Adv. Appl. Prob. **14,** 56–67 (1982)
Printed in N. Ireland
0001-8678/82/010056-12$01.45

STEINHAUS'S GEOMETRIC LOCATION PROBLEM FOR RANDOM SAMPLES IN THE PLANE

DORIT HOCHBAUM,* *Carnegie-Mellon University*
J. MICHAEL STEELE,** *Stanford University*

Abstract

Let

$$M_n = \min_{S:|S|=[n\alpha]} \sum_{i=1}^{n} \min_{j\in S} \|X_i - X_j\|^p$$

where X_i, $1 \leqq i \leqq n$, are i.i.d. and uniformly distributed in $[0, 1]^2$. It is proved that $M_n \sim cn^{1-p/2}$ a.s. for $1 \leqq p < 2$. This result is motivated by recent developments in the theory of algorithms and the theory of subadditive processes as well as by a well-known problem of H. Steinhaus.

LOCATION PROBLEM; k-MEDIAN PROBLEM; PROBABILISTIC ALGORITHM; SUBADDITIVE PROCESS; SUBADDITIVE EUCLIDEAN FUNCTIONAL

1. Introduction

The work of Steinhaus (1956) was apparently the first explicit treatment of the natural question 'How should one choose n points from a mass distributed in the plane so as to best represent the whole?' The main objective of this article is to treat a stochastic analogue of Steinhaus's problem.

One principal motivation for this stochastic analogue comes from developments in the theory of algorithms. The first of these is the discovery by Karp (1977) of an efficient probabilistic algorithm for solving the traveling salesman problem. The second development was the proof of Papadimitriou (1981) of the conjecture of Fisher and Hochbaum (1980) that the 'Euclidean k-median location problem' is NP-complete.

More will be said about these algorithmic considerations in a later section, but we should first state our results. For any $x_i \in \mathbb{R}^2$ and integers k and n we define

$$(1.1) \qquad M(k; x_1, x_2, \cdots, x_n) = \min_{S:|S|=k} \sum_{i=1}^{n} \min_{j\in S} \|x_i - x_j\|.$$

Received 11 November 1980; revision received 31 March 1981.

* Postal address: Graduate School of Industrial Administration, Carnegie–Mellon University, Pittsburgh, PA 15213, U.S.A.

** Postal address: Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

Research supported in part by the Office of Naval Research Contract N00014-76-C-0475.

(Here the minimum is over all $S \subset \{x_1, x_2, \cdots, x_n\}$ such that S has cardinality $|S| = k$.)

In words, one views S as a set of k 'centers', and each 'site' x_i is served by its nearest center $x_j \in S$ at a cost equal to $\|x_i - x_j\|$. The quantity $M(k; x_1, x_2, \cdots, x_n)$ is therefore the minimal cost attainable by an optimal choice of k centers. This language is chosen in sympathy with the applications which have been suggested in Bollobás (1973), Cornuejols, Fisher and Nemhauser (1978), and Starret (1974).

Our main result is the following.

Theorem 1. If X_i, $1 \leqq i < \infty$, are independent and uniform on $[0, 1]^2$, then for any $0 < \alpha < 1$ one has

$$\lim_{n \to \infty} n^{-1/2} M([n\alpha]; X_1, X_2, \cdots, X_n) = C_\alpha \tag{1.2}$$

with probability 1 for some constant $0 < C_\alpha < \infty$.

For reasons which will be discussed in the last section it will be useful to generalize this result slightly. For any $1 \leqq p < \infty$ we define

$$M_p(k; x_1, x_2, \cdots, x_n) = \min_{S:|S|=k} \sum_{i=1}^{n} \min_{j \in S} \|x_i - x_j\|^p. \tag{1.3}$$

Under the same hypotheses as in Theorem 1 we shall prove the following.

Theorem 2. For any $1 \leqq p < 2$ and $0 < \alpha < 1$ we have with probability 1

$$\lim_{n \to \infty} M_p([\alpha n]; X_1, X_2, \cdots X_n)/n^{1-p/2} = C_{\alpha,p}$$

for some constant $0 < C_{\alpha,p} < \infty$.

