論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Some Problems for Applications of Markov Chains
著者(和文)	高橋幸雄
Author(English)	YUKIO TAKAHASHI
出典(和文)	学位:理学博士, 学位授与機関:東京工業大学, 報告番号:甲第515号, 授与年月日:1972年3月26日, 学位の種別:課程博士, 審査員:森村英典
Citation(English)	Degree:Doctor of Science, Conferring organization: Tokyo Institute of Technology, Report number:甲第515号, Conferred date:1972/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

SOME PROBLEMS FOR APPLICATIONS OF MARKOV CHAINS

BY YUKIO TAKAHASHI

PREFACE

Markov chain is one of the simplest stochastic processes and has many applications in various fields, not only in natural science but also in social and cultual sciences. The basic concepts of Markov chains were introduced by A. A. Markov in 1907. Since that time, properties of Markov chains have been studied by many mathematicians, and at present almost all elementary problems in the theory of Markov chains have been solved. These results were published in many books such as [1], [2], [3] and in many related books such as [4], [5], [6], [7], [8], [9], [10], [11].

However when we want to apply the Markov chain models to practical problems, new difficulties arise. In practical problems, the fundamental conditions for stationary Markov chains may be slightly violated. In such cases, what will happen if we apply the Markov chain models to practical problems assuming that all the conditions for the stationary Markov chains are satisfied? Can we expect that our modelings cause no serious errors? To answer these questions, we have to develope a new theory for applications of Markov chains.

The needs for developing such theories for applications also arise in almost all branches for modelings. However it was only recently that the needs were recognized and studies for applications were begun. So, only a few results have been obtained.

Studies for applications of a model may require discussions without some of original conditions which simplify the model. So, such studies may be very difficult and some problems cannot be solved without experiences through practical examples or simulation experiments.

In this thesis, some problems for applications of Markov chains will be exposed and solved. The author wishes that the theory for applications of Markov chains will be greately developed and Markov chains will be correctly applied in many fields.

In Chapter 1 of this thesis, problems for applications of Markov chains are classified by conditions being lacking, and a survey of related works is made. Three problems are solved in Chapters 2,3 and 4. In Chapter 2, the range of eigenvalues of stochastic matrices is treated and the effects of lumpings of states on the eigenvalues of transition matrices are discussed. In Chapter 3, the effects of small deviations in the transition matrices are discussed. In Chapter 4, an iterative method for obtaining the numerical values of the limiting vectors, or the stationary distributions, of Markov chains is proposed and in Chapter 5, an example of the application of the method is exhibited.

The articlein Appendix 1 is the author's thesis of the master's degree, and it treats a problem arising when the stationarity of a Markov chain is lost. It is within our stream of consideration. The article in Appendix 2 is concerning to a queueing model. It is out of our stream of consideration. However since it is an application of a Markov chain, it is inserted in this thesis as an appendix.

In this thesis we principally consider problems arising when we try to apply stationary Markov chains with finite state spaces and discrete time parameters. As a rule, we will use the same terms as in [1] except for Appendix 1. In Appendix 1, we will use the terms introduced in [4].

The author wishes to express his sincere thanks to Professor H. Morimura for valuable suggestions and encouragements.

References

[1] Kemeny, J.G. & Snell, J.L., Finite Markov Chains. Van Nostrand Comp., Princeton (1960).

[2] Kemeny, J.G., Snell, J.L. & Knapp, A.W., Denumerable Markov Chains. Van Nostrand Comp., Princeton (1966).

[3] Chung, K.L., Markov Chains with Stationary Transition Probabilities (2nd ed.). Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen, Band 104. Springer-Verlag, Heidelgerg (1967).

[4] Feller, W., An Introduction to Probability Theory and Its Applications. vol. 1 (3rd ed.). John Wiley & Sons, Inc., New York (1968).

[5] Doob, J.L., Stochastic Processes. John Wiley & Sons, Inc., New York (1953).

[6] Spitzer, F., Principles of Random Walk. Van Nostrand Comp., Princeton (1964).

[7] Takács, L., Stochastic Processes, Problems and Solutions. Methuen, London (1962).

[8] Parzen, E., Stochastic Processes. Holden-Day, San Francisco (1965).

[9] Cox, D.R. & Miller, H.D., The Theory of Stochastic Processes. Methuen, London (1965).

[10] Karlin, S., A First Course in Stochastic Processes. Academic Press, New York (1966).

[11] Cohen, J.W., The Single Server Queue. North Holland Publishing Company, Amsterdam (1969).

CONTENTS

CHAPTER 1 PROBLEMS FOR APPLICATIONS OF MARKOV CHAINS

1. Introduction

Markov chains have been applied in many fields. Recently in biology and social and cultual science, they have been recognized as a new tool for analyses, and many applications have been made. At the same time, demands for a new theory for application of Markov chains were brought through these applications. In these applications, it is very rare that all the conditions for stationary Markov chains are satisfied. In such cases, nevertheless, we cannot help applying stationary Markov chains to the problems with mere expectations that these approximations cause no serious discrepancies. At present, we have only a few assurances for such expectations. We should study more and expand the range of such assurances.

Thus we should develope a new theory for applications of Markov chains. Main problems of it will be studies of the cases where there exist slight lacks in the conditions for stationary Markov chains. So far only a few works have been done. Since it was only recently that the needs for the theory were recognized, studies for it have just begun.

2. Theoretical applications and practical applications

Applications of Markov chains can be divided into two groups. One of them is the group of applications which use Markov chains as a tool for analyses in another stochastic models with special structures, e.g., random walks, queueing models, inventory models and so on. These applications will be called as theoretical applications. Another group consists of applications in which we define the states, estimate the transition probabilities and then guess the manners of transitions or obtain the limiting vectors. These applications will be called as practical applications. We should sharply distinguish these two groups.

