The Annals of Probability
1977, Vol. 5, No. 5, 814–817

A GENERALIZATION OF THE KARLIN–McGREGOR THEOREM ON COINCIDENCE PROBABILITIES AND AN APPLICATION TO CLUSTERING

BY F. K. HWANG

Bell Laboratories

Karlin and McGregor calculated the coincidence probabilities for n particles independently executing a Markov process of a certain class. This note extends their result by allowing the particles to have different stopping times. Applied to a one-dimensional clustering problem, this gives a new solution computationally simpler than previous ones.

1. Introduction. Consider a continuous-time Markov process whose state-space is the set of integers and whose s states are all stable. Suppose that n labeled particles start out in states $\alpha_1 > \cdots > \alpha_n$ respectively and execute the process simultaneously and independently under the restriction that whenever a transition occurs, the particle moves from a given state into only one of the two neighboring states. Karlin and McGregor [2] proved a theorem which gives the coincidence probabilities for these particles. In this note we generalize their theorem by allowing the particles to have different stopping times. The generalized theorem is applied to a one-dimensional clustering problem.

2. A generalization of the Karlin–McGregor theorem. Let $S_i(t)$ denote the state in which particle i is found at time t. Then $S_i(0) = \alpha_i$. Let E be the event that $S_i(t_i) = \beta_i$ for $i = 1, \cdots, n$ with $\beta_1 > \beta_2 > \cdots > \beta_2$, where t_i is a given stopping time for particle i, without any two particles ever having been coincident during the intervening time. Our problem is to find $\Pr(E)$, the probability of the event E.

To compute $\Pr(E)$, we need to consider a larger ensemble of events. Let σ be a permutation of the set $\{1, \cdots, n\}$. Let E_{ij} be the event that particle i starts in state α_i and stops at a given time t_j in state β_j under the condition C_{ij} (a generalization of the condition used in [4]) which prescribes that for every $t_k < t_j$

$$S_i(t_k) > \beta_k \quad \text{if} \quad j < k\,,$$
$$S_i(t_k) < \beta_k \quad \text{if} \quad j > k\,.$$

Let p_{ij} denote the probability of E_{ij}.

GENERALIZED KARLIN–McGREGOR THEOREM. *Suppose either* $t_1 \geqq t_2 \geqq \cdots \geqq t_n$ *or* $t_1 \leqq t_2 \leqq \cdots \leqq t_n$. *Then* $\Pr(E) = \det|p_{ij}|$.

(The special case that all the stopping times are identical, hence $\{C_{ij}\}$ a vacuous set, is known as the Karlin–McGregor theorem.)

Received January 3, 1977.

AMS 1970 *subject classifications.* Primary 60J05; Secondary 60E05.

Key words and phrases. Coincidence probabilities, Markov process, stopping time, cluster, generalized birthday problem.

PROOF. Let σ be a permutation of the set $\{1, \cdots, n\}$. Define

$$E_\sigma = \bigcap_{j=1}^n E_{\sigma(j)j} .$$

Then

$$\Pr(E_\sigma) = \prod_{j=1}^n p_{\sigma(j)j} .$$

Therefore

$$\det |p_{ij}| = \sum_\sigma \text{sign}(\sigma) \Pr(E_\sigma)$$

where sign $(\sigma) = 1$ or -1 according to whether σ is an even or odd permutation.

We first note that if σ is not the identity permutation, then E_σ can be realized only when a coincidence state has occurred. This conclusion is of course forced by the conditions $\{C_{ij}\}$.

Consider any realization θ (of the n joint executions) which has a coincidence state. Let t^0 be the first time a coincidence occurs in θ, say between particle i and particle j with $i < j$ (our argument can be easily modified for the case that more than two particles coincide at t^0). Let R be the closed region formed by the path of particle i, the path of particle j and the line $t = 0$. Then no stopping time can lie in R. Suppose the contrary, that t_x is such a stopping time. Then either particle x has a coincidence (with particle i or particle j) before t^0 or necessarily $i < x < j$ and $t_x < \min\{t_i, t_j\}$. But the former possibility is a contradiction to our definition of t^0 and the latter a contradiction to our assumption on the stopping times.

