SIAM J. COMPUT.
Vol. 19, No. 4, pp. 666-672, August 1990

GREEDY MATCHING ON THE LINE*

ALAN FRIEZE,† COLIN McDIARMID,‡ AND BRUCE REED§

Abstract. The problem of finding a perfect matching of small total length in a complete graph whose vertices are points in the interval [0, 1] is considered. The greedy heuristic for this problem repeatedly picks the two closest unmatched points x and y, and adds the edge xy to the matching. It is shown that if $2n$ points are randomly chosen uniformly in [0, 1], then the expected length of the matching given by the greedy algorithm is $\theta(\log n)$. This compares unfavourably with the length of the shortest perfect matching, which is always less than 1.

Key words. greedy, matching, Euclidean, line, average case, worst case

AMS(MOS) subject classifications. 68Q25, 90C27, 68R10, 05C70, 60C05, 60D05

1. Introduction. We are interested in finding a perfect matching of small length on a set of points drawn from the interval [0, 1]. That is, given a set A of $2n$ numbers $\{x_1, \cdots, x_{2n}\}$, each of which is between 0 and 1, we want to partition A into n pairs $\{y_1, z_1\}, \{y_2, z_2\}, \cdots, \{y_n, z_n\}$ so that we minimize $\sum_{i=1}^{n} |y_i - z_i|$. We can solve this problem by reordering the elements of A so that $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(2n)}$ and then set $y_i = x_{(2i-1)}$ and $z_i = x_{(2i)}$. It is easy to see that this method always gives the optimal matching and, furthermore, the length of this matching is at most 1.

It is natural to ask how well the greedy approach performs in this simple setting. The greedy matching algorithm first finds distinct elements a, b in A such that $|a-b| = \min\{|c-d|: c, d \in A, c \neq d\}$ and selects $\{a, b\}$ as one pair of the partition. The remaining pairs are found by using the same procedure on $A - \{a, b\}$.

It is not difficult to see that the greedy matching algorithm selects a matching of length at most $O(\ln(n))$ when applied to a set of $2n$ points. Also, one may construct examples to show that the worst-case weight of a greedy matching is $\Omega(\ln(n))$. Indeed, in § 5, we identify the worst-case behaviour rather precisely. (For related work see Avis [1] and Rheingold and Tarjan [4].) However, we are more interested in average-case behaviour. We prove that if $2n$ points are chosen at random from the uniform distribution on [0, 1], then the expected length of the resulting matching is $\Omega(\ln(n))$. This settles a question raised by Avis, Davis, and Steele in [2], where results are given on greedy Euclidean matching in d-dimensional spaces for $d \geqq 2$. (Note that the greedy algorithm may be of practical use when $d \geqq 2$, though it is only of theoretical interest in the case $d = 1$ considered here.)

Since this is our main result, we now state it again. Given n numbers $x_1, \cdots, x_n$ (n even) in the interval [0, 1], let $G[x_1, \cdots, x_n]$ denote the length of the corresponding greedy matching (break ties arbitrarily).

THEOREM. *Let $X_1, \cdots, X_n$ be n independent random variables, each uniformly distributed on* [0, 1]. *Then $E[G(X_1, \cdots, X_n)] \geqq \frac{1}{12} \ln(n)$ for n sufficiently large.*

2. Bags, sticks, and entropy. Rather than considering our points directly we shall focus on the distances between them. To this end, we begin with some definitions. Given an n-tuple $\mathbf{x} = (x_1, \cdots, x_n)$ of reals in [0, 1], let $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n)}$ denote the numbers rearranged in nondecreasing order and let $z_k = x_{(k+1)} - x_{(k)}$ for $k = 0, \cdots, n$

* Received by the editors September 28, 1988; accepted for publication (in revised form) October 9, 1989.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
‡ Department of Statistics, Oxford University, Oxford, England.
§ Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada.

where $x_{(0)} = 0$, $x_{(n+1)} = 1$. We shall refer to $z_1, \cdots, z_{n-1}$ as the (nearest neighbour) *sticks* and z_0, z_n as the *endsticks*. Let $L(\mathbf{x}) = [z_1, \cdots, z_{n-1}]$ and let $B(\mathbf{x})$ be the unordered (multi)set of these values. Thus $L(\mathbf{x})$ is the *l*ist of sticks and $B(\mathbf{x})$ the *b*ag of sticks. Note that the endsticks are not included here.