Naturally, it will suffice to prove just Theorem 2, and this proof will occupy the next three sections. In Section 2 we concentrate on the key combinatorial observations which will make the theorem possible. Section 3 contains a Tauberian argument which partially parallels that used in Steele (1981b), but with the significant difference that the present problem lacks the monotonicity previously relied on in the theory of sub-additive Euclidean functionals (Steele (1981a,b)). The elementary Lemma 3.4 on the 'differentiation' of an asymptotic series is one of the devices introduced here which may prove useful in other non-monotone problems.

The proof of Theorem 2 is completed in Section 4 with the help of the Efron–Stein jackknife inequality teamed with the combinatorial lemmas of Section 2 and classical arguments.

The last section briefly discusses the algorithmic application of this result. We also discuss the extension of our results to non-uniformly distributed random variables, and comment briefly on some open problems.

2. Combinatorial and geometric lemmas

Most of the observations which are special to the location problem are contained in this section, but one can find some hidden generality since most subadditive Euclidean functionals have properties analogous to those that follow.

Lemma 2.1. There is a constant $\beta=\beta_p$ such that for all $x_i\in[0,1]^2$, $1\leqq i\leqq n$, and $k\geqq 2$

$$M_p(k;x_1,x_2,\cdots,x_n)\leqq \beta nk^{-p/2}. \tag{2.1}$$

Proof. Let an integer s be chosen so that $s^2\leqq k\leqq (s+1)^2$, and divide $[0,1]^2$ into s^2 cells of side s^{-1}. For each occupied cell C_i choose one element $x_i'\in C_i$ and set $S'=\{x_i':1\leqq i\leqq s^2\}$. Since each cell has diameter $\sqrt{2}s^{-1}$ and since $|S'|\leqq k$, we have the generous bound

$$M_p(k;x_1,x_2,\cdots,x_n)\leqq n(\sqrt{2}s^{-1})^p\leqq 2^{p/2}n(k^{\frac{1}{2}}-1)^{-p}$$

which yields the lemma.

The next lemma will provide the combinatorial linchpin required in our variance bounds.

Lemma 2.2. There is a constant $\beta'=\beta_p'$ such that for all $x_i\in\mathbb{R}^2$, $1\leqq i\leqq n$, and $k\geqq 2$ the difference

$$M_p(k;x_1,x_2,\cdots,x_n)-M_p(k+1;x_1,x_2,\cdots,x_n)\equiv\Delta_p$$

satisfies the bounds

$$0\leqq\Delta_p\leqq\beta' nk^{-1-p/2}. \tag{2.2}$$

Proof. Without loss we can assume that all of the $\|x_i-x_j\|^p$ are different. For any set of optimal centers S associated with $M_p(k+1;x_1,x_2,\cdots,x_n)$ we define for each $i\in S$ the set

$$N(i)=\{j;\|x_j-x_i\|<\|x_j-x_k\|,\ \forall k\in S,\ k\neq i\}.$$

This gives the representation

$$M_p(k+1;x_1,x_2,\cdots,x_n)=\sum_{i\in S}\sum_{j\in N(i)}\|x_i-x_j\|^p.$$

We now note that

$$\#\{i\in S:\sum_{j\in N(i)}\|x_i-x_j\|^p>4M_p(k+1;x_1,x_2,\cdots,x_n)/(k+1)\}\leqq(k+1)/4 \tag{2.3}$$

and also that

$$\#\{i\in S:|N(i)|\geqq 4n/(k+1)\}\leqq(k+1)/4. \tag{2.4}$$

This implies that for H defined by

$$H=\left\{i\in S: \sum_{j\in N(i)} \|x_i-x_j\|^p \leqq 4\beta n(k+1)^{-p/2}/(k+1) \textit{ and } |N(i)|\leqq 4n/(k+1)\right\}.$$

We have by (2.3), (2.4), and Lemma 2.1 that $\#H\geqq(k+1)/2$.

This time we choose s so that $s^2<(k+1)/2\leqq(s+1)^2$ and again divide $[0,1]^2$ into s^2 cells of side s^{-1}. Some cell must contain two elements of H, say x_{i_1} and x_{i_2}.

