Randomized Concentration Inequalities

Qi-Man Shao

Southern University of Science and Technology

Outline

1. Introduction

Let $\xi_1, \xi_2, \cdots, X_n$ be independent random variables with $E\xi_i = 0$ and

$$\sum_{i=1}^{n} E\xi_i^2 = 1.$$

Let $W_n = \sum_{i=1}^{n} \xi_i$. Consider the non-linear statistic

$$T_n = W_n + \Delta,$$

where $\Delta = \Delta(\xi_i, 1 \leq i \leq n)$.

Assume that $\Delta \to 0$ in probability. Then

$$T_n \xrightarrow{d.} Z,$$

provided the Lindeberg condition is satisfied, where Z is the standard normal random variable.

- Question: What is the error of the approximation?

$$\sup_x |P(T_n \geq x) - P(Z \geq x)| = ?$$

and

$$\frac{P(T_n \geq x)}{P(Z \geq x)} = 1 + ??$$

- The error of normal approximation for W_n is well understood. Can we establish the error of

$$P(T_n \geq x) - P(W_n \geq x)?$$

- If $E|\Delta|^p < \infty$, $p > 0$, then it is easy to see that

$$\begin{aligned} &\sup_x |P(T_n \geq x) - P(Z \geq x)| \\ &\leq \sup_x |P(W_n \geq x) - P(Z \geq x)| + 2\left(E|\Delta|^p\right)^{1/(1+p)}. \end{aligned}$$

- Remark: The bound is best possible.

- Observe that

$$P(T_n \geq x) - P(W_n \geq x) \leq P(x - \Delta \leq W_n < x),$$

$$P(T_n \geq x) - P(W_n \geq x) \geq -P(x \leq W_n < x - \Delta).$$

- Aim: Establish Randomized Concentration Inequality

$$P(\Delta_1 \leq W_n \leq \Delta_2),$$

where Δ_1 and Δ_2 are measurable functions of $\{\xi_i, 1 \leq i \leq n\}$.

2. Randomized Concentration Inequalities

Recall $\xi_i,\ 1 \le i \le n$ are independent random variables with $E\xi_i = 0$ and $\sum_{i=1}^n E\xi_i^2 = 1$. Let $W = \sum_{i=1}^n \xi_i$, Δ_1 and Δ_2 be measurable functions of $\{\xi_j, 1 \le j \le n\}$.

- Chen and Shao (2007):

$$\begin{aligned} & P(\Delta_1 \le W \le \Delta_2) \\ \le\ & 2\sum_{i=1}^n E|\xi_i|^3 + E|W(\Delta_2 - \Delta_1)| \\ & + \sum_{i=1}^n E|\xi_i(\Delta_1 - \Delta_1^{(i)})| + \sum_{i=1}^n E|\xi_i\,(\Delta_2 - \Delta_2^{(i)})|, \end{aligned}$$

where $\Delta_1^{(i)}$ and $\Delta_2^{(i)}$ are measurable functions of $\{\xi_j, j \neq i\}$.

The term $E|W(\Delta_2 - \Delta_1)|$ can be replaced by $E|\Delta_2 - \Delta_1|$, which makes it possible to establish a sharp Cramér type moderate deviation for self-normalized non-linear statistics in Shao and Zhou (2016).

- Shao and Winxin Zhou (2016):

$$\begin{aligned} & P(\Delta_1 \le W \le \Delta_2) \\ \le \quad & 21 \sum_{i=1}^{n} E|\xi_i|^3 + 6\, E|\Delta_2 - \Delta_1| \\ & + 4 \sum_{i=1}^{n} E|\xi_i(\Delta_1 - \Delta_1^{(i)})| + 4 \sum_{i=1}^{n} E|\xi_i\, (\Delta_2 - \Delta_2^{(i)})|, \end{aligned}$$

where $\Delta_1^{(i)}$ and $\Delta_2^{(i)}$ are measurable functions of $\{\xi_j, j \neq i\}$.

- Remark: In the above inequalities, it is presumed that $\Delta_1 \le \Delta_2$. Recall that our original aim is to bound

$$P(W \le \Delta_2) - P(W \le \Delta_1).$$

When $E\Delta_1 = E\Delta_2$, one would expect to have a better bound.

