Probability in the Engineering and Informational Sciences, 2, 1988, 143–156. Printed in the U.S.A.

PROBABILISTIC ANALYSIS OF A GREEDY HEURISTIC FOR EUCLIDEAN MATCHING

DAVID AVIS[1]
School of Computer Science
McGill University
Montréal, Québec
H3A 2K6, Canada

BURGESS DAVIS[2]
Department of Mathematics and Statistics
Purdue University
West Lafayette, Indiana 47907

J. MICHAEL STEELE[3]
Program in Statistics and Operations Research
Princeton University
Princeton, New Jersey 08544

Given a collection of n points in the plane, the Euclidean matching problem is the task of decomposing the collection into matched pairs connected by line segments in such a way as to minimize the sum of all the segment lengths. The greedy heuristic provides an approximate solution to the Euclidean matching problem by successively matching the two closest unmatched points. We study the behavior of G_n, the sum of the lengths of the segments produced by the greedy heuristic. In particular, it is proved that if the points are realized as n independent observations from a common distribution with compact support, then $G_n/\sqrt{n}$ converges to a constant with probability one.

[1]Research support by N.S.E.R.C. Grant A3013 and F.C.A.C. Grant EQ-1678.

[2]Research support in part by NSF Grant DMS-8500998.

[3]Research supported in part by NSF Grant DMS-8414069.

 0269-9648/88 $5.00 + .00

1. INTRODUCTION

For any set $\{x_1, x_2, \ldots, x_n\}$ of n points in the Euclidean plane, a *matching* is a collection of $\lfloor n/2 \rfloor$ disjoint pairs of points. By the *weight* of a matching we will denote the sum of the Euclidean distances between the elements of the pairs in the matching.

The best known methods for finding a minimal weight matching are implementations of Edmonds' algorithm and have running times proportional to n^3 (see, e.g., Papadimitriou and Steiglitz [10]). Because of the relative slowness of the Edmonds' algorithm, substantial attention has been given to heuristic methods for obtaining matchings which are almost optimal (see, e.g., Avis [2], Iri, Murota, and Matsui [7], Reingold and Supowit [12], and Reingold and Tarjan [13]).

One particularly appealing heuristic is the so-called *greedy algorithm* which successively matches the two closest unmatched pairs of points. Even naive implementations of the greedy heuristic are faster than Edmond's algorithm, but by making use of specialized data structure one can do much better. In particular, Bentley and Saxe [4] provided an implementation of the greedy algorithm which has a worst case running time of $0(n^{3/2}\log n)$.

If G_n and OPT_n denote the respective weights of the greedy and optimal matchings of an n-set $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^2$, there are several known relations between G_n and OPT_n (see Avis [1]). Among the most interesting is the ratio bound of Reingold and Tarjan [13],

$$G_n/\mathrm{OPT}_n \leq (4/3)n^{\log_2 1.5}. \tag{1.1}$$

By providing an explicit construction, Reingold and Tarjan [13] also established that inequality (1.1) cannot be improved. Still, as one might suspect, (1.1) is often too pessimistic, and in many problems one finds that G_n and OPT_n are of the same order.

In Papadimitriou [11], the behavior of OPT_n was studied for points $\{X_i\}$ which were chosen at random from $[0,1]^2$, and that paper outlined how the technique used in Beardwood, Halton, and Hammersley [3] to study the traveling salesman problem, could be modified to prove

$$\mathrm{OPT}_n \sim c_{\mathrm{OPT}}\sqrt{n} \tag{1.2}$$

with probability one.

One consequence of the main theorem proved here is that one also has

$$G_n \sim c_G\sqrt{n}$$

with probability one. The result to be proved provides convergence which is stronger than convergence with probability one and deals with random variables with general distributions in $\mathbb{R}^d$.

THEOREM 1.1: *For each integer $d \geq 2$, there is a positive constant k_d such that if $X_1, X_2, \ldots$ are independent and identically distributed random variables*

with values in $\mathbb{R}^d$ *and bounded support, and if* G_n *denotes the Euclidean edge weight of the matching attained by the greedy algorithm applied to* $\{X_1, X_2, \ldots, X_n\}$, *then given* $\epsilon > 0$

$$\sum_{n=1}^{\infty} P\left(\left| n^{(1-d)/d} G_n - k_d \int_{\mathbb{R}^d} f(x)^{(d-1)/d} dx \right| > \epsilon \right) < \infty.$$

Here, f *is the density with respect to d-dimensional Lebesgue measure of the absolutely continuous part of the distribution of the* X_i.

