The Annals of Applied Probability
1993, Vol. 3, No. 3, 802–818

A SELECTION–REPLACEMENT PROCESS ON THE CIRCLE

BY E. G. COFFMAN, JR., E. N. GILBERT AND P. W. SHOR

AT & T Bell Laboratories

Given N points on a circle, a selection–replacement operation removes one of the points and replaces it by another. To select the removed point, an extra point P, uniformly distributed, arrives at random and starts moving counterclockwise around the circle; P removes the first point it encounters. A new random point, uniformly distributed, then replaces the removed point. The quantity of interest is $d = d(N)$, the distance that the searching point P must travel to select a point. After many repeated selection–replacements, the joint probability distribution of the N points tends to a stationary limit. We examine the mean of d in this limit. Exact means are found for $N \leq 3$. For large N, the mean grows like $(\log^{3/2} N)/N$. These means are larger than the means $1/(N+1)$ that would be obtained with N independent uniformly distributed points because the selection mechanism tends to cluster the N points into clumps.

In a computer application, the circle represents a track on a disk memory, P is a read–write head, the N points mark the beginnings of N files and d determines the time wasted as the head moves from the end of the last file processed to the beginning of the next. N is a parameter of the service rule (the next service goes to one of the N customers waiting the longest).

1. Introduction. Given $N \geq 1$ initial points marked by crosses on a circle, iterate the following two steps (see Figure 1):

1. Choose a point P uniformly at random on the circle and delete the first cross encountered in a counterclockwise scan from point P.
2. Choose another point P' uniformly at random on the circle and mark it as a new cross.

After many iterations of these steps, the joint distribution of the N crosses approaches a stationary limit, as verified later. Let $d_n = d_n(N)$ be the length of the scan in Step 1 of the nth iteration, as illustrated in Figure 1. The problem is to determine

$$E[d] = \lim_{n \to \infty} E[d_n], \tag{1.1}$$

where $d = d(N)$ is the distance from a random point P to the counterclockwise nearest of N crosses with the stationary distribution. For convenience, take the circumference of the circle to be the unit of distance.

Because the new cross P' in Step 2 is uniformly distributed, one might think the N crosses would become uniformly and independently distributed

Received July 1992; revised November 1992.
AMS 1991 *subject classifications.* 60F99, 05C70, 60J10.
Key words and phrases. Selection-replacement process, matching, probabilistic analysis of algorithms.

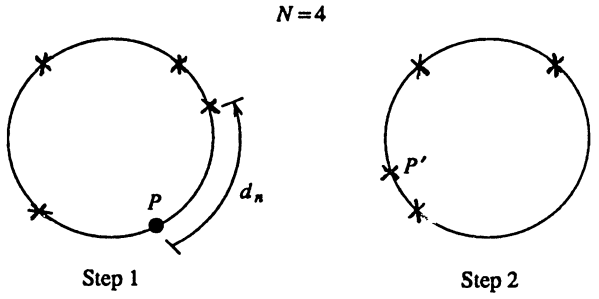


FIG. 1. *Selection–replacement.*

over the circle. For that distribution, the distance $d^* = d^*(N)$ from a random point P to the counterclockwise nearest cross would have $P(d^* > y) = (1 - y)^N$ and

$$E[d^*] = \int_0^1 (1 - y)^N \, dy = 1/(N + 1). \tag{1.2}$$

In fact, d is distributed like d^* if $N = 1$ or 2, but not if $N \geq 3$. If the deleted point in Step 1 had been equally likely to be any of the N crosses, then d and d^* would have been distributed alike. However, by picking P uniformly, Step 1 will usually delete an endpoint of one of the longer intervals. This paper derives an explicit formula for $E[d]$ when $N = 3$, and gives an asymptotic estimate of $E[d]$ for large N. Before getting into these and other results, the source of the problem will be briefly described.

Selection–replacement sequences occur in a model of computer disk scheduling, where the circle represents a track on the disk. The disk actually rotates at constant speed past a fixed read/write (R/W) head, but it will be simpler to imagine the head rotating around a stationary track. Requests for R/W operations arrive by some process and join a queue. Each request requires the head to start at a given point A and to read or write a file until it arrives at another given point B. A and B are assumed to be independently and uniformly distributed on the circle. An equivalent assumption is that the starting location A and the length on the track from A to B are independent uniform samples from $[0, 1]$. These assumptions are simplifications, made to render the disk scheduling problem tractable.

If the head served requests in order of arrival, it would waste half a revolution, on the average, just traveling to A. The head would be more efficient if it served the request whose point A it first encountered. That would be unfair to older customers in the queue, so as a compromise, we require the head to serve one of the N oldest customers; it chooses the one with the counterclockwise nearest point A. This A is the cross that Step 1 deletes; B is point P of the next Step 1. When a R/W operation begins, a new request in the queue becomes the Nth oldest; its A is point P' of Step 2. Taking $N = 1$ gives service in order of arrival but larger N achieves shorter

delays. For other disk models, see Fuller and Baskett (1975) and Coffman and Hofri (1986).

The service rate is greatest when the system is saturated, so that at least N requests always await service. Under that condition, the head moves a mean distance $E[d]$ finding A and 0.5 reading or writing. The maximum service rate is then $\mu = 1/(E[d] + 0.5)$. The queue is sure to be unstable if the arrival rate exceeds μ. The queue can be shown to be stable if the arrival process is Poisson with rate less than μ.

2. Results. The *state* just after the nth iteration of Step 2 is the unordered set $\omega_n^+ = \{x_1, \ldots, x_N\}$, where $x_1, \ldots, x_N$ are the coordinates of the N crosses relative to some fixed origin on the circle. The coordinates x_i are fractions of the circle circumference; the x_i will lie in $[0, 1]$. It is easy to see that the process $\{\omega_n^+\}$, to be called the *selection–replacement* (SR) process, is a Markov chain. To verify that the SR process is ergodic, it is enough to observe that the state space is compact and that the transition from any state $\{x_1, \ldots, x_N\}$ to any other state $\{x_1', \ldots, x_N'\}$ in r iterations has a strictly positive density for all $r \geq N$ [e.g., see Friedman (1970)]. Note also that the stationary distribution of $\{\omega_n^+\}$ must have a density p invariant under permutations of the coordinates and under relocations of the origin. With $\oplus$ denoting addition mod 1,

$$p(x_1, \ldots, x_N) = p(x_1 \oplus a, \ldots, x_N \oplus a) \tag{2.1}$$

for all a.