Let θ' be the realization obtained from θ by interchanging the paths of particle i and particle j after t^0. Then clearly $\Pr(\theta) = \Pr(\theta')$. Furthermore, if there exists a stopping time t_k such that condition C_{ij} is satisfied (on t_k) under θ but not under θ', then t_k must lie in the region R. Since no such t_k exists, $\theta \in E_\sigma$ implies $\theta' \in E_{\sigma'}$ for some σ'. It is also clear that sign $(\sigma) = -$sign (σ'). Therefore $\Pr(\theta)$ and $\Pr(\theta')$ cancel each other out. Thus only those realizations from E_I, where I is the identity permutation, which do not have coincidence states contribute to the det $|p_{ij}|$. Since they all have plus signs, we have proved:

$$\det |p_{ij}| = \Pr(E) .$$

3. A one-dimensional clustering problem. Consider $N (\geqq 2)$ points distributed independently and uniformly in $[0, 1)$. Let n_p denote the maximum number of points contained in a subinterval of size p. The problem is to find the probability distribution function $\Pr(n_p < n)$ for all p, n, and N. Wallenstein and Naus [5] gave a formula for $\Pr(n_p \leqq n)$ in the case that p is rational. Huntington and Naus [1] gave a computationally simpler formula which imposes no restrictions on the parameters. In this section, we apply the generalized Karlin–McGregor theorem to give a formula for $\Pr(n_p < n)$ (with no restrictions) which achieves further computational simplification. Our approach is very similar to what Saperstein [4] did for a discrete clustering problem known as the generalized birthday problem.

Define $L = [p^{-1}]$, the largest integer not exceeding p^{-1}. Assume $p^{-1} > L$, since otherwise we can compute $\Pr(n_p < n)$ by the formula given by Naus [3]. Divide the interval $[0, 1)$ into $L + 1$ disjoint half-open intervals $I_1, \cdots, I_{L+1}$ where the first L intervals are of length p and the last one is of length $p' = 1 - Lp$. Let n_i, $i = 1, \cdots, L + 1$, denote the number of points in I_i and let $\pi(n, L + 1, N)$ be the set of all partitions of N objects into $L + 1$ ordered parts such that no part contains n objects or more. Let $y_i(t)$ be the number of points in the subinterval $[(i - 1)p, (i - 1)p + t)$ of I_i. Define

$$\begin{aligned} p_i &= p \qquad \text{for} \quad 1 \leqq i \leqq L , \\ &= p' \qquad \text{for} \quad i = L + 1 . \end{aligned}$$

Then $y_i(p_i) = n_i$.

Let F be the event that a given $(L + 1)$-tuple $(n_1, \cdots, n_{L+1}) \in \pi(n, L + 1, N)$. Define

$$\alpha_i = \textstyle\sum_{j=1}^{i-1} n_j - (i - 1)n$$

and

$$\beta_i = \alpha_i + y_i(p_i) = \alpha_i + n_i .$$

Then

$$\begin{aligned} &\Pr(n_p < n, F) \\ &\quad = \Pr(n_i + y_{i+1}(t) - y_i(t) < n \;\text{ for all }\; i = 1, \cdots, L \;\text{ and }\; 0 \leqq t \leqq p_i, F) \\ &\quad = \Pr(n_i - n + y_{i+1}(t) < y_i(t) \;\text{ for all }\; i = 1, \cdots, L \;\text{ and }\; 0 \leqq t \leqq p_i, F) \\ &\quad = \Pr(\alpha_{i+1} + y_{i+1}(t) < \alpha_i + y_i(t) \\ &\qquad\qquad \text{for all }\; i = 1, \cdots, L \;\text{ and }\; 0 \leqq t \leqq p_i, F) . \end{aligned}$$