In a theoretical application, the fundamental conditions for a stationary Markov chain, e.g., the Markov property, the stationarity, and so on, will be derived from the assumptions of the original model. For example, the imbedded Markov chain method for a queueing model of type M/G/1 is a typical theoretical application. In the method, the Markov property can be derived from the assumptions of the Poisson input and mutually independent and identically distributed service times. On the other hand, in a practical application , we should first examine whether the conditions for a stationary Markov chain are satisfied. The brand switching model in the marketing research is a typical practical application. For the model, many discussions about methods for determining the states of a Markov chain and methods for the estimation of transition probabilities have been made.

Thus in a theoretical application all the conditions have been assumed, but in a practical application we should examine the appropriateness of the conditions. In the sense, practical applications are in the higher level than theoretical applications.

3. Four main problems

We first consider practical applications. Markov chains being used in practical applications are, in most cases, stationary Markov chains with finite state spaces and discrete time parameters. So, we restrict our considerations in applications of such Markov chains. In a practical application, we should examine the appropriateness of the conditions. In this case, the following two conditions should be examined.

(1) Markov property

(2) Stationarity of the transition probabilities

As stated later (in Section 4), the Markov property is deeply related to the states being adopted.

If the above two conditions are satisfied, then we should estimate the transition probabilities. In the estimation, we should pay attention to

(3) Accuracy of the transition probabilities

If the estimated values of the transition probabilities are not so accurate, we cannot believe the results derived from them.

There is an approach to these problems with statistical methods (see [1] and [2]). Unfortunately, these methods require large size samples, and ordinarily we have not samples of so large size that we can obtain sufficiently accurate judgements. So we shall proceed with another approach. Namely we shall study (1) the cases in which the Markov property is slightly lost, (2) the cases in which the stationarity in a long period cannot be expected, and (3) the cases in which the accuracy of the transition probabilities is not enough.

Now we consider theoretical applications. As mentioned in the preceding section, in these applications the fundamental conditions for Markov chains have been assumed. So the above three problems for practical applications need not be considered. However in theoretical applications, the number of states is generally very large, and sometimes it is infinitely large. In such cases, it is very difficult to obtain the numerical values of the characteristic quantities of the chains such as the limiting vectors, the mean first passage times, taboo probabilities and so on.

We have many numerical methods for the solutions of systems of equations. However we will probably be able to develope another and more suitable numerical methods for obtaining the limiting vectors if we skillfully use special structures of transition matrices. Thus our fourth problem is to develope

(4) Numerical methods for Markov chains

In the following sections, we shall discuss about these four problems and survey related works which have been done.

4. On the Markov property

The most important concept for Markov chains is the Markov property. This property is briefly mentioned as follows. " Only the knowledge of the present state of the chain affects on the next transition of the chain, and the knowledge of the past states does not affects on it." Let $S = \{ s_1, s_2, \cdots, s_s \}$ be the set of all possible states (the state space) of the process and denote by f_n the state of the process at time n $(n = 0,1,2, \cdots)$. Then we say that the process has the Markov property, or the process is a Markov chain, if

$$P \{ f_{n+1} = s_j \mid f_0 = s_{i_0}, f_1 = s_{i_1}, \cdots, f_n = s_{i_n} \}$$
$$= P \{ f_{n+1} = s_j \mid f_n = s_{i_n} \}$$

for every $n = 1,2,3, \cdots$ and $s_j, s_{i_0}, s_{i_1}, \cdots, s_{i_n} \in S$. If a process has the Markov property, then all the probability laws of the process are uniquely determined by the initial distribution $P \{ f_0 = s_i \}$, $s_i \in S$, and the transition probabilities $P \{ f_{n+1} = s_j \mid f_n = s_i \}$ for $n = 1,2,3, \cdots$ and $s_i, s_j \in S$. If the Markov chain is stationary, all the probability laws of the chain is determined by the initial vector $\pi = \{ \pi_i = P \{ f_0 = s_i \} ; s_i \in S \}$ and the stationary transition matrix $P = \{ p_{ij} = P \{ f_1 = s_j \mid f_0 = s_i \} ; (s_i, s_j) \in S \times S \}$.

In a practical application, it is difficult to define the states so that the process has the Markov property. One approach to the problem is using the concept of multiple Markov chains.

We consider a process $\{f_n\}$ with finite state space S. The second order process $\{h_n\}$ derived from the process $\{f_n\}$ is the process defined by the following relation:

$$h_n = (s_i, s_j) \text{ if and only if } f_{n-1} = s_i \text{ and } f_n = s_j ,$$

where f_n and h_n represent the states at time n of the original and the derived processes respectively. Thus the second order process has the state space $S \times S$. In like manner we can define the t-th order process derived from $\{f_n\}$. A process will be called as a t-th order process or a multiple process if it is the t-th order process derived from a process.

Any process $\{f_n\}$ with finite state space and discrete time parameter can be considered as a Markov chain with the infinite state space $\overline{S} = S \cup (S \times S) \cup (S \times S \times S) \cup \cdots$. Because we can define a process $\{h_n\}$ whose state at time n is the (n+1)-th order vector $h_n = (f_0, f_1, \cdots, f_n)$, where f_n represents the state of the original process at time n. Then the derived process $\{h_n\}$ clearly has the Markov property. This construction of a Markov chain is a trivial one. However it suggests an approach to the problem concerning to the Markov property.

We consider a process $\{f_n\}$ on a finite state space S. The knowledge of the states at early times $f_0, f_1, f_2, \cdots$ have ordinarily little influence on the state f_n for large n. Hence we may expect that if we take sufficiently large t, then the process (or, strictly speaking, the t-th order process derived from it) can be approximated by a t-th order Markov chain. Hence we are satisfied if we obtain the smallest one of such t . In order to obtain such t, we may proceed as follows. If the observed process has not the Markov property, then we examine whether the second order process derived from it has the Markov property or not. If the second order process has not, then we examine the third order process. If the third order process has not the Markov property, then we examine the fourth order process. •••••••

In order adopt this procedure, we need a method for deciding whether a process has the Markov property or not. T. W. Anderson & Leo A. Goodman [1] gave a statistical test for the Markov property. However, as mentioned in the preceding section, statistical tests require samples of very large size. But ordinarily we have not samples of so large size that we can obtain sufficiently accurate judgements.