$N = 3$. Section 3 derives the following stationary density of $N = 3$ crosses under the SR process:

$$p(x_1, x_2, x_3) = f(|x_1 - x_2|) + f(|x_1 - x_3|) + f(|x_2 - x_3|), \tag{2.2}$$

where

$$f(\vartheta) = \frac{1 - \vartheta(1 - \vartheta)}{\left[\frac{1}{2} + \vartheta(1 - \vartheta)\right]^2}. \tag{2.3}$$

To determine $E[d]$, let $\vartheta_1, \vartheta_2, \vartheta_3$ denote the lengths of the gaps between the x_i. In terms of the ϑ_i, (2.2) is $p(x_1, x_2, x_3) = f(\vartheta_1) + f(\vartheta_2) + f(\vartheta_3)$. A state $\omega^+ = \{x_1, x_2, x_3\}$ could equally well have been represented as $\{x, x \oplus \vartheta_1, x \oplus \vartheta_1 \oplus \vartheta_2\}$, where x is uniform on $(0, 1)$ and ϑ_1, ϑ_2 are gaps satisfying $0 < \vartheta_1 + \vartheta_2 < 1$. Each state has three such representations, corresponding to three choices for the point x. The joint density for ϑ_1 and ϑ_2 is then $[f(\vartheta_1) + f(\vartheta_2) + f(1 - \vartheta_1 - \vartheta_2)]/3$.

Now consider the gap in which point P of Step 1 falls. $E[d]$ is half the mean size of this gap. For given gap lengths, P is in gap i with probability ϑ_i, so

$$E[d] = \tfrac{1}{6}\int_0^1 \int_0^{1-\vartheta_2} \left[\vartheta_1^2 + \vartheta_2^2 + (1 - \vartheta_1 - \vartheta_2)^2\right] \times \left[f(\vartheta_1) + f(\vartheta_2) + f(1 - \vartheta_1 - \vartheta_2)\right] d\vartheta_1\, d\vartheta_2.$$

Evaluating this integral is tedious, but elementary. The result is

$$E[d] = \frac{\sqrt{3}}{4}\log(2-\sqrt{3}) + \frac{5}{6} = 0.263074$$

(all logarithms in this paper will be to base e). When x_1, x_2, x_3 are independent, as well as uniformly distributed, (1.2) gives the slightly smaller $E[d^*] = \frac{1}{4}$. The comparison suggests a greater clumping of crosses under the SR process. The comparisons that follow also indicate clumping.

By (2.3), $f(\vartheta)$ takes its maximum at $\vartheta = 0$ or $\vartheta = 1$, where $f(\vartheta) = 4$. Then $p(x_1, x_2, x_3) \le 12$ and achieves this maximum when $x_1 = x_2 = x_3$ (extreme clumping). Also, if ω^+ has gaps $\vartheta_1, \vartheta_2, \vartheta_3$ $(\vartheta_1 + \vartheta_2 + \vartheta_3 = 1)$ between the crosses then, because $f(\vartheta)$ is convex,

$$p(\omega^+) = f(\vartheta_1) + f(\vartheta_2) + f(\vartheta_3) \ge 3f\left(\frac{\vartheta_1 + \vartheta_2 + \vartheta_3}{3}\right) = 3f\left(\frac{1}{3}\right)$$
$$\ge 4.4733724.$$

This minimum is achieved with $\vartheta_1 = \vartheta_2 = \vartheta_3 = \frac{1}{3}$, a state with least clumping. For a comparison, consider the probability density $p^*(\omega^+)$ for three crosses chosen independently and uniformly from the unit circumference. An unordered triple $\{x_1, x_2, x_3\}$ is then represented by a point (ξ_1, ξ_2, ξ_3) (the x_i in numerical order) distributed over the tetrahedron $0 \le \xi_1 \le \xi_2 \le \xi_3 \le 1$ with constant density $p^*(\omega^+) = 6$.

Finally, consider the expected gap lengths under the SR process. Routine but lengthy calculations give

$$E[\text{longest gap}] = \int_0^1 \int_0^{1-\vartheta_2} \max\{\vartheta_1, \vartheta_2, 1-\vartheta_1-\vartheta_2\}$$
$$\times [f(\vartheta_1) + f(\vartheta_2) + f(1-\vartheta_1-\vartheta_2)]\, d\vartheta_1\, d\vartheta_2$$
$$= \frac{5}{4} + \sqrt{3}\log(2-\sqrt{3}) - \frac{9\sqrt{3}}{8}\log\frac{19-8\sqrt{3}}{13} - \frac{3}{8}\log\frac{13}{9}$$
$$= 0.637762,$$

$$E[\text{shortest gap}] = -1 + \frac{\sqrt{3}}{4}\log(2-\sqrt{3}) - \frac{9\sqrt{3}}{8}\log\frac{19-8\sqrt{3}}{13} - \frac{3}{8}\log\frac{13}{9}$$
$$= 0.098541,$$

$$E[\text{middle gap}] = 1 - 0.637762 - 0.098541 = 0.263697.$$

These means may be compared with the values that are obtained with three independently distributed points [obtained by replacing $f(\vartheta)$ by 2]. In this case, the longest, shortest and middle gaps have mean lengths $11/18 = 0.611111$, $1/9 = 0.111111$ and $5/18 = 0.277777$. The larger mean longest gap and smaller mean shortest gap obtained with (2.3) indicate clumping.

Large N. The estimates for $E[d]$ will use the following standard asymptotic notation: $f(N) = O(g(N))$ will mean that $f(N)/g(N)$ remains less than

some constant for all sufficiently large N; $f(N) = \Omega(g(N))$ will mean that $g(N) = O(f(N))$; and $f(N) = \Theta(g(N))$ will mean that both $f(N) = O(g(N))$ and $f(N) = \Omega(g(N))$.