▶ A refined randomized concentration inequality

- Lei and Shao (2021):

$$|P(W \le \Delta) - E\Phi(\Delta)| \le 300 \sum_{i=1}^{n} E|\xi_i|^3 + 4 \sum_{i=1}^{n} E|\xi_i(\Delta - \Delta^{(i)})|$$
$$+11 \sum_{i=1}^{n} E\xi_i^2 E|\Delta - \Delta^{(i)}| + 25 \sum_{i=1}^{n} \sum_{j=1}^{n} E\xi_j^2 E|\xi_i(\Delta - \Delta^{(j)})|,$$

where $\Delta^{(i)}$ is any function of $\{\xi_j, j \ne i\}$.

- It's not clear if the red part could be removed.

▶ A randomized exponential concentration inequality

- Shao (2010): Let $\gamma = \sum_{i=1}^n E|\xi_i|^3$. Then for $\lambda \geq 0$,

$$
\begin{aligned}
& Ee^{\lambda(W+\Delta)} I(\Delta_1 \leq W + \Delta \leq \Delta_2) \\
\leq \quad & (Ee^{2\lambda(W+\Delta)})^{1/2} \exp(-\frac{1}{64\gamma^2}) \\
& +4e^{\lambda\delta} \Big\{ Ee^{\lambda(W+\Delta)} |W| (|\Delta_2 - \Delta_1| + 2\gamma) \\
& +2 \sum_{i=1}^n Ee^{\lambda(W^{(i)}+\Delta^{(i)})} |\xi_i| (|\Delta_1 - \Delta_1^{(i)}| + |\Delta_2 - \Delta_2^{(i)}|) \\
& + \sum_{i=1}^n E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|)(3 + \lambda(|\Delta_2 - \Delta_1| + 2\gamma) \\
& \qquad \max\left(e^{\lambda(W+\Delta)}, e^{\lambda(W^{(i)}+\Delta^{(i)})}\right) \Big\}
\end{aligned}
$$

where $W^{(i)} = W - \xi_i$.

In particular, we have

- For $\lambda = 0$,

$$
\begin{aligned}
& P(\Delta_1 \le W + \Delta \le \Delta_2) \\
& \le \quad 64\gamma + E|W||\Delta_2 - \Delta_1| \\
& \qquad +2\sum_{i=1}^{n} E|\xi_i|(|\Delta_1 - \Delta_1^{(i)})| + |\Delta_2 - \Delta_2^{(i)}|) \\
& \qquad +3\sum_{i=1}^{n} E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|).
\end{aligned}
$$

- For $\Delta_1 \ge x$ and $\lambda = x$, we could establish an exponential inequality for $P(\Delta_1 \le W + \Delta \le \Delta_2)$.

- The proof is based on the Stein method. The inequality may provide a useful tool to prove the Cramér type moderate deviation for Studentized statistics.

▶ An application to U-statistic

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed random variables, and let $h(x, y)$ be a real-valued Borel measurable symmetric function, i.e., $h(x, y) = h(y, x)$. Define the U-statistic with the kernel h by

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j).$$

Let $g(x) = Eh(x, X_2)$ and $\sigma_1^2 = Eg^2(X_1)$. Assume $\sigma_1 > 0$. It is known that

$$\frac{\sqrt{n} U_n}{2\sigma_1} = W + \Delta,$$

where $W = \frac{1}{\sqrt{n}\sigma_1} \sum_{i=1}^{n} g(X_i)$,

$$\Delta = \frac{\sqrt{n}}{n(n-1)\sigma_1} \sum_{1 \le i < j \le n} \{h(X_i, X_j) - g(X_i) - g(X_j)\}.$$

Let

$$\Delta^{(l)} = \frac{\sqrt{n}}{n(n-1)\sigma_1} \sum_{1 \le i < j \le n, i \ne l, j \ne l} \{h(X_i, X_j) - g(X_i) - g(X_j)\}.$$