Now, let $X_1, \cdots, X_n$ be independently and identically distributed uniform on $[0, 1]$. We shall use two easy properties of the corresponding random sticks.

Property 1. Let B be any bag of $n-1$ sticks (considered distinct). Then, conditional on $B(\mathbf{X}) = B$ the distribution of the corresponding list $L(\mathbf{X})$ is uniform on the $(n-1)!$ orderings of the $n-1$ sticks in B.

This result is intuitively obvious, or see Feller [3], pp. 74-76.

Property 2. With probability $\to 1$ as $n \to \infty$,

$$\max\{Z_k : k = 0, \cdots, n\} \leqq 2 \ln(n)/n.$$

To check this result, note that for each k

$$\text{Prob}(Z_k > t) = (1-t)^n \quad \text{if } 0 < t < 1$$

(see, for example, Feller [3], p. 22). Thus

$$\text{Prob}(\max Z_k > t) \leqq (n+1)(1-t)^n \leqq (n+1)/n^2 \quad \text{if } t = 2\ln(n)/n.$$

To begin, we use the first property to obtain a lower bound on the conditional expected value $E(G[\mathbf{X}] \mid B[\mathbf{X}] = B)$ that depends on the length of the sticks in B. Then we point out that the second property ensures that this lower bound gives the desired result.

Thus, we now focus our attention on a fixed bag B of n sticks (where n is odd). By Property 1 above, the distribution of $G(\mathbf{X})$ conditional on $B(\mathbf{X}) = B$ is the same as that of the value CHOOSE(B) returned by the following randomized recursive algorithm.

Let x be a minimum element in B
if $|B| = 1$ *then return* x
else
with probability $2/n$ choose y uniformly from $B - \{x\}$ and *return* $x + \text{CHOOSE}(B - \{x, y\})$ (this corresponds to x being the leftmost or rightmost stick, with neighbour y)
with probability $(n-2)/n$ choose y uniformly from $B - \{x\}$ and z uniformly from $B - \{x, y\}$ and *return* $x + \text{CHOOSE}(B - \{x, y, z\} \cup \{x + y + z\})$ (this corresponds to x having left neighbour y and right neighbour z)

Now let

$$F(B) = E[\text{CHOOSE}(B)] = E[G(\mathbf{X}) \mid B(\mathbf{X}) = B].$$

By considering the algorithm CHOOSE(B) we obtain a recurrence for $F(B)$. Let $B = \{a_1, a_2, \cdots, a_n\}$ be a bag of n sticks where n is odd, $n > 1$, and a_1 is a minimum element. Then

$$F(B) = a_1 + \frac{2}{n}\frac{1}{n-1}\sum_{j=2}^{n} F(B - \{a_1, a_j\}) + \frac{n-2}{n}\frac{2}{(n-1)(n-2)}\sum_{j=2}^{n-1}\sum_{k=j+1}^{n} F(B - \{a_1, a_j, a_k\} \cup \{a_1 + a_j + a_k\}).$$

This recurrence is the key to our analysis. It will allow us to prove a lower bound on $F(B)$ in terms of the entropy of the stick lengths.