There are several observations one should make concerning Theorem 1.1 before attending to the proof. First, one should note that since the X_i have compact support, we always have

$$\int f(x)^{(d-1)/d} \, dx < \infty.$$

Second, the Borel–Cantelli Lemma applied to the conclusion of Theorem 1.1 implies that $n^{(1-d)/d} G_n$ converges with probability one. The stronger version of convergence guaranteed by Theorem 1.1 is called *complete convergence*, and it is precisely such convergence that is most relevant to the probabilistic analysis of algorithms like the partitioning method introducing by R.M. Karp for the traveling salesman problems (cf, Karp [8] and Karp and Steele [9]).

The third observation concerns the distributional assumptions on the $\{X_i\}$. If the support of the $\{X_i\}$ is purely singular (i.e., $P(X_i \in A) = 1$, where A is a set of Lebesque measure zero), then f is identically zero and Theorem 1.1 tells us that $G_n = o(n^{(d-1)/d})$ with probability one. Also, there is a small technical point which should not be ignored. If the $\{X_i\}$ have a nontrivial singular part and a nontrivial absolutely continuous part, then the associated density f is not a probability density, since it does not integrate to 1.

Finally, since the essential ideas behind the proof of Theorem 1.1 are already present when $d = 2$, we will restrict ourselves to that case in order to keep the notation uncluttered.

2. PRELIMINARY LEMMAS

Several nonprobabilistic results about the greedy matching algorithm will be needed. The best of these, Lemma 2.4, spells out a basic combinatorial smoothness of the greedy functional, but first we need some bound on the lengths of the edges chosen by the greedy heuristic.

LEMMA 2.1: *There is a constant* c_1 *such that for any square S of side s and any n points* $\{x_1, x_2, \ldots, x_n\} \subset S$, *we have* $|x_i - x_j| \leq c_1 s n^{-1/2}$ *for some* $1 \leq i < j \leq n$.

PROOF: If each x_i were covered by a disc with center at x_i and radius r, and, if all of the discs were nonintersecting, then each disc would cover at least

$\pi r^2/4$ of the square. The total area of the square covered by the discs then would be at least $n\pi r^2/4$, hence we have $n\pi r^2/4 \leq 1$ and establish the lemma with a crude $c_1 = 4\pi^{-1/2}$ (see also Few [6]). ■

LEMMA 2.2: *There is a constant c_2 such that any k edges of a greedy matching which lie completely in a subsquare S of $[0,1]^2$ of side s will have total weight at most $c_2\sqrt{k}\, s$.*

PROOF: List the edges $e_1, e_2, \ldots, e_k$ which lie in S in the reverse order from which they were chosen for the greedy matching. Note for any $1 \leq j \leq k$ that when edge e_j was chosen for the greedy matching, there were at least $2(k-j)$ of the X_i, namely, the endpoints of the rest of the e_i, which were yet unmatched in S. Thus, by Lemma 2.1, the length of e_j is bounded by $c_1 s(\sqrt{2(k-j)+2})^{-1}$, and the lemma follows, since $\sum_{j=1}^{k} (\sqrt{2(k-j)+2})^{-1} < 2\sqrt{k}$. ■

LEMMA 2.3: *For any $\delta > 0$, a greedy matching of $\{x_1, x_2, \ldots, x_n\} \subset [0,1]^2$ has at most $c_3\delta^{-2}$ edges with length as great as δ.*

PROOF: If there are τ edges as large as δ, then the total length of these edges is at least $\tau\delta$. Now, by Lemma 2.2, this total length of any τ edges of the greedy matching is bounded by $c_2\tau^{1/2}$. Hence, we have $\tau \leq c_2^2/\delta^2$, giving $c_3 \leq c_2^2$. ■

We now have the tools in place to prove the main result of this section.

LEMMA 2.4: *There is a constant c_4 such that $|G(x_1, x_2, \ldots, x_n) - G(x_1, x_2, \ldots, x_{n+k})| \leq c_4\sqrt{k}$.*

PROOF: First let $e_1, e_2, \ldots, e_s$ be a list of edges of the complete graph on $\{x_1, x_2, \ldots, x_{n+k}\}$, and suppose that the e_i are listed in order of increasing length. An algorithm which simultaneously and inductively constructs a greedy matching A of $\{x_1, x_2, \ldots, x_{n+k}\}$ and a greedy matching B of $\{x_1, x_2, \ldots, x_n\}$ can be given as follows: for $i = 1$ to s, do

(i) if e_i has no endpoint in common with the edges of A, add e_i to A,

(ii) if e_i has no endpoint in common with the edges of B *and* no endpoint in $D = \{x_{n+1}, x_{n+2}, \ldots, x_{n+k}\}$, add e_i to B.

