Non-Asymptotic Analysis of Short Asymptotic Expansions for Integer-valued Sums with Applications

V.ULYANOV (HSE University and Lomonosov MSU)

.

joint results with S.Bobkov, Z.Su and X.Wang

$X_1, \ldots, X_n$ – independent r.v., $\mathbf{E}|X_k|^3 < \infty$
$S_n = X_1 + \cdots + X_n$. with $\mu = \mathbf{E}S_n$ and $\sigma^2 = \mathbf{Var}S_n$.
Third order Lyapunov ratio

$$L_3 = \frac{1}{\sigma^3} \sum_{k=1}^{n} \mathbf{E}|X_k - \mathbf{E}X_k|^3$$

Berry–Esseen inequality:

$$\sup_x \left| \mathbf{P}\{S_n \leq x\} - \Phi\left(\frac{x-\mu}{\sigma}\right) \right| \leq cL_3.$$

Here Φ denotes the standard normal distribution function with probability density function $\varphi(x)$.

Necessarily $L_3 \geq 1/\sqrt{n}$ which may be reversed in the i.i.d. case $X_k = \xi_k/\sqrt{n}$ up to a factor depending on ξ_1. Then

$$\mathbf{P}\{S_n \leq x\} = \Phi\left(\frac{x-\mu}{\sigma}\right) + O\left(\frac{1}{\sqrt{n}}\right), \qquad n \to \infty.$$

For more precise statement, put

$$\Phi_3(x) = \Phi(x) - \frac{l_3}{6}(x^2-1)\varphi(x), \qquad x \in \mathbf{R},$$

where

$$l_3 = \frac{1}{\sigma^3}\mathbf{E}(S_n - \mathbf{E}S_n)^3 = \frac{1}{\sigma^3}\sum_{k=1}^{n}\mathbf{E}(X_k - \mathbf{E}X_k)^3.$$

The idea goes back to Chebyshev's work of 1887.

Problem is to study whether or not it is possible to improve approximation by replacing Φ with Φ_3. Such a replacement would not deteriorate this bound in view of the relation $|l_3| \le L_3$. On the other hand, comparison of smooth linear functionals (for example, characteristic functions) of F_n and Φ_3 suggests that an improvement is indeed possible in various natural cases.

In particular, in the i.i.d. case, if X_1 has a non-lattice distribution, Esseen (1945) derived

$$\mathbf{P}\{S_n \leq x\} = \Phi_3\left(\frac{x-\mu}{\sigma}\right) + o\left(\frac{1}{\sqrt{n}}\right), \qquad n \to \infty, \quad (1)$$

The remainder term may be improved to $O(1/n)$, provided that $\mathbf{E}X_1^4 < \infty$ and assuming that the Cramér continuity condition

$$\limsup_{t\to\infty} \left|\mathbf{E}e^{itX_1}\right| < 1,$$

is met, see Cramér (1937).

It was also shown by Esseen that a representation similar to (1) remains to hold for lattice distributions by adding to Φ_3 a discontinuous periodic function with a factor of the order $1/\sqrt{n}$. To make the statement more transparent, suppose X_1 takes integer values with $h = 1$ being the maximal step (span), so that the distribution of X_1 is not supported on $h\mathbf{Z}$ for $h > 1$.
Then, see Esseen (1945)

$$\mathbf{P}\{S_n \leq x\} = \Phi_3\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{\sigma}\psi(x)\varphi\left(\frac{x-\mu}{\sigma}\right) + o\left(\frac{1}{\sqrt{n}}\right), \tag{2}$$

with $\psi(x) = x - [x] + \frac{1}{2}$

$$\mathbf{P}\{S_n \leq x\} = \Phi_3\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{\sigma}\psi(x)\varphi\left(\frac{x-\mu}{\sigma}\right) + o\left(\frac{1}{\sqrt{n}}\right), \tag{3}$$

Since S_n is integer-valued, we are concerned with $\mathbf{P}\{S_n \leq k\}$ for integers k only. Then, applying (2) to $x = k + \frac{1}{2}$, the additional term is vanishing, and we obtain

$$\mathbf{P}\{S_n \leq k\} = \Phi_3\left(\frac{k+1/2-\mu}{\sigma}\right) + o\left(\frac{1}{\sqrt{n}}\right),$$

This well-known phenomenon was a focus of investigations in the scheme of Bernoulli trials, see e.g. Bernstein (1945), Feller (1943) and Uspensky (1937). If X_k takes 1 and 0 with probabilities p and $q = 1 - p$, we have $\mu = np$, $\sigma^2 = npq$, and

$$\sup_{0 \le k \le n} \left| \mathbf{P}\{S_n \le k\} - \Phi\left(\frac{k - np}{\sqrt{npq}}\right) \right| \le \frac{c}{\sqrt{npq}}.$$

In (1937), Uspensky established a two-term approximation

$$\sup_{0\le k\le n}\left|\mathbf{P}\{S_n \le k\} - \Phi_3\left(\frac{k+1/2-np}{\sqrt{npq}}\right)\right| \le \frac{c}{npq},$$

here the Chebyshev-Edgeworth correction is

$$\Phi_3(x) = \Phi(x) - \frac{p-q}{6\sqrt{npq}}(x^2-1)\,\varphi(x).$$

This approach was adapted by to get a Poissonian analog.

Next step has been made by Deheuvels, Puri, and Ralescu (1989) who extended the inequality to independent Bernoulli random variables X_k taking the values 0 and 1 with not necessarily equal probabilities $p_k = \mathbf{P}\{X_k = 1\}$. Namely, we similarly have

$$\sup_{0 \le k \le n} \left| \mathbf{P}\{S_n \le k\} - \Phi_3\left(\frac{k + 1/2 - \mu}{\sigma}\right) \right| \le \frac{c}{\sigma^2},$$

where

$$\mu = p_1 + \cdots + p_n, \quad \sigma^2 = p_1 q_1 + \cdots + p_n q_n \quad (q_k = 1 - p_k).$$

Our aim is to extend the bound to general independent integer-valued random variables under the 4-th moment condition

Definition

Given an integer-valued random variable ξ with characteristic function $v(t) = \mathbf{E}\, e^{it\xi}$ $(t \in \mathbf{R})$, put

$$\mathrm{V}(\xi) = - \sup_{0<t<2\pi} \frac{\ln |v(t)|}{1 - \cos t}. \tag{4}$$

One important feature of this functional:

Proposition. *If ξ is integer-valued, then $0 \leq \mathsf{V}(\xi) < \infty$. Moreover, $\mathsf{V}(\xi) > 0$ if and only if the distribution of ξ is non-degenerate and has the maximal step $h = 1$.*

$\mathsf{V}(\xi)$ quantifies the “strength" of the property that the maximal step of the lattice distribution of ξ is exactly $h = 1$.

We have

$$\mathsf{V}(\xi) \leq \mathbf{Var}\xi.$$

We have

$$\mathrm{V}(\xi) \leq \mathbf{Var}\xi.$$

Example. Suppose that $\mathbf{P}\{\xi = \pm 1\} = (1-\varepsilon)/2$ and $\mathbf{P}\{\xi = 0\} = \varepsilon$ for some $\varepsilon \in (0,1)$. Then $h = 1$, while $\mathrm{V}(\xi) \to 0$ as $\varepsilon \to 0$. It is not surprising that the limit distribution has a larger maximal step $h = 2$.

On the other hand, we have an equality

$$\mathrm{V}(\xi) = \mathbf{Var}\xi = pq,$$

when ξ has a Bernoulli distribution with parameters $p = \mathbf{P}\{\xi = 1\}$ and $q = \mathbf{P}\{\xi = 0\}$.

Main Result

... **Theorem.** (Bobkov and U.(2022) *Let the integer-valued random variables* $X_1, \dots, X_n$ *be independent and have finite 4-th moments. For sum* $S_n = X_1 + \dots + X_n$, *put* $\mu = \mathbf{E}S_n$, $\sigma^2 = \mathbf{Var}(S_n)$, *and* $\mathrm{V} = \sum_{k=1}^n \mathsf{V}(X_k)$. *Then we have with absolute constant* $c > 0$

$$\sup_{k \in \mathbf{Z}} \left| \mathbf{P}\{S_n \le k\} - \Phi_3\left(\frac{k + 1/2 - \mu}{\sigma}\right) \right| \le \frac{c\sigma^2}{\mathrm{V}} L_4. \qquad (5)$$

... **Corollary.** *Suppose that the integer-valued random variables ξ_k are independent and have a common non-degenerate distribution with maximal step $h = 1$ and such that $\mathbf{E}\xi_1^4 < \infty$. Putting $\mu = n\,\mathbf{E}\xi_1$ and $\sigma^2 = n\,\mathbf{Var}(\xi_1)$, we have*

$$\mathbf{P}\{S_n \le k\} = \Phi_3\left(\frac{k + 1/2 - \mu}{\sigma}\right) + O\left(\frac{1}{n}\right)$$

as $n \to \infty$ uniformly over all $k \in \mathbf{Z}$.

Moreover, the involved constant in the remainder term does not exceed, up to a numerical factor, the next quantity

$$\frac{\mathbf{E}(\xi_1 - \mathbf{E}\xi_1)^4}{n\,\mathrm{V}(\xi_1)\mathbf{Var}\xi_1}.$$

Smoothing operation

Let η denote a random variables which is independent of all X_k and has a uniform distribution U on the interval $(-1/2, 1/2)$. If all X_k are integer-valued, the random variable $\widetilde{S}_n = S_n + \eta$ has an absolutely continuous distribution and

$$\mathbf{P}\left\{\widetilde{S}_n \le k + \frac{1}{2}\right\} = \mathbf{P}\{S_n \le k\}, \qquad k \in \mathbf{Z}.$$

- The object of study

Let J be a finite index set, $\{X_i, i \in J\}$ a sequence of random variables taking values in $\mathbb{Z}_+$. For a given set $A \subset J$,

$$X_A = \{X_i, i \in A\}.$$

Call $\{X_i, i \in J\}$ a classic locally dependent structure if

(LD1): For each $i \in J$, there exists $A_i \subset J$ such that
X_i is independent with $\{X_j : j \notin A_i\}$.

(LD2): For each $i \in J$, there exists $B_i \supset A_i$ such that
X_{A_i} is independent with $\{X_k : k \notin B_i\}$.

(LD3): For each $i \in J$, there exists $C_i \supset B_i$
X_{B_i} is independent with $\{X_l : l \notin C_i\}$.

Note B_i is bigger than A_i, C_i is bigger than B_i.

- Example: $\xi_1, \xi_2 \cdots, \xi_n$ i.i.d. $Be(p)$. $X_i = \xi_i \xi_{i+1}$

X_i is independent with X_j when
$|j - i| > 1 \Rightarrow A_i = \{i-1, i, i+1\}$.

$\{X_{i-1}, X_i, X_{i+1}\}$ is independent with X_j when $|j-i| > 2$,
$\Rightarrow B_i = \{i-2, i-1, i, i+1, i+2\}$.

$\{X_{i-2}, X_{i-1}, X_i, X_{i+1}, X_{i+2}\}$ is independent with X_j when
$|j-i| > 3$
$\Rightarrow C_i = \{i-3, i-2, i-1, i, i+1, i+2, i+3\}$.

- The object of our study:

$$W = \sum_{i \in J} X_i.$$

Such locally dependent structure can be found in Chen and Shao 2004, Röllin 2008 etc.

1 Chen L H Y, Shao Q M. Normal approximation under local dependence[J]. 2004.
2 Röllin A. Symmetric and centered binomial approximation of sums of locally dependent random variables[J]. 2008.

- The approximating variables:

$$M_1 = B(n,p) * \mathcal{P}(\lambda);$$

$$M_2 = NB(r,p) * \mathcal{P}(\lambda).$$

- The total variation distance

$$d_{TV}(W, M_i) = \sup_{k \in \mathbb{Z}} |P(W = k) - P(M_i = k)|.$$

- Question: $d_{TV}(W, M_i) \leq ?$

- Two questions naturally arise.

Q 1: How to determine these parameters like $n, p, r, \bar{p}, \lambda$?

A basic idea: The first three moments match each other

$$\mathbf{E} M_i = \mathbf{E} W, \quad \mathbf{E} M_i^2 = \mathbf{E} W^2, \quad \mathbf{E} M_i^3 = \mathbf{E} W^3.$$

Q 2: Which is a better variable for approximating W?

Put $\quad \mu = \mathbf{E} W, \quad \sigma^2 = \textbf{\textit{Var}}\, W$

Consider two different cases, separately.

- Case 1: $\mu > \sigma^2$, we use M_1 to approximate W.
- Case 2: $\mu < \sigma^2$, we use M_2 to approximate W.

2. Main results

To state our main results, we introduce some additional notation.

$$\theta_1 = \frac{\lambda}{\lfloor n + \lambda/p \rfloor q}, \ \theta_2 = \frac{\lambda q}{rq + \lambda p},$$

$$\mathcal{S}_2(W) := \sup_{h:\|h\| \le 1} \left| \mathbf{E} \Delta^{(2)} h(W) \right|,$$

$$X_{C_i} = \{X_l, l \in C_i\}, \ \mathcal{S}_2(W|X_{C_i}) := \sup_{h:\|h\| \le 1} \left| \mathbf{E}(\Delta^{(2)} h(W)|X_{C_i}) \right|$$

where $\Delta^{(2)}$ denotes second-order difference

$$\Delta^{(2)} h(W) = h(W+2) - 2h(W+1) + h(W),$$

$$\mathcal{S}_i(W) := \operatorname{esssup} \mathcal{S}_2(W|X_{C_i}), \quad S(W) = \sup_i \mathcal{S}_i(W),$$

$$\gamma_i = \sum_{(\mathbf{E})} \left[\mathbf{E}X_i(\mathbf{E})X^*_{A_i}(\mathbf{E})X^*_{B_i}(\mathbf{E})X^*_{C_i} + \mathbf{E}X_i(\mathbf{E})X^*_{B_i}(\mathbf{E})X^*_{C_i}) \right],$$
$$\gamma = \sum_{i \in J} \gamma_i,$$

where $\sum_{(\mathbf{E})}$ denotes the sum over all possible **E** in front of each X_i.

Theorem

(Su, Wang and U.(2025)) (i) Assume $\mu > \sigma^2$ *and* $2\lambda < q\lfloor n + \lambda/q \rfloor$*, where* $\{n, p, q, \lambda\}$ *was given in Case 1, then we have*

$$d_{TV}(W, M_1) \leq C\left[\frac{\gamma S(W)}{(1-2\theta_1)q\mu} + \frac{1}{\mu}\right];$$

(ii) Assume $\mu < \sigma^2$ *and* $2\lambda q < rq + \lambda p$*, where* $\{r, p, q, \lambda\}$ *was given in Case 2, then we have*

$$d_{TV}(W, M_2) \leq C\left(1 \vee \frac{q}{p}\right)\frac{\gamma S(W)}{(1-2\theta_2)\mu}.$$

Denote by $\mathbf{\Phi}$ the standard normal distribution function, ϕ the standard normal density function

$$\mathbf{\Phi}_{3,j}(x) = \mathbf{\Phi}(x) - \frac{l_{3,j}}{6}(x^2 - 1)\phi(x), \quad x \in \mathbf{R},$$

$$V_j = -\sum_{i=1}^{|J|} \sup_{0 \le t \le 2\pi} \frac{\ln \left|\mathbf{E} e^{it\xi_{i,j}}\right|}{1 - \cos t}, \quad j = 1, 2.$$

From Bobkov and Ulyanov 2022,

$$\sup_{k \in \mathbf{Z}} \left| \mathbf{P}(M_j \le k) - \mathbf{\Phi}_{3,j}\left(\frac{k + 1/2 - \mu}{\sigma}\right) \right| \le \frac{\sigma^2 L_{4,j}}{V_j}.$$

Theorem
Denote $K_J = \sup_{i\in J} |A_i||B_i||C_i|$. *If* $\lim_{|J|\to\infty} K_J$ *and all* $\mathbf{E}X_i^3$ *are bounded by a constant, then*

$$\sup_{k\in\mathbf{Z}}\left|\mathbf{P}(W \le k) - \mathbf{\Phi}_3\left(\frac{k+1/2-\mu}{\sigma}\right)\right| \le C\frac{|J|}{\sigma^4 \wedge (\sigma^2\mu)}.$$

$\xi_1, \cdots, \xi_n$ i.i.d. $Be(p)$. A (k_1, k_2)-event if there occurs $k_1 > 0$ consecutive $0'$s followed by k_2 consecutive $1'$s
Define

$$X_j = (1 - \xi_j) \cdots (1 - \xi_{j+k_1-1}) \, \xi_{j+k_1} \cdots \xi_{j+k_1+k_2-1},$$

$$W = \sum_{i=1}^{n+1-k_1-k_2} X_i$$

Theorem

(Su, Wang and U.(2025)) Let $m \geq 1$ *and* $\{p_i, i \in J\}$ *are identical to* p*, assume*

$$b := (1-p)^{k_1} p^{k_2} < \frac{2(2m+1)}{2(2m+1)^2 + 3m(m+1)} := c_m.$$

Then we have

$$d_{TV}(W, M_1) = O(N^{-1});$$
$$\sup_{k \in \mathbf{Z}} \left| \mathbf{P}(W \leq k) - \mathbf{\Phi}_3\left(\frac{k + 1/2 - \mu}{\sigma}\right) \right| = O(N^{-1}).$$

Other Applications

(i) k – runs

(ii) Birthday problem

(iii) Counting monochromatic edges in uniformly colored graphs

(iv) Triangles in the Erdős-Rényi random graph

References

1. Bobkov S G, Ulyanov V V. The Chebyshev–Edgeworth Correction in the Central Limit Theorem for Integer-Valued Independent Summands. Theory Probab. Appl., 2022, 66(4): 537-549.
2. Zhonggen Su, Vladimir V. Ulyanov, Xiaolin Wang, On Approximation of Sums of Locally Dependent Random Variables via Perturbation of Stein Operator, Theory Probab. Appl., 2025, 70(1): 24-36.
3. F Götze, A Naumov, V Ulyanov, Asymptotic analysis of symmetric functions. - Journal of Theoretical Probability, 2017