Section 4 proves

$$E[d] = \Theta\left(\frac{\log^{3/2} N}{N}\right), \tag{2.4}$$

a measure of clumping, as it exceeds (1.2) by a factor of $\Theta(\log^{3/2} N)$. This result is proved by a novel application of up-right matching theory [see Coffman and Lueker (1991), Chapter 3]. In fact, the theory provides the stronger result that, for all N sufficiently large and for $n \geq N^2$, $E[d_n] = \Theta(\log^{3/2} N/N)$.

Simulations of the SR process have given the values of $E[d]$ shown in Table 1. These results have been used to test the convergence of $E[d]$ to a function proportional to $\log^{3/2} N/N$, as $N \to \infty$. The excellent fit given by the empirical formula

$$E[d] \approx \frac{[\log(2.2N)]^{3/2}}{4.9(N-1)}$$

TABLE 1

Simulated values of $E[d]$ *and approximations* $b = ([\log(2.2N)]^{3/2})/(4.9(N-1))$ *and* $E[d^*] = 1/(N+1)$

N	$E[d]$	b	$E[d^*]$
2	0.33333	0.368046	0.333333
3	0.26305	0.264518	0.250000
4	0.22185	0.218171	0.200000
5	0.18606	0.189448	0.166667
6	0.17385	0.169168	0.142857
7	0.15810	0.153794	0.125000
8	0.14560	0.141596	0.111111
9	0.13505	0.131607	0.100000
10	0.12615	0.123231	0.090909
12	0.11235	0.109876	0.076923
15	0.09730	0.0953076	0.062500
20	0.08045	0.0790696	0.047619
30	0.06110	0.0603494	0.032258
50	0.04288	0.0424445	0.019608
100	0.02588	0.0258220	0.009901
200	0.015241	0.0154004	0.004975
500	0.007579	0.00757941	0.001996
1000	0.004266	0.00436168	0.000999
2000	0.002431	0.00248076	0.000500
5000	0.001155	0.00115889	0.000200
10000	0.000646	0.000645311	0.000100

shown in Table 1 suggests that there is an asymptotic constant of proportionality close to $1/4.9$. The last column of Table 1 shows that $E[d^*]$ of (1.2) disagrees greatly with $E[d]$ when N is large; then the cross locations are not even approximately independent.

A variant. An interesting problem arises when Step 1 is changed to the greedy rule: The cross eliminated in Step 1 is the one nearest to the point P, allowing either clockwise or counterclockwise scanning. For the greedy process the scan distance $\hat{d}_n$ is defined in analogy with d_n. The greedy SR process seems to be more difficult to analyze; for example, up-right matching theory does not appear to be a useful approach. The only result currently known, planned for a forthcoming paper, is an estimate of the second moment of $\hat{d}$: $E[\hat{d}^2] = \Omega(1/N)$. This contrasts with the second moment of the SR process, which satisfies $E[d^2] < (E[d])^2 = \Theta(\log^3 N/N^2)$ by (2.4).

Simulations show that the greedy process produces greater clumping, which compensates for the time saved in scanning to the nearest point. Surprisingly, the data give evidence of a threshold N_0, near 1500, such that $E[\hat{d}] \le E[d]$, $1 \le N \le N_0$, and $E[\hat{d}] > E[d]$, $N > N_0$.

3. Derivation of (2.2)–(2.3). The SR process for $N = 3$ will be analyzed together with the ergodic Markov chain $\{\omega_n^-\}$, where $\omega_n^- = \{x_1, x_2\}$ is the unordered set of coordinates of the $N - 1 = 2$ crosses just before the nth iteration of Step 2; that is, just after the deletion in Step 1 of the nth iteration. Note that the invariance (2.1) also applies to the stationary density of $\{\omega_n^-\}$, which we denote by $q(x_1, x_2)$. For example, the origin can be relocated so as to coincide with either x_1 or x_2, so that

$$q(x_1, x_2) = q(0, |x_1 - x_2|) = q(0, 1 - |x_1 - x_2|). \tag{3.1}$$

This shows that $q(x_1, x_2)$ can be expressed as a function $f(\vartheta(x_1, x_2))$ of a single variable $\vartheta(x_1, x_2) = |x_1 - x_2|$, with $f(1 - \vartheta) = f(\vartheta)$.

State $\omega_n^+ = \{x_1, x_2, x_3\}$ can occur if and only if $\omega_n^- = \{x_1, x_2\}$, $\{x_2, x_3\}$ or $\{x_1, x_3\}$. In the transition from ω_n^- to ω_n^+, the new coordinate is a uniform random sample from $[0, 1]$, so

$$p(x_1, x_2, x_3) = q(x_1, x_2) + q(x_2, x_3) + q(x_1, x_3). \tag{3.2}$$

This proves (2.2).

Conversely, $\omega_{n+1}^- = \{x_1, x_2\}$ if and only if, for some point ξ, $\omega_n^+ = \{x_1, x_2, \xi\}$ (the cross deleted in the nth iteration was at ξ). Then

$$q(x_1, x_2) = f(\vartheta(x_1, x_2)) = \int_0^1 p(\xi, x_1, x_2) P(\xi; x_1, x_2)\, d\xi,$$

where $P(\xi; x_1, x_2)$ is the conditional probability that the last deletion was at ξ, given the three coordinates ξ, x_1 and x_2. Now for a deletion to have occurred at ξ, Step 1 must have chosen P between ξ and the point x_1 or x_2

