X}_1'\|\|\|_G^2\log(\tfrac{1}{\delta})]^{1/2}\} \text{ or}$$

$$\Big\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{X}_i - \mathrm{E}\boldsymbol{X}_i)\Big\| \le$$
$$\frac{1}{\sqrt{n}}\{[\mathrm{E}\|\boldsymbol{X}_1 - \mathrm{E}\boldsymbol{X}_1\|^2]^{1/2} + [2\|\|\|\boldsymbol{X}_1 - \boldsymbol{X}_1'\|\|\|_E^2\log(\tfrac{1}{\delta})]^{1/2}\} + \frac{\|\|\|\boldsymbol{X}_1 - \boldsymbol{X}_1'\|\|\|_E}{n}\log(\tfrac{1}{\delta}).$$

Median-of-mean estimators for $\|\|\|\boldsymbol{X}_1 - \boldsymbol{X}_1'\|\|\|_R$

- For data with outliers, $\|\|\|\boldsymbol{X}_1 - \boldsymbol{X}_1'\|\|\|_R$, ($R = G$ or E) in Theorem 2.8 is substituted by MOM U-statistics estimator in Joly & Lugosi (2016).

- Let $B_1, \ldots, B_b$ be the partition of $[n]$ as previous such that $n = mb$, then the U-statistic of $\left\|\left\|\|\boldsymbol{X} - \boldsymbol{X}'\|\right\|\right\|_E$ and $\left\|\left\|\|\boldsymbol{X} - \boldsymbol{X}'\|\right\|\right\|_G$ coming from block B_s and B_t are

$$U_{\theta_1, B_s, B_t}(\|\boldsymbol{X} - \boldsymbol{X}'\|; k) = \left[\frac{1}{!k}\frac{2}{m(m-1)}\sum_{i\in B_s,\, j\in B_t,\, i<j}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^k\right]^{1/k},$$

$$U_{\theta_2, B_s, B_t}(\|\boldsymbol{X} - \boldsymbol{X}'\|; k) = \left[\frac{1}{(2k-1)!!}\frac{2}{m(m-1)}\sum_{i\in B_s,\, j\in B_t,\, i<j}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^{2k}\right]^{1/(2k)},$$

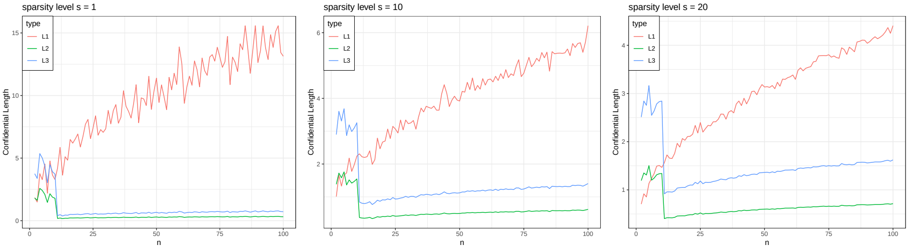
$$\left\|\left\|\|\widehat{\boldsymbol{X} - \boldsymbol{X}'}\|\right\|\right\|_{n,E} := \max_{2\le k\le k(\theta_1)} \operatorname*{med}_{s,t\in[b]} U_{\theta_1, B_s, B_t}(\|\boldsymbol{X} - \boldsymbol{X}'\|; k),$$

$$\left\|\left\|\|\widehat{\boldsymbol{X} - \boldsymbol{X}'}\|\right\|\right\|_{n,G} := \max_{1\le k\le k(\theta_2)} \operatorname*{med}_{s,t\in[b]} U_{\theta_2, B_s, B_t}(\|\boldsymbol{X} - \boldsymbol{X}'\|; k) \text{ respectively.}$$

Joly, E. and Lugosi, G.(2016). Stochastic Processes and their Applications, 126(12):3760–3773.

- Example under high-dimensional X to show that the simulation performance of CIs by (2), our **Theorems 3.7 and 3.9** under the sub-Gaussian norm.
- Suppose the i.i.d. Gaussian observation $\{X_i\}_{i=1}^{n}$ are $d = 2000$ dimensional vectors with sparsity level $s = 1, 10, 20$, i.e.

$$X_i(s) = \big(\underbrace{N(0, 1/s), \ldots, N(0, 1/s)}_{\text{s times}}, 0, \ldots, 0\big)^{\top} \in \mathbb{R}^d.$$



The confidential length when $\delta = 0.05$ under (2) (denoted as $\mathrm{L1}$), Theorem 3.8 (denoted as $\mathrm{L2}$), and Theorem 3.9 (denoted as $\mathrm{L3}$) is shown as sparsity level s varies.

