J. Appl. Prob. **23**, 524–528 (1986)
Printed in Israel

BOUNDARY DOMINATION AND THE DISTRIBUTION OF THE LARGEST NEAREST-NEIGHBOR LINK IN HIGHER DIMENSIONS

J. MICHAEL STEELE,* *Princeton University*
LUKE TIERNEY,** *University of Minnesota*

Abstract

For a sample of points drawn uniformly from either the d-dimensional torus or the d-cube, $d \geqq 2$, we give limiting distributions for the largest of the nearest-neighbor links. For $d \geqq 3$ the behavior in the torus is proved to be different from the behavior in the cube. The results given also settle a conjecture of Henze (1982) and throw light on the choice of the cube or torus in some probabilistic models of computational complexity of geometrical algorithms.

COMPUTATIONAL GEOMETRY; SPIRAL SEARCH; EXTREME-VALUE DISTRIBUTION; GUMBEL DISTRIBUTION

1. Introduction

If $p_1, p_2, \cdots, p_n$ are n given points of $\mathbb{R}^d$, it is a basic problem of computational geometry to determine the set of nearest-neighbor linkages, i.e. to determine for each p_i which element of $\{p_1, p_2, \cdots, p_{i-1}, p_{i+1}, \cdots, p_n\}$ is nearest to p_i.

The work done on this problem from the point of view of computational complexity is quite extensive, but the works of Friedman et al. (1975), Lee et al. (1976), Friedman et al. (1977), and Bentley et al. (1980) provide a tracing of the basic development of the areas in terms of average-case behavior. From the point of view of worst-case behavior, basic contributions are made in Shamos (1978), Lipton and Tarjan (1977), and Zolnowsky (1978).

The length of the largest of the nearest-neighbor links is defined formally by

$$Z(p_1, p_2, \cdots, p_n) = \max_{1 \leqq i \leqq n} \min_{j: j \neq i} \|p_i - p_j\| \tag{1.1}$$

Received 19 July 1984; revision received 13 May 1985.

* Postal address: Department of Statistics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544, USA.

** Postal address: School of Statistics, 270 Vincent Hall, 206 Church St SE, University of Minnesota, Minneapolis, MN 55455, USA.

Research supported in part by NSF Contract DMS–8414069.

where $\|p-q\|$ denotes the usual Euclidean distance. This quantity comes up in almost all discussion of nearest-neighbor computations although its appearance is not always explicit.

The object of principal interest here is the sequence of random variables Z_n defined by

$$Z_n = Z(X_1, X_2, \cdots, X_n) \tag{1.2}$$

where the X_i are independent and uniformly distributed on either the d-cube $[0,1]^d$ or the d-torus obtained by identifying opposite faces of the d-cube. This random variable is closely related to a similar weighted nearest-neighbor variable which has been studied in Henze (1981), (1982), (1983).

There are three limit results which will be given and which provide an asymptotic understanding of Z_n. Because of the proximity of Henze's work only the last limit result will be proved in detail. Theorem 2 answers the conjecture of Henze (1982), cf. p. 354, item 5.

One practical implication of Theorem 2 is that in the modeling applications of computational geometry one would be well advised in many cases to work in the d-torus as opposed to the d-cube. To do otherwise, one risks having to be seriously concerned with counterintuitive boundary effects whenever $d \geqq 3$. The value of this advice might be particularly felt by those who would attempt to appraise the computational complexity of a nearest-neighbor procedure by means of simulation.

Theorem 1. For X_i, $1 \leqq i < \infty$, independent and uniformly distributed on the d-torus, one has

$$\lim_{n\to\infty} P(Z_n^d > (t + \log n)/n\omega_d) = 1 - \exp(-e^{-t}) \tag{1.3}$$

where ω_d is the volume of the unit sphere in $\mathbb{R}^d$, $d \geqq 2$.

For the d-cube the boundary begins to play a role for $d \geqq 3$, as the following results illustrate.

Theorem 2. For X_i, $1 \leqq i < \infty$ independent and uniform on $[0,1]^2$, one has

$$\lim_{n\to\infty} P(Z_n^2 \geqq (t + \log n)/\pi n) = 1 - \exp(-e^{-t}) \tag{1.4}$$

but for X_i, $1 \leqq i < \infty$, independent and uniform on $[0,1]^d$, $d \geqq 3$, one has

$$\lim_{n\to\infty} P(Z_n^d \geqq (t + \log n)/\omega_d n) = 1. \tag{1.5}$$

The weighted nearest-neighbor random variables studied by Henze are given by

$$Z'_n = \max_{1 \leqq i \leqq n} \min \left(\min_{j: j \neq i} \| X_i - X_j \|, \| X_i - \partial S \| \right)$$

where ∂S is the boundary of the d-cube. The proofs of (1.3) and (1.4) can be obtained by modification of the results of Henze (1982), and this modification will not be given here. The main goal is the proof of (1.5) which we give in Section 2.