We now define a suboptimal choice for the k centers by using $S'=S\backslash\{x_{i_1}\}$. We continue to serve all $\bigcup_{i\neq i_1}\{x_j : j\in N(i)\}$ as before but now serve the elements of $\{x_j : j\in N(i_1)\}$ by x_{i_2}. The cost of serving $\{x_j : j\in N(i_1)\}$ by x_{i_2} is

$$\begin{aligned}\sum_{j\in N(i_1)} \|x_j-x_{i_2}\|^p &\leqq \sum_{j\in N(i_1)} 2^p(\|x_j-x_{i_1}\|^p+\|x_{i_1}-x_{i_2}\|^p)\\ &\leqq 2^p\{4\beta n(k+1)^{-p/2-1}+(\sqrt{2}s^{-1})^p(4n)/(k+1)\}.\end{aligned} \tag{2.5}$$

For $k\geqq 2$ the last expression can be bounded by $\beta' nk^{-1-p/2}$; and, since (2.5) majorizes Δ_p, this proves the second inequality in (2.2). The first inequality is trivial, so the lemma is complete.

One often finds it useful to know that in an optimal allocation no site is very far from a center.

Lemma 2.3. There is a constant $\beta''=\beta''_p$ with the property that for any S with $|S|=k$ which minimizes $\sum_{j=1}^n \min_{i\in S}\|x_i-x_j\|^p$ we have

$$\delta_p\equiv\delta_p(k;x_1,x_2,\cdots,x_n)\equiv\max_{1\leqq j\leqq n}\min_{i\in S}\|x_i-x_j\|^p\leqq\beta''_p nk^{-1-p/2}.$$

Proof. If we take $S'=S\cup\{x'\}$ where x' is the element of $\{x_1,x_2,\cdots,x_n\}$ farthest from any element of S then we have

$$M_p(k+1;x_1,x_2,\cdots,x_n)\leqq M_p(k;x_1,x_2,\cdots,x_n)-\delta_p.$$

By Lemma 2.2 we then see that $\delta_p\leqq\beta' nk^{-1-p/2}$, so we can take $\beta''=\beta'$.

The next lemma will be useful in applying jackknife methods to obtain variance bounds. By the notation $\hat{x}_i$ we mean that x_i is to be omitted from the sample, thus $\{x_1,x_2,\cdots,\hat{x}_i,\cdots,x_n\}=\{x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_n\}$. Also, we let $1(y\leqq\delta)$ be 1 or 0 accordingly as $y\leqq\delta$ holds or not.

Lemma. Setting $\delta=\delta_p(k;x_1,x_2,\cdots,x_n)$ and $m_i=\min_{j:j\neq i}\|x_j-x_i\|^p$, we have

$$\begin{aligned}M_p(k;x_1,x_2,\cdots,\hat{x}_i,\cdots,x_n)\leqq{}&M_p(k;x_1,x_2,\cdots,x_n)\\ &+2^p\delta\sum_{j=1}^n 1(\|x_j-x_i\|^p\leqq\delta)\end{aligned} \tag{2.6}$$

and

$$M_p(k;x_1,x_2,\cdots,x_n)\leqq M_p(k;x_1,x_2,\cdots,\hat{x}_i,\cdots,x_n)+2^p(m_i+\delta). \tag{2.7}$$

Proof. To prove the first inequality let S be an optimal choice of k centers for $\{x_1, x_2, \cdots, x_n\}$. If $x_i \notin S$ the inequality is trivial, so suppose $x_i \in S$. We now choose a sub-optimal set S' of k centers for $\{x_1, x_2, \cdots, \hat{x}_i, \cdots, x_n\}$. First let

$$N = \{x_j : \|x_j - x_i\| < \|x_j - x_{i'}\|, \forall i' \in S, i' \neq i\}.$$

If $N \neq \{x_i\}$ take any $x' \in N \backslash \{x_i\}$ and set $S' = (S \backslash \{x_i\}) \cup \{x'\}$, but if $N = \{x_i\}$ just take any $x' \in \{x_1, x_2, \cdots, \hat{x}_i, \cdots, x_n\}$ and define S' as before.