4.1 The UCB in Bandit Problem

The sequential decisions: a player in a casino, choosing between K different slot machines (a K-armed bandit), each with a different unknown reward r.vs.

$$\{Y_k\}_{k=1}^K \in \mathbb{R} \text{ (may be unbounded, non-Gaussian or negative).}$$

- **Don't know prior information** in the casino, assume

$$\|Y_k - \mu_k\|_R \le C < \infty, \ (R = G \text{ or } E), \ k \in [K]. \tag{3}$$

Aim: Find the index $t^\star$ with the optimal mean reward $\mu_{t^\star} = \max_{k\in[K]} \mu_k$

without frequently choosing sub-optimal arms (i.e.,reward r.vs.$\{Y_k\}_{k\neq t^\star}$).

- **A dilemma** is collecting new information by *exploring sub-optimal arms* (exploration) and selecting the best action (exploitation) on collected data.
- **Enough money** in round $t \in [T]$, the player pull an arm $A_t \in [K]$. Conditioning on the action $\{A_t\}_{t\in[T]}$, the observed reward $Y_{A_t} \sim P_{A_t}$ ind.
- **The criterion** for $\{A_t\}_{t\in[T]}$ is to minimize the *cumulative regret*

$$\mathrm{Reg}_T(Y, A) := \sum_{t=1}^T (\mu_{t^\star} - \mu_{A_t}).$$

The Upper Confidence Bound from concentration

Let **the number of pull for arm k until time t**:

$$T_k(t) := \text{card}\{1 \le \tau \le t : A_\tau = k\} \text{ during the bandit process.}$$

We define **the running average of the rewards of arm k at time t**:

$$\overline{Y}_{T_k}(t) := \frac{1}{T_k(t)} \sum_{\tau \le t, A_\tau = k} Y_k(\tau) \text{ (those instances arm } k \text{ was selected)}$$

- If we have **tight concentration-based CIs**

$$\left[\overline{Y}_{T_k(t)} - c_k(t), \overline{Y}_{T_k(t)} + c_k(t)\right] \text{ (length } 2c_k(t) \text{ decreases with time } t)$$

- Bootstrapped threshold $\widehat{c}_k(t)$ by Hao B. et al. (2019), and

we confidently reckon that the reward of arm k is $\overline{Y}_{T_k(t)} + \widehat{c}_k(t)$.

Regret $\text{Reg}_T(Y, A)$ achieves minimax rate $O(\sqrt{KT})$ up to a log T.

Hao, B., Yadkori, Y. A., Wen, Z., and Cheng, G. (2019). Bootstrapping upper confidence bound. Advances in Neural Information Processing Systems, volume 32, pages 12123–12133.

Bootstrapped UCB with estimated sub-R norms

For any reward vector $\mathrm{Z}_{T_k(t)} = (Z_1, \ldots, Z_{T_k(t)})^\top$ with i.i.d. $\{Z_k\}_{k=1}^{T_k(t)}$ as the observation of a fixed arm, define $\widehat{\varphi}_R(\mathrm{Z}_{T_k(t)})$ satisfying $\mathrm{P}_{\mathrm{Z}_n}(|\overline{Z}_{T_k(t)} - \mathrm{E}Z| \ge \widehat{\varphi}_R(\mathrm{Z}_{T_k(t)})) \le \alpha$ with

$$\widehat{\varphi}_R(\mathrm{Z}_{T_k(t)}) := \sqrt{\frac{2\log(4/\alpha)}{T_k(t)}} \frac{\|\widehat{Z_1 - \mathrm{E}Z_1}\|_R}{1 - T_k^{-1/2}(t)} \quad \text{or} \quad \left[\sqrt{\frac{2}{T_k(t)}\log(\frac{4}{\alpha})} + \frac{2}{T_k(t)}\log(\frac{4}{\alpha})\right] \frac{\|\widehat{Y - \mu}\|_E}{1 - T_k^{-1/2}(t)}.$$

Algorithm 1: Bootstrapped UCB

Input: the number of bootstrap repetitions B; $R = G$ or R.

for $t = 1, \ldots, K$ **do**

Pull each arm once to initialize the algorithm.

end

for $t = K+1, \ldots, T$ **do**

Set confidence level $\alpha \in (0, 1)$.