Before quitting the introduction, it is worth noting that nearest-neighbor statistics have recently been studied from a different point of view in Bickel and Breiman (1983) and Shilling (1983a, b). These authors provide much information about *sums* of functions of nearest-neighbor link lengths and the application of such sums to the theory of goodness-of-fit tests.

2. Boundary behavior

We now give the proof of the limit relation (1.5) for the interesting case of $d \geqq 3$. We fix $\varepsilon > 0$ and choose for each n a sequence of $M(n)$ points y_i on the one-dimensional faces (i.e. the edges) of $[0,1]^d$ such that

$$\| y_i - y_j \| \geqq 2(1+\varepsilon) z_n \quad \text{for } i \neq j \tag{2.1}$$

where $z_n^d = (t + \log n)/n\omega_d$, and

(2.2) all of the y_i are at least a distance $(1+\varepsilon)z_n$ from the corners of the cube.

It is easy to check that we may choose $M(n)$ such that $M = M(n) \sim \alpha / z_n$ where α depends only on d.

We next let $C_{i,n}$ be the event that the ball $B(y_i, \varepsilon z_n) = \{x : \| x - y_i \| \leqq \varepsilon z_n\}$ contains exactly one of the points of $\{X_1, X_2, \cdots, X_n\}$ and the remainder of the ball $B(y_i, (1+\varepsilon)z_n)$ contains no further such points.

On setting $D_n = \bigcup_{i=1}^M C_{i,n}$ we see that (1.5) follows at once if we show $P(D_n) \to 1$. This will be done with help from the Poisson process.

We denote probabilities which are calculated with respect to a homogeneous Poisson process with rate n by a subscript π, and we calculate

$$P_\pi(D_n^c) = (1 - P_\pi(C_{i,n}))^M \tag{2.3}$$

and

$$\begin{aligned} P_\pi(C_{i,n}) &= n\varepsilon^d \omega_d z_n^d 2^{1-d} \exp(-n(1+\varepsilon)^d \omega_d z_n^d 2^{1-d}) \\ &= \varepsilon^d (\log n + t) 2^{1-d} \exp(-2^{1-d}(1+\varepsilon)^d (\log n + t)). \end{aligned} \tag{2.4}$$

Since $d \geqq 3$, we now see that $\varepsilon > 0$ can be chosen such that $(1+\varepsilon)^d < 2^{d-1}/d$, so using the fact that $M(n) \sim \alpha \omega_d^{1/d} (\log n)^{-1/d} n^{1/d}$ we obtain $P_\pi(D_n^c) \to 0$.

It remains to show that this last relationship also holds under the uniform model.

For each $1 \leqq i \leqq M$ we let K_i and L_i denote the number of sample points in $B(y_i, \varepsilon z_i)$ and $B(y_i, (1+\varepsilon)z_i) \setminus B(y_i, \varepsilon z_i)$ respectively. The event D_n depends only on these counts. We complete the proof of (1.5) by establishing the following result.

Lemma. For any event E_n which depends only on $\{K_i, L_i : 1 \leqq i \leqq M\}$ we have $P(E_n) - P_\pi(E_n) \to 0$ as $n \to \infty$.

Proof of the lemma. We denote the probability mass function of $K_1, K_2, \cdots, K_M$, $L_1, L_2, \cdots, L_M$ by g_n or by h_n accordingly as one uses the Poisson or the uniform model. The likelihood ratio is given by

$$R_n = \frac{h_n(K_1, K_2, \cdots, K_M, L_1, L_2, \cdots, L_M)}{g_n(K_1, K_2, \cdots, K_M, L_1, L_2, \cdots, L_M)}. \tag{2.5}$$

In order to show $P(E_n) - P_\pi(E_n) \to 0$, it will suffice to show that $R_n \to 1$ in probability under the uniform model. (For this reduction see Weiss (1969), pp. 261–262, or Weiss (1965), pp. 219–220.)