3. A lemma.

LEMMA. *Let* $A = \{a_1, \cdots, a_N\}$ *be an odd cardinality set of positive numbers that sum to* 1. *Let*

$$H(A) = \alpha \sum_{i=1}^{N} a_i \ln\left(\frac{1}{a_i}\right) - \beta \sum_{i=2}^{N} \frac{1}{i} - 2\alpha \sum_{i=2}^{N} \frac{\ln(i)}{i^2},$$

where $\alpha = \dfrac{1}{\ln(12)} \approx 0.40243$ *and* $\beta = \alpha \ln(24/11) \approx 0.31396$. *Then* $F(A) \geq H(A)$.

Proof. We note that if $N = 1$, then $H(A) = 0$ and $F(A) = 1$, so the inequality holds. We shall assume it holds for $N < n$ (where $n > 1$) and prove it for $N = n$. Furthermore, we may assume that a_1 is a minimal element of A. Then, as we noted earlier:

$$F(A) = a_1 + \frac{2}{n(n-1)} \sum_{j=2}^{n} F(A - \{a_1, a_j\}) + \frac{2}{n(n-1)} \sum_{j=2}^{n-1} \sum_{k=j+1}^{n} F(A - \{a_1, a_j, a_k\} \cup \{a_1 + a_j + a_k\}).$$

By the induction hypothesis,

$$F(A) \geq a_1 + \frac{2}{n(n-1)} \sum_{j=2}^{n} H(A - \{a_1, a_j\}) + \frac{2}{n(n-1)} \sum_{j=2}^{n-1} \sum_{k=j+1}^{n} H(A - \{a_1, a_j, a_k\} \cup \{a_1 + a_j + a_k\}).$$

By the definition of H,

$$F(A) \geq H(A) + a_1 + \frac{\beta}{n} + \frac{\beta}{n-1} + \frac{2\alpha \ln(n)}{n^2} + \frac{2\alpha \ln(n-1)}{(n-1)^2} + \frac{2\alpha}{n(n-1)} \sum_{j=2}^{n} \left(-a_j \ln\left(\frac{1}{a_j}\right) - a_1 \ln\left(\frac{1}{a_1}\right)\right) + \frac{2\alpha}{n(n-1)} \sum_{j=2}^{n-1} \sum_{k=j+1}^{n} \left((a_j + a_k + a_1) \ln\left(\frac{1}{a_j + a_k + a_1}\right) - a_j \ln\left(\frac{1}{a_j}\right) - a_k \ln\left(\frac{1}{a_k}\right) - a_1 \ln\left(\frac{1}{a_1}\right)\right).$$

So,

$$F(A) \geq H(A) + a_1 + \frac{2\beta}{n} + \frac{2\alpha \ln(n)}{n^2} + \frac{2\alpha \ln(n-1)}{(n-1)^2} - \alpha a_1 \ln\left(\frac{1}{a_1}\right) - \frac{2\alpha}{n(n-1)} \sum_{j=2}^{n} a_j \ln\left(\frac{1}{a_j}\right) + \frac{2\alpha}{n(n-1)} \sum_{j=2}^{n-1} \sum_{k=j+1}^{n} \left\{(a_j + a_k + a_1) \ln\left(\frac{1}{a_j + a_k + a_1}\right) - a_j \ln\left(\frac{1}{a_j}\right) - a_k \ln\left(\frac{1}{a_k}\right)\right\}.$$

We note that by the convexity of $x \ln(x)$, the single sum is at most $\ln(n-1)$. We shall show that, given that $a_1 = a$, the double sum is minimized when $a_2 = \cdots = a_n = (1-a)/(n-1) = b$, say.

CLAIM.

$$S(A) = \sum_{j=2}^{n-1} \sum_{k=j+1}^{n} \left\{ (a_j + a_k + a_1) \ln\left(\frac{1}{a_j + a_k + a_1}\right) - a_j \ln\left(\frac{1}{a_j}\right) - a_k \ln\left(\frac{1}{a_k}\right) \right\}$$
$$\geqq \frac{1}{2}(n-1)(n-2)\left((a+2b) \ln\left(\frac{1}{a+2b}\right) - 2b \ln\left(\frac{1}{b}\right)\right).$$

Proof of Claim. Suppose $a_2 \neq a_3$. Let $a_2' = a_3' = (a_2 + a_3)/2$ and let $a_i' = a_i$ for $i = 1, 4, 5, \cdots, n$. It will suffice to show that $S(A') < S(A)$. To this end let $f_k(x) = x \ln(x) - (x + a_1 + a_k) \ln(x + a_1 + a_k)$ for $k = 4, \cdots, n$.