We now serve $\{x_1, x_2, \cdots, \hat{x}_i, \cdots, x_n\}$ by S' as follows. If $N = \{x_i\}$ then each point is served by the same center as it was served by in S. If $N \neq \{x_i\}$ then the elements of $N \backslash \{x_i\}$ are served by x', and the others are served as before. The cost of this sub-optimal choice is bounded by

$$M_p(k; x_1, x_2, \cdots, x_n) + \sum_{j=1}^{n} \|x_j - x'\|^p 1(x_j \in N \backslash \{x_i\})$$
$$\leqq M_p(k; x_1, x_2, \cdots, x_n) + 2^p \delta \sum_{j=1}^{n} 1(\|x_j - x_i\|^p \leqq \delta).$$

The second inequality (2.7) is easier. Let S be an optimal choice of centers for $\{x_1, x_2, \cdots, \hat{x}_i, \cdots, x_n\}$ and note that x_i can be served by an element of S at a cost less than

$$2^p \Big(\min_{j : j \neq i} \|x_i - x_j\|^p + \max_{j : j \neq i} \min_{t \in S} \|x_t - x_j\|^p \Big) \leqq 2^p (m_i + \delta).$$

3. Regular expectations

For brevity we set $M_n = M_p([\alpha n]; X_1, X_2, \cdots, X_n)$. Our first objective is to show that $EM_n \sim cn^{p/2}$. The method begins as in the classical approach taken by Beardwood, Halton, and Hammersley (1959) in the study of the traveling salesman problem. As noted in the introduction, the main novelty here is due to the necessity of overcoming the fact that M_n fails to have the monotonicity $M_n \leqq M_{n+1}$. The impact of this non-monotonicity is even more strongly felt in the next section. (The desire to understand a subadditive Euclidean functional which failed to be monotone provided the second principle motivating this work.)

We now let Π denote a Poisson point process on $\mathbb{R}^2$. For any Borel $A \subset \mathbb{R}^2$, $\Pi(A)$ will consist of a set of N_A points uniformly distributed in A, where N_A is itself a Poisson random variable with mean $\lambda(A)$, the Lebesgue measure of A.

Lemma 3.1. Let $A = [0, t]^2$ and set $\phi(t) = EM_p([\alpha N_A]; \Pi(A))$, then for all integers $m \geqq 1$ we have

$$\phi(t) \leqq m^2 \phi(t/m). \tag{3.1}$$

Proof. Let A be divided into m^2 cells Q_i of side t/m, then by the suboptimality of local optimization we have

$$M_p([\alpha N_A]; \Pi(A)) \leqq \sum_{i=1}^{m^2} M_p([\alpha N_{Q_i}]; \Pi(Q_i)).$$

On taking expectations and using the homogeneity of Π we obtain (3.1).

Lemma 3.2. If $\phi(t)$ is any continuous function which satisfied (3.1) for all m then

$$\lim_{t\to\infty} \phi(t)/t^2 = \lim_{t\to\infty} \inf \phi(t)/t^2 \equiv C_{\alpha,p}. \tag{3.2}$$

Proof. By the continuity of $\phi(t)$ and the definition $C_{\alpha,p} \equiv \lim_{t\to\infty} \inf \phi(t)/t^2$ we can choose an interval (a, b) such that

$$\phi(t)/t^2 \leqq C_{\alpha,p} + \varepsilon \tag{3.3}$$

for all $t \in (a, b)$. By (3.1) we can conclude that also we have (3.3) for all $t \in \bigcup_{m=1}^{\infty} (ma, mb)$. Since $I_m \equiv (ma, mb)$ and I_{m+1} intersect for all $m \geqq a(b-a)^{-1}$, we see $\bigcup_{m=1}^{\infty} (ma, mb)$ contains $(m_0 a, \infty)$; and, therefore,

$$\limsup_{t\to\infty} \phi(t)/t^2 \leqq C_{\alpha,p} + \varepsilon,$$

which completes the proof.