Calculate the boostrapped quantile $q_{\alpha/2}(\mathrm{Y}_{T_k(t)} - \overline{Y}_{T_k(t)}, \mathrm{w}^B)$ with the bootstrapped Rademacher weights w^B independent with any Y.

Pull the arm

$$A_t = \operatorname*{argmax}_{k \in [K]} \mathrm{UCB}_k(t) := \operatorname*{argmax}_{k \in [K]} \{\overline{Y}_{T_k(t)} + q_{\alpha/2}(\mathrm{Y}_{T_k(t)} - \overline{Y}_{T_k(t)}, \mathrm{w}^B) + \widehat{\varphi}_R(\mathrm{Y}_{T_k(t)})\}.$$

Receive reward Y_{A_t}.

end

Consider a stochastic K-armed sub-Gaussian or -exponential bandit under (3).

Theorem 3.1

For any round T, according to Theorem 1.9, choosing $\widehat{\varphi}_R(\mathrm{Y}_{T_k(t)})$ by re-scaled MOM estimator with blocked sample size:

$m_2 = \inf\{m \in \mathbb{N}_+ : \max_{1\le\kappa\le k_{Y_k,R}/2} \overline{g}_{\kappa,m}(\sigma_\kappa) \wedge \underline{g}_{\kappa,m}(\sigma_\kappa) \ge 1/\sqrt{T_k(t)}\}$ *for* $R = G$,

$m_1 = \inf\{m \in \mathbb{N}_+ : \max_{1\le\kappa\le k_{Y_k,R}} \overline{h}_{\kappa,m}(\sigma_\kappa) \wedge \underline{h}_{\kappa,m}(\sigma_\kappa) \ge 1/\sqrt{T_k(t)}\}$ *for* $R = E$,

and block number $b \ge 8\log(T^2 k_{Y_k,G}/4)$ *or* $b \ge 8\log(T^2 k_{Y_k,E}/2)$.

Let μ_1^* be the optimal mean reward. Fix a confidence level $\alpha = 4/T^2$, then:

(i) *for $R = G$, the problem-independent regret*

$$\mathrm{Reg}_T \le 8(2+\sqrt{2})C\sqrt{TK\log T} + (4T^{-1} + 2T^{-25-16\sqrt{2}} + 8)K\mu_1^*.$$

(ii) *for $R = E$, the problem-independent regrets of our method is*

$$\mathrm{Reg}_T \le 8(3+2\sqrt{2})C\sqrt{2KT\log T} + 16C^2\log T + \left(\frac{4}{T} + \frac{1}{T^{(4(2+\sqrt{2})^2C-1)\vee 1}} + 8\right)K\mu_1^*.$$

Optimal algorithm:

Regret bounds achieve minimax rate $O(\sqrt{KT})$ up to a log T.

Simulation for 3 Bootstrapping UCBs

- The machine has $K = 5$ arms with each arm give a Gaussian reward, whose mean is $0.1, 0.05, 0.02, 0.01, 0.01$ and standard variance is $0.1^2, 0.05^2, 0.02^2, 0.01^2, 0.01^2$.
- (a). $\widehat{c}_k(t)$ (denoted by *Estimated Norm*);(b) the asymptotic estimated $c_k(t)$ by CLT; (c) Algorithm 1 the non-asymptotic estimated $c_k(t)$ (use bounded Hoeffding's inequality for Gaussian r.v.). Average over 200 replications.

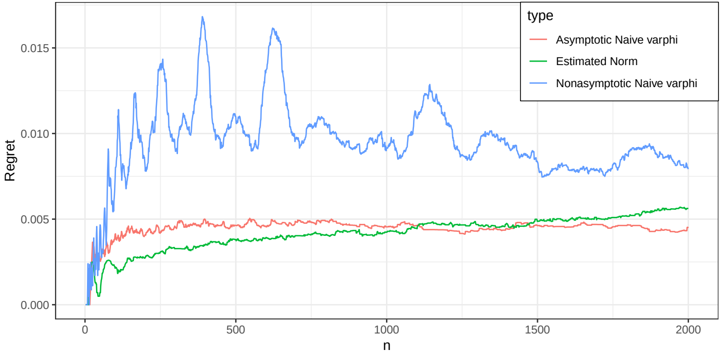


Figure: Regret under three methods.