Now,

$$S(A') - S(A) = \{a_2' \ln(a_2') + a_3' \ln(a_3') - a_2 \ln(a_2) - a_3 \ln(a_3)\}$$
$$+ \sum_{i=4}^{n} (f_i(a_2') + f_i(a_3') - f_i(a_2) - f_i(a_3)).$$

Here, the first term comes from the term in the double sum where $j = 2$ and $k = 3$, while the ith term in the sum comes from the terms in the double sum where $k = i$ and $j = 2$ or $j = 3$. Since the functions $x \ln(x)$ and $f_k(x)$ are strictly convex, we have $S(A') - S(A) < 0$ as required. □

Applying the claim, we see that

$$F(A) \geqq H(A) + a + \frac{2\beta}{n} + 2\alpha \frac{\ln(n)}{n^2} - \alpha a \ln\left(\frac{1}{a}\right)$$
$$+ \left(\frac{n-2}{n}\right) \alpha \left((a+2b) \ln\left(\frac{1}{a+2b}\right) - 2b \ln\left(\frac{1}{b}\right)\right)$$
$$\geqq H(A) + \frac{2\beta}{n} + 2\alpha \frac{\ln(n)}{n^2} - \frac{2\alpha}{n} a \ln\left(\frac{1}{a}\right)$$
$$+ \frac{(n-2)\alpha}{n}\left(\frac{a}{\alpha} + a \ln\left(\frac{a}{a+2b}\right) + 2b \ln\left(\frac{b}{a+2b}\right)\right).$$

Setting $r = b/a$ and noting that $a \ln(1/a) \leqq \ln(n)/n$ since $a \leqq 1/n$, we get

$$F(A) \geqq H(A) + \frac{2\beta}{n} + \left(\frac{n-2}{n}\right) \alpha b f(r),$$

where $f(r) = (1/\alpha r) + (1/r) \ln(1/(1+2r)) + 2 \ln(r/(1+2r))$. Now $f'(r) = (1/r^2)(-(1/\alpha) + \ln(1+2r))$, so $f(r)$ is minimized when $\ln(1+2r) = 1/\alpha$ and thus $f(r) \geqq -2 \ln(2 + (2/e^{1/\alpha} - 1))$. Hence

$$F(A) \geqq H(A) + \frac{2\beta}{n} - \left(\frac{n-2}{n}\right) 2\alpha b \ln\left(2 + \frac{2}{e^{1/\alpha} - 1}\right).$$

Next we note that $b \leqq 1/(n-1)$, so

$$F(A) \geqq H(A) + \frac{1}{n}\left(2\beta - 2\alpha \ln\left(2 + \frac{2}{e^{1/\alpha} - 1}\right)\right).$$

Thus, by our choice of α and β, we have $F(A) \geqq H(A)$ as required.

4. Completing the proof of the Theorem. Consider n points $x_1, \cdots, x_n$ in $[0,1]$ (n even) such that each of the corresponding $(n+1)$ sticks (including endsticks) has length at most $d(<\frac{1}{2})$. Thus the corresponding bag B of $n-1$ sticks has sum of lengths $\sigma \geqq 1-2d$. From the lemma (by scaling by $1/\sigma$)

$$F(B) \geqq \sigma\left[\alpha \sum_{x\in B} \frac{x}{\sigma} \ln\left(\frac{\sigma}{x}\right) - \beta \sum_{i=2}^{n-1} 1/i - 2\alpha \sum_{i=2}^{n-1} \frac{\ln(i)}{i^2}\right]$$