One can prove that

$$E\Delta^2 \le \frac{\sigma^2}{2(n-1)\sigma_1^2}$$

and

$$E|\Delta - \Delta^{(l)}|^2 \le \frac{\sigma^2}{n(n-1)\sigma_1^2}.$$

Therefore, we have

- Assume that $Eh(X_1, X_2) = 0$ and $\sigma^2 = Eh^2(X_1, X_2) < \infty$. Then

$$\sup_z |P(\frac{\sqrt{n}U_n}{2\sigma_1} \leq z) - \Phi(z)| \leq \frac{2\sigma}{(n-1)^{1/2}\sigma_1} + \frac{9E|g(X_1)|^3}{n^{1/2}\sigma_1^3}.$$

- Applications: Multisample U-statistics, L-statistics, Random sums, functional of non-linear statistics, the Cramér type moderate deviation for self-normalized non-linear statistics, ...

Let $\{\xi_i : 1 \le i \le n\}$ be a family of R^d-valued independent random vectors satisfying $E\xi_i = 0$ and $\sum_{i=1}^n E\xi_i\xi_i^T = I_d$. Let

$$W = \sum_{i=1}^n \xi_i.$$

For any convex set $A \subset R^d$ and any $\epsilon > 0$, let

$$A^\epsilon = \{y \in R^d : \|y - x\| \le \epsilon, x \in A\}.$$

▶ A concentration inequality

Let

$$\gamma = \sum_{i=1}^{n} E\|\xi_i\|^3.$$

- Chen and Fang (2011): For any $\epsilon > 0$,

$$P(W \in A^{4\gamma+\epsilon} \setminus A^{4\gamma}) \leq C d^{1/2}(\epsilon + \gamma).$$

▶ A randomized concentration inequality

- Shao and Zhang (2021)

  Let Δ be a nonnegative random variable. Then for any convex set $A \subset R^d$,

$$
\begin{aligned}
& P\big(W \in A^{4\gamma+\Delta} \setminus A^{4\gamma}\big) \\
\leq \quad & 19 d^{1/2}\gamma + 2E\big\{\|W\|\Delta\big\} + 2\sum_{i=1}^{n} E\{\|\xi_i\||\Delta - \Delta^{(i)}|\},
\end{aligned}
$$

  where $\Delta^{(i)}$ is any random variable independent of ξ_i.

- Remark: It would be interesting if $d^{1/2}$ can be replaced by $d^{1/4}$. and $E\big\{\|W\|\Delta\big\}$ by $E|\Delta|$.

4. Berry–Esseen Bounds for Multivariate Nonlinear Statistics

Let $\xi_1, \ldots, \xi_n$ be R^d-valued random element satisfying $E\xi_i = 0$ and $\sum_{i=1}^n E\xi_i\xi_i^T = I_d$ and let $W = \sum_{i=1}^n \xi_i$. Let $\mathcal{C}$ be the class of convex sets in R^d.

- Bentkus (1986):

$$\sup_{A\in\mathcal{C}} \big| P(W \in A) - P(Z_{0,I_d} \in A) \big| \leq C d^{1/4} \sum_{i=1}^n E\|\xi_i\|^3,$$

  where $Z_{\mu,\Sigma} \sim N(\mu, \Sigma)$ and C is an absolute constant.

- Remark: It is believed that the above bound is best possible.

Let T be a non-linear statistic

$$T = W + D, \quad \text{where} \quad D = D(\xi_1, \ldots, \xi_n).$$

- Shao and Zhang (2021):

$$\sup_{A \in \mathcal{C}} \big| P(T \in A) - P(Z \in A) \big|$$

$$\leq \quad 259 d^{1/2} \gamma + 2E\big\{ \|W\| \Delta \big\} + 2 \sum_{i=1}^{n} E\big\{ \|\xi_i\| |\Delta - \Delta^{(i)}| \big\},$$

  for any random variables Δ and $(\Delta^{(i)})_{1 \leq i \leq n}$ such that $\Delta \geq \|D\|$ and $\Delta^{(i)}$ is independent of ξ_i.

- Remark: It seems challenging to replace $d^{1/2}$ by $d^{1/4}$.
- The result provides a convergence rate of order $O(n^{-1/2})$ for a wide class of non-linear statistics.