Thus the problem how to define the states so that the process has the Markov property, cannot solved by statistical methods alone. We should study more about the properties of the observed processes from the point of view of the probability theory.

We saw that ayn process with a finite state space can be approximated by a multiple Markov chain. Hence we should study processes with finite state spaces, from each of which we can derive a multiple Markov chain. However the class of these processes is contained in the class of lumped processes (defined below) derived from finite Markov chains. So we shall study the latter class.

We consider a process $\{f_n\}$ with a state space $S = \{s_1, s_2, \cdots, s_s\}$. We let $A = \{A_1, A_2, \cdots, A_r\}$ be a partition of S, and define a new process $\{g_n\}$ on A as follows. We let $g_n = A_j$ if and only if $f_n \in A_j$. The new process will be called as a lumped process derived from the process $\{f_n\}$. This process will also be called as a function of the process $\{f_n\}$. Because, we can define a function F on S onto A such that $F(s_i) = A_j$ if and only if $s_i \in A_j$, and then we can write as $g_n = F(f_n)$.

E. J. Gilbert [3], S. W. Dharmadhikari [4], [5], A. Heller [6] and others obtained conditions for a process being a function of a finite Markov chain, or a lumped process derived from a finite Markov chain. C. J. Burke & M. Rosenblatt [7] obtained conditions for a lumped process derived from a finite Markov chain being again a Markov chain. In Chapter 2 of this thesis, the range of eigen-values of stochastic matrices are examined, and it is shown that using the results of the examination, we can obtain an information from a sample of the lumped process about the number of states of the original Markov chain.

In these studies, methods for constructing the original Markov chain from the observed lumped process are treated, and hence the complete knowledge about the probability laws of the observed process is assumed. However in practical applications, we cannot obtain the complete knowledge. Hence studies from another point of view are also necessary. One of them is a study about the situation where a finite Markov chain is used as an approximation of a lumped process derived from another Markov chain with large number of states. In Chapter 2, the eigenvalues of the original Markov chain and of the approximate Markov chain are compared by a Monte Carlo method, and it is conjectured that the rate of convergence to the steady state of the approximate Markov chain is in most cases faster than that of the original Markov chain.

5. On the stationarity

If the transition probabilities of a Markov chain are constant in time, we say that the Markov chain is stationary. In most practical approximations, we use stationary Markov chains. The reason of it is that the estimations of transition probabilities are very difficult in the non-stationary cases. However, as A. S. C. Ehrenberg [8] asserted for the brand switching models, in most applications in social science, the assumption of the stationarity is doubtful. Hence it is necessary to study the situation where a stationary Markov chain is used as an approximation of a non-stationary Markov chain. Unfortunately, the author has never seen such a study.

Some studies have been done for non-stationary Markov chains by Yu. V. Linnik, N. A. Sapogov, T. A. Sarymsakov, R. L. Dobrushin, J. Hajnal, J. L. Mott and others (see the bibliographies of T. A. Sarymsakov [9] and F. Harary, B. Lipstein & G. P. H. Styan [10]). Their main theme is the study of conditions for the ergodicity of non-stationary Markov chains.

F. Harary, B. Lipstein & G.P.H. Styan [10] studied non-stationary Markov chains from a different point of view. They defined a causative matrix by $C_n = P_n^{-1} P_{n+1}$ where P_n is the transition matrix from time n to time $n+1$ of a non-stationary Markov chain. B. Lipstein [11] proposed an application of this causative matrix. He used the differences of the eigen-values of the causative matrix from 1 as measures for changes of the structure of transition matrices. This is a facinating application of a non-stationary Markov chain. But it does not seem to assist to solve our main problem stated above.

In Appendix 1, another generalization is shown. The motivation of the study was in the Markov brand switching model. The Markov brand switching model is a typical practical application of a Markov chain. In the model, the probability that the process is in state s_i is considered as the purchasing level (i.e., the rate of buyers) of Brand s_i, and the transition probability from state s_i to s_j is considered as the proportion of switchers from Brand s_i to Brand s_j to buyers of Brand s_i. For this model, as A. S. C. Ehrenberg [8] asserted, the assumption of the stationarity of the transition probabilities cannot be expected in most cases. We should rather treat the transition probabilities, i.e., the proportions of switchers, as random variables. This indicates a new model, the Markov chain model with random transition matrices.

In Appendix 1, a Markov chain with random transition matrices (a M.C. with R.T.M.) is defined as a Markov process on the space of the s-th order stochastic vectors, and some properties of it are examined. Furthermore, stationary M.C.'s with R.T.M. (i.e., M.C.'s with R.T.M. having mutually independent and identically distributed transition matrices) are classified into three groups; ergodic chains, aperiodic and non-ergodic chains, and periodic chains.

6. On the accuracy of the transition probabilities.

The next problem is on the accuracy of the transition probabilities. Estimating methods for transition probabilities of a stationary Markov chain have been studied by many authors (see [1] , [2] , [12] , [13] and also see books [14] , [15]). When informations about the transitions of the chain (micro data) are available, we can estimate the transition probabilities with the maximum likelihood estimates (see [1]). When only informations about the distributions of the chain in a long time (macro data) are available, we estimate them with some of restricted least squares estimates, weighted restricted least squares estimates, maximum absolute deviations estimates (see [13]), Chorafas's estimates (see [16]) and others. However, in most cases, these estimates have errors which cannot be disregarded. If we calculate estimates of other characteristic quantities of the chain, such as the limiting vector, the mean first passage times, taboo probabilities and so on, from these estimates of the transition probabilities, then they might have larger errors.