We now need to recall some of the structure which comes with a union of two matchings like A and B. In the first place, an edge can be placed in both A and B (such edges will be called *double edges* of $A \cup B$). Also, the union of A and B may have an important structure called an *alternating path*. Here, an alternating path is a path whose edges are alternately in the matchings A and B. For example, the path through the vertices $v_1, v_2, \ldots, v_k$ would be called an alternating path in $(v_i, v_{i+1}) \in A$ for odd i and $(v_i, v_{i+1}) \in B$ for even i (or vice

versa). By default, any path having only one edge will be considered as a *bona fide* alternating path. (Alternating paths are a basic tool of matching theory, see, e.g., Bondy and Murty ([5], pp. 70–79.)

In order to make the structure of $A \cup B$ explicit, we will verify by induction the following hypothesis.

At iteration i, all *non-double* edges of $A \cup B$ are either

(a) edges with an endpoint in D, or

(b) edges which belong to at least one alternating path terminating at a vertex in D. Moreover, this alternating path has edges with monotone decreasing length as one traverses the path toward the edge which meets D.

We will now verify the induction process. The hypothesis is valid for $i = 1$, so we assume the hypothesis is valid for i and consider edge e_{i+1} which we can assume without loss to be a non-double edge contained in $A \cup B$. If e_{i+1} has an endpoint in D, condition (a) holds and the induction hypothesis is verified for $i + 1$. We are left with two cases:

Case 1: Suppose $e_{i+1} \in A$ but $e_{i+1} \notin B$, and e_{i+1} has no endpoint in D. Since $e_{i+1} \notin B$, one endpoint of e_{i+1} must meet an edge e_j, $j \leq i$, which is already in B but which is not in A. Since $j \leq i$ and the edges are ordered, we have $|e_j| \leq |e_i|$. Also, by the induction hypothesis, e_j belongs to an alternating path of $A \cup B$ having monotone edge lengths and terminating with an edge which has a vertex in D. Therefore, in this case, the induction hypothesis is verified.

Case 2: This concerns the possibility that $e_{i+1} \in B$ but $e_{i+1} \notin A$ and e_{i+1} does not meet D. The induction step is identical to that of Case 1.

To complete the proof of the lemma, we have to bound the absolute value of the difference $\sum_{e \in A} |e| - \sum_{e \in B} |e|$. Since the edges which are in both A and B will cancel, we only need to estimate the difference over those edges which belong to an alternating chain (of length ≥ 1) with termination in D. If C is any such chain, we have that the difference

$$\Delta_C = \sum_{e \in A \cap C} |e| - \sum_{e \in B \cap C} |e|$$

is bounded in absolute value by $\max \{|e| : e \in C\}$, since the edge lengths of the edges in C are monotone. Moreover, each chain C must end in D and each vertex of D is in at most one edge of A, so the total number of alternating chains C is at most $|D|$. Since no two chains share an edge, we have that the sum $\sum_C |\Delta_C|$ is bounded by the sum of $M \leq |D| = k$ edges. If M_A of these edges belong to A and M_B belong to B, then we see by Lemma 2.2 that

$$\sum_C |\Delta_C| \le c_2(M_A^{1/2} + M_B^{1/2}). \tag{2.1}$$

Since $M_A + M_B \le k$, we see that Eq. (2.1) completes the proof with $c_4 = 2c_2$. ∎

With this basic lemma established, we are in a position to relate local properties of the greedy matching to global properties of the matching.

In the subsequent analysis, it will be essential to know how breaking the unit square $Q = [0,1]^2$ into m^2 equal subsquares Q_i, $1 \le i \le m^2$, changes a greedy matching. For any set $K \subset [0,1]^2$, we let ∂K denote the boundary of K, i.e., the points in the closure of K and K^c. Next we let $F = \bigcup_{i=1}^{m^2} \partial Q_i - \partial Q$, i.e., F is the interior *grating* of the partition of Q given by the Q_i. For $\delta > 0$, we will let F^δ denote the set of points of Q which are within δ of F, i.e., F^δ is an interior grating thickened by δ. Similarly, we let $Q_i^\delta = Q_i - F^\delta$.

The same method used in Lemma 2.4 can be used to relate the greedy matchings on the decomposition of Q. By simultaneously constructing all $m^2 + 1$ matchings of the Q_i^δ, $1 \le i \le m^2$, and Q, we obtain the following:

LEMMA 2.5: *Let A denote the greedy matching of $\{x_1, x_2, \ldots, x_n\} \subset Q$ and let B be the union of the m^2 greedy matchings of $\{x_1, x_2, \ldots, x_n\} \cap Q_i^\delta$, $1 \le i \le m^2$. The union $A \cup B$ consists of (a) double edges, i.e., edges belonging to both A and B, and (b) edges which are part of a monotone alternating path which contains an edge of A which either (i) has an endpoint in F^δ or (ii) has endpoints in two different Q_i^δ.*

Here, one should recall that some of the so-called alternating chains may be of length one, i.e., chains consisting of a single edge are still regarded as honest alternating chains.