$$\geqq \sigma\left[\alpha\left(\sum_{x\in B} \frac{x}{\sigma}\right) \ln\frac{1-2d}{d} - \beta \sum_{i=2}^{n-1} 1/i - 2\alpha \sum_{i=2}^{n-1} \frac{\ln(i)}{i^2}\right]$$

$$= (1-2d)\left[\alpha \ln\frac{1-2d}{d} - \beta \sum_{i=2}^{n-1} 1/i - 2\alpha \sum_{i=2}^{n-1} \frac{\ln(i)}{i^2}\right].$$

Now set $d = 2\ln(n)/n$ and use Property 2. We find that

$$E[G(X_1, \cdots, X_n)] \geqq (1+o(1))E[G(X_1, \cdots, X_n) \mid \max_{0\leq j\leq n} Z_j \leqq 2\ln(n)/n]$$

$$\geqq (1+o(1))((\alpha+o(1))\ln(n) - \beta\ln(n) + O(1))$$

$$= (\alpha - \beta + o(1))\ln(n) \approx (0.088)\ln(n).$$

This completes the proof of the theorem. □

5. Worst-case results. In this section we consider lists of points on which the greedy algorithm performs particularly badly.

For any nonnegative integer k consider the list $\mathbf{x}(k) = \{i/3^k : i = 0, 1, \cdots, 3^k\}$ of 3^k+1 points in $[0,1]$. For $k \geqq 1$ the greedy algorithm applied to $\mathbf{x}(k)$ can pick 3^{k-1} intervals of length 3^{-k}, namely, the intervals $[(3j+1)3^{-k}, (3j+2)3^{-k}]$ for $j = 0, 1, \cdots, 3^{k-1}-1$, and then be left with the points $\mathbf{x}(k-1)$. Hence $\mathbf{x}(k)$ has a greedy matching of weight $k/3+1$. It follows that for any even n, there is a list of n points that has a greedy matching of weight $\frac{1}{3}\lfloor \log_3(n-1)\rfloor + 1$.

On the other hand, we shall show that any greedy matching on a list of n points has weight at most $\frac{1}{3}\log(n-1)+1$. Thus for n of the form 3^k+1 the above examples are worst possible and for all other n we are not far off. From now on we shall just use $\log(x)$ to denote $\log_3(x)$.

PROPOSITION. *For a bag $A = \{a_1, \cdots, a_N\}$ of sticks (with N odd), let $W(A)$ be the length of the longest greedy matching of any n-tuple $\mathbf{x}$ such that the corresponding list $L(\mathbf{x})$ of sticks is a permutation of A. Then*

$$W(A) \leqq \frac{1}{3}\sum_{i=1}^{N} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=1}^{N} a_i.$$

Proof. We prove this by induction on the cardinality of A. If $|A| = 1$, then $W(A) = a_1$ and the theorem is true. So we suppose the theorem holds for $N < n$, where $n \geqq 3$, and prove it for $N = n$. Let $\mathbf{x}$ be a list of points such that $L(\mathbf{x})$ is a permutation of A and such that $G(\mathbf{x}) = W(A)$. We may assume that a_1 is the minimal element of A chosen in the first iteration when the greedy matching is applied to $\mathbf{x}$.

Case 1. a_1 is the rightmost or leftmost stick of $L(\mathbf{x})$. In this case, let a_2 be the neighbour of a_1 in $L(\mathbf{x})$. Then

$$W(A) = G(\mathbf{x}) \leqq a_1 + W(A - \{a_1, a_2\}).$$

Now, by our induction hypothesis:

$$W(A) \leqq a_1 + \frac{1}{3}\sum_{i=3}^{n} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=3}^{n} a_i$$

$$\leqq \frac{1}{3}\sum_{i=1}^{n} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=1}^{n} a_i.$$

Case 2. a_1 is an interior stick of $L(\mathbf{x})$. In this case we may assume that a_2 and a_3 are the neighbours of a_1 in $L(\mathbf{x})$ and that $a_2 \leqq a_3$.