In practical applications, we are very interested in the magnitudes of the errors of the estimates. The magnitudes of the errors of the estimates of the transition probabilities can be obtained by ordinary statistical methods. But when we calculate the other quantities from the transition probabilities, it is not easy to obtain the magnitudes of errors of estimates of them. We need further examination about relations between the transition probabilities and them. As an example, we shall formulate the problem for the limiting vectors.

Let $P = \{ p_{ij} ; i,j = 1,2, \cdots ,s \}$ be the transition matrix of a stationary regular Markov chain and $\alpha = \{ a_i ; i = 1,2, \cdots ,s \}$ be its limiting vector. Suppose that there is another stationary Markov chain with transition matrix $P' = \{ p'_{ij} \}$ and the limiting vector $\alpha' = \{ a'_i \}$. If P' is close to P, we can expect that α' is also close to α. Then how close are they?

P.J. Schweitzer [17] gave an answer of this problem by giving a perturbation series expansion of α' in powers of the matrix $U = (P' - P)(I - P + A)^{-1}$, where A is the matrix with α in each row. However, since U is defined as the product of $(P' - P)$ and an inverse of a matrix, we cannot easily guess it when we only know that P' is close to P.

J. L. Smith [18] showed that if $p_{ij} - \Delta p^-_{ij} \leqq p'_{ij} \leqq p_{ij} + \Delta p^+_{ij}$ for all i and j, then α' lies in a convex cone in the s-th order Eucledian space bounded by at most $3s$ hyperplanes. His method gives us a precise information about the range of α', but we can hardly guess it without calculating a linear programming problem.

In Chapter 3, a simple bound for α' is obtained using special properties of the matrix (I - P). If $|p'_{ij} - p_{ij}| \leqq p_{ij}\epsilon$ for all i and j, then

$$|a'_k - a_k| \leq 2(s-1)(1-a_k)a_k\epsilon + O(\epsilon^2).$$

This idea can be easily adopted for obtaining simple bounds for some other characteristic quantities.

7. Numerical methods for Markov chains

In theoretical applications of Markov chains, we sometimes need the numerical values of characteristic quantities of the chains such as the limiting vector, mean first passage times, absorbing probabilities, taboo probabilities and so on. These values can be obtained by ordinary methods for numerical calculations. For example, the limiting vector of a Markov chain can be calculated by ordinary methods for systems of linear equations such as Gaussian Elimination, Gauss-Seidel iterative method and so on.

However in most theoretical applications, the transition matrices have special structures. For example, the imbeded Markov chains used in the analyses of queueing systems of type M/G/1, have transition matrices of the form

$$P = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & p_0 & p_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Hence it is desiable to develope new suitable methods of numerical calculations for Markov chains with transition matrices with special structures.

In Chapter 4, an iterative method is proposed for the calculations of the limiting vectors, and it is used for the analysis of the speed class sequencing in the air traffic control in Chapter 5. It is available for all regular Markov chains, and in particular, it is effective for Markov chains with the following structures. The states of a Markov chain are grouped into several blocks in a natural way. The probabilities of transitions from one block to another are very little, and hence if once it enters in a block, it stays in the block for a long time. For such a chain, it may be adequate to separate the transitions within the same block and the transitions between different blocks. This idea is related to considering the lumped processes of the Markov chain. In our method, the transitions in several lumped processes are followed separately. This coresponds to partitioning the system of equations into several systems of equations. If we want to calculate the limiting vector of such a chain by the Gauss-Seidel iterative method, the rate of convergence is very slow. In our method, we avoid this difficulty by partitioning the system of equations into several smaller systems of equations. For, as a rule, we may expect that the smaller is the number of equations in the system, the faster is the rate of convergence in an iterative method.

References

[1] Anderson, T.W. & Goodman, L.A., "Statistical inference about Markov chains", Ann. Math. Stat. 28 (1957), 89-109.

[2] Billingsley, P., "Statistical methods in Markov chains", Ann. Math. Stat. 32 (1961), 12-40.

[3] Gilbert, E.J., "On the identifiability problem for functions of finite Markov chains," Ann. Math. Stat. 30 (1959), 688-697.

[4] Dharmadhikari, S.W., "Functions of finite Markov chains," Ann. Math. Stat. 34 (1963), 1022-1032.

[5] Dharmadhikari, S.W., "A characterisation of a class of functions of finite Markov chains," Ann. Math. Stat. 36 (1965), 524-528.

[6] Heller, A., "On stochastic processes derived from Markov chains," Ann. Math. Stat. 36 (1965), 1286-1291.

[7] Burke, C.J. & Rosenblatt, M., "A Markovian function of a Markov chain," Ann. Math. Stat. 29 (1958), 1112-1122.

[8] Ehrenberg, A.S.C., "An appraisal of Markov brand-switching models," Journ. of Marketing Research 2 (1965), 347-362.

[9] Sarymsakov, T.A., "Inhomogeneous Markov chains," Theory Prob. Appl. 6 (1961), 178-185.

[10] Harary, F., Lipstein, B. & Styan, G.P.H., "A matrix approach to nonstationary chains," Jorn. Ops. Res. Soc. Amer. 18 (1970), 1168-1181.

[11] Lipstein, B., "Test marketing : a perturbation in the market place," Management Sci. Ser. B 14 (1968), 437-448.

[12] Theil, H. & Guido Rey, "A quadratic programming approach to the estimation of transition probabilities," Management Sci. 12 (1966), 714-721.

[13] Lee, T.C., Judge, G.G. & Cain, R.L., "A sampling study of the properties of estimators of transition probabilities," Management Sci. 15 (1969), 374-398.

[14] Billingsley, P., Statistical Inference for Markov Processes. Univ. Chicago Press, Chicago (1961).

[15] Lee, T.C., Judge, G.G. & Zellner, A., Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data. North-Holland Publishing Company, Amsterdam (1970).