Lemma 3.3. For $1 \leqq p < 4$ we have for $x \uparrow 1$ that

$$\sum_{n=1}^{\infty} (EM_n)x^n \sim C_{\alpha,p}\Gamma(2-p/2)/(1-x)^{2-p/2} \tag{3.4}$$

and

$$\sum_{k=1}^{n} EM_k \sim C_{\alpha,p}(2-p/2)^{-1}n^{2-p/2}. \tag{3.5}$$

Proof. Calculating $\phi(t)$ by conditioning and making a change of scale from $[0, t]^2$ to $[0, 1]^2$ shows that (3.2) can be written out as

$$\phi(t) = \sum_{n=1}^{\infty} t^p(EM_n)e^{-t^2}t^{2n}/n! \sim C_{\alpha,p}t^2. \tag{3.6}$$

By changing variables $t^2 = u$ we see that

$$\sum_{n=1}^{\infty} (EM_n)e^{-u}u^n/n! \sim C_{\alpha,p}u^{1-p/2}. \tag{3.7}$$

Now by the Abelian theorem for Borel summability (e.g. Doetsch (1943), p.

191) and the fact that $1-p/2>-1$ we have as $x \to 1$ that

$$\sum_{n=1}^{\infty} (EM_n - EM_{n-1})x^n \sim C_{\alpha,p}\Gamma(2-p/2)/(1-x)^{1-p/2}. \tag{3.8}$$

Multiplying by $(1-x)^{-1}$ then completes the proof of (3.4). Since $EM_n \geqq 0$ we get (3.5) from (3.4) by an immediate application of the Karamata Tauberian theorem (Feller (1971), p. 447).

We should now like to 'differentiate' (3.5) in order to obtain the asymptotics of EM_n. Fortunately, the next lemma shows that this is (just barely) legitimate.

Lemma 3.4. If $\sum_{k=1}^{n} m_k \sim cn^{\gamma}$ for $\gamma>1$ and $m_{k+1} \geqq m_k - Bk^{\gamma-2}$ for some B and all $k \geqq 1$ then

$$m_n \sim c\gamma n^{\gamma-1}. \tag{3.9}$$

Proof. Let $y>1$ be chosen and note that

$$\sum_{n \leqq k \leqq yn} m_k \geqq \sum_{n \leqq k \leqq yn} \left(m_n - B\sum_{j=n}^{k} j^{\gamma-2}\right) \geqq n(y-1)m_n - B\sum_{k=n}^{yn}\sum_{j=n}^{k} j^{\gamma-2}.$$

Dividing by n^{γ} and using the Euler–Maclaurin summation formula to handle the double sum gives

$$(y^{\gamma}-1)c \geqq (y-1)\limsup\,(m_n/n^{\gamma-1}) - (\gamma-1)^{-1}\{\gamma^{-1}(y^{\gamma}-1)-(y-1)\}B.$$

Next dividing by $y-1$ and letting $y \downarrow 1$ shows

$$\gamma c \geqq \limsup\,(m_n/n^{\gamma-1}).$$

In a completely analogous way one can show that $\limsup\,(m_n/n^{\gamma-1}) \geqq \gamma c$ by estimating the sum $\sum_{yn \leqq k \leqq n} m_k$ where $y<1$.

The next lemma is the main consequence of this section.

Lemma 3.5. For $1 \leqq p < 2$ we have $EM_n \sim C_{\alpha,p} n^{1-p/2}$ as $n \to \infty$.

Proof. We already have $\sum_{l=1}^{n} EM_l \sim C_{\alpha,p}(2-p/2)^{-1} n^{2-p/2}$. By Lemma 3.4 with $1<\gamma=2-p/2$ it suffices to show

$$EM_{l+1} \geqq EM_l - Bl^{-p/2} \tag{3.10}$$

for some B and all l. By Lemma 2.4 (with sample size $n+1$ and $\hat{X}_i = \hat{X}_{n+1}$) and by Lemma 2.2 (if $[(l+1)\alpha]>[l\alpha]$) we have

$$M_l \geqq M_l - 2^p\delta \sum_{j=1}^{l+1} 1(\|X_j - X_i\|^p \leqq \delta) - ([(l+1)\alpha]-[(l\alpha)])\Delta_p. \tag{3.11}$$

Here by Lemma 2.3, $\delta \leqq \beta_p'' l[l\alpha]^{-1-p/2}$; and by Lemma 2.2, $\Delta_p \leqq \beta' l[l\alpha]^{-1-p/2}$.

By elementary estimates we then see that

$$E\sum_{j=1}^{l+1} 1(\|X_j - X_i\|^p < \delta)$$

is bounded, so taking expectations in (3.11) yields (3.10).