The next lemma shows that despite the *a priori* global nature of the greedy matching process, it can be approximately localized. To express this precisely, we let $G(S)$ denote length of the greedy matching of $\{X_1, X_2, \ldots, X_n\} \cap S$ and allow the dependence of $G(S)$ on n to be implicit.

LEMMA 2.6: *Given $\delta > 0$, let τ denote the number of edges in the greedy matching of Q which have either (i) an endpoint within δ of F, or (ii) length at least as great as 2δ. One then has the bound,*

$$\left| G(Q) - \sum_{i=1}^{m^2} G(Q_i) \right| \le c_5 \tau^{1/2}. \tag{2.2}$$

PROOF: We let A and B denote the matchings given in Lemma 2.5, and we note that by the same considerations detailed in Lemma 2.4 the absolute value of the difference $\sum_{e \in A} |e| - \sum_{e \in B} |e|$ is majorized by

$$\sum_C \max\{|e| : e \in C\}. \tag{2.3}$$

The total number of distinct edges in the sum given above is bounded by τ, so arguing as in Eq. (2.1) of Lemma 2.4, we can bound the sum given by Eqs. (2.3) by $c_2\tau^{1/2}$, establishing the lemma with $c_5 = c_2$. ∎

3. UNIFORMLY DISTRIBUTED CASE

For any two sequences of random variables $\{A_n\}$ and $\{B_n\}$, which are defined on the same probability space, we say $\{A_n\}$ and $\{B_n\}$ are *equivalent* and write $A_n \cong B_n$ provided that for any $\epsilon > 0$, we have $\sum_{n=1}^{\infty} P(|A_n - B_n| > \epsilon) < \infty$. This relation is a *bona fide* equivalence, and we will particularly make use of transitivity, i.e., $A_n \cong B_n$ and $B_n \cong C_n$ imply $A_n \cong C_n$.

In this section, $X_1, X_2, \ldots$, will stand for independent and identically distributed random variables uniformly distributed in the unit square of $\mathbb{R}^2$.

The main result of this section can be stated quite succinctly.

THEOREM 3.1: *There is a constant c, $0 < c < \infty$, such that*

$$G(X_1, X_2, \ldots, X_n)/\sqrt{n} \cong c.$$

The proof of Theorem 3.1 requires several lemmas. The first of these spells out two simple properties of triangular arrays which prove to be very handy.

LEMMA 3.1:

(i) *Let Y_{ki}, $1 \le k < \infty$, $1 \le i \le n_k$, be a triangular array of random variables. For each k, we assume the n_k random variables Y_{ki}, $1 \le i \le n_k$, are independent and identically distributed. We let μ_k denote the mean of Y_{ki} and set $E(Y_{ki} - \mu_k)^4 = m_k$. If we have liminf $n_k/k > 0$ and limsup $m_k < \infty$, then*

$$\sum_{i=1}^{n_k} Y_{ki}/n_k \cong \mu_k.$$

(ii) *If $Z_1, Z_2, \ldots$, is a sequence of Poisson random variables with any joint distribution, and if $EZ_k = a_k$, then lim sup $a_k/k < \infty$ implies*

$$(Z_k - a_k)/k \cong 0.$$

PROOF: To check part (i), we first center the series $W_{ki} = Y_{ki} - \mu_k$ and calculate the fourth moment,

$$E\left[\left(\sum_{i=1}^{n_k} W_{ki}\right)^4\right] = \sum_{i=1}^{n_k} EW_{ki}^4 + 3n_k(n_k - 1)EW_{k1}^2 W_{k2}^2.$$

By Schwartz's inequality, the second summand is bounded by $n_k(n_k - 1)m_k$, so

$$E\left[\left(\sum_{i=1}^{n_k} W_{ki}/n_k\right)^4\right] \le 3n_k^{-2}m_k.$$

Now for any $\epsilon > 0$

$$P\left(\left|\sum_{i=1}^{n_k} W_{ki}/n_k\right| > \epsilon\right) \le \epsilon^{-4} E\left[\left(\sum_{i=1}^{n_k} W_{ki}/n_k\right)^4\right] \le \epsilon^4 3 n_k^{-2} m_k,$$

and $\sum_{i=1}^{\infty} m_k n_k^{-2}$ converges because liminf $n_k/k > 0$ and limsup $m_k < \infty$.