[16] Chorafas, D.N., Systems and Simulation. Academic Press, New York, (1965).

[17] Schweitzer, P.J.,"perturbation theory and finite Markov chains," J. Appl. Prob. 5 (1968), 401-413.

[18] Smith, J.L., "Approximate stationary probability vectors of a finite Markov chain," SIAM J. Appl. Math. 20 (1971), 612-618.

CHAPTER 2 ON APPROXIMATIONS OF LUMPED PROCESSES BY MARKOV CHAINS AND THE RANGE OF EIGENVALUES OF STOCHASTIC MATRICES

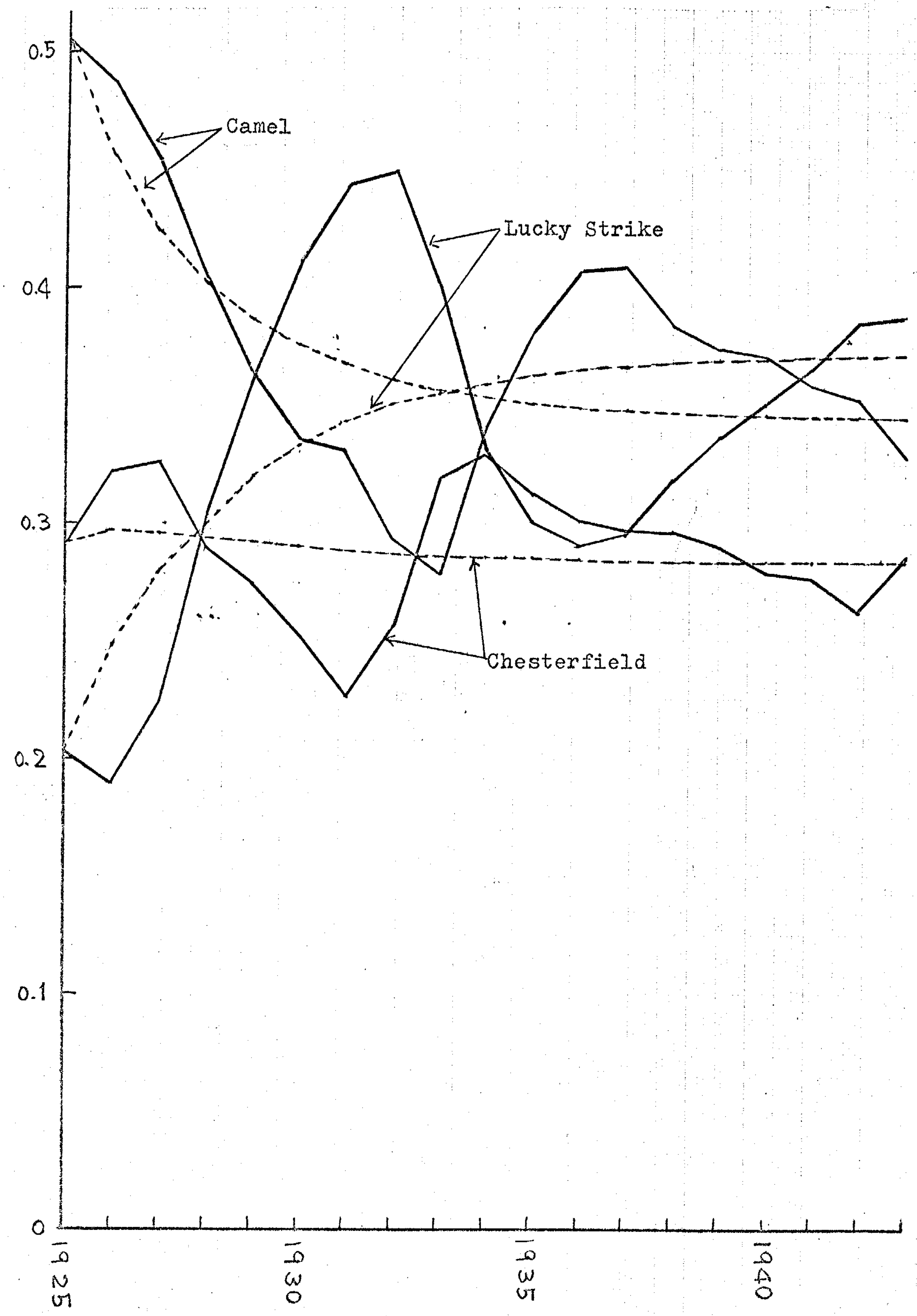
1. Introduction

The curves in solid lines in Fig. 1 represent the market shares of three cigarette brands in the United States during the period 1925 - 1943 . L. G. Telser [10] and M. Teil & G. Rey [11] used this data for explaining the problem of estimating transition probabilities in Markov chains from a set of observations on the distributions. In [11] , it was shown that the transition matrix is estimated from the data as

$$(1.1) \qquad P = \begin{pmatrix} .6686 & .1423 & .1891 \\ .0 & .8683 & .1317 \\ .4019 & .0 & .5981 \end{pmatrix}$$

by the restricted least squares estimation. The curves in broken lines in Fig. 1 are the estimated market shares calculated from the same initial distribution (.5056 .2028 .2916) and the estimate transition matrix P in (1.1). They almost monotonically tend to limits, but on the other hand, curves in solid lines rather oscillate with damping. The period of the oscillation is about ten years and the amplitudes become about halfs in a period, i.e., in ten years.

Fig. 1. Market shares of three cigarette brands



Thus the estimation is not successful in the manners of convergences. The cause of the ill success is not in the estimating method but in the attitude to try to approximate the original process by a Markov chain with three states. The aim of this paper is to clarify the reason of it.