4.2 Bounds of Excess risk in deep networks

Consider the feedforward NN as the compositional function class indexed by parameter β: $f(x;\beta) : \mathbb{R}^p \to \mathbb{R}$ for the function of p-dimensional data x

$$\mathcal{NN}(N, L) := \left\{ f(x;\beta) = W_L\sigma_L\left(W_{L-1}\sigma_{L-1}\left(\cdots W_1\sigma_1\left(W_0x\right)\right)\right) \in \mathbb{R} \,|\, \beta := (W_0, \ldots, W_L) \right\}.$$

where $W_l \in \mathbb{R}^{N_{l+1}\times N_l}$ for $l = 0, 1, \ldots, L$ with $N_0 = p$, $\{\sigma_j\}_{j=1}^{L} : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$, and width $N = \max\{N_1, \ldots, N_L\}$.

- For independent $\{(X_i, Y_i)\}_{i=1}^{n}$ and loss function $l(\cdot, \cdot)$, true β_n^* satisfies

$$f^* := f(x;\beta^*) = \underset{f \text{ is measurable}}{\operatorname{argmin}} R_l^n(f) \text{ with } R_l^n(f) := \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} l\left(Y_i, f(X_i;\beta)\right)\right].$$

- Define the DNNs approximated parameter $\beta_{\mathcal{N}}^*$

$$f_{\mathcal{N}}^* := f(x;\beta_{\mathcal{N}}^*) = \underset{f\in\mathcal{NN}(N,L),\beta\in\Theta}{\operatorname{argmin}} R_l^n(f), \tag{4}$$

where the function class $\mathcal{NN}(N, L)$ consists of feedforward NN under Θ.

Estimate f^* by ERM estimator $\hat{f}_{\mathcal{N}}$ under $\mathcal{N}(N, L)$, i.e.

$$\hat{f}_{\mathcal{N}} := f(x; \hat{\beta}) = \underset{f \in \mathcal{NN}(W,L), \beta \in \Theta}{\operatorname{argmin}} \hat{R}_l^n(f), \text{ with } \hat{R}_l^n(f) := \frac{1}{n} \sum_{i=1}^{n} l\big(Y_i, f(X_i; \beta)\big). \quad (5)$$

- **(N.1)**: The parameter set:

$$\beta \in \Theta_2 := \{\gamma := (W_0, \ldots, W_L) : \max_{0 \le l \le L} \sigma_{\max}(W_l) \le M\} \cap \{\|\gamma - \beta_n^*\|_{\mathrm{F}} \le K\} \subseteq \mathbb{R}^{\sum_{l=0}^{L} N_{l+1} N_l},$$

for $M > 0$, where $\|\gamma\|_{\mathrm{F}} := \sqrt{\sum_{l=0}^{L} \|W^l\|_{\mathrm{F}}^2}$ with $\left\|W^l\right\|_{\mathrm{F}} := \sqrt{\sum_{k=1}^{N_{l+1}} \sum_{j=1}^{N_l} (W_{kj}^l)^2}$.

- **(N.2)**: The loss $l(y, f(x; \beta))$ has Lipschitz function $L(x, y)$

$$|l(y, f(x; \beta_1)) - l(y, f(x; \beta_2))| \le L(x, y)|f(x; \beta_2) - f(x; \beta_1)|, \ \beta_1, \ \beta_2 \in \Theta.$$

(N.2b): Bounded Lipschitz constant $\max_{1 \le i \le n} |L(x, y)| \le B_f$ for a constant B_f.

(N.2g): $\max_{1 \le i \le n} \|L(X_i, Y_i)\|_G < \infty$ and $\max_{i \in [n]} \mathrm{E}[L^2(X_i, Y_i)\|X_i\|_{\ell_2}^2] < \infty$.

Main results

Theorem 3.2 (Excess risk bounds in deep networks)

Let $\varepsilon_{\mathcal{N}\mathcal{N}} = \inf_{f \in \mathcal{N}\mathcal{N}(N,L), \beta \in \Theta} |R_l^n(f_{\mathcal{N}}^*) - R_l^n(f^*)|$. *For estimator* $\hat{f}_{\mathcal{N}}$ *in* (5) *and* $f_{\mathcal{N}}^*$ *in* (4) *with* $\Theta = \Theta_2$ *and* $\{\sigma_k\}_{k=1}^L$ *is the* 1*-Lipschitz positive homogeneous activation.* (a). *Under (N.1),(N.2b) and (L.3), with probability at least* $1-\delta$, *one has*

$$R_l^n(\hat{f}_{\mathcal{N}}) - R_l^n(f^*) - \varepsilon_{\mathcal{N}\mathcal{N}} \le 2KB_f M^L \left(\sqrt{\tfrac{32L}{n} \log(\tfrac{1}{\delta}) \max_{i \in [n]} \left\| \|X_k\|_{\ell_2} \right\|_G^2} + \sqrt{\tfrac{L}{n} \mathrm{E} \max_{i \in [n]} \|X_k\|_{\ell_2}^2} \right);$$