The second part of Lemma 3.1 follows from the first part by noting that each of the Poisson random variables Z_k can be represetned as the sum of k independent Poisson variables Y_{ki}, each having expectation a_k/k. The $\{Y_{ki}\}$ then form a triangular array to which Lemma 3.1(i) can be applied. ■

Returning to the proof of Theorem 3.1, we let $\Gamma(t)$ be a Poisson random variable with parameter t which is independent of the sequence $\{X_i\}$. By Lemma 2.4, we have the bound

$$|G(X_1, X_2, \ldots, X_n) - G(X_1, \ldots, X_{\Gamma(n)})| \le c_4 |\Gamma(n) - n|^{1/2}$$

so, for any $\epsilon > 0$, we have

$$\sum_{k=1}^{\infty} P\left(\left|\frac{G(X_1, X_2, \ldots, X_n)}{\sqrt{n}} - \frac{G(X_1, \ldots, X_{\Gamma(n)})}{\sqrt{n}}\right| > \epsilon\right)$$
$$\le \sum_{k=1}^{\infty} P\left(\left|\frac{\Gamma(n) - n}{n}\right| > \epsilon^2 c_4^{-2}\right) < \infty,$$

where the last inequality comes from applying Lemma 3.1(ii). Thus, to prove Theorem 3.1, it suffices to prove

$$G(X_1, \ldots, X_{\Gamma(n)})/\sqrt{n} \cong c. \tag{3.1}$$

Now, if R is any subregion of $[0,1]^2$, and $\{x_1, x_2, \ldots, x_n\} \subset [0,1]^2$, let $G_R(x_1, x_2, \ldots, x_n)$ be the greedy matching of the x_i which are elements of R. If S is a subsquare of $[0,1]^2$ of side length s, then the number of elements of $\{X_1, X_2, \ldots, X_{\Gamma(t)}\} \cap S$ is Poisson with parameter $s^2 t$, so by scaling we see $G_S(X_1, \ldots, X_{\Gamma(t)})$ is equal in distribution to $sG(X_1, \ldots, X_{\Gamma(s^2 t)})$.

Now, if we let $\phi(t) = EG(X_1, \ldots, X_{\Gamma(t)})$ and $\delta(t) = E[(G(X_1, \ldots, X_{\Gamma(t)}) - \phi(t))^4]$, then from Lemma 2.4, we can see both ϕ and δ are continuous and hence bounded on compact intervals. Finally, we recall that if $\{S_i\}$, $1 \le i < N$ are any subsets of $[0,1]^2$ such that $S_i \cap S_j$ has measure zero for each i and j with $i \ne j$, then for any variables Z_i, $1 \le i \le N$, such that Z_i is a function of the random point set $\{X_1, X_2, \ldots, X_{\Gamma(n)}\} \cap S_i$, then the variables $\{Z_i\}$ are independent.

Next, let M be an integer, let $\theta(n)$ be the largest integer satisfying $n/\theta(n)^2 > M$, and note that $n/\theta(n)^2 \to M$ as $n \to \infty$. When we divide $[0,1]^2$ into $\theta(n)^2$

equal squares $S_1, S_2, \ldots, S_{\theta(n)^2}$, we see by Lemma 3.1(i), the boundedness of δ on compact intervals, and the continuity of ϕ, that we have

$$\sum_{i=1}^{\theta(n)^2} G_{S_i}(X_1, \ldots, X_{\Gamma(n)})/\theta(n) \cong \phi(n/\theta(n)^2) \cong \phi(M). \tag{3.2}$$

We now need to relate this Poisson limit result to our original problem.

LEMMA 3.2: *Given $\epsilon > 0$ there exists M_0 such that $M \geq M_0$ implies*

$$\sum_{n=1}^{\infty} P\left(\left| \sum_{i=1}^{\theta(n)^2} G_{S_i}(X_1, \ldots, X_{\Gamma(n)})/\sqrt{n} - G(X_1, X_2, \ldots, X_n)/\sqrt{n} \right| > \epsilon \right) < \infty. \tag{3.3}$$

PROOF: We first fix $\lambda > 0$, and put $\delta(n) = \lambda/2\theta(n)$; so, in the notation of Lemma 2.5, the area of $F^{\delta(n)}$ does not exceed λ. Next, let $u(n)$ denote the number of edges of the greedy matching of $\{X_1, \ldots, X_{\Gamma(n)}\}$ which have length as great as $2\delta(n)$, and let $v(n)$ denote the number of points lying in $F^{\delta(n)}$. By Lemma 2.3, we have $u(n) < c_3(2\delta(n))^{-2}$, so since $\phi(n)^2 \leq n/M$, we have $u(n) < c_3 n/M\lambda^2$. Also, $v(n)$ is a Poisson random variable with mean not exceeding $2\lambda n$, so by Lemma 3.1(ii), we have $\Sigma P(v(n) > 3\lambda n) < \infty$. Thus, if M is chosen so large that $c_3/M\lambda^2 < \lambda$, we have

$$\sum_{n=1}^{\infty} P\{(u(n) + v(n))^{1/2} > (4\lambda n)^{1/2}\} < \infty.$$

In view of Lemma 2.6 and the definitions of u and v, we see the proof of Lemma 3.2 is complete. ■