2. Expansion of a Stochastic Matrix by Eigenvalues

We consider a stochastic matrix $P = (p_{ij})$ of order n. The entries of P are nonnegative and the row sums are equal to 1. One of the eigenvalues of P is equal to 1 and the absolute values of others are less than or equal to 1. To simplify the discussions, we assume that the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ are different with each other and that $\lambda_1 = 1$ and $|\lambda_k| < 1$ for $k = 2, 3, \cdots, n$. Then P can be represented as

$$(2.1) \qquad P = T \Lambda T^{-1}$$

where Λ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \cdots$ $\cdots, \lambda_n$, and T is a square matrix whose k-th column is a right eigenvector $t^k = (t_j^k)$ of P corresponding to λ_k. It follows that

$$(2.2) \qquad P = H_1 + \lambda_2 H_2 + \cdots + \lambda_n H_n$$

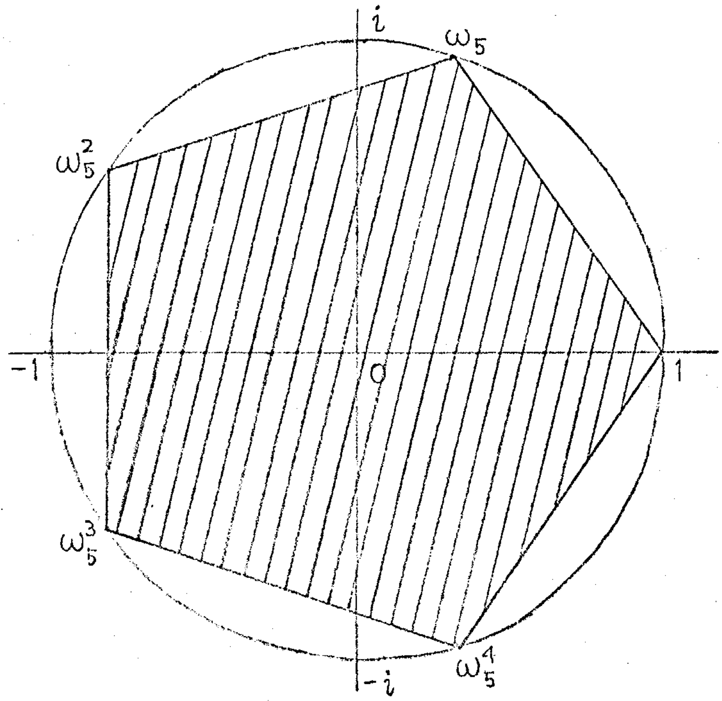
where H_k is a square matrix of order n of the form

3. The Range of Eigenvalues of Stochastic Matrices (I)

In this and the next sections we shall treat the range of eigenvalues of stochastic matrices. By the well known Frobenius's theorem, the eigenvalues of stochastic matrices lie in the unit circle in the complex plane or on the boundary of it. However, not every point in the circle can be an eigenvalue of a stochastic matrix of order n. We denote by D_n the range of eigenvalues of stochastic matrices of order n, i.e., the set of all z such that there exists a stochastic matrix of order n with an eigenvalue z. We shall study about D_n.

We denote by ω_n the complex root $\cos\frac{2\pi}{n} + i\sin\frac{2\pi}{n}$ of the equation $x^n = 1$. Then the n roots of the equation are 1, ω_n, $\omega_n^2, \cdots, \omega_n^{n-1}$. We denote by C_n the domain with its boundary bounded by the regular polygon whose vertices are the roots 1, ω_n, $\omega_n^2, \cdots, \omega_n^{n-1}$ (see Fig. 2).

Fig. 2. C_5



We prepare four elementary lemmas. We consider a stochastic matrix $P = (p_{ij})$ of order n and denote the eigenvalues of it by $\lambda_1, \lambda_2, \cdots, \lambda_n$.

Lemma 1. $\lambda_1 + \lambda_2 + \cdots + \lambda_n = p_{11} + p_{22} + \cdots + p_{nn}$.

Proof. The lemma can be easily proved by comparing the coefficients of λ^{n-1} in both sides of the equation $|P - \lambda I| = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$.

Lemma 2. If the last column of P is the zero vector, i.e., $p_{in} = 0$ for $i = 1,2, \cdots, n$, then one of the eigenvalues is zero and others coincide with eigenvalues of a stochastic matrix Q formed by deleting the n-th row and the n-th column from P.

Proof. The proof of this lemma is trivial.

Lemma 3. Let $Q_\alpha = \alpha P + (1 - \alpha) I$ for $0 \leqq \alpha \leqq 1$. Then Q is a stochastic matrix and has eigenvalues

$$(3.1) \qquad \lambda_k' = \alpha \lambda_k + 1 - \alpha \qquad (k = 1,2, \cdots, n).$$

Proof. It is clear that Q_α is a stochastic matrix. Since

$$(3.2) \qquad |Q_\alpha - \lambda_k' I| = |\alpha P + (1 - \alpha) I - (\alpha \lambda_k + 1 - \alpha) I|$$

$$= |\alpha| \cdot |P - \lambda_k I| = 0 ,$$

λ_k' is an eigenvalue of Q_α.

Note: We can easily show that Q_α is expanded in the form in (2.2) with the same H_k's as P. The lemma reflects the fact that the eigenvalues of a matrix formed by a linear combination of two matrices with the same H_k's are given by the same linear combinations of corresponding eigenvalues of two matrices. We shall use this fact in the proof of Theorem 4.

We will abbreviate the segment with end points z and z' of a line in the complex plane as the segment $[z, z']$.

Lemma 4. If z is in D_n then the segment $[1, z]$ is also contained in D_n.

Proof. This lemma is a trivial corollary of Lemma 3.

Now we shall prove some theorems for D_n.

Theorem 1. $D_n \supset D_{n-1}$.