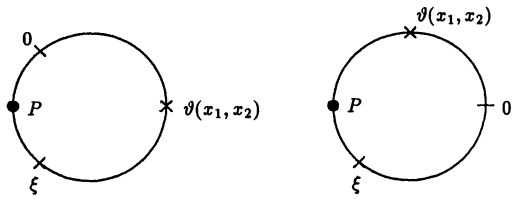


FIG. 2. *Point P of the preceding iteration.*

nearer to ξ in the clockwise direction (see Figure 2). In terms of the state $\{0, \xi, |x_1 - x_2|\}$, obtained by placing the origin at x_1 or x_2, $P(\xi; x_1, x_2)$ is either ξ or $\xi - \vartheta(x_1, x_2)$, so the transition equation becomes

$$q(x_1, x_2) = \int_0^{\vartheta(x_1, x_2)} p(0, \xi, \vartheta(x_1, x_2))\xi \, d\xi + \int_{\vartheta(x_1, x_2)}^1 p(0, \xi, \vartheta(x_1, x_2))(\xi - \vartheta(x_1, x_2)) \, d\xi. \tag{3.3}$$

A change of variables in the second integral converts (3.3) to

$$q(x_1, x_2) = \int_0^{\vartheta(x_1, x_2)} p(0, \xi, \vartheta(x_1, x_2))\xi \, d\xi + \int_0^{1-\vartheta(x_1, x_2)} p(0, \xi + \vartheta(x_1, x_2), \vartheta(x_1, x_2))\xi \, d\xi. \tag{3.4}$$

Substitute (3.2) and introduce the function $f(\vartheta)$ to obtain

$$f(\vartheta) = \int_0^{\vartheta} [f(\xi) + f(\vartheta) + f(\vartheta - \xi)]\xi \, d\xi + \int_0^{1-\vartheta} [f(\vartheta + \xi) + f(\vartheta) + f(\xi)]\xi \, d\xi.$$

Changes of variables, use of $f(\xi) = f(1 - \xi)$ and easy manipulations give the integral equation

$$(\tfrac{1}{2} + \vartheta - \vartheta^2)f(\vartheta) = (1 - \vartheta)\int_0^1 f(\xi) \, d\xi + (2\vartheta - 1)\int_0^{\vartheta} f(\xi) \, d\xi. \tag{3.5}$$

Define the new function $u(\vartheta) = \int_0^{\vartheta} f(\xi) \, d\xi$ and the constant $K = \int_0^{1/2} f(\xi) \, d\xi = \frac{1}{2}\int_0^1 f(\xi) \, d\xi$, where the last equality follows from the symmetry of f. Then (3.5) becomes

$$\left(\frac{1}{2} + \vartheta - \vartheta^2\right)\frac{du}{d\vartheta} = 2(1 - \vartheta)K + (2\vartheta - 1)u. \tag{3.6}$$

Rewrite (3.6) as

$$\frac{d}{d\vartheta}\left\{\left(\frac{1}{2} + \vartheta - \vartheta^2\right)u\right\} = 2(1 - \vartheta)K,$$

so that

$$\left(\tfrac{1}{2} + \vartheta - \vartheta^2\right)u = -(1-\vartheta)^2 K + \alpha, \tag{3.7}$$

with α a constant of integration. But $u\left(\tfrac{1}{2}\right) = K$ and (3.7) show that $\left(\tfrac{3}{4}\right)K = -\left(\tfrac{1}{4}\right)K + \alpha$, so $\alpha = K$ and (3.7) becomes

$$u = \frac{\left[1 - (1-\vartheta)^2\right]}{\tfrac{1}{2} + \vartheta - \vartheta^2} K. \tag{3.8}$$

Finally, substitution of (3.8) into (3.5) gives

$$f(\vartheta) = \frac{(1 - \vartheta + \vartheta^2)K}{\left(\tfrac{1}{2} + \vartheta - \vartheta^2\right)^2} = \frac{(1 - \vartheta(1-\vartheta))K}{\left[\tfrac{1}{2} - \vartheta(1-\vartheta)\right]^2}, \tag{3.9}$$

in which the symmetry about $\tfrac{1}{2}$ is obvious.

To evaluate K, note that

$$\int_0^1 \int_0^1 q(x_1, x_2)\, dx_1\, dx_2 = 2$$

because every unordered pair $\{x_1, x_2\}$, $x_1 \neq x_2$, appears twice in the integral over the unit square. Then

$$\begin{aligned} 2 = \int_0^1 \int_0^1 f(\vartheta(x_1, x_2))\, dx_1\, dx_2 &= 2\int_0^1 \int_0^{x_2} f(\vartheta(x_1, x_2))\, dx_1\, dx_2 \\ &= 2\int_0^1 \int_0^{x_2} f(x_2 - x_1)\, dx_1\, dx_2. \end{aligned}$$

A change of variables and substitution of (3.8) gives

$$1 = \int_0^1 u(x)\, dx = K\int_0^1 \frac{2\vartheta - \vartheta^2}{\tfrac{1}{2} + \vartheta - \vartheta^2}\, d\vartheta = K.$$

Substitution into (3.9) then proves (2.3):

$$f(\vartheta) = \frac{1 - \vartheta(1-\vartheta)}{\left[\tfrac{1}{2} + \vartheta(1-\vartheta)\right]^2}.$$

4. Proof of (2.4). The asymptotics in (2.4) come from the following stronger result.

THEOREM 1. *Let $N \to \infty$ and $n \to \infty$, keeping $n \geq N^2$. Then*

$$E[d_n] = \Theta\left(\frac{\log^{3/2} N}{N}\right).$$

The proof of Theorem 1 interprets the SR process as a special way of matching random points (dots) to other random points (crosses). The cylinder in Figure 3(a) is a product space of the circle in Figure 1 and a time axis (time increases moving down the cylinder) that shows the locations and arrival

