The Annals of Probability
1989, Vol. 17, No. 4, 1646–1650

ON NORMAL APPROXIMATIONS OF DISTRIBUTIONS IN TERMS OF DEPENDENCY GRAPHS

BY PIERRE BALDI AND YOSEF RINOTT

University of California, San Diego and Hebrew University

Bounds on the error in the normal approximation of sums of dependent random variables introduced by Stein are interpreted in terms of dependency graphs. This leads to improvements on a central limit theorem of Petrovskaya and Leontovich and recent applications by Baldi and Rinott. In particular, bounds on rates of convergence are obtained. As an application we study the normal approximation to the number of local maxima of a random function on a graph.

1. Introduction.

Stein (1986), page 110, proves the following:

Let V be a finite set and for each $i \in V$, let X_i be a real random variable and S_i a subset of V such that $EX_i = 0$, $EX_i^4 < \infty$ and $E(\sum_{i \in V} X_i \sum_{j \in S_i} X_j) = 1$. Let $W = \sum_{i \in V} X_i$ and $\sigma_{ij} = EX_iX_j$. Then for all real w,

$$\begin{aligned} |P(W \le w) - \Phi(w)| \le (\pi/2)^{1/2} E \sum_{i \in V} |E(X_i | X_j,\ j \notin S_i)| \\ + 2^{3/4}\pi^{-1/4} \sqrt{E \sum_{i \in V} |X_i| \Big(\sum_{j \in S_i} X_j \Big)^2} \\ + 2\sqrt{E\Big[\sum_{i \in V} \sum_{j \in S_i} (X_iX_j - \sigma_{ij}) \Big]^2}. \end{aligned} \tag{1}$$

Consider now the case of a set of random variables $\{X_i,\ i \in V\}$ indexed by the vertices of a graph $G = (V, E)$. G is said to be a *dependency graph* if for any pair of disjoint sets A_1, A_2 in V such that no edge in E has one endpoint in A_1 and the other in A_2, the sets of random variables $\{X_i,\ i \in A_1\}$ and $\{X_i,\ i \in A_2\}$ are independent. A useful version of Stein's result is obtained by reducing his conditions to conditions on the dependency graph.

THEOREM 1. *Let $\{Z_i,\ i \in V\}$ be random variables having a dependency graph $G = (V, E)$. For $i \in V$, let $L_i^{(k)}$ denote the number of connected subsets of V of cardinality at most k which contain i. Let $W = \sum_{i \in V} Z_i$ and $\sigma^2 = \operatorname{Var} W < \infty$. Set*

$$\frac{1}{\sigma^k} \sum_{i \in V} L_i^{(k)} E|Z_i - EZ_i|^k = A_k < \infty, \qquad k = 3, 4.$$

Received January 1988; revised November 1988.

AMS 1980 *subject classifications.* Primary 60F05; secondary 05C99.

Key words and phrases. Central limit theorem, dependent variables, rates of convergence, random local maxima.

Then for all real w,

$$\left|P\left(\frac{W-EW}{\sigma} \le w\right) - \Phi(w)\right| \le c\left(\sqrt{A}_3 + \sqrt{A}_4\right)$$

for some constant $0 \le c < 8$ *which does not depend on* $\{Z_i\}$.

All proofs are deferred to Section 2.

In the case of a sequence $\{Z_{in},\ i \in V_n\}$, $n = 1, 2, \dots,$ with $W_n = \sum_{i \in V_n} Z_{in}$, σ_n^2, A_{kn} defined accordingly, we conclude that $(W_n - EW_n)/\sigma_n \to N(0,1)$ provided $A_{kn} \to 0$ as $n \to \infty$ for $k = 3$ and 4. Theorem 1 improves a result of Petrovskaya and Leontovich (1982) who showed that $(W_n - EW_n)/\sigma_n \to N(0,1)$ if $A_{kn} \to 0$, *for all* $k \ge 3$, by convergence of all moments. Rates of convergence to normality were not discussed by Petrovskaya and Leontovich at all.

The next result improves Corollary 2 in Baldi and Rinott (1989); see also Janson (1988).

COROLLARY 2. *Let* $\{Z_i,\ i \in V\}$ *be random variables having a dependency graph* $G = (V, E)$. *Set* $W = \sum_{i \in V} Z_i$ *and* $\sigma^2 = \operatorname{Var} W$. *Let* D *denote the maximal degree of* G *and suppose* $|Z_i| \le B$ *a.s. Define*

$$Q = \frac{|V|D^2B^3}{\sigma^3}.$$

Then

$$\left|P\left(\frac{W-EW}{\sigma} \le w\right) - \Phi(w)\right| \le 32(1 + \sqrt{6})Q^{1/2}.$$

Note that the case $D \equiv m$ corresponds to m-dependence.

Example of application: *The number of local maxima on a graph.* Consider a sequence $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ of graphs. (Script letters are used to indicate that these are *not* dependency graphs.) Let Y_i, $i \in \mathscr{V}$, be i.i.d. continuous random variables. For $i \in \mathscr{V}$ define the 0-1 indicator variable Z_i by $Z_i = 1$ iff $Y_i > Y_j$ for all $j \in \mathscr{V}$ such that $d(i, j) = 1$, where $d(i, j)$ denotes the shortest path distance between the vertices i and j of $\mathscr{V}$. Note that $d(i, j) = 1$ iff i and j are neighbors (i.e., connected by an edge in $\mathscr{E}$), so $Z_i = 1$ indicates that Y_i is a "local maximum." (The Y_i's are evidently used only to induce a random ranking on the vertices.) Let d_i denote the degree of vertex i in $\mathscr{G}$ and let $T = |\{(i, j)\colon i, j \in \mathscr{V}, d(i, j) = 2\}|$ and W be the random number of local maxima $W = \sum_{i \in \mathscr{V}} Z_i$.

PROPOSITION 3. *Assume that*:

(a) $d_i \le d$ *for every* $i \in \mathscr{V}$.

(b) $T \ge cd^{\alpha}|\mathscr{V}|$ *for some positive* c *and* α.

If

$$R = d^{(17-3\alpha)/2}|\mathscr{V}|^{-1/2},$$

then

$$\left|P\left(\frac{W-EW}{(\operatorname{Var} W)^{1/2}} \le w\right) - \Phi(w)\right| \le 64(1+\sqrt{6})12^{3/4}c^{-3/4}R^{1/2}.$$

EXAMPLE. Consider the graph where $\mathscr{V} = \{0,\dots,m-1\}^n$ and $(u,v) \in \mathscr{E}$ if $u, v \in \mathscr{V}$ agree on all but one coordinate. We have $|\mathscr{V}| = m^n$, $d = n(m-1)$ and $T = |\mathscr{V}|\binom{n}{2}(m-1)^2 \ge 4^{-1}d^2|\mathscr{V}|$ for $n \ge 2$, so that we can take $\alpha = 2$ and $c = 4^{-1}$. Then $R = R_{m,n} = [n(m-1)]^{5.5}/m^{n/2}$. Note that for any m, $\lim_{n\to\infty} R_{m,n} = 0$ and for $n \ge 12$, $\lim_{m\to\infty} R_{m,n} = 0$. In these cases, Proposition 3 implies asymptotic normality of W. Similarly, Theorem 1 or Corollary 2 can be applied to obtain asymptotic normality for various problems, such as those in Baldi and Rinott (1989).

2. Proofs.

PROOF OF THEOREM 1. Define $X_i = (Z_i - EZ_i)/\sigma$ and let $S_i = \{j \in V: d(i,j) \le 1\}$ so that S_i consists of i and all the vertices connected to i. Since Z_i, Z_j are independent if $j \notin S_i$, we have

$$\sigma^2 = E\left[\sum_{i\in V}(Z_i - EZ_i)\right]^2 = E\left[\sum_{i\in V}(Z_i - EZ_i)\sum_{j\in V}(Z_j - EZ_j)\right]$$

$$= E\left[\sum_{i\in V}(Z_i - EZ_i)\sum_{j\in S_i}(Z_j - EZ_j)\right]$$

and dividing by σ^2 we get $E\sum_{i\in V}X_i\sum_{j\in S_i}X_j = 1$. Clearly $EX_i = 0$ and $EX_i^4 < \infty$ is subsumed by $A_4 < \infty$. Therefore we can apply Stein's result and calculate bounds for the right-hand side of (1).