However, $\alpha_1 > \alpha_2 > \cdots > \alpha_{L+1}$ and $\beta_1 > \beta_2 > \cdots > \beta_{L+1}$ under F. Therefore the event $(n_p < n, F)$ can be interpreted as the event that $L + 1$ particles with stopping times $t_i = p_i$ jointly execute a Poisson process without coincidence. According to the generalized Karlin–McGregor theorem, the probability of this event is

$$\det |p_{ij}|$$

where C_{ij} is the condition

$$\alpha_i + y_j(p') > \beta_{L+1} \qquad \text{for} \quad i = 1, \cdots, L + 1 \quad \text{and} \quad j = 1, \cdots, L ,$$

and

$$\begin{aligned} p_{ij} &= \textstyle\sum_{x=\beta_{L+1}+1}^{\beta_j-\alpha_i} \dfrac{(\lambda p')^x e^{-\lambda p'}}{x!} \cdot \dfrac{[\lambda(p_j - p')]^{\beta_j-\alpha_i-x} e^{-\lambda(p_j-p')}}{(\beta_j - \alpha_i - x)!} \\ &= \dfrac{\lambda^{\beta_j-\alpha_i} e^{-\lambda p_j}}{(\beta_j - \alpha_i)!} \textstyle\sum_{x=\beta_{L+1}+1}^{\beta_j-\alpha_i} \binom{\beta_j-\alpha_i}{x} (p')^x (p_j - p')^{\beta_j-\alpha_i-x} , \end{aligned}$$

which we abbreviate as $(\lambda^{\beta_j-\alpha_i} e^{-\lambda p_j}/(\beta_j - \alpha_i)!) g_{ij}$. Therefore

$$\begin{aligned} \Pr(n_p < n, F \mid (n_1, \cdots, n_{L+1})) &= \frac{\det |p_{ij}|}{\prod_{j=1}^{L+1} ((\lambda p_j)^{n_j} e^{-\lambda p_j}/n_j!)} \\ &= \det \left| \frac{n_j!\, g_{ij}}{(\beta_j - \alpha_i)!} \right| . \end{aligned}$$

Finally,

$$\Pr(n_p < n) = \sum_{(n_1,\cdots,n_{L+1})\in\pi(n,L+1,N)} \Pr(n_1, \cdots, n_{L+1} \mid N) \times \Pr(n_p < n, F \mid (n_1, \cdots, n_{L+1}))$$
$$= N! \sum_{(n_1,\cdots,n_{L+1})\in\pi(n,L+1,N)} \det\left|\frac{g_{ij}}{(\beta_j - \alpha_i)!}\right| .$$

The above formula involves computations of the determinant of a $(L+1) \times (L+1)$ matrix as many times as there are partitions in $\pi(n, L+1, N)$. The formula given by Huntington and Naus [1] involves similar computations but the number of matrices involved equals the number of ways of partitioning N objects into $2L+1$ ordered parts such that the number of objects in two adjacent parts is always less than n. It is clear that the former collection contains many fewer elements than the latter. The fact that g_{ij} is a sum has negligible effect on the computing work since the main effort is spent in inverting the $(L+1) \times (L+1)$ matrices.

REFERENCES

[1] HUNTINGTON, R. J. and NAUS, J. I. (1975). A simpler expression for *k*th nearest neighbor coincidence probabilities. *Ann. Probability* **3** 894–896.

[2] KARLIN, S. and MCGREGOR, J. (1959). Coincidence probabilities. *Pacific J. Math.* **9** 1141–1164.

[3] NAUS, J. I. (1966). Some probabilities, expectations, and variances for the size of largest clusters and smallest intervals. *J. Amer. Statist. Assoc.* **61** 1191–1199.

[4] SAPERSTEIN, B. (1975). Note on a clustering problem. *J. Appl. Probability* **12** 629–632.

[5] WALLENSTEIN, S. R. and NAUS, J. I. (1973). Probabilities for a *k*th nearest neighbor problem on the line. *Ann. Probability* **1** 188–190.

BELL LABORATORIES
MURRAY HILL, NEW JERSEY 07974