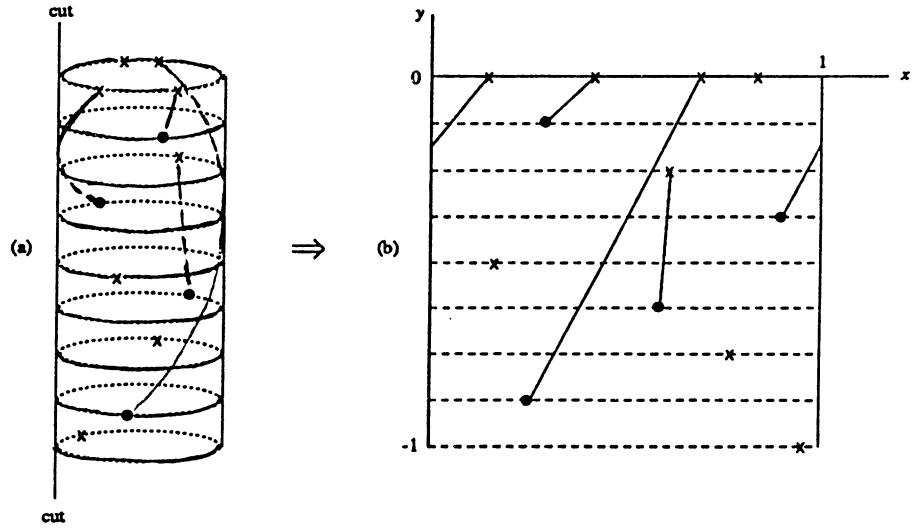


FIG. 3. *An* SR *matching*; $N = 4$, $n = 4$.

times of dots and crosses in the SR process. In each Step 1 of the process a dot arrives and an earlier cross is deleted. The dot and cross are then considered matched. Figure 3(a) represents the match by a geodesic path moving counterclockwise on the cylinder, backward in time, from the dot to the cross. Figure 3(b) shows the same cylinder, cut and flattened into a rectangle and then stretched linearly into a unit square. In Figure 3(b), some of the paths become straight line segments, to be called *matching lines* or simply *lines*. However, other paths, which crossed the cut in the cylinder, become broken *wrap-around lines*. The set of lines will be called an *SR matching*. Because of their special properties, SR matchings are not covered by existing results. Theorem 1 is proved by relating SR matchings to others that are better understood; namely, those in the standard model of planar matching theory.

In general, one is given a set $I_n = \{C_1, \ldots, C_n; D_1, \ldots, D_n\}$ of n crosses and n dots in the unit square. There need be no initial set of crosses at the top of the square, nor do the dots and crosses have to alternate as in SR matching problems. A set M of lines connecting cross–dot pairs in I_n is a *matching* if no cross or dot belongs to more than one pair. Each line is a straight line segment between the matched points; there are no wrap-around lines.

M is an *up-right* (UR) *matching*, if every line in M moves up and to the right, that is, if a dot at (x, y) is matched to a cross at (x', y'), then $x < x'$ and $y < y'$. Figure 4 shows a UR matching. Note that some points cannot be paired. A *maximum* UR matching pairs as many points as possible, but may leave some number U_m of points unmatched. Let H_n denote the sum of the horizontal components of the matching. The following theorem applies to U_n and H_n when the $2n$ points are independently and uniformly distributed at random.

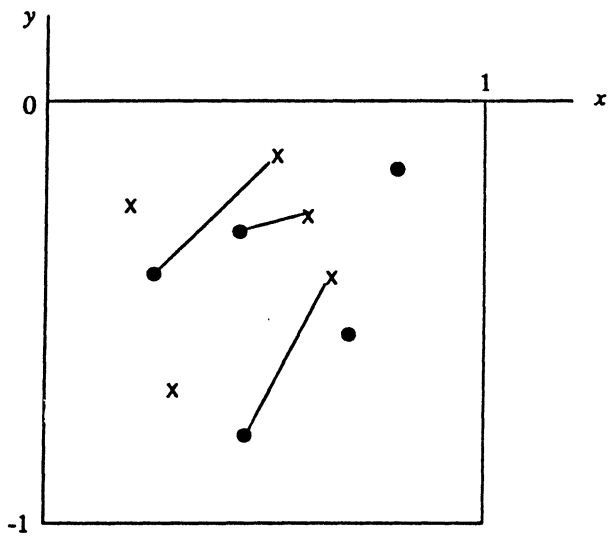


FIG. 4. *A* UR *matching*; $n = 5$, $U_n = 4$.

THEOREM 2 [Leighton and Shor (1989), Shor (1986) and Rhee and Talagrand (1988)]. *A maximum UR matching for n crosses and n dots, distributed uniformly and independently in a square, has*

$$E[U_n] = \Theta(\sqrt{n}\log^{3/4} n), \qquad E[H_n] = \Theta(\sqrt{n}\log^{3/4} n).$$

In addition, there exists a constant $\beta > 0$ *such that*

$$P(U_n \le \beta\sqrt{n}\log^{3/4} n) \ge 1 - n^{-\sqrt{\log n}}$$

for all n sufficiently large.

A UR matching remains a UR matching if the $2n$ points are displaced vertically, as long as they stay in the same vertical order. Then Theorem 2 also applies to points with independent x coordinates, uniformly distributed on $[0, 1]$, and with any distribution of y coordinates that keeps all $(2n)!$ orderings equally likely. For several other ways of stating the hypotheses without affecting the result, see Coffman and Lueker [(1991), page 53].

Theorem 2 does not apply directly to SR matchings because, as already noted, N initial crosses $C_1^0, \ldots, C_N^0$ of an SR matching lie on the x axis; SR matchings can contain wrap-around lines and the dots and crosses of SR matchings alternate. However, the proof of Theorem 1 adapts Theorem 2 to SR matchings in such a way that the result for $E[H_n]$ gives the desired bounds on $E[d_n]$. The proof requires three preliminary lemmas. The first shows that the random problems of Theorem 2, although lacking the property that dots and crosses must alternate, have a similar but weaker property.

Define $m = \lfloor \sqrt{n \log n} \rfloor$ and let $I_{n+m} = \{C_1, \ldots, C_{n+m}; D_1, \ldots, D_{n+m}\}$ be a random UR matching problem of Theorem 2. Let Z be a list of the C_i's and D_i's in decreasing order of their y coordinates. Define the random variable $z_i = +1$ or -1, $1 \le i \le 2(n+m)$, according as the ith point listed in Z is a cross or a dot, respectively.

