The Annals of Applied Probability
1993, Vol. 3, No. 3, 794–801

A MATCHING PROBLEM AND SUBADDITIVE EUCLIDEAN FUNCTIONALS[1]

BY WANSOO T. RHEE

Ohio State University

A classical paper by Steele establishes a limit theorem for a wide class of random processes that arise in problems of geometric probability. We propose a different (and arguably more general) set of conditions under which complete convergence holds. As an application of our framework, we prove complete convergence of $M(X_1, \ldots, X_n)/\sqrt{n}$, where $M(X_1, \ldots, X_n)$ denotes the shortest sum of the lengths of $\lfloor n/2 \rfloor$ segments that match $\lfloor n/2 \rfloor$ disjoint pairs of points among $X_1, \ldots, X_n$, where the random variables $X_1, \ldots, X_n, \ldots$ are independent and uniformly distributed in the unit square.

1. Introduction. By L, we denote a real-valued function of the finite subsets of $\mathbb{R}^d$ (including $\varnothing$), where $d \geq 2$ will be fixed throughout the paper. Under natural conditions, Steele [4] proves that if $(X_i)_{i \geq 1}$ are independent and uniformly distributed in $[0,1]^d$, then there is a constant $\beta(L)$ such that

$$\lim_{n \to \infty} \frac{L(\{X_1, \ldots, X_n\})}{n^{(d-1)/d}} = \beta(L). \tag{1}$$

This result provides a unified proof for a number of theorems in geometric probability. Steele's result suffers, however, from two drawbacks. The first one is that the conditions on L are often not satisfied. The second is that it does not imply complete convergence, that is, that for all $\varepsilon > 0$,

$$\sum_{n \geq 1} P\left(\left| \frac{L(\{X_1, \ldots, X_n\})}{n^{(d-1)/d}} - \beta(L) \right| \geq \varepsilon \right) < \infty.$$

In most cases, neither drawback is too serious. Steele's argument can be adapted to a variety of conditions (but the charm of having a general theorem is lost in the process), and complete convergence can often be obtained through a straightforward use of martingales, as used, for example, in [2]. There are, however, cases (and in particular, the matching problem stated in the abstract) where complete convergence is not so immediate (as will be explained in Section 4). While developing an approach that works for this example, we realized that this approach would work under very general conditions and that it might be worthwhile to spell these out.

In the present note, we point out general conditions that are slightly different from Steele's (and in practice seem weaker), that imply complete

Received August 1992; revised February 1993.
[1]Supported in part by NSF Grant CCR-90-00611 and Dean's Fellowship, College of Business, Ohio State University.
AMS 1991 *subject classifications*. Primary 60D05; secondary 60G17
Key words and phrases. Matching problem, subadditive functionals, complete convergence.

convergence (and much more) and that hold for all the examples that this author is aware of.

We first list a number of conditions on L. (The names of these conditions have been chosen in order to be consistent with Steele's notation.)

A0. $L(\varnothing) = 0$.

A1. For every $\alpha > 0$ and every finite subset F of $\mathbb{R}^d$, we have $L(\alpha F) = \alpha L(F)$, where $\alpha F = \{\alpha x;\ x \in F\}$.

A2. For every $x \in \mathbb{R}^d$ and every finite subset F of $\mathbb{R}^d$, we have $L(F + x) = L(F)$, where $F + x = \{y + x;\ y \in F\}$.

The following is a (rather formal) weakening of condition A5 of [4].

A5′. There exists a constant C_1 with the following property. For $m = 2$ or $m = 3$, consider the partition $(Q_i)_{i \le m^d}$ of $[0,1]^d$ into m^d equal cubes. Then, for every finite subset F of $[0,1]^d$ that does not meet the boundary of the cubes Q_i, we have

$$L(F) \le \sum_{i \le m^d} L(F \cap Q_i) + C_1.$$

B1. There exists a constant C_2 such that for all finite subsets F, G of $[0,1]^d$, we have

$$|L(F \cup G) - L(F)| \le C_2 (\operatorname{card} G)^{(d-1)/d}.$$

We consider random variables $X_1, X_2, \ldots$ that are independent and uniformly distributed on $[0,1]^d$. We assume that $L(\{X_1, \ldots, X_n\})$ is measurable. Our main result is as follows.