(b). *Under (N.1),(N.2g) and (L.3), with probability at least* $1-\delta$,

$$\begin{aligned} R_l^n(\hat{f}_{\mathcal{N}}) - R_l^n(f^*) - \varepsilon_{\mathcal{N}\mathcal{N}} \le\; & 4KM^L \max_{i \in [n]} \|L(X_i, Y_i)\|_G \left\| \|X_i\|_{\ell_2} \right\|_G [\sqrt{\frac{2L}{n} \log(\frac{1}{\delta})} \\ & + \frac{\sqrt{L}}{n} \log(\frac{1}{\delta})] + 2KM^L \sqrt{\frac{L}{n} \mathrm{E}[\max_{i \in [n]} L^2(X_i, Y_i) \|X_k\|_{\ell_2}^2]}, \end{aligned}$$

where K, M *are constants given in (N.1), and* B_f *is defined in (N.2b).*

Consider max-F norm parameter space in Golowich's et al. (2020):

$$\Theta_1 := \{\beta := (W_0, \dots, W_L) | \max_{0 \le l \le L} \|W_l\|_F \le B\}.$$

Corollary 3.3

If $\{X_i\}_{i=1}^n$ are independent with $\max_{i \in [n]} \big\| \|X_i\|_{\ell_2} \big\|_G < \infty$ and $\{Y_i\}_{i=1}^n$ are bounded. For estimator $\hat{f}_{\mathcal{N}}$ in (5) and $f^*_{\mathcal{N}}$ in (4) with $\Theta = \Theta_1$ and $\{\sigma_k\}_{k=1}^L$ is the 1-Lipschitz positive homogeneous activation. (a) *For $\mathcal{F} = \mathcal{NN}(N, L)$, we have*

$$\mathrm{Rad}(\mathcal{F}|S) \le \frac{4B^L\sqrt{L\log 2}}{\sqrt{n}}[\tfrac{1}{n}\textstyle\sum_{i=1}^n \big\|\|X_i\|_{\ell_2}\big\|_G^2]^{1/2} + \frac{B^L}{\sqrt{n}}\sqrt{\textstyle\sum_{i=1}^p(\tfrac{1}{n}\sum_{j=1}^n \mathrm{E}X_{ij}^2)}.$$

(b) *Let $l_{f^*}(Z_i) := l\big(Y_i, f(X_i; \beta^*)\big)$. If data $\{(Y_i, X_i)\}_{i=1}^n$ is i.i.d. and $l(y, \mathcal{F})$ is L_f-Lipschitz as $|l(y, f(x;\beta)) - l(y, f(x;\beta')))| \le L_f |f(x;\beta) - f(x;\beta'))|$, then $R_l^n(\hat{f}_{\mathcal{N}}) - R_l^n(f^*) - \varepsilon_{\mathcal{NN}}$ has following bounds with probability at least $1-\delta$,*

$$\frac{4B^L}{\sqrt{n}}\left[4\sqrt{L\log 2}[\frac{1}{n}\sum_{i=1}^n \big\|\|X_i\|_{\ell_2}\big\|_G^2]^{1/2} + [\sum_{i=1}^p\sum_{j=1}^n \frac{\mathrm{E}X_{ij}^2}{n}]^{1/2}\right] + 8L_f \max_{i\in[n]}\sup_{f\in\mathcal{F}} \|f(X_i)\|_G \sqrt{\frac{2\log(1/\delta)}{n}}.$$

If $B := \max_{0 \le l \le L} \|W_l\|_F$ in Θ_1 and $M := \max_{0 \le l \le L} \sigma_{\max}(W_l)$ in Θ_2, then

M^L in Theorem 3.2 is smaller than B^L in Corollary 3.3 since $M < B$.

Golowich, N., Rakhlin, A., and Shamir, O. (2020). Information and Inference: A Journal of the IMA, 9(2):473–504.

Thanks.