We can now return to the completion of the proof of Theorem 3.1. By the equivalence relation (3.2), Lemma 3.2, the continuity of ϕ, and the fact that $n/\theta(n)^2 \to M$, we see that for any $\epsilon > 0$ there is an M_ϵ such that $M \geq M_\epsilon$ implies

$$\sum_{n=1}^{\infty} P(|G(X_1, X_2, \ldots, X_n)/\sqrt{n} - \phi(M)/\sqrt{M}| > \epsilon) < \infty. \tag{3.4}$$

One final (and slippery) observation is that Eq. (3.4) implies that $\phi(M)/\sqrt{M}$ converges to a limit as $M \to \infty$. To check this, note that if $\alpha = \liminf \phi(M)/\sqrt{M} \neq \limsup \phi(M)/\sqrt{M} = \beta$ one could not have Eq. (3.4) for $\epsilon < (\beta - \alpha)/3$. Therefore, $\lim M^{-1/2}\phi(M) = c < \infty$ exists, and $n^{-1/2}G_n \cong c$, completing the proof of Theorem 3.1.

4. GENERAL EXTENSION

We now extend the results on uniform distributions to random variables with compact support. The essence of the problem continues to be visible with $d =$

2, so we will restrict attention to that case. This time we let $X_1, X_2, \ldots$, be an infinite sequence of independent, identically distributed random variables with compact support in $\mathbb{R}^2$ and denote the common distribution of the X_i by ν.

We can assume without loss of generality that ν has support in $[0,1]^2$, and we claim there is also no loss of generality in assuming

$$\nu\{(x,y): x \text{ or } y \text{ is rational}\} = 0. \tag{4.1}$$

To check this, first note that there is an uncountable set $A \subset \mathbb{R}$ such that the difference of any two elements of A is irrational. One way to build A is to let the singleton $\{1\}$ be extended to a basis B for $\mathbb{R}$ viewed as a vector space over $\mathbb{Q}$ and take the required set A to be $B - \{1\}$.

The uncountability of A and the fact that any differences of elements of A are irrational guarantees that there is an $a \in A$ such that $\nu\{(x,y): x = a + r, r \text{ rational}\}$ is zero and a $b \in A$ such that $\nu\{(x,y): y = b + r, r \text{ rational}\}$ equals zero. The translation of ν by the pair (a,b) then satisfies Eq. (4.1). Moreover, any further translation of ν by rational numbers, or dilation of ν by rational numbers, still satisfies Eq. (4.1); in particular, any translation and dilation that maps the support of ν to $[0,1]^2$.

We next denote the singular part of ν by μ. Let f stand for the density of the absolutely continuous part of μ, and define our target limit constant k by

$$k = c \int_{[0,1]^2} f^{1/2}(x)dx,$$

where c is the constant of Theorem 3.1.

The main result of this section is the following.

THEOREM 4.1: $G(X_1, X_2, \ldots, X_n)/\sqrt{n} \cong k$.

PROOF: As before, it suffices to show that

$$G(X_1, \ldots, X_{\Gamma(n)})/\sqrt{n} \cong k. \tag{4.2}$$

We begin the proof of Eq. (4.2) by decomposing the square just as was done in Lemma 2.5. Next, we define a piecewise constant approximation to f on $Q = [0,1]^2$ by letting $f_m(x) = m^2 \int_{Q_i} f$ for x in the interior of Q_i and letting $f_m(x) = 0$ for any x not in the interior of some Q_i. Also, for any finite set S, we let $\#S$ denote the number of elements of S.

Now, choose $0 < \eta < 1$ and pick an integer m (depending on η) which is sufficiently large to guarantee the following four approximation conditions:

(i) There exists a sub-collection $\{Q_i, i \in E\}$ of $Q_i, \ldots, Q_{m^2}$ such that $\#E/m^2 < \eta$ and $\mu([0,1]^2 - H) < \eta$, where $H = \bigcup_{i \in E} Q_i$.

(ii) $\left| \int_{[0,1]^2} f^{1/2} - \int_{[0,1]^2 - H} f_m^{1/2} \right| < \eta.$

(iii) $\int_{[0,1]^2} |f - f_m| < \eta.$

(iv) $m > 1/\eta.$

We let $\tilde{\nu}_m = \tilde{\nu}$ be the probability measure with singular part μ and absolutely continuous part f_m, and note that for any compact set K we have $|\tilde{\nu}(K) - \nu(K)| < 2\eta$. By the coupling lemma of Strassen [16], there will then exist a pair of random vectors (Z_1, W_1) such that Z_1 has the distribution ν, W_1 has the distribution $\tilde{\nu}$, and $P(Z_1 \neq W_1) < 2\eta$. We next let $(Z_1, W_1), \ldots,$ be independent copies of these vectors, and for each integer n, we let $\Gamma(n)$ be a Poisson random variable which is independent of the sequence $\{(Z_i, W_i)\}$.