The first term vanishes by $EX_i = 0$ and X_i, X_j independent if $j \notin S_i$. Next consider the second term:

$$E\sum_{i\in V}|X_i|\left(\sum_{j\in S_i}X_j\right)^2 \le \sum_{i\in V}\sum_{j,k\in S_i}E|X_i||X_j||X_k|$$

$$\le \tfrac{1}{3}\sum_{i\in V}\sum_{j,k\in S_i}\left(E|X_i|^3 + E|X_j|^3 + E|X_k|^3\right)$$

by the arithmetic-geometric mean inequality. Now $\sum_{j\in S_i}\sum_{k\in S_i}1 \le 2L_i^{(3)}$ and for j fixed, $\sum_{i:\, j\in S_i}\sum_{k\in S_i}1 \le 2L_j^{(3)}$. So $E\sum_{i\in V}|X_i|(\sum_{j\in S_i}X_j)^2 \le 2A_3$.

Next consider the last term on the right-hand side of (1). Since $\sum_{i\in V}\sum_{j\in S_i}\sigma_{ij} = E\sum_{i\in V}X_i\sum_{j\in S_j}X_j = 1$ we have

$$E\left[\sum_{i\in V}\sum_{j\in S_i}(X_iX_j - \sigma_{ij})\right]^2$$

$$= E\left(\sum_{i\in V}\sum_{j\in S_i}X_iX_j - 1\right)^2 = E\left(\sum_{i\in V}\sum_{j\in S_i}X_iX_j\right)^2 - 1.$$

Now

$$\begin{aligned}E\left(\sum_{i\in V}\sum_{j\in S_i}X_iX_j\right)^2 &= E\sum_{i\in V}\sum_{j\in S_i}X_i^2X_j^2 + E\sum_{i\in V}\sum_{j,\,k\in S_i,\ j\neq k}X_i^2X_jX_k \\ &\quad + E\sum_{i,\,k\in V,\ i\neq k}\ \sum_{j\in S_i,\ l\in S_k}X_iX_jX_kX_l.\end{aligned} \tag{2}$$

Using the arithmetic-geometric mean inequality, the first term is bounded by $\sum_{i\in V}L_i^{(2)}EX_i^4$ and the second term by $2\sum_{i\in V}L_i^{(3)}EX_i^4$. Since $L_i^{(k)}$ counts sets of cardinality at most k, $L_i^{(4)} \geq L_i^{(3)} \geq L_i^{(2)}$. So the sum of the first two terms is bounded by $3A_4$. Finally,

$$E\sum_{i,\,k\in V,\ i\neq k}\ \sum_{j\in S_i,\ l\in S_k}X_iX_jX_kX_l = \sum_{\{i,\,j,\,k,\,l\}}EX_iX_jX_kX_l + \sum_{\{i,\,j\}\{k,\,l\}}EX_iX_jX_kX_l,$$

where in both sums $i \neq k$, $j \in S_i$ and $l \in S_k$, but in the first $\{i, j, k, l\}$ form a connected set whereas in the second $\{i, j\}$ is disconnected from $\{k, l\}$. The first sum on the right-hand side is again bounded by $6A_4$ and the second term is equal to $\sum EX_iX_jEX_kX_l$ with summation over all independent set $\{i, j\}\{k, l\}$ with $i \neq k$, $j \in S_i$ and $l \in S_k$. If in the latter sum we allowed summation over all i, k and $j \in S_i$, $l \in S_k$, we would be adding only sums over connected sets so the added part would be no larger than $6A_4$ and the increased sum is $\sum_{i\in V,\ j\in S_i}EX_iX_j\sum_{k\in V, l\in S_k}EX_kX_l = 1$. Combining all these calculations we obtain

$$|P(W \leq w) - \Phi(w)| \leq 2^{3/4}\pi^{-1/4}\sqrt{2A_3} + 2\sqrt{15A_4} < 8(\sqrt{A}_3 + \sqrt{A}_4). \qquad \square$$

PROOF OF COROLLARY 2. Note that

$$L_i^{(k)} \leq (k-1)!D^{k-1} \quad\text{and}\quad E|Z_i - EZ_i|^k \leq 2^kB^k.$$

So $\sum_{i\in V}L_i^{(k)}E|Z_i - EZ_i|^k/\sigma^k \leq 2^k(k-1)!|V|D^{k-1}B^k/\sigma^k$ and from Theorem 1 we obtain

$$\begin{aligned}|P((W - EW)/\sigma \leq w) - \Phi(w)| &\leq 8\left(\sqrt{A_3} + \sqrt{A_4}\right) \\ &\leq 32Q^{1/2} + 32\sqrt{6}\,Q^{2/3} \quad (\text{using } |V| \geq D) \\ &\leq 32(1 + \sqrt{6})Q^{1/2} \quad \text{when } Q < 1\end{aligned}$$

(the case $Q \geq 1$ is trivial). □

PROOF OF PROPOSITION 3. Observe that Z_i and Z_j are dependent if and only if $d(i, j) \leq 2$ in $\mathscr{G}$. Therefore we can construct a dependency graph G from $\mathscr{G}$ by adding to $\mathscr{E}$ a new edge for any pair of vertices of $\mathscr{G}$ at distance 2. With the notation of Corollary 2, we have $D =$ maximal degree of dependency graph $\leq d^2 + d$. Let h_{ij} be the number of common neighbors in $\mathscr{G}$ to both vertices i and j. A direct calculation shows that the variance of W is given by

$$\begin{aligned} &\sum \operatorname{Var} X_i + \sum \operatorname{Cov}(X_i, X_j) \\ &\quad = \sum_{i \in \mathscr{V}} d_i(d_i + 1)^{-2} - \sum_{i,\, j:\, d(i,\, j) = 1} (d_i + 1)^{-1}(d_j + 1)^{-1} \qquad (3) \\ &\qquad + \sum_{i,\, j:\, d(i,\, j) = 2} (h_{ij})(d_i + 1)^{-1}(d_j + 1)^{-1}(d_i + d_j - h_{ij} + 2)^{-1}. \end{aligned}$$

The first two terms of (3) are equal to $\frac{1}{2}\Sigma_{i,\, j:\, d(i,\, j) = 1}((d_i + 1)^{-1} - (d_j + 1)^{-1})^2$. This sum is always positive and vanishes if and only if the graph $\mathscr{G}$ has constant degree. Noting that $d(i, j) = 2$ implies $h_{ij} \geq 1$, we have

$$\operatorname{Var} W \geq T(d + 1)^{-2}(2d + 1)^{-1} \geq cd^{\alpha}|\mathscr{V}|/12d^3.$$

We then complete the proof by applying Corollary 2 with $B = 1$. □

REFERENCES

BALDI, P. and RINOTT, Y. (1989). Asymptotic normality of some graph related statistics. *J. Appl. Probab.* **26** 171–175.

JANSON, S. (1988). Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs. *Ann. Probab.* **16** 305–312.

PETROVSKAYA, M. B. and LEONTOVICH, A. M. (1982). The central limit theorem for a sequence of random variables with a slowly growing number of dependencies. *Theory Probab. Appl.* **27** 815–825.

STEIN, C. (1986). *Approximate Computation of Expectations*. IMS, Hayward, Calif.

JPL 198-330
CALIFORNIA INSTITUTE OF TECHNOLOGY
4800 OAK GROVE DRIVE
PASADENA, CALIFORNIA 91109

DEPARTMENT OF MATHEMATICS C-012
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093