4. Completion of the proof

The results of the preceding sections will now be brought together to prove Theorem 2. The only new tool required is the recent result of Efron and Stein (1981) which says that the jackknife estimate of variance is positively biased. Explicitly, we first suppose that $S(x_1, x_2, \cdots, x_{n-1})$ is any symmetric function of $n-1$ vectors x_i. For each i we set $S_i = S(x_1, x_2, \cdots, \hat{x}_i, \cdots, x_n)$ and also set $S_. = 1/n \sum_{i=1}^n S_i$. If X_i are any independent and identically distributed random vectors, the Efron–Stein jackknife inequality says that

$$\text{Var}\, S(X_1, X_2, \cdots, X_{n-1}) \leqq E\sum_{i=1}^{n} (S_i - S_.)^2. \tag{4.1}$$

We shall now apply this inequality with the aid of the combinatorial bounds of Section 2.

Lemma 4.1. If X_i, $1 \leqq i < \infty$ are independent and uniformly distributed on $[0, 1]^2$, then for a constant C not depending on n we have

$$\text{Var}\, M_p([\alpha n]; X_1, X_2, \cdots, X_n) \equiv \text{Var}\, M_n \leqq Cn^{1-p}. \tag{4.2}$$

Proof. We first note that if S is replaced by any other variable, the right side of (4.1) is only increased. Using (4.1) and Lemma 2.4 we now calculate (with $\delta = \delta_p([n\alpha]; X_1, X_2, \cdots, X_{n+1})$):

$$\begin{aligned}\text{Var}\, M_n &\leqq E\sum_{i=1}^{n+1} (M_p([\alpha n]; X_1, X_2, \cdots, \hat{X}_i, \cdots, X_{n+1}) \\ &\qquad - M_p([\alpha n]; X_1, X_2, \cdots, X_{n+1}))^2 \\ &\leqq E\sum_{i=1}^{n+1} \left(2^p\delta \sum_{j=1}^{n+1} 1(\|X_j - X_i\|^p \leqq \delta) + 2^p\delta + 2^p \min_{j:j\neq i} \|X_j - X_i\|^p\right)^2.\end{aligned} \tag{4.3}$$

Replacing δ by $\beta_p'' n[\alpha n]^{-1-p/2} = \rho_n$ will by Lemma 2.3 only increase the right side, so using Vinogradov's symbol to ignore irrelevant constants we have

$$\text{Var}\, M_n \ll nE\left(\rho_n^2\left(\sum_{j=1}^{n+1} 1(\|X_j - X_1\|^p \leqq \rho_n)\right)^2 + \rho_n^2 + \min_{j:j\neq 1} \|X_j - X_1\|^{2p}\right). \tag{4.4}$$

Now, since $\rho_n^2 \ll n^{-p}$, it is an elementary calculation to show

$$E\left(\sum_{j=1}^{n+1} 1(\|X_j - X_1\|^p \leqq \rho_n)\right)^2 \ll E\left(\sum_{j=1}^{n+1} 1(\|X_j - X_1\| \leqq n^{-\frac{1}{2}})\right)^2 \ll 1$$

and

$$E \min_{j:j\neq 1} \|X_j - X_1\|^{2p} \ll n^{-p}.$$

These bounds and (4.4) imply (4.2) and complete the lemma.

We now note that $\operatorname{Var}(M_n/n^{1-p/2}) \ll 1/n$ and that this bound is not sharp enough to automatically imply complete convergence. It is therefore necessary to resort to a subsequence and maximal argument to prove Theorem 2.

By the bound (4.2), Lemma (3.5), the Borel–Cantelli lemma and Chebyshev's inequality one can easily show

$$\lim_{n_k \to \infty} M_{n_k}/n_k^{1-p/2} = C_{\alpha,p} \quad \text{a.s.} \tag{4.5}$$

for the subsequence $n_k = [k^\gamma]$ for any $\gamma > 1$.

We now set

$$D_k = \max_{n_k \leqq n < n_{k+1}} |M_n/n^{1-p/2} - M_{n_k}/n_k^{1-p/2}|$$

and note that to complete the proof of the theorem it suffices to show $D_k \to 0$ a.s. For this it certainly suffices to show $E\sum_{k=1}^{\infty} D_k^2 < \infty$.