LEMMA 1. *Let*

$$\delta_C(n+m) = \max_{1 \le j \le 2(n+m)} \sum_{i=1}^{j} z_i \tag{4.1}$$

denote the maximum excess of crosses over dots encountered in a scan of the list Z. Then

$$P(\delta_C(n+m) > m) = O\left(\frac{1}{n}\right). \tag{4.2}$$

By symmetry,

$$P(\delta_D(n+m) > m) = O\left(\frac{1}{n}\right), \tag{4.3}$$

where $\delta_D(n+m)$ is the maximum excess of dots over crosses in a scan of Z.

The Appendix proves Lemma 1.

A problem $\{C_1, \ldots, C_n; D_1, \ldots, D_n\}$ is *alternating* if the crosses and dots have independent x coordinates, uniformly distributed on $[0,1]$, but the points in order of decreasing y coordinate are $D_1, C_1, D_2, \ldots, D_n, C_n$. Lemma 2 uses Lemma 1 to prove that Theorem 2 holds even when restricted to alternating problems. In Lemma 2, M_n denotes a *maximum* UR matching for the problems of Theorem 2; then U_n denotes the number of points left unmatched by M_n. The notation M_n^*, U_n^*, H_n^* is defined for alternating problems in analogy with M_n, U_n, H_n.

LEMMA 2. *An alternating UR matching problem for n crosses and n dots, with all x coordinates independently and uniformly distributed on* $[0,1]$, *has*

$$E[U_n^*] = \Theta(\sqrt{n} \log^{3/4} n), \qquad E[H_n^*] = \Theta(\sqrt{n} \log^{3/4} n).$$

Moreover, there exists a constant $\beta > 0$ such that

$$P(U_n^* \le \beta \sqrt{n} \log^{3/4} n) \ge 1 - n^{-\sqrt{\log n}}$$

for all n sufficiently large.

PROOF OF $E[U_n^*] = O(\sqrt{n} \log^{3/4} n)$. Define $m = \lfloor \sqrt{n \log n} \rfloor$ and the random problem I_{n+m} as in Lemma 1. Consider a maximum UR matching M_{n+m} of I_{n+m}. The proof constructs from M_{n+m} a UR matching for a random alternating problem and then shows that the new matching satisfies the upper bound. Define the subset $I_n = \{C_{m+1}, \ldots, C_{n+m}; D_1, \ldots, D_n\}$ and let K_n

be the matching in M_{n+m} restricted to points in I_n, that is, a point of I_n is matched in K_n if and only if it is matched in M_{n+m}, and to a point also in I_n. I_n and I_{n+m} differ by $2m$ points, so the number V_n of unmatched points in K_n is at most $2m + U_{n+m}$. Because $m = \lfloor \sqrt{n \log n} \rfloor$, Theorem 2 shows that $E[U_{n+m}] = O(\sqrt{n}\ \log^{3/4} n)$, so

$$E[V_n] = E[U_{n+m}] + O\left(\sqrt{n \log n}\right) = O\left(\sqrt{n}\ \log^{3/4} n\right). \tag{4.4}$$

Next, change the y coordinates of points of I_n to obtain the alternating problem $I_n^* = \{C_1^*, \ldots, C_n^*; D_1^*, \ldots, D_n^*\}$, where the x coordinates of C_i^* and D_i^* are, respectively, those of C_{i+m} and D_i, $1 \le i \le n$. To make I_n^* alternating, the y coordinates of C_i^* and D_i^* can be, respectively, $-i/n$ and $-(2i - 1)/(2n)$, $1 \le i \le n$.

Construct the UR matching K_n^* of points in I_n^* such that D_i^* is matched to C_j^* if and only if (a) D_i is matched to C_{j+m} in K_n and (b) C_j^* is above D_i^*, $j \le i$. Let W_n be the number of points matched in K_n but not K_n^*, so that $V_n^* = V_n + W_n$ gives the number of unmatched points in K_n^*. The goal is now to show that $E[W_n]$ is so small that $E[V_n^*]$ and $E[V_n]$ have the same asymptotic bound in (4.4). The upper bound will then follow, because a maximum UR matching of I_n^* has at most as many unmatched points as K_n^*; that is, $U_n^* \le V_n^*$.

To estimate $E[W_n]$, suppose D_i and C_{j+m} are matched in K_n, but the corresponding points D_i^* and C_j^* are not matched in K_n^*. Then C_j^* is below D_i^* in I_n^* (i.e., $j > i$), but C_{j+m} is above D_i in I_n. This means that in I_{n+m} the crosses $C_1, \ldots, C_{j+m}$ are all above D_i, $i < j$. This in turn implies that $\delta_C(I_{n+m}) > j + m - i \ge m$, an event having probability $O(1/n)$ by Lemma 1. The bound $W_n \le n$ holds trivially, so $E[W_n] \le n \cdot O(1/n) = O(1)$ and

$$E[V_n^*] = E[V_n] + E[W_n] = O\left(\sqrt{n}\ \log^{3/4} n\right)$$

by (4.4). □

PROOF OF $E[U_n^*] = \Omega(\sqrt{n}\ \log^{3/4} n)$. The proof here is complementary to the preceding proof. Here, define $I_n = \{C_1, \ldots, C_n; D_{m+1}, \ldots, D_{n+m}\}$ and the alternating problem $I_n^* = \{C_1^*, \ldots, C_n^*; D_1^*, \ldots, D_n^*\}$, where C_i^* and D_i^* have the same x coordinates as C_i and D_{i+m}, respectively, $1 \le i \le n$. Now consider the maximum UR matching M_n^* of I_n^* leaving U_n^* points unmatched. Construct a matching K_n for I_n so that D_{i+m} and C_j are matched in K_n if and only if (a) D_i^* and C_j^* are matched in M_n^* and (b) C_j is up and to the right of D_{i+m}. By Theorem 2 the expected number of unmatched points in K_n has the estimate

$$E[V_n] = \Omega\left(\sqrt{n}\ \log^{3/4} n\right). \tag{4.5}$$

Let W_n be the number of points matched in M_n^* but unmatched in K_n, so that $V_n = U_n^* + W_n$. Suppose that D_{i+m} and C_j are unmatched in K_n, but D_i^* and C_j^* are matched in M_n^*, and hence $j \le i$. Then in I_{n+m} the dots $D_1, \ldots, D_{i+m}$ are all above C_j. Together with $j \le i$, this implies that $\delta_D(I_{n+m}) \ge m$. As before, this leads to $E[W_n] = O(1)$ by Lemma 1.