Now, $W(A) = G(\mathbf{x}) \leqq a_1 + W(A - \{a_1, a_2, a_3\} \cup \{a_1 + a_2 + a_3\})$. By our induction hypothesis,

$$W(A) \leqq a_1 + \frac{1}{3}\sum_{i=4}^{n} a_i \log\left(\frac{1}{a_i}\right) + \frac{1}{3}(a_1 + a_2 + a_3) \log\left(\frac{1}{a_1 + a_2 + a_3}\right)$$

$$+ \sum_{i=4}^{n} a_i + (a_1 + a_2 + a_3)$$

$$= \frac{1}{3}\sum_{i=1}^{n} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=1}^{n} a_i + a_1 + \frac{1}{3}(a_1 + a_2 + a_3) \log\left(\frac{1}{a_1 + a_2 + a_3}\right)$$

$$- \frac{1}{3} a_1 \log\left(\frac{1}{a_1}\right) - \frac{1}{3} a_2 \log\left(\frac{1}{a_2}\right) - \frac{1}{3} a_3 \log\left(\frac{1}{a_3}\right).$$

Fixing a_1 and $a_1 + a_2 + a_3$, since $a_1 \leqq a_2 \leqq a_3$, we know that $-a_2 \log(1/a_2) - a_3 \log(1/a_3)$ is maximized when $a_2 = a_1$. Furthermore, fixing a_1 and fixing $a_1 = a_2$ we see that

$$\frac{1}{3}(a_1 + a_2 + a_3) \log\left(\frac{1}{a_1 + a_2 + a_3}\right) - \frac{1}{3} a_3 \log\left(\frac{1}{a_3}\right)$$

is maximized when $a_3 = a_1$. Thus,

$$W(A) \leqq \frac{1}{3}\sum_{i=1}^{n} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=1}^{n} a_i + a_1 + \frac{1}{3}(3a_1) \log\left(\frac{1}{3a_1}\right) - 3\left(\frac{1}{3} a_1 \log\left(\frac{1}{a_1}\right)\right)$$

$$= \frac{1}{3}\sum_{i=1}^{n} a_i \log\left(\frac{1}{a_i}\right) + \sum_{i=1}^{n} a_i. \qquad \square$$

Now, for any bag $A = \{a_1, \cdots, a_{n-1}\}$ of sticks,

$$\frac{1}{3}\sum_{i=1}^{n-1} a_i \log\left(\frac{1}{a_i}\right) \leqq \frac{1}{3}\log(n-1) \quad \text{and} \quad \sum_{i=1}^{n-1} a_i \leqq 1.$$

Thus, from our proposition, the greedy matching applied to a set of n points constructs a matching of length at most $\frac{1}{3}\log(n-1) + 1$. We note that this implies our examples are the worst possible for n of the form $3^k + 1$. Indeed, it is easy to see from our proposition that they are unique such examples.

Acknowledgments. Our thanks to Mike Saks and Bill Steiger for helpful discussions.

REFERENCES

[1] D. Avis, *Worst case bounds for the Euclidean matching problem*, Comput. Math. Appl., 7 (1981), pp. 251-257.

[2] D. AVIS, B. DAVIS, AND M. STEELE, *Probabilistic analysis of a greedy heuristic for Euclidean matching*, Probab. in the Eng. Inform. Sci., 2 (1988), pp. 143–156.

[3] W. FELLER, *An Introduction to Probability Theory and Its Applications, Vol. II*, Second edition, John Wiley, New York, 1971.

[4] E. M. RHEINGOLD AND R. E. TARJAN, *On a greedy heuristic for complete matching*, SIAM J. Comput., 10 (1981), pp. 676–681.