We set $a_n = |M_{n+1}/(n+1)^{1-p/2} - M_n/n^{1-p/2}|$ and note

$$a_n \ll |M_{n+1} - M_n|/n^{1-p/2} + M_n/n^{2-p/2}.$$

By Lemma 2.1 we have $M_n \ll n^{1-p/2}$, and if $[\alpha(n+1)] = [\alpha n]$ the same estimates used in (4.4) will show $E(M_{n+1} - M_n)^2 \ll n^{-p}$. If $[\alpha(n+1)] = [\alpha n] + 1$ we first note $M_{n+1} \leqq M_n$.

Now we can also check that M_n cannot be much bigger than M_{n+1}. Setting $k = [\alpha n]$ we have by Lemma 2.2 that

$$M_n \leqq M(k+1; X_1, X_2, \cdots, X_n) + \beta' n k^{-1-p/2}$$

and by Lemma (2.4) (and (2.3)) that

$$M(k+1; X_1, X_2, \cdots, X_n) \leqq M(k+1; X_1, X_2, \cdots, X_n, X_{n+1}) + 2^p \delta \sum_{i=1}^{n} (1\, \|X_i - X_{n+1}\|^p \leqq \delta)$$

where $\delta \equiv \beta_p''(n+1)(k+1)^{-1-p/2}$. Together these bounds and elementary calculations show in the case that $[\alpha(n+1)] = [\alpha n] + 1$ that one again has $E(M_{n+1} - M_n)^2 \ll n^{-p}$, and hence $Ea_n^2 \ll 1/n^2$.

The final calculation is that

$$ED_k^2 \leqq E\Big(\sum a_n\Big)^2 \leqq (n_{k+1}-n_k)\Big(\sum Ea_n^2\Big) \ll k^{\gamma-1}\sum n^{-2} \ll k^{-2},$$

where the three sums are each over the range $n_k \leqq n < n_{k+1}$.

This verifies $E\sum_{k=1}^{\infty} D_k^2 < \infty$ and completes the proof of Theorem 2, except for verifying that indeed $C_{\alpha,p} > 0$. To show this last fact we set

$$Z_n = \frac{1}{n}\sum_{1\leqq i\leqq j\leqq n} 1(\|X_i - X_j\| \leqq \beta/\sqrt{n})$$

and note that easy calculations show

$$EZ_n \to \beta^2\pi/2 \quad \text{as} \quad n\to\infty, \tag{4.6}$$

and

$$\operatorname{Var} Z_n \to 0 \quad \text{as} \quad n\to\infty. \tag{4.7}$$

Since M_n is the sum of $n-[n\alpha]$ elements of $S=\{\|X_i - X_j\|^p : 1\leqq i<j\leqq n\}$ we have

$$M_n \geqq (\beta n^{-\frac{1}{2}})^p \,.\, 1(nZ_n < (n-[n\alpha])/2)\,.\,(n-[n\alpha])/2. \tag{4.8}$$

Taking expectations in (4.8) we have

$$n^{p/2-1}EM_n \geqq \beta^p P(Z_n < (1-\alpha)/2)\,.\,(1-\alpha)/2. \tag{4.9}$$

For $\beta^2 < (1-\alpha)/\pi$, Equations (4.6), (4.7) and Chebyshev's inequality will suffice to show that the right side of (4.9) is bounded away from 0. This shows $C_{\alpha,p} > 0$ and completes the proof.

5. Algorithmic implications

The fact that the K-median problem has been proved by Papadimitriou (1981) to be NP-hard means that it is extremely unlikely that there is an efficient algorithm for calculating the optimal choice of αn centers from n sites (cf. Karp (1972)). Therefore, since the K-median problem occurs in a variety of practical contexts, it seems quite desirable to find efficient algorithms which are capable of providing approximate optimal center selections.