Finally, we can lay out notation for the variables of basic interest:

$$U_n = \{Z_1, \ldots, Z_{\Gamma(n)}\},$$
$$V_n = \{W_1, \ldots, W_{\Gamma(n)}\},$$
$$V_n^k = V_n \cap Q_k, \qquad 1 \leq k \leq m^2,$$
$$W_n^k = V_n^k \cap \{Q_k - \text{support } \mu\}, \text{ and}$$
$$\Delta_n = U_n \Delta V_n \qquad \text{where } \Delta \text{ denotes the symmetric difference.}$$

Since the Z_i have the same distribution as the X_i, the complete convergence of $n^{-1}|\Gamma(n) - n|$ to zero and Lemma 2.6 lead us to

$$G(U_n)/\sqrt{n} \cong G(X_1, X_2, \ldots, X_n)/\sqrt{n}, \tag{4.3}$$

by the same path used in Section 3. Now, since Δ_n has the Poisson distribution with parameter not greater than $2n\eta$, we see from Lemma 2.4 and Lemma 3.1(ii) that there is a constant β_0 not depending on m such that

$$\sum_{n=1}^{\infty} P(|G(U_n)/\sqrt{n} - G(V_n)/\sqrt{n}| > \beta_0 \eta) < \infty. \tag{4.4}$$

Next, we claim that there is a constant β_1 not depending on m such that

$$\sum_{n=1}^{\infty} P\left(\left| G(V_n)/\sqrt{n} - \sum_{k=1}^{m^2} G(V_n^k)/\sqrt{n} \right| > \beta_1 \eta \right) < \infty. \tag{4.5}$$

To prove Eq. (4.5), we first note that by the assumption made in Eq. (4.1) we can choose $\delta(\eta) = \delta$ to be so small that $\nu(F_m^\delta) < \eta$. The number of points of V_n in F_m^δ then has the Poisson distribution with parameter $n\nu(F_m^\delta) < n\eta$,

and by Lemma 2.3 the number of edges in the greedy matching which have length exceeding 2δ does not exceed $c_3(2\delta)^{-2}$. The proof of Eq. (4.5) is thus completed by applying Lemma 2.6 and arguing again just as in the proof of Lemma 3.2.

We next claim that there is a constant β_2 not depending on m such that

$$\sum_{n=1}^{\infty} P\left(\left|\sum_{k=1}^{m^2} G(V_n^k)/\sqrt{n} - \sum_{k=1}^{m^2} G(W_n^k)/\sqrt{n}\right| \geq \beta_2 \eta^{1/2}\right) < \infty. \tag{4.6}$$

To prove Eq. (4.6), we first note that $\#V_n^k$ is a Poisson random variable with mean $n\tilde{\nu}(Q_k)$ and for each fixed n the variables $\#V_n^k$, $1 \leq k \leq m^2$, form an independent collection. By Lemma 2.2, scaling, and the elementary inequality $\sqrt{n_1} + \cdots + \sqrt{n_s} \leq \sqrt{s}\sqrt{n_1 + \cdots + n_s}$, we have,

$$\begin{aligned}\sum_{k\in E} G(V_n^k) &\leq c_2 m^{-1} \sum_{k\in E} \sqrt{\#V_n^k} \\ &\leq c_2 m^{-1} \sqrt{\#E} \sqrt{\sum_{k\in E} \#V_n^k}.\end{aligned}$$

Now, $\Sigma \#V_n^k$ is a Poisson random variable with parameter less than or equal to n, and the approximation condition (i) says $(\#E)^{1/2}/m < \eta$, so Lemma 3.1(ii) gives

$$\sum_{n=1}^{\infty} P\left(\left|n^{-1/2} \sum_{k\in E} G(V_n^k)\right| > 2c_2\eta\right) < \infty. \tag{4.7}$$

By a completely similar analysis, we also have

$$\sum_{n=0}^{\infty} P\left(\left|n^{-1/2} \sum_{k\in E} G(W_n^k)\right| > 2c_2\eta\right) < \infty. \tag{4.8}$$

Lemmas 2.2 and 2.4 show that

$$\begin{aligned}\sum_{k\notin E} |G(V_n^k) - G(W_n^k)| &\leq c_4 m^{-1} \sum_{k\notin E} \sqrt{\#(V_n^k - W_n^k)} \\ &\leq c_4 m^{-1} \sqrt{m^2} \sqrt{\sum_{k\notin E} \#(V_n^k - W_n^k)}.\end{aligned}$$

Now, since the sum $\sum_{k\notin E} \#(V_n^k - W_n^k)$ is a Poisson random variable with mean bounded by $n\eta$, Lemma 3.1(ii) gives

$$\sum_{n=0}^{\infty} P\left(\left|\sum_{k\notin E} G(V_n^k)/\sqrt{n} - \sum_{k\notin E} G(W_n^k)/\sqrt{n}\right| > 2c_4\eta^{1/2}\right) < \infty.$$

Together with Eqs. (4.7) and (4.8), this last inequality completes the proof of Eq. (4.6).