▶ Stochastic Gradient Decent Algorithms (SGD)

Let $f : \Theta \to R$ be a smooth function, where $\Theta \subset R^d$. Consider the problem of searching for the minimum point θ^*. Assume that

$$f(\theta) = E\{F(\theta, X)\}.$$

- SGD:

  Let $\theta_0 \in \mathbb{R}^d$ be an initial value (might be random). For $n \geq 1$, we update θ_n by

$$\theta_n = \theta_{n-1} - \ell_n \nabla F_n(\theta_{n-1}) = \theta_{n-1} - \ell_n(\nabla f(\theta_{n-1}) + \zeta_n),$$

  where ℓ_n is the learning rate, $F_i(\theta) = F(\theta, X_i)$ and $\zeta_n = \nabla F_n(\theta_{n-1}) - \nabla f(\theta_{n-1})$.

- Consider the averaged version

$$\bar{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1}\theta_i.$$

- Write

$$\begin{aligned}\zeta_n &= \nabla F_n(\theta_{n-1}) - \nabla f(\theta_{n-1}) \\ &= \underbrace{\nabla F_n(\theta^*) - \nabla f(\theta^*)}_{\xi_n} \\ &\quad + \underbrace{\left\{\nabla F_n(\theta_{n-1}) - \nabla F_n(\theta^*)\right\} - \left\{\nabla f(\theta_{n-1}) - \nabla f(\theta^*)\right\}}_{\eta_n := g(\theta_{n-1}, X_n)}.\end{aligned}$$

It follows that

$$\sqrt{n}(\bar{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i \xi_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} Q_i \eta_i + D_2,$$

and D_2 is a remainder term and Q_i is a nonrandom matrix depending only on $\nabla^2 f(\theta^*)$.

Regularity conditions

(i) $\|\theta_0 - \theta^*\|_4 \leq \tau_0$

(ii) $\max_{1 \leq i \leq n} \mathbb{E} \|\xi_i\|^4 \leq \tau^4, \quad \sup_x \|g(\theta, x)\| \leq c_1 \|\theta - \theta^*\|.$

(iii) The function f is *L-smooth* and *strongly convex* with convexity constant $\mu > 0$. That is, f is twice differentiable and for all $\theta \in R^d$,

$$\mu I_d \preccurlyeq \nabla^2 f(\theta) \preccurlyeq L I_d. \tag{1}$$

(iv) There exist positive constants c_2 and β such that for all θ such that $\|\theta - \theta^*\| \leq \beta$,

$$\left\|\nabla^2 f(\theta) - \nabla^2 f(\theta^*)\right\|_S \leq c_2 \left\|\theta - \theta^*\right\|. \quad (2)$$

Here, $\|A\|_S = \sqrt{\lambda_{\max}(A^T A)}$ is the spectral norm.

- CLT for SGD

  - Polyak and Juditsky (1992): If $\ell_n = a_0 n^{-\alpha}$, then

$$\sqrt{n}(\bar{\theta}_n - \theta^*) \xrightarrow{d.} N(0, \Sigma) \quad \text{for some } \Sigma > 0.$$

▶ Berry-Esseen bound for SGD

- Shao and Zhang (2021) :

Let $\ell_n = a_0 n^{-\alpha}$ where $1/2 < \alpha \le 1$. Assume that the regular conditions are satisfied. Then we have for $\alpha \in (1/2, 1)$,

$$\sup_{A\in\mathcal{C}}\Big|P\big[\sqrt{n}\Sigma_n^{-1/2}(\bar\theta_n - \theta^*) \in A\big] - P[Z \in A]\Big| \le C\big(d^{3/2} + \tau^3 + \tau_0^3\big)\big(d^{1/2}n^{-1/2} + n^{-\alpha+1/2}\big).$$

If $\ell_n = a_0 n^{-1}$ with $a_0\lambda_{\min}(\nabla^2 f(\theta)) \ge 1$ for all $\theta \in \Theta$, we have

$$\sup_{A\in\mathcal{C}}\Big|P\big[\sqrt{n}\Sigma_n^{-1/2}(\bar\theta_n - \theta^*) \in A\big] - P[Z \in A]\Big| \le C d^{1/2} n^{-1/2}(\log n)^3(d^{3/2} + \tau^3 + \tau_0^3).$$

▶ Application to M-estimators

Let $X, X_1, \ldots, X_n$ be i.i.d. random variables that take values in a space $\mathcal{X}$.