The results of this article take a step toward this by providing an estimate for the *value* of an optimal selection. This value can be used in the construction of approximately optimal probabilistic algorithms for the K-median in a manner which is completely parallel to the way the asymptotic optimal *value* provided by Beardwood, Halton, and Hammersley (1959) has been used by Karp (1977) in the study of the traveling salesman problem. One algorithm of this type for the K-median problem (but with $K<\log n$) has already been constructed in Fisher and Hochbaum (1981).

6. Concluding remarks and open problems

One of the motives for investigating the functional

$$M_p([\alpha n]; X_1, X_2, \cdots, X_n) = \min_{S:|S|=[n\alpha]} \sum_{i=1}^{n} \min_{j\in S} \|X_i - X_j\|^p$$

for general p is the trite observation that as $p\to\infty$ we have

$$M_p^{1/p} \to \min_{S:|S|=[n\alpha]} \max_{1\leqq i\leqq n} \min_{j\in S} \|X_i - X_j\| = H_n.$$

The functional H_n is of independent interest and it was hoped that the present methods might throw some light on its probabilistic behavior. We now believe that $n^{-\frac{1}{2}}H_n$ converges in distribution, but we have no idea how this might be proved. Since our methods seem to require $1\leqq p<2$ and are more pertinent to strong laws, an entirely new technique may be needed.

There are also basic open problems directly concerning $M_n = M_p([\alpha n]; X_1, X_2, \cdots, X_n)$. In particular, it seems almost certain that a result analogous to our Theorem 2 must hold when the X_i are independent, identically distributed, and bounded. The methods used in Steele (1981a,c) seem to fail to help in the location problem because of the difficulty of establishing the intermediate result for step densities.

Finally, there is the question of determining $C_{\alpha,p}$. This is usually hopeless, but perhaps not in this case. Fejes-Tóth (1959) was able to determine the analogous constant in the original Steinhaus problem, and McClure (1976) has been able to extend the work of Fejes-Tóth to other functionals and extremal problems.

References

Beardwood, J., Halton, H. J. and Hammersley, J. M. (1959) The shortest path through many points. *Proc. Camb. Phil. Soc.* **55,** 299–327.

Bollobás, B. (1973) The optimal arrangement of producers. *J. London Math. Soc.* (2) **6,** 417–421.

Cornuejols, G., Fisher, M. L. and Nemhauser, G. L. (1978) Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms. *Management Sci.* **23,** 789–810.

Doetsch, G. (1943) *Theorie und Anwendung der Laplace-Transformation.* Dover, New York.

Efron, B. and Stein, C. (1981) The jackknife estimate of variance. *Ann. Statist.* **9,** 586–596.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications,* Vol. II, 2nd edn. Wiley, New York.

Fejes-Tóth, L. (1959) Sur la représentation d'une population infinie par un nombre fini d'éléments. *Acta Math. Acad. Sci. Hungar.* **10,** 299–304.

FISHER, M. L. AND HOCHBAUM, D. S. (1980) Probabilistic analysis of the K-median problem. *Math. Operat. Res.* **5,** 27–34.

KARP, R. M. (1972) Reducibility among combinatorial problems. In *Complexity of Computer Computations*, ed. R. E. Miller and J. W. Thatcher, Plenum Press, New York, 85–104.

KARP, R. M. (1977) Probabilistic analysis of partitioning algorithms for the traveling salesman problem in the plane. *Math. Operat. Res.* **2,** 209–224.

MCCLURE, D. E. (1976) Characterization and approximation of optimal plane partitions. Technical Report, Division of Applied Mathematics, Brown University.

PAPADIMITRIOU, C. H. (1981) Worst case and probabilistic analysis of a geometric location problem. *SIAM J. Computing* **10,** 542–557.

STARRET, D. A. (1974) Principles of optimal location in a large homogeneous area. *J. Econom. Theory* **9,** 418–448.

STEELE, J. M. (1981a) Subadditive Euclidean functionals and non-linear growth in geometric probability. *Ann. Prob.* **9,** 365–376.

STEELE, J. M. (1981b) Optimal triangulation of random samples in the plane. *Ann. Prob.*

STEELE, J. M. (1981c) Growth rates of minimal spanning trees of random samples in space. *Z. Wahrscheinlichkeitsth.*

STEINHAUS, H. (1956) Sur la division de corps matériels en parties. *Bull. Acad. Polon. Sci.* **4,** 801–804.