Finally, we note that

$$\sum_{k=1}^{m^2} G(W_n^k) n^{-1/2} \cong c \int f m^{1/2}$$

follows from

$$G(W_n^k) \cong (c/m^2)\alpha_k^{1/2}, \tag{4.9}$$

where α_k is the value of f_m on Q_k. Given that $X_i \in Q_k$, we know X_i is uniformly distributed on Q_k, so the relation (4.9) follows from scaling, the fact that $\#W_k^n$ has a Poisson distribution with parameter $n\alpha_k/m^2$, Theorem 3.1, and Lemma 2.4. ∎

5. OPEN PROBLEM AND RELATED WORK

The limit theory for greedy matchings has been established with most of the generality and precision one might expect. Still, there are some tenacious open problems.

It is not difficult to extend the restriction of Theorem 1.1 a little way beyond compact support, but it does not seem easy to provide a broad sufficient condition even if one restricts attention to absolutely continuous random variables.

A second problem of interest concerns matching in one dimension. By the natural one-dimensional analogue of Lemma 2.1, one can show that $G_n = 0(\log(n))$, and it is reasonable to conjecture that for X_i independent and identically distributed on $U[0,1]$, one has $G_n \cong c \log n$ with probability 1. The tools used here are not sharp enough to deal with this question, and it is not even known if $EG_n \to \infty$ and $n \to \infty$. For a related problem concerning optimal triangulation, one may consult Steele [15].

These last two problems seem solvable, but there are other open issues associated with G_n which are less likely to be resolved. First, we do not know any analytical method to determine (or even estimate) the value of the constant k_d which appears in our limit theorem, although one can get some idea of k_d by Monte-Carlo experiments. Second, we strongly expect that the approximate representation of $G(Q)$ as a sum of independent and identically distributed random variables $G(Q_i)$ should be strong enough to permit the development of a central limit theorem. Still, it is not yet possible to extract the central theorem from Lemma 2.6, and the central limit problem is likely to be much more difficult than that lemma suggests. For related problems for a number of different functionals, one may consult Steele [14].

References

1. Avis, D. (1981). Worst case bounds for the Euclidean matching problem. *Computers and Mathematical Applications* 7:251–257.
2. Avis, D. (1983). A survey of heuristics for the weighted matching problem. *Networks* 13:475–493.

3. Beardwood, J., Halton, J.H., & Hammersley, J.M. (1959). The shortest path through many points. *Proceedings of the Cambridge Philosophical Society* 55:299–327.
4. Bentley, J.L. & Saxe, J.B. (1980). Decomposable searching problems 1: static to dynamic transformations. *Journal of Algorithms* 1:301–358.
5. Bondy, J.A. & Murty, U.S.R. (1976). *Graph Theory with Applications.* New York: American Elsevier Publishing Co., Inc.
6. Few, L. (1955). The shortest road through n points in a region. *Mathematika* 2, 141–144.
7. Iri, M., Murota, K., & Matsui, S. (1981). Linear time heuristics for the minimum weight perfect matching problem on a plane. *Information Proceedings Letters* 12:206–209.
8. Karp, R.M. (1977). Probabilistics analysis of partitioning algorithms for the traveling salesman problem in the plane. *Mathematics of Operations Research* 2:209–224.
9. Karp, R.M. & Steele, J.M. (1985). Probabilistic analysis of heuristics. In E.L. Lawler et al. (eds.), *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* NY: John Wiley and Sons, pp. 181–206.
10. Papadimitriou, C.H. & Steiglitz, K. (1982). *Combinatorial Optimization: Algorithms and Complexity.* Englewood Cliffs, NJ: Prentice-Hall.
11. Papadimitriou, C.H. (1977). The probabilistic analysis of matching heuristics. *Proceedings of the 15 Allerton Conference on Communication Control and Computing*, pp. 368–378.
12. Reingold, E.M. & Supowit, K.J. (1983). Probabilistic analysis of divide and conquer heuristics for minimum weighted Euclidean matchings. *Networks* 13:49–66.
13. Reingold, E.M. & Tarjan, R.E. (1981). On a greedy heuristic for complete matching. *SIAM Journal of Computing* 10:676–681.
14. Steele, J.M. (1981). Subadditive Euclidean functionals and nonlinear growth in geometric probability. *Annals of Probability* 9:365–376.
15. Steele, J.M. (1982). Optimal triangulation of random samples in the plane. *Annals of Probability* 10:548–553.
16. Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics* 36:423–439.