Let $\Theta \subset R^d$ be a parameter space. For each $\theta \in \Theta$, let $m_\theta(\cdot) : \mathcal{X} \to \mathbb{R}$ be a loss function. Assume that $\theta \mapsto m_\theta(x)$ is twice differentiable with respect to θ for every $x \in \mathcal{X}$. Denote

$$\mathbb{M}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i), \quad M(\theta) = E m_\theta(X).$$

Let θ^* be the unique point that minimizes the function $M(\theta)$. $\hat{\theta}_n$ is called an M-estimator of θ^* if it minimizes the function $\mathbb{M}_n(\theta)$.

▶ Asymptotic properties of $\hat{\theta}_n - \theta^*$

- Under some regularity conditions, it is known that (see, e.g., Van der Vaart (1998))

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d.} N(0, \Sigma).$$

- Bentkus et al. (1997) proved a Berry–Esseen bound of order $n^{-1/2}$ for M-estimators under some regularity conditions and a consistency condition

$$P(|\hat{\theta}_n - \theta^*| \geq \delta) \leq a\, n^{-1/2}.$$

- Here we apply our general result to prove a Berry–Esseen bound under some simpler conditions.

- Regularity conditions
  (i) The function $\theta \mapsto m_\theta(x)$ is twice differentiable for all $x \in \mathcal{X}$, and

$$\begin{aligned} M(\theta) - M(\theta^*) &\geq \mu \|\theta - \theta^*\|^2, \quad \text{(Convexity)} \\ |m_\theta(x) - m_{\theta^*}(x)| &\leq m_1(x)\|\theta - \theta^*\|, \quad \forall x \in \mathcal{X}, \quad \text{(Lipschitz)} \\ \|\ddot{m}_\theta(x) - \ddot{m}_{\theta^*}(x)\| &\leq m_2(x)\|\theta - \theta^*\|, \quad \forall x \in \mathcal{X}. \quad \text{(Lipschitz)} \\ \ddot{m}_{\theta^*}(x) &\preccurlyeq m_3(x) I_d, \quad \text{(Boundedness at } \theta^*) \end{aligned}$$

  (ii) For m_1, m_2, m_3,

$$\|m_1(X)\|_9 \leq c_1, \quad \|m_2(X)\|_4 \leq c_2, \quad \|m_3(X)\|_4 \leq c_3,$$

  where $\|Y\|_p = (\mathbb{E}\,|Y|^p)^{1/p}$.

  (iii) Let $\xi_i = \dot{m}_{\theta^*}(X_i)$, $\Sigma = E\left\{\xi_i \xi_i^T\right\}$ and $V = E\{\ddot{m}_{\theta^*}(X)\}$. Moreover, assume that

$$\lambda_1 \leq \lambda(\Sigma) \leq \lambda_2 \quad \lambda(V) \geq \lambda_3, \quad \|\xi_1\|_4 \leq c_4 d^{1/2}.$$

- Berry–Esseen bounds for M - estimators
  - Shao and Zhang (2021):

$$\sup_{A\in\mathcal{C}}\left|P\left[n^{1/2}\Sigma^{-1/2}V(\hat{\theta}_n-\theta^*)\in A\right]-P(Z\in A)\right|\le Cd^2n^{-1/2},$$

where $C>0$ is a constant depending only on $c_1, c_2, c_3, c_4, \mu, \lambda_1, \lambda_2$ and λ_3.

Main References

- L.H.Y. Chen, L. Goldstein and Q.M. Shao (2011). Normal Approximation by Stein's Method. Springer .
- Q.M. Shao and W.X. ZHou (2016). Cramér type moderate deviation theorems for self-normalized processes. *Bernoulli* **22**, 2029 - 2079.
- Q.M. Shao and Z.S. Zhang (2021), Berry - Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms. *Bernoulli (to appear).*


Thank You