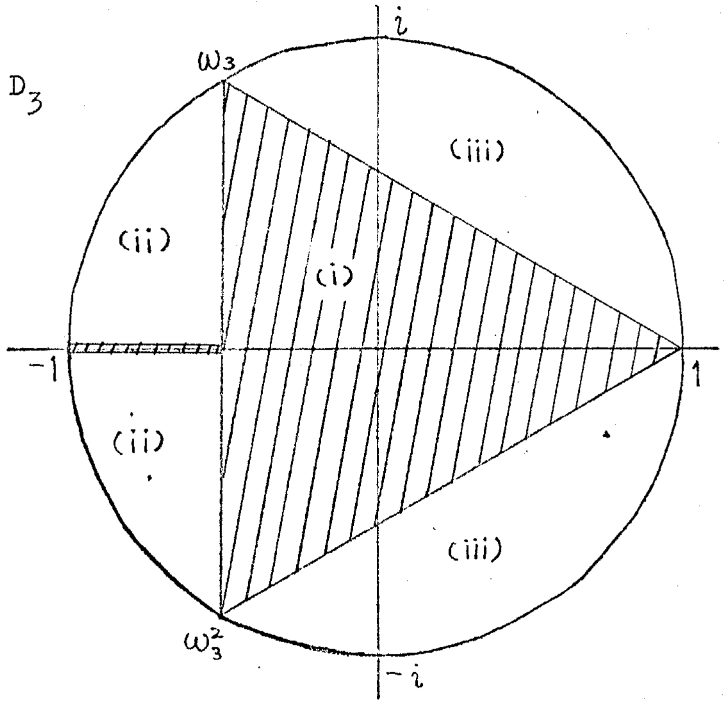
Proof. This proof is obvious from Lemma 2.

Theorem 2. $D_2 = C_2$, i.e., D_2 is the segment $[-1, 1]$ of the real axis.

Proof. When $n = 2$, the eigenvalues of P are $\lambda_1 = 1$ and $\lambda_2 = p_{11} + p_{22} - 1$ by Lemma 1. Hence λ_2 is a number on $[-1, 1]$. Conversely, for any number z on $[-1, 1]$, we can easily construct a stochastic matrix of order 2 having an eigenvalue z.

<u>Theorem 3</u>. $D_3 = C_2 \cup C_3$.

<u>Fig. 3</u>. D_3



<u>Proof</u>. From Theorems 1 and 2, it is clear that $C_2 \subset D_3$. So we are sufficient to prove that $C_3 \subset D_3$ and that no number out of $C_2 \cup C_3$ can be an eigenvalue of a stochastic matrix of order 3 .

(i) That $C_3 \subset D_3$ is a special case of Theorem 4. But here we prove it in another way. We first prove that any number on the segment $[\omega_3, \omega_3^2]$ is in D_3 . Such a number is represented as $z = -1/2 + bi$ $(-\sqrt{3}/2 \leq b \leq \sqrt{3}/2)$. The stochastic matrix

$$(3.3) \qquad P = \begin{pmatrix} 0 & 1/4 + b^2 & 3/4 - b^2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

has eigenvalues 1 , $z = -1/2 + bi$, $\bar{z} = -1/2 - bi$. Hence z is in D_3 . By Lemma 4, we can show that any point on the segment $[1, z]$ is in D_3 . Thus $C_3 \subset D_3$.

(ii) We prove that any complex number $z = a + bi$ such that $a < -1/2$ and $b \neq 0$, is not in D_3. If z is an eigenvalue of a stochastic matrix P of order 3, then the complex conjugate $\bar{z}$ of z is also another eigenvalue of P. For, if $|P - zI| = 0$, then the complex conjugate $|P - \bar{z}I|$ of $|P - zI|$ is also equal to zero. Hence in this case three eigenvalues of P are 1, z and $\bar{z}$. By Lemma 1, we have

$$p_{11} + p_{22} + p_{33} = 1 + z + \bar{z} = 1 + 2a . \tag{3.4}$$

Since p_{11}, p_{22} and p_{33} are nonnegative, it follows that $a \geq -1/2$. Thus any complex number $z = a + bi$ such that $a < -1/2$ and $b \neq 0$, is not contained in D_3.

(iii) We prove that any complex number $z = a + bi$ such that $|b/(1-a)| < 1/\sqrt{3}$, is not in D_3. This is a special case of Theorem 5. However here we shall prove it directly.

Solving the equation $|P - \lambda I| = 0$, we have

$$\begin{aligned} a &= \frac{1}{2}(p_{11} + p_{22} + p_{33} - 1) \\ &= 1 - \frac{1}{2}(p_{12} + p_{13} + p_{21} + p_{23} + p_{31} + p_{32}) \end{aligned} \tag{3.5}$$

and

$$b^2 = \frac{1}{4}\left[- (p_{11} + p_{22} + p_{33} - 1)^2 + |P| \right] \tag{3.6}$$

$$= \frac{1}{4}\Big[- (p_{12}^2 + p_{13}^2 + p_{21}^2 + p_{23}^2 + p_{31}^2 + p_{32}^2)$$

$$- 2(p_{12}p_{13} + p_{21}p_{23} + p_{31}p_{32} + p_{12}p_{21} + p_{13}p_{31} + p_{23}p_{32})$$

$$+ (p_{12}p_{23} + p_{12}p_{31} + p_{12}p_{32} + p_{13}p_{21} + p_{13}p_{23} + p_{13}p_{32}$$

$$+ p_{21}p_{31} + p_{21}p_{32} + p_{23}p_{31})\Big] .$$

It follows that

(3.7) $$\left(\frac{b}{1-a}\right)^2 - 1$$

$$= -2\Big[(p_{12}^2 + p_{13}^2 + p_{21}^2 + p_{23}^2 + p_{31}^2 + p_{32}^2) + 2(p_{12}p_{13} + p_{21}p_{23} + p_{31}p_{32}$$

$$+ p_{12}p_{21} + p_{13}p_{31} + p_{23}p_{32})\Big] / (p_{12} + p_{13} + p_{21} + p_{23} + p_{31} + p_{32})^2$$

$$\leqq -2\Big[(p_{12} + p_{13})^2 + (p_{21} + p_{23})^2 + (p_{31} + p_{32})^2\Big] /$$

$$(p_{12} + p_{13} + p_{21} + p_{23} + p_{31} + p_{32})^2$$

$$\leqq -2/3 .$$

Hence we have

(3.8) $$\left|\frac{b}{1-a}\right| \leqq \frac{1}{\sqrt{3}} .$$

Thus $z = a + bi$ such that $|b/(1-a)| > 1/\sqrt{3}$, is not contained in D_3.