Then (4.5) and $E[V_n] = E[U_n^*] + O(1)$ imply the lower bound $E[U_n^*] = \Omega(\sqrt{n}\ \log^{3/4} n)$. □

Karp, Luby, and Marchetti-Spaccamela (1984) prove that the following simple algorithm gives maximum UR matchings for problems I_n: Scan the dots in order of decreasing y coordinate, matching each new dot D to the leftmost unmatched cross, if any, that lies above and to the right of D. Our final lemma proves an analogous optimality result for SR matchings.

SR matchings are special instances of *full matchings* that would be obtained from Figure 1 if the dot (point P) still moved counterclockwise, but did not necessarily stop at the first cross encountered. Like SR matchings, full matchings map into a unit square with N crosses on the x axis, an alternating problem I_n and n matching lines. A line still moves upward and to the right with a gap if it is a wrap-around line, but it can end at any unused cross.

LEMMA 3. *Among full matchings, the SR matching has the smallest sum of horizontal components.*

The Appendix gives a short proof.

The ground is now prepared for a proof of Theorem 1. Apart from initial sets of crosses, the proof deals only with alternating problems; the asterisk notation of Lemma 2 is no longer needed. The proof requires asymptotic upper and lower bounds on the mean of $H^S = d_1 + \cdots + d_n$ in an SR matching. Both begin by deriving a bound for a special limiting case with $N \to \infty$ and $n \to \infty$, keeping

$$n = n(\kappa, N) = \lfloor \kappa N^2/\log^{3/4} N \rfloor,$$

where κ is a positive constant to be determined. All bounds are independent of the configuration of initial crosses.

Lower bound. Consider $n = n(\kappa_1, N)$ steps of the SR process with κ_1 unspecified for the moment. Cut the cylinder of Figure 3(a) by a vertical line at an angle uniformly distributed around the cylinder and independent of the N initial crosses. Then construct Figure 3(b). Define $\eta(x)$ to be the number of matching lines that cross a vertical line at coordinate x, $0 \le x \le 1$. Then

$$E[H^S] = \int_0^1 E[\eta(x)]\, dx. \tag{4.6}$$

The random way that the cylinder of Figure 3(a) was cut ensures that $E[\eta(x)]$ is independent of x and of the configuration of crosses at the top of the cylinder. Then $E[\eta(x)] = E[\eta(1)]$ and by (4.6), $E[H^S] = E[\eta(1)]$; that is, $E[H^S]$ is the expected number of wrap-around lines.

Now remove from the SR matching all wrap-around lines, the N crosses C_i^0 of the initial state and all lines incident to the C_i^0. What remains is a UR matching with an expected number of unmatched dots, $E[V_n]$, that is at most

the expected number $E[\eta(1)]$ of removed wrap-around lines plus the number N of initial crosses:

$$E[V_n] \le E[\eta(1)] + N. \tag{4.7}$$

By Lemma 2, $E[V_n] = \Omega(\sqrt{n}\log^{3/4} n)$, so with κ_1 sufficiently large, the substitution $n = n(\kappa_1, N)$ shows that there exists a $\gamma > 1$ such that $E[V_n] > \gamma N$ for all N sufficiently large. Then (4.7) implies that $E[\eta(1)] = \Omega(N)$. Finally, the average of the d_k over the $n(\kappa_1, N)$ iterations is then

$$\frac{E[H_n^S]}{n} = \frac{E[\eta(1)]}{n} = \Omega\left(\frac{\log^{3/2} N}{N}\right).$$

To obtain the lower bound of Theorem 1, run the SR process for at least N^2 steps. Break this sequence into a large number (of order $\log^{3/2} N$, at least) of consecutive trials of $n(\kappa_1, N)$ steps each. The bound just derived applies to each trial, so the lower bound of Theorem 1 must apply to the entire sequence.

Upper bound. Now begin with a sequence of $n = n(\kappa_2, N)$ steps of the SR process. The argument uses Lemma 3, which shows that the expected sum of horizontal components in any full matching is an upper bound on $E[H^S]$. An algorithm for constructing a suitable full matching follows.

ALGORITHM A. N initial crosses and an alternating problem I_n form the input. As a preliminary step, construct a maximum UR matching M_n for I_n and let U_n be the number of points M_n fails to match. As noted prior to Lemma 3, M_n can be found by the algorithm of Karp, Luby and Marchetti-Spaccamela (1984). If M_n leaves more than N dots unmatched, that is, if $U_n > 2N$, then discard M_n and let the SR matching itself be the output of the algorithm. But if $U_n \le 2N$, then extend M_n to a full matching by pairing the $U_n/2$ unmatched dots with $U_n/2$ of the N initial crosses in any way (this step may introduce wrap-around lines). The resulting matching is then the output of the algorithm.

Let H_n^A denote the sum of horizontal components in the full matching constructed by Algorithm A. H_n is defined similarly for M_n. If $U_n \le 2N$, then $H_n^A \le H_n + N$. The bound $H_n^A \le n$ holds trivially, so

$$E[H_n^A] \le E[H_n] + N + nP(U_n > 2N). \tag{4.8}$$

Now with $n = n(\kappa_2, N)$ and N sufficiently large, $\kappa_2 > 0$ can be chosen so that $U_n \le 2N$ with probability at least $1 - n^{-\sqrt{\log n}}$. With β as given in Lemma 2, a simple calculation shows that any κ_2, $0 < \kappa_2 < \sqrt{2}/\beta^2$, will do. Then because $E[H_n] = O(\sqrt{n}\log^{3/4} n)$ by Lemma 2, (4.8) gives

$$E[H_n^A] \le O(\sqrt{n}\log^{3/4} n) + N + n \cdot n^{-\sqrt{\log n}} = O(N)$$

and

$$\frac{E[H_n^S]}{n} \le \frac{E[H_n^A]}{n} = O\left(\frac{\log^{3/2} N}{N}\right).$$

Again, the proof for any number $n > N^2$ of steps follows by breaking the sequence of n steps into consecutive trials of length $n(\kappa_2, N)$ and applying the bound just derived. □

APPENDIX

Proofs of Lemmas 1 and 3.