THEOREM 1. *Under conditions* A0, A1, A2, A5′ *and* B1 *the following occurs*: *For some constant* $\beta(L)$, *we have*

$$\lim_{n \to \infty} \frac{EL(\{X_1, \ldots, X_n\})}{n^{(d-1)/d}} = \beta(L). \tag{2}$$

For some universal constant C, *we have, for all* $n \ge 1$, *and all* $t \ge 0$,

$$P\big(|L(\{X_1, \ldots, X_n\}) - EL(\{X_1, \ldots, X_n\})| \ge t\big) \le C \exp\left(- \frac{1}{Cn} \left(\frac{t}{C_2} \right)^{2d/(d-1)} \right). \tag{3}$$

It does not seem possible to obtain a rate of convergence in (2) without extra hypotheses. However, as a by-product of Theorem 1, we obtain complete convergence.

The paper is organized as follows. In Section 2, we compare our conditions with those of [4] and we explain why our conditions are essentially weaker. In Section 3, we prove Theorem 1. In Section 4, we analyze the matching problem that motivated these abstract results.

2. Comparing with Steele's conditions. By A2, $a = L(\{x\})$ is independent of $x \in \mathbb{R}^d$. A basic observation is as follows.

PROPOSITION 1. *Assume conditions A0, A1, A2, and A5′. Set* $a_2 = C_1/(2^{d-1} - 1)$. *Set* $a_1 = a + d2^{d-1}a_2$. *Then for each nonempty finite subset* F *of* $[0,1]^d$, *we have*

$$L(F) \le a_1(\operatorname{card} F)^{(d-1)/d} - a_2 \le a_1(\operatorname{card} F)^{(d-1)/d}. \tag{4}$$

PROOF. This is proved by induction over card F. First, we observe that the result holds when card $F = 1$, by definition of a_1, a.

Before we perform the induction step, we observe that, by concavity of the function $f(x) = x^{(d-1)/d}$, we have

$$\frac{1}{\operatorname{card} I}\sum_{i\in I} n_i^{(d-1)/d} \le \left(\sum_{i\in I}\frac{1}{\operatorname{card} I}n_i\right)^{(d-1)/d}$$

for all integers $(n_i)_{i\in I}$. Thus

$$\sum_{i\in I} n_i^{(d-1)/d} \le (\operatorname{card} I)^{1/d}\left(\sum_{i\in I} n_i\right)^{(d-1)/d}. \tag{5}$$

Assume now that (4) has been established for all sets G with card $G < n$, and consider $F \subset [0,1]^d$ with card $F = n$. Consider the partition $(Q_i)_{i\le 2^d}$ of $[0,1]^d$ into 2^d equal cubes. First, we observe that we can assume that F is not a subset of a certain Q_i. Indeed, by A1 and A2 it suffices to prove (4) for F' rather than F, where $F' = \alpha F + x$ and where $\alpha \ge 1$ is chosen as large as possible so that we can find $x \in \mathbb{R}^d$ for which $F' \subset [0,1]^d$. Thus, for $i \le 2^d$, we have $n_i = \operatorname{card}(F \cap Q_i) < n$, so that by A1 and A2 and the induction hypothesis, we have

$$L(F \cap Q_i) \le \tfrac{1}{2}\left(a_1 n_i^{(d-1)/d} - a_2\right)$$

whenever $n_i > 0$. Because $L(F \cap Q_i) = 0$ when $n_i = 0$ (by A0), by A5′ and (5) we have

$$L(F) \le \tfrac{1}{2}\left(\sum_{i\in I}\left(a_1 n_i^{(d-1)/d} - a_2\right)\right) + C_1,$$

where $I = \{i \le 2^n;\ n_i > 0\}$. Thus, by (5) we have, setting card $I = m$,

$$L(F) \le \frac{m^{1/d}}{2} n^{(d-1)/d} a_1 - \frac{m}{2}a_2 + C_1.$$

To conclude, it suffices to prove that the right-hand side is at most $a_1 n^{(d-1)/d} - a_2$ or, equivalently, replacing C_1 by its value $(2^{d-1} - 1)a_2$, that

$$\frac{a_2}{2}(2^d - m) \le a_1 n^{(d-1)/d}\left(1 - \frac{m^{1/d}}{2}\right).$$

Because $n \ge 1$, $a_1 \ge d2^{d-1}a_2$, it suffices to show that

$$2^d - m \le d2^d\left(1 - \frac{m^{1/d}}{2}\right).$$

Setting $m = x2^d$, this reduces to the elementary fact that $1 - x \le d(1 - x^{1/d})$ for $0 \le x \le 1$. □

REMARK. A consequence of Proposition 1 is that condition A6 of [4] (and thus condition A4) follows from A0, A1, A2 and A5′.

We consider now the following conditions:

A3. For every finite set $F \subset \mathbb{R}^d$ and every $x \in \mathbb{R}^d$, we have $L(F) \le L(F \cup \{x\})$.

A7. There exists a constant C_3 such that, for all finite subsets F_1, F_2 of $[0, 1]^d$, we have

$$L(F_1 \cup F_2) \le L(F_1) + L(F_2) + C_3.$$

PROPOSITION 2. *Conditions* A1, A2, A3, A5′ *and* A7 *imply condition* B1.

PROOF. By A3, A7 and Proposition 1, we have

$$\begin{aligned} L(F) &\le L(F \cup G) \le L(F) + L(G) + C_3 \\ &\le L(F) + a_1(\operatorname{card} G)^{(d-1)/d} + C_3 \\ &\le L(F) + (a_1 + C_3)(\operatorname{card} G)^{(d-1)/d}. \end{aligned}$$

□

We know of no natural example where A5′ holds but A7 fails. On the other hand, there are numerous examples where A3 fails [e.g., if $L(F)$ is the length of the minimum spanning true tree through F]. Thus, in practice, when conditions A1, A2 and A5′ hold, condition B1 is weaker than condition A3.

3. Proof of Theorem 1. The proof of (2) follows; for example, the argument of the proof of [3] Theorem 3.1 (which itself is a routine variation on the arguments of [4]) and we do not reproduce it here. To prove (3), the temptation is to use the information

$$|L(F \cup \{x\}) - L(F)| \le C \tag{6}$$

and martingale difference sequences as in [2]. For $d = 2$, this fails to give complete convergence and for $d \ge 3$ this gives a result weaker than (3) [more

precisely, as in (4), the exponent $2d/(d-1)$ has to be replaced by 2]. The basic observation is that the situation becomes much clearer if one thinks in an abstract way. The basic fact is as follows.

PROPOSITION 3. *Consider the set* $\Omega = ([0,1]^d)^n$ *and the uniform measure* μ *on* Ω. *Consider a subset* A *of* Ω, *with* $\mu(A) \geq \frac{1}{2}$. *For a point* $\bar{y} = (y_1, \ldots, y_n) \in \Omega$ (*where* $y_1, \ldots, y_n \in [0,1]^d$), *consider the Hamming distance*

$$\phi(\bar{y}) = \inf\{k; \exists\, (x_1, \ldots, x_n) \in A, \operatorname{card}\{i \leq n;\, x_i \neq y_i\} \leq k\}$$

of $\bar{y}$ *to* A.

Then, we have

$$\mu(\{\phi(\bar{y}) \geq t\}) \leq 4\exp\left(-\frac{t^2}{8n}\right). \tag{7}$$

COMMENT. This statement has been known for a long time by the specialists. The proof (that will be sketched below) is contained in [1], Section 7.9, page 36). Our contribution here is the recognition that this statement is the heart of the matter (rather than trying to use martingales directly).

SKETCH OF PROOF. (No attempt is made at obtaining sharp numerical constants.) Applying Azuma's inequality to the function $\phi(\bar{y})$ yields

$$\mu\left(\left\{\left|\phi(\bar{y}) - \int \phi\, d\mu\right| \geq u\right\}\right) \leq 2\exp\left(-\frac{u^2}{2n}\right). \tag{8}$$

Taking $u = \int \phi\, d\mu$, we see that because $\phi(\bar{y}) = 0$ for $\bar{y} \in A$, the right-hand side of (8) is greater than or equal to $\frac{1}{2}$. Thus

$$\tfrac{1}{2} \leq 2\exp\left(-\frac{\left(\int \phi\, d\mu\right)^2}{2n}\right)$$

that is, $\int \phi\, d\mu \leq \sqrt{2n \log 4}$. Thus, from (8),

$$\mu\left(\left\{\phi(\bar{y}) \geq u + \sqrt{2n \log 4}\right\}\right) \leq 2\exp\left(-\frac{u^2}{2n}\right)$$

so that, for $t \geq 2\sqrt{2n \log 4}$,

$$\begin{aligned}\mu(\{\phi(\bar{y}) \geq t\}) &\leq 2\exp\left(-\frac{\left(t - \sqrt{2n \log 4}\right)^2}{2n}\right) \\ &\leq 2\exp\left(-\frac{t^2}{8n}\right).\end{aligned}$$

Thus, for all t,

$$\mu(\{\phi(\bar{y}) \geq t\}) \leq 4\exp\left(-\frac{t^2}{8n}\right). \qquad \square$$

To apply Proposition 3, we denote by M a median of the random variable $Z = L(\{X_1, \dots, X_n\})$, so that $P(Z \le M) \ge \frac{1}{2}$. For an n-tuple $\bar{x} = (x_1, \dots, x_n)$ of Ω, we set $L(\bar{x}) = L(\{x_1, \dots, x_n\})$.

Consider the subset A of Ω that consists of the n-tuples $\bar{x} = (x_1, \dots, x_n)$ such that $L(\bar{x}) \le M$. The condition $P(Z \le M) \ge \frac{1}{2}$ means that $\mu(A) \ge \frac{1}{2}$.

We observe that

$$L(\bar{y}) \le M + 2C_2 \phi(\bar{y})^{(d-1)/d}. \tag{9}$$

Indeed, if $\bar{x} = (x_1, \dots, x_n) \in A$ and $\bar{y} = (y_1, \dots, y_n)$, setting

$$F = \cup\{x_i;\ x_i = y_i\},$$

by B1 we have

$$|L(\bar{x}) - L(F)| \le C_2 k^{(d-1)/d},$$
$$|L(\bar{y}) - L(F)| \le C_2 k^{(d-1)/d},$$

where $k = \operatorname{card}\{i;\ x_i \ne y_i\}$. Because $L(\bar{x}) \le M$, (9) follows. It follows from (9) that

$$P(Z \ge M + t) = \mu(\{\bar{y};\ L(\bar{y}) \ge M + t\})$$
$$\le \mu\left(\left\{\bar{y};\ \phi(\bar{y}) \ge \left(\frac{t}{2C_2}\right)^{d/(d-1)}\right\}\right).$$

From (7) we have

$$P(Z \ge M + t) \le 4 \exp\left(-\frac{t^{2d/(d-1)}}{8n(2C_2)^{2d/(d-1)}}\right).$$

A similar inequality holds for $P(Z \le M - t)$, as is seen by replacing A by the set of $\bar{x}$ such that $L(\bar{x}) \ge M$, so that

$$P(|Z - M| \ge t) \le 8 \exp\left(-\frac{t^{2d/(d-1)}}{8n(2C_2)^{2d/(d-1)}}\right). \tag{10}$$

Using the identity

$$E(h) = \int_0^\infty P(h \ge t)\, dt$$

whenever $h \ge 0$, we get, by a routine computation that $E|Z - M| \le Kn^{(d-1)/2d}C_2$ (where K is a universal constant), so that $|EZ - M| \le Kn^{(d-1)/2d}C_2$. Combining with (10), we have

$$P\left(|Z - EZ| \ge t + Kn^{(d-1)/2d}C_2\right) \le 8 \exp\left(-\frac{t^{2d/(d-1)}}{8n(2C_2)^{2d/(d-1)}}\right),$$

from which (3) follows.

It should be pointed out that (3) simply requires that the random variables $X_1, \dots, X_n$ be independent. On the other hand, to extend (2) to the nonuni-

form case, other conditions on L are apparently necessary (like, e.g., condition A8 of [4]).

4. Application to the matching problem. Consider a subset $F = \{x_1, \dots, x_n\}$ of $\mathbb{R}^d$. Set $m = \lfloor n/2 \rfloor$. Consider m disjoint pairs (= subsets of card 2) of F, say $(x_{i(1)}, x_{j(1)}), \dots, (x_{i(m)}, x_{j(m)})$ and the sum

$$\sum_{l \le m} d(x_{i(l)}, x_{j(l)}) \tag{11}$$

of the distance of the elements of these pairs. Define $L(F)$ as the infimum of the quantities (11) under all choices of pairs. Define $L(\emptyset) = 0$. It is obvious that A0, A1, A2 and A5′ hold (with $C_2 = 3^d\sqrt{d}$). It is also obvious that A3 does not hold, as is shown by the example $F = \{(0,0),(0,1)\}$, $x = (0, \varepsilon)$, so that $L(F) = 1$, $L(F \cup \{x\}) = \varepsilon$.

We now prove that B1 holds. First, we observe the inequality

$$L(F \cup G) \le L(F) + L(G) + \sqrt{d}$$

that is obtained by considering the union of an optimal matching of F and an optimal matching of G, and if both card F and card G are odd, by matching the two leftover points together. Combining with (4), we get

$$\begin{aligned} L(F \cup G) &\le L(F) + \sqrt{d} + a_1(\operatorname{card} G)^{(d-1)/d} \\ &\le L(F) + \left(\sqrt{d} + a_1\right)(\operatorname{card} G)^{(d-1)/d} \end{aligned} \tag{12}$$

For the reverse inequality, consider an optimal matching of $F \cup G$. Consider the set F_1 of points of F that are matched to points of G and let $F' = F \setminus F_1$. Surely we have card $F_1 \le$ card G, and obviously $L(F') \le L(F \cup G)$ [by considering the restriction of the optimal matching of $L(F \cup G)$ to F']. Using (12) for F' rather than F, F_1 rather than G completes the proof of B1.

We now explain why the usual martingale arguments apparently do not succeed in proving complete convergence. The reason is that it is very difficult to improve the trivial inequality

$$\begin{aligned} &\left| L(\{x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n\}) \right. \\ &\left. -L(\{x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n\}) \right| \le d(x_i, y_i). \end{aligned}$$

This relation is obtained by matching y_i to the same point x_j as x_i (and matching all the other points the same way) and using the triangle inequality

$$\left| d(x_i, x_j) - d(y_i, x_j) \right| \le d(x_i, y_i).$$

Presumably x_i is close to x_j; when $d(y_i, x_j) \gg 1/\sqrt{n}$, it is very inefficient to match y_i to x_j. However, if one wants to match y_i to a point x_l, then the point to which x_l was matched has to find a new partner (and so on). It is not clear how to control this chain reaction and indeed, there are configurations where this is impossible (e.g., where the points in each matched pair in the optimal matching of $\{x_1, \dots, x_n\}$ are very close, and the pairs at distance of order at least $n^{-1/2}$ to each other). While controlling the "chain reaction"

might be possible in the general case, an analysis of this situation is certainly going to be considerably more difficult than the present approach, and for the time being, we do not know how to improve upon (3).

Concerning (2), it does not seem possible to obtain a rate of convergence unless A5′ is replaced by a two-sided control. In the present case, such a two-sided estimate can be obtained by repeating the arguments of reference 3 (Theorem 2.6). Along the lines of Theorem 3.1 of reference 3, one then obtains (when $d = 2$) that if $F \subset [0,1]^2$ is generated by an homogeneous Poisson point process of intensity λ, then

$$\left| \frac{EL(F)}{\sqrt{\lambda}} - \beta(L) \right| \leq K \frac{\log \lambda}{\sqrt{\lambda}},$$

where K is a numerical constant and "de-Poissonization" using (3) yields

$$\left| \frac{EL(\{X_1, \ldots, X_n\})}{\sqrt{n}} - \beta(L) \right| \leq \frac{K}{n^{1/4}}$$

(for another numerical constant K). Combined with (3), this shows that for $d = 2$, all $n \geq 1$ and $u \geq 0$,

$$P\left(\left| L(\{X_1, \ldots, X_n\}) - \sqrt{n}\, \beta(L) \right| \geq u n^{1/4} \right) \leq K \exp\left(- \frac{u^4}{K} \right),$$

where K is (yet another) numerical constant.

REFERENCES

[1] MILMAN, V. and SCHECHTMAN, G. (1986). *Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Math.* **1200**. Springer, New York.

[2] RHEE, W. and TALAGRAND, M. (1987). Martingale inequalities and NP-complete problems. *Math. Oper. Res.* **12** 177–181.

[3] RHEE, W. (1992). Probabilistic analysis of a capacitated vehicle routing problem II. Unpublished manuscript.

[4] STEELE, J. M. (1981). Subadditive Euclidean functionals and nonlinear growth in geometric probability. *Ann. Probab.* **9** 365–376.

FACULTY OF MANAGEMENT SCIENCES
OHIO STATE UNIVERSITY
301 HAGERTY HALL
1775 COLLEGE ROAD
COLUMBUS, OHIO 43210-1399