<u>Theorem 4</u>. $D_n \supset C_2 \cup C_3 \cup \cdots \cup C_n$.

<u>Proof</u>. If $D_n \supset C_n$, then the theorem is proved by Theorem 1 using the mathematical induction on n . In order to prove that $D_n \supset C_n$, we use a similar idea to Lemma 3. We will abbreviate ω_n as ω .

Since C_n is the regular polygon with vertices $1, \omega, \omega^2, \cdots, \omega^{n-1}$, each point in C_n can be represented as a convex linear combination

$$(3.9) \qquad z = \alpha_0 + \alpha_1\omega + \alpha_2\omega^2 + \cdots + \alpha_{n-1}\omega^{n-1}$$

where $\alpha_k \geqq 0$ and $\alpha_0 + \alpha_1 + \cdots + \alpha_{n-1} = 1$. We shall construct a stochastic matrix Q with an eigenvalue z.

We consider the stochastic matrix of order n

$$(3.10) \qquad P = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The eigenvalues of P are $1, \omega, \omega^2, \cdots, \omega^{n-1}$. So P can be expanded as

$$(3.11) \qquad P = H(1) + \omega H(\omega) + \omega^2 H(\omega^2) + \cdots + \omega^{n-1} H(\omega^{n-1}),$$

where $H(x)$ is a matrix-valued function

$$(3.12) \qquad H(x) = \frac{1}{n}\begin{pmatrix} 1 & x^{n-1} & x^{n-2} & \cdots & x \\ x & 1 & x^{n-1} & \cdots & x^2 \\ x^2 & x & 1 & \cdots & x^3 \\ \vdots & \vdots & \vdots & & \vdots \\ x^{n-1} & x^{n-2} & x^{n-3} & \cdots & 1 \end{pmatrix}.$$

Hence the k-th power of P is represented as

(3.13) $$P^k = H(1) + \omega^k H(\omega) + \omega^{2k} H(\omega^2) + \cdots + \omega^{(n-1)k} H(\omega^{n-1}) .$$

This relation is also satisfied for $k=0$, i.e., for $P^0 = I$. Hence it is available for every $k = 0,1,2,\cdots$. Now we define a stochastic matrix

(3.14) $$Q = \alpha_0 I + \alpha_1 P + \alpha_2 P^2 + \cdots + \alpha_{n-1} P^{n-1} .$$

From (3.13), Q is represented as

(3.15) $$Q = H(1) + f(\omega) H(\omega) + f(\omega^2) H(\omega^2) + \cdots$$
$$\cdots + f(\omega^{n-1}) H(\omega^{n-1}) .$$

where

(3.16) $$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{n-1} x^{n-1} .$$

Hence Q has eigenvalues 1, $f(\omega)$, $f(\omega^2)$, $\cdots$, $f(\omega^{n-1})$. Since $z = f(\omega)$, z is an eigenvalue of Q and hence it is in D_n. This completes the proof.

4. The Range of Eigenvalues of Stochastic Matrices (II)

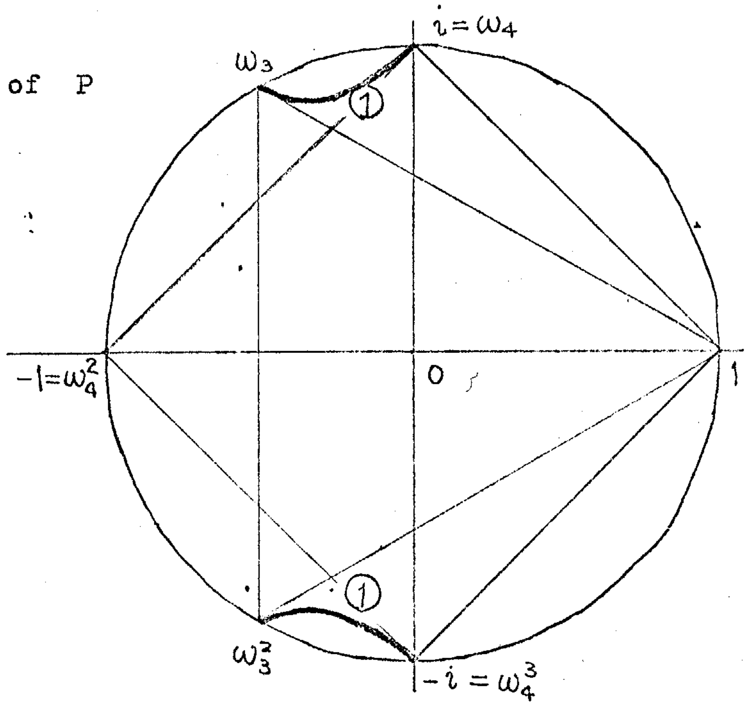
In Theorem 4, we saw that $D_n \supset C_2 \cup C_3 \cup \cdots \cup C_n$. Then is it true that $D_n = C_2 \cup C_3 \cup \cdots \cup C_n$? The answer is "no". For example the stochastic matrix

(4.1) $$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \alpha & 0 & 0 & 1-\alpha \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

has an eigenvalue z which is not in $C_2 \cup C_3 \cup C_4$. The two complex eigenvalues $z = a \pm bi$ of P are on the hyperbolic curve

$\frac{b^2}{3} - (a + \frac{1}{3})^2 = \frac{2}{9}$ (see Fig. 4) .

Fig. 4. Eigