J. Appl. Prob. **26**, 171–175 (1989)
Printed in Israel

ASYMPTOTIC NORMALITY OF SOME GRAPH-RELATED STATISTICS

PIERRE BALDI,* *University of California, San Diego*
YOSEF RINOTT,** *Hebrew University*

Abstract

Petrovskaya and Leontovich (1982) proved a central limit theorem for sums of dependent random variables indexed by a graph. We apply this theorem to obtain asymptotic normality for the number of local maxima of a random function on certain graphs and for the number of edges having the same color at both endpoints in randomly colored graphs. We briefly motivate these problems, and conclude with a simple proof of the asymptotic normality of certain U-statistics.

CENTRAL LIMIT THEOREM; DEPENDENT VARIABLES; MAXIMA OF RANDOM FUNCTIONS; RANDOM COLORINGS; U-STATISTICS; NEURAL NETWORKS

1. Introduction: background and applications

Sums of dependent 0–1 (indicator) or bounded random variables arise often in statistics, random graphs theory and other areas of applied probability. Since such variables have all moments, it is natural to study the limiting distributions of their sums by computing moments, a task which may however lead to tedious calculations.

In this paper we study the asymptotic distribution of the number of maxima (or minima) of certain random functions on a hypercube, a problem which arises in connection with combinatorial optimization (Tovey (1985)), the study of neural networks (Baldi (1988)), and certain models in statistical mechanics (Derrida (1980), (1981), Gross and Mezard (1984)). An interpretation in terms of game theory will be briefly sketched. We also prove a central limit theorem for the distribution of the number of edges which have the same color at both endpoints in a randomly colored graph. This result can be applied to prove asymptotic normality of certain two-sample statistics based on nearest neighbor or minimal spanning tree graphs. Finally, we present a very simple proof of the asymptotic normality of U-statistics.

To these problems we apply the moment approach via the following theorem which is a variant of a recent result of Petrovskaya and Leontovich (1982). Rates of convergence will be studied elsewhere.

Received 9 October 1987; revision received 11 December 1987.

* Postal address: Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA.

** Postal address: Department of Statistics, Hebrew University, Jerusalem.

Work done while visiting the Department of Mathematics at UCSD.

Definition. The graph $G=(V,E)$ is said to be a *dependency graph* for the random variables $\{X_\alpha, \alpha \in V\}$ 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 $\{X_\alpha, \alpha \in A_1\}$ and $\{X_\alpha, \alpha \in A_2\}$ are independent.

Theorem 1. Let $\{X_{\alpha n}, \alpha \in V_n\}$ be random variables having a dependency graph $G_n=(V_n, E_n)$, $n=1,2,\cdots$. For α in V_n, let $L_{\alpha n}^{(k)}$ denote the number of connected subsets of V_n of cardinality at most k which contain α. Let $S_n = \Sigma_{\alpha \in V_n} X_{\alpha n}$, $\sigma_n^2 = \operatorname{Var} S_n < \infty$ and $|V_n| \to \infty$. Suppose that for all $k \geqq 3$

$$\sum_{\alpha \in V_n} L_{\alpha n}^{(k)} E\,|X_{\alpha n}|^k = o(\sigma_n^k). \tag{1}$$

Then

$$\frac{S_n - ES_n}{\sigma_n} \to \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

Remark. The proof in [14] yields a stronger result, namely convergence of all moments.

Corollary 2. Under the conditions of Theorem 1, let D_n denote the maximal degree of G_n and suppose $|X_{\alpha n}| \leqq B_n$ a.s. If instead of (1) we have

$$\frac{|V_n| D_n^2 B_n^3}{\sigma_n^3} \to 0 \quad \text{as } n \to \infty \tag{2}$$

then

$$\frac{S_n - ES_n}{\sigma_n} \to \mathcal{N}(0,1).$$

The proof of Corollary 2 and of the applications below will be given in Section 2.

Remark. We learnt while preparing this manuscript that this result can also be obtained as a special case of a theorem of Janson (1988). His proof is based on semi-invariants rather than moments.

Application A. Consider the n-dimensional cube $C_n = \{0,1\}^n$ and let $\{Y_{\alpha n}, \alpha \in C_n\}$ be i.i.d. continuous random variables. For $\alpha \in C_n$ define the 0–1 indicator variable $X_{\alpha n}$ by $X_{\alpha n}=1$ if and only if $Y_{\alpha n} > Y_{\beta n}$ for all $\beta \in C_n$ such that $H(\alpha,\beta)=1$, where $H(\alpha,\beta)$ denotes the Hamming distance between the vertices α, β of C_n. Thus $X_{\alpha n}=1$ if $Y_{\alpha n}$ is a 'local maximum'. (Note that the $Y_{\alpha n}$'s are used only to induce a random ranking on the vertices.) Let $S_n = \Sigma_{\alpha \in C_n} X_{\alpha n}$, that is S_n counts the number of local maxima on the cube.

Proposition 3.

$$\frac{S_n - ES_n}{(\operatorname{Var} S_n)^{1/2}} \to \mathcal{N}(0,1) \quad \text{as } n \to \infty. \tag{3}$$

More generally, we can replace C_n by $C_n^* = \{0,1,\cdots,s\}^n$ and (3) holds where now for $\alpha, \beta \in C_n^*$, $H(\alpha,\beta)$ denotes the number of coordinates (out of n) in which α and β differ. In game theory, this problem arises when we consider a game in which each of n players chooses one of the strategies $\{0, \cdots, s\}$, thus forming a vertex $\alpha \in C_n^*$. Then each player is paid $Y_{\alpha n}$. A local maximum in this case is called a Nash equilibrium of the (stochastic) game, hence S_n counts the number of equilibria. In a similar setting, Powers (1986) obtained Poisson limit theorems when the number of strategies goes to infinity.

Application B. Let $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{E}_n)$ be a sequence of multigraphs (allowing more than one edge between a pair of vertices). Assume that each vertex is independently assigned one of two colors with probability p and $q = 1 - p$ respectively. Let T_n denote the number of edges whose endpoints have the same color. Let Δ_n denote the maximal degree of $\mathcal{G}_n$. Poisson limit theorems for such models are given in Janson (1986).

Proposition 4. If

$$\Delta_n^4 = o(|\mathcal{E}_n|) \tag{4}$$

then

$$\frac{T_n - ET_n}{(\mathrm{Var}\, T_n)^{1/2}} \to \mathcal{N}(0, 1).$$

In the case where $\mathcal{G}_n$ is a k-nearest-neighbor multigraph or a minimal spanning tree, Proposition 4 is closely related to results of Schilling (1986), Henze (1988) and Friedman and Rafsky (1979) on the asymptotic normality of two-sample test statistics. For these graphs (4) holds because both the k-nearest neighbor and minimal spanning tree graphs in R^m have a bounded maximal degree (with bound depending on m).

Application C. Let $X_1, \cdots, X_n$ be i.i.d. random variables. Given a symmetric (invariant under permutations of the arguments) function $h(x_1, \cdots, x_m)$ of $m \leqq n$ variables, the corresponding U-statistics are defined as

$$U_n = U(X_1, \cdots, X_n) = \binom{n}{m}^{-1} \sum h(X_{i_1}, \cdots, X_{i_m})$$

where the sum extends over all $\binom{n}{m}$ subsets of indices from $\{1, \cdots, n\}$. U-statistics and their asymptotic distribution were introduced and studied by Hoeffding (1948). We demonstrate the utility of Corollary 2 by giving a very simple proof of the asymptotic normality of bounded U-statistics. This approach can be extended to obtain easy proofs of asymptotic normality for generalized U-statistics and V-statistics (see Serfling (1980), Chapter 5), and to the cases when h is not bounded or is allowed to depend on n. For an example of the latter type see Abramson and Goldstein (1987).

Proposition 5. Suppose $X_1, \cdots, X_n$ are i.i.d. random variables and h is a symmetric bounded function satisfying $\mathrm{Var}\, E\{h(X_1, \cdots, X_m) \mid X_1\} > 0$. Then for any fixed m

$$\frac{U_n - EU_n}{(\mathrm{Var}\, U_n)^{1/2}} \to \mathcal{N}(0, 1) \quad \text{as } n \to \infty.$$

2. Proofs of applications

Proof of Corollary 2. First observe that $L_{\alpha n}^{(k)} \leqq (k-1)! D_n^{k-1}$. To see this construct a connected set containing α inductively by choosing at the mth step a vertex from the $m < k$ vertices already in the set and adding an adjacent vertex. For each such step there are at most mD_n possibilities, $m = 1, \cdots, k-1$. Therefore

$$\frac{1}{\sigma_n^k} \sum_{\alpha \in V_n} L_{\alpha n}^{(k)} E\,|X_{\alpha n}|^k \leqq \frac{(k-1)!}{\sigma_n^k}\,|V_n|\,D_n^{k-1} B_n^k. \tag{5}$$

For $k = 3$, the right-hand side converges to 0 by (2). Also $D_n \leqq |V_n|$ implies $(D_n B_n/\sigma_n)^3 \leqq |V_n| D_n^2 B_n^3/\sigma_n^3 \to 0$, so that $D_n B_n/\sigma_n \to 0$, implying together with (2) that the right-hand side of (5) converges to 0 for all $k \geqq 3$ and application of Theorem 1 completes the proof.

Proof of Proposition 3. We first prove Proposition 3 for the cube C_n. For vertices $\alpha, \beta \in C_n$ satisfying $H(\alpha, \beta) > 2$, $X_{\alpha n}$ and $X_{\beta n}$ are independent, and in applying Corollary 2 we take $V_n = C_n$ and the edge set E_n consists of edges between vertices α, β such that $H(\alpha, \beta) \leqq 2$. Hence $D_n = n + \binom{n}{2} \leqq n^2$. Notice that $X_{\alpha n} = 1$ requires $Y_{\alpha n} > Y_{\beta n}$ for n other vertices β and since the Y's are i.i.d. (any exchangeable Y's would give the same result), $P(X_{\alpha n} = 1) = 1/(n+1)$. If $H(\alpha, \beta) = 1$ we have $EX_{\alpha n} X_{\beta n} = 0$. Hence $\operatorname{Cov}(X_{\alpha n}, X_{\beta n}) = -1/(n+1)^2$. If $H(\alpha, \beta) = 2$ then an elementary calculation yields $EX_{\alpha n} X_{\beta n} = 1/n(n+1)$ and combining these results we obtain $E\,\Sigma_{\alpha \in C_n} X_{\alpha n} = 2^n/(n+1)$, and $\sigma_n^2 = \operatorname{Var} \Sigma_{\alpha \in C_n} X_{\alpha n} = 2^{n-1}(n-1)/(n+1)^2$. Since $|V_n| = |C_n| = 2^n$ and $|X_{\alpha n}| \leqq 1$, condition (2) of Corollary 2 follows from the fact that $2^n n^4[(2^{n-1}(n-1)/(n+1)^2)]^{-3/2} \to 0$.

For the case of C_n^* we obtain by similar calculations $E\,\Sigma_{\alpha \in C_n^*} X_{\alpha n} = (s+1)^n/(sn+1)$ and $\sigma_n^2 = \operatorname{Var} \Sigma_{\alpha \in C_n^*} X_{\alpha n} = (s+1)^n s(n-1)/2(sn+1)^2$ while $D_n \leqq (sn)^2$ and Corollary 2 applies.

Proof of Proposition 4. In order to apply Corollary 2, we construct a dependency graph $G_n = (V_n, E_n)$. The vertex set of G_n is the edge set of $\mathscr{G}_n$, i.e., $V_n = \mathscr{E}_n$. For $\alpha = (v, w) \in \mathscr{E}_n$, let $X_{\alpha n}$ be the 0–1 indicator variable defined by $X_{\alpha n} = 1$ if and only if v and w have the same color, whence $T_n = \Sigma_{\alpha \in V_n} X_{\alpha n}$. Then for $\alpha = (v, w)$, $\beta = (r, s) \in V_n$, $X_{\alpha n}$ and $X_{\beta n}$ are independent unless α and β have a common vertex in $\mathscr{V}_n$, i.e., $\{v, w\} \cap \{r, s\} \neq \varnothing$. Thus, for a given $\alpha = (v, w)$ there are at most $2(\Delta_n - 1)$ pairs of the form (v, s) or (r, w), and if we construct E_n by connecting all these pairs to α we obtain a dependency graph $G_n = (V_n, E_n)$ for the variables $\{X_{\alpha n}\}$ with maximal degree $D_n = 2(\Delta_n - 1)$. In order to apply Corollary 2, we now compute $\sigma_n^2 = \operatorname{Var} \Sigma_{\alpha \in V_n} X_{\alpha n}$. First notice that $EX_{\alpha n} = p^2 + q^2$ and that $\operatorname{Var} X_{\alpha n} = (p^2 + q^2)2pq$. In the multigraph $\mathscr{G}_n$, suppose $\alpha, \beta \in \mathscr{E}_n$ are two edges connecting the same pair of vertices. Then $X_{\alpha n} = X_{\beta n}$ so that $\operatorname{Cov}(X_{\alpha n}, X_{\beta n}) = (p^2 + q^2)2pq$. Let $m(n)$ be the number of such (unordered) pairs α, β in $\mathscr{E}_n$. Next let $\alpha, \beta \in \mathscr{E}_n$ have exactly one common vertex in $\mathscr{V}_n$. Then $EX_{\alpha n} X_{\beta n} = p^3 + q^3$. There are $\Sigma_{v \in \mathscr{V}_n} \binom{d_v}{2} - 2m(n)$ such pairs, where for $v \in \mathscr{V}_n$, d_v denotes the degree of the vertex v in the multigraph $\mathscr{G}_n$. If $\alpha, \beta \in \mathscr{E}_n$ have no common vertex then $\operatorname{Cov}(X_{\alpha n}, X_{\beta n}) = 0$. Combining these facts $\sigma_n^2 = |\mathscr{E}_n|(p^2 + q^2)2pq + 2m(n)(p^2 + q^2)2pq + [\Sigma_{v \in \mathscr{V}_n} d_v(d_v - 1) - 4m(n)][(p^3 + q^3) - (p^2 + q^2)^2]$. Since all the terms in the latter expression are non-negative, we obtain $\sigma_n^2 \geqq |\mathscr{E}_n|(p^2+q^2)2pq = c|\mathscr{E}_n|$ for some $c > 0$. Therefore $|V_n|D_n^2/\sigma_n^3 \leqq |\mathscr{E}_n|(2\Delta_n - 2)^2/\sigma_n^3 = O(\Delta_n^2/|\mathscr{E}_n|^{1/2})$. The right-hand side converges to 0 by (4), and the desired result follows from Corollary 2.

Proof of Proposition 5. We construct a dependency graph for the $\binom{n}{m}$ variables $\binom{n}{m}^{-1}h(X_{i_1},\cdots,X_{i_m})$. The vertices of the graph are subsets $\{i_1,\cdots,i_m\}$ of $\{1,\cdots,n\}$, hence $|V_n| = \binom{n}{m}$. Two vertices are connected by an edge if and only if their corresponding subsets have a non-empty intersection. Therefore the maximal degree of the graph satisfies $D_n \leqq m\binom{n-1}{m-1}$. Also, Var $U_n \geqq c/n$, for some constant $c > 0$ (see Serfling (1980), p. 183). If $|h(x_1,\cdots,x_m)| \leqq B$ then $B_n \leqq \binom{n}{m}^{-1}B$ and thus

$$|V_n|D_n^2(B_n/\sigma_n)^3 \leqq O\left(\binom{n}{m}\left[m\binom{n-1}{m-1}\right]^2\left[\binom{n}{m}^{-1}Bn^{1/2}\right]^3\right) \leqq O(n^{-1/2}) \to 0.$$

Proposition 5 follows now from Corollary 2.

References

Abramson, I and Goldstein L. (1987) Efficient nonparametric testing by functional estimation. Submitted for publication.

Baldi, P. (1988) Neural networks, orientations of the hypercube and algebraic threshold functions. *IEEE Trans. Information Theory* **34**(3).

Derrida, B. (1980) Random energy model; limit of a family of disordered models. *Phys. Rev. Letters* **45**, 79–82.

Derrida, B. (1981) Random energy model: an exactly solvable model of disordered systems. *Phys. Rev.* **B24**, 2613–2626.

Friedman, J. H. and Rafsky, L. C. (1979) Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**, 697–717.

Gross, D. J. and Mezard, M. (1984) The simplest spin glass. *Nuclear Phys.* **B240** [FS12], 431–452.

Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences, *Ann. Statist.* **16**, 772–783.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **18**, 293–325.

Janson, S. (1986) Birthday problems, randomly coloured graphs and Poisson limits of sums of dissociated variables. Uppsala University Department of Mathematics, Report 16.

Janson, S. (1988) Normal convergence by higher semi-invariants with applications to sums of dependent random variables and random graphs. *Ann. Prob.*

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 Prob. Appl.* **27**, 815–825.

Powers, I. Y. (1986) Three Essays on Game Theory. Ph. D. Dissertation, Yale University.

Schilling, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81**, 799–806.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Tovey, C. (1985) Hill climbing with multiple local optima. *SIAM J. Algebraic and Discrete Methods* **6**, 386–393.