To write an explicit formula for the likelihood ratio R_n, we introduce the following notations:

$$p_n = \varepsilon^d \omega_d z_n^d 2^{1-d}, \quad q_n = [(1+\varepsilon)^d - \varepsilon^d]\omega_d z_n^d 2^{1-d}, \qquad r_n = p_n + q_n \tag{2.6}$$

and

$$U = U_n = \sum_{i=1}^{M} K_i, \quad V = V_n = \sum_{i=1}^{M} L_i, \qquad W = W_n = U_n + L_n. \tag{2.7}$$

This notation permits us to write

$$R_n = (n)_W n^{-W}(1 - Mr_n)^{n-W} \exp(nMr_n) \tag{2.8}$$

where $(n)_s$ denotes the falling factorial $n(n-1)(n-2)\cdots(n-s+1)$. The leading term $(n)_W n^{-W}$ can be most easily estimated by first noting

$$(n)_W n^{-W} = \prod_{k=1}^{W} (1 - k/n) \geqq 1 - W(W+1)/(2n). \tag{2.9}$$

Under the uniform model $W = W_n$ is just a binomial random variable with sample size n and success probability $M(1+\varepsilon)^d \omega_d z_n^d 2^{1-d} \beta (\log n/n)^{(d-1)/d}$ for a constant β. One can thus easily check that $W_n^2/n \to 0$ in probability, so $(n)_W^n n^{-W} \to 1$ in probability.

One can similarly express the remaining factors of R_n as

$$(1 - Mr_n)^{n-W} \exp(nMr_n) = \exp(n(Mr_n + \log(1 - Mr_n)) - W \log(1 - Mr_n)). \tag{2.10}$$

Since $Mr_n \to 0$ and $nM^2 r_n^2 \to 0$, and $W_n \log(1 - Mr_n) \to 0$ in probability we have completed the proof that $R_n \to 1$ in probability. This was all we needed to establish the main result expressed in Equation (1.5).

Acknowledgement

We should like to thank Jon Bentley for bringing this problem to our attention. We should also like to thank the referee for careful comments which have helped the exposition enormously.

References

BENTLEY, J. L. (1976) Divide and Conquer Algorithms for Closest Point Problems in Multidimensional Space. Unpublished Ph.D. Thesis, University of North Carolina.

BENTLEY, J. L., WEIDE, B. W. AND YAO, A. G. (1980) Optimal expected-time algorithms for closest point problems. *ACM Trans. Math. Software* **6**, 563–580.

BICKEL, P. J. AND BREIMAN, L. (1983) Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Prob.* **11**, 185–214.

FRIEDMAN, J. H., BASKETT, F. AND SHUSTEK, L. J. (1975) An algorithm for finding nearest neighbors. *IEEE Trans. Computers* **24**, 1000–1006.

FRIEDMAN, J. H., BENTLEY, J. L. AND FINKEL, R. A. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3**, 209–226.

HENZE, N. (1981) Ein asymptotischer Satz uber den maximalen Minimalabstand von unabhängigen Zufallsvektoren mit Anwendung auf einen Anpassungstest im $\mathbb{R}^p$ und auf der Kugel. Unpublished Doctoral Dissertation, University of Hannover.

HENZE, N. (1982) The limit distribution for maxima of 'weighted' r-th nearest-neighbor distances. *J. Appl. Prob.* **19**, 344–354.

HENZE, N. (1983) Ein asymptotisher Satz uber den maximalen Minimlabstand von unabhängigen Zufallsvektoren mit Anwendung auf einen Anpassungstest im $\mathbb{R}^p$ und auf der Kugel. *Metrika* **30**, 245–259.

LEE, R. C. T., CHIN, I. I. AND CHANG, S. C. (1976) Application of principal component analysis to multikey searching. *IEEE Trans. Software Engineering* **2**, 185–193.

LIPTON, R. AND TARJAN, R. E. (1977) Applications of a planar separator theorem. *18 Symp. Foundations of Computer Science*, IEEE, 162–170.

PAPADIMITRIOU, C. H. AND BENTLEY, J. L. (1980) Worst-case analysis of nearest neighbor searching by projection. Technical Report CMU–CS–80–109, Department of Computer Science, Carnegie-Mellon University.

SCHILLING, M. F. (1983a) Goodness of fit testing in $\mathbb{R}^m$ based on the weighted empirical distribution of nearest neighbor statistics. *Ann. Statist.* **11**, 1–12.

SCHILLING, M. F. (1983b) An infinite-dimensional approximation for nearest neighbor goodness of fit tests. *Ann. Statist.* **11**, 13–24.

SHAMOS, M. I. (1978) Computational Geometry. Unpublished Ph.D. Thesis, Yale University.

WEIDE, B. W. (1978) Statistical Methods in Algorithm Design and Analysis. Ph.D. Thesis, Carnegie-Mellon University.

WEISS, L. (1965) On asymptotic sampling theory for distributions approaching the uniform distribution. *Z. Wahrscheinlichkeitsth.* **4**, 217–221.

WEISS, L. (1969) The asymptotic joint distribution of an increasing number of sample quantiles. *Ann. Inst. Statist. Math.* **21**, 257–263.

YUVAL, G. (1976) Finding nearest neighbors. *Inf. Process. Lett.* **5**, 63–65.

ZOLNOWSKY, J. E. (1978) Topics in Computational Geometry. Ph.D. Thesis, Stanford University.