PROOF OF LEMMA 1. The proof can be put in terms of random walks that start at level 0 and make $2r$ steps z_i, equally likely to be $+1$ or -1. Let $u(k)$ be the probability that such a walk ends at level $2k$ and suppose $k \ge 0$. The walk has $r + k$ positive steps and $r - k$ negative steps, so

$$u(k) = \binom{2r}{r+k} \Big/ 2^{2r}.$$

All such walks ending at $2k$ visit level k at some step. Take any such walk and reflect the part of the walk following the last visit to level k about level k; the reflected walk ends at level 0. The reflection principle [Feller (1957), Section III.2] then shows that walks starting at level 0, reaching level k and ending at level 0, all in $2r$ steps, also have probability $u(k)$. The conditional probability that a walk reaches level k, given that it starts and ends at level 0, is then $u(k)/u(0)$.

In (4.2), $P(\delta_C(n + m) > m)$ is the probability that a random walk, beginning at level 0 and ending at level 0 after $2r = 2(n + m)$ steps, reaches level $k = m + 1$. Then

$$P(\delta_C(r) \ge k) = \frac{u(k)}{u(0)} = \frac{r!r!}{(r+k)!(r-k)!} = \frac{r(r-1)\cdots(r+1-k)}{(r+1)(r+2)\cdots(r+k)}.$$

Combine the jth factors of numerator and denominator into

$$\frac{r+1-j}{r+j} = 1 - \frac{2j-1}{r+j} < 1 - \frac{2j-1}{r+k} < \exp\left(-\frac{2j-1}{r+k}\right),$$

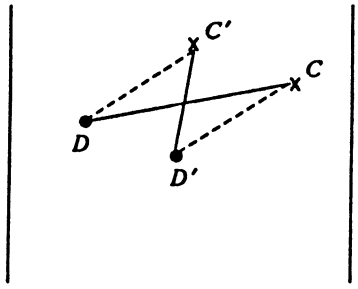
so that

$$P(\delta_C(r) \ge k) < \exp\left\{-\frac{1+3+\cdots+(2k-1)}{r+k}\right\} = \exp\left(-\frac{k^2}{r+k}\right).$$

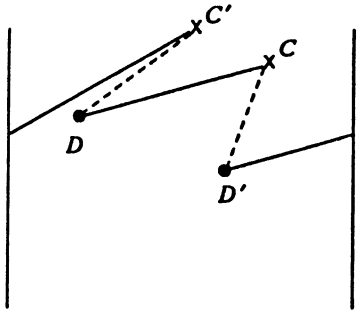
To obtain (4.2), substitute $n + m$ for r and $m + 1$ for k, with $m = \lfloor \sqrt{n} \log n \rfloor$. □

PROOF OF LEMMA 3. Let M^S and M^F be, respectively, the SR matching and an arbitrary, but different full matching on the same problem. Let D be the highest dot matched to different crosses, to C in M^F and C' in M^S. Assume without loss of generality that the line between D and C is not a wrap-around line (a cylinder having the matching M^F can always be cut along a vertical so that this property holds). C' is closer on the right of D than C because D is the first dot where M^F and M^S disagree. Either or both of C and C' may be one of the initial crosses.

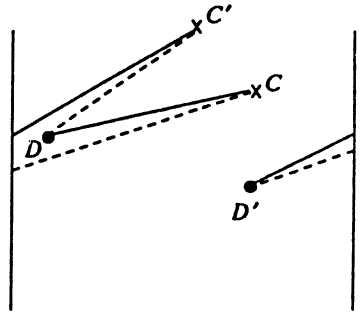
Change M^F by pairing D with C'. If C' is unmatched in M^F, then the new matching is a valid full matching with a smaller sum of horizontal edge components. On the other hand, if a dot D' is paired with C' in M^F, then pair D' with C. The result is shown in Figure 5, in which D' is to the left of C', between C' and C, or to the right of C. In each case, the sum of the horizontal components has not increased.



(a) D' to left of C'; sum of horizontal edge components unchanged



(b) D' between C' and C; sum of horizontal edge components is reduced



(c) D' to right of C; sum of horizontal edge components is unchanged

FIG. 5. *Transforming a full matching. Dashed and solid lines are, respectively, new and old lines.*

Repeating the preceding transformation at most n times converts M^F into M^S without increasing the sum of horizontal components. □

REFERENCES

COFFMAN, E. G., JR. and LUEKER, G. S. (1991). *Probabilistic Analysis of Packing and Partitioning Algorithms*. Wiley, New York.

COFFMAN, E. G., JR. and HOFRI, M. (1986). Queueing models of secondary storage devices. *Queueing Syst.* **2** 129–168.

FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications* **1**, 2nd ed. Wiley, New York.

FRIEDMAN, N. A. (1970). *Introduction to Ergodic Theory*. Van Nostrand Reinhold, New York.

FULLER, S. H. and BASKETT, F. (1975). An analysis of drum storage units. *J. Assoc. Comput. Mach.* **22** 83–105.

KARP, R. M., LUBY, M. and MARCHETTI-SPACCAMELA, A. (1984). A probabilistic analysis of multidimensional bin packing problems. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing* 289–298, Assoc. Comput. Mach., New York.

LEIGHTON, F. T. and SHOR, P. W. (1989). Tight bounds for minimax grid matching with applications to the average-case analysis of algorithms. *Combinatorica* **9** 161–187.

RHEE, W. T. and TALAGRAND, M. (1988). Exact bounds for the stochastic upward matching problem. *Trans. Amer. Math. Soc.* **307** 109–125.

SHOR, P. W. (1986). The average-case analysis of some on-line algorithms for bin packing. *Combinatorica* **6** 179–200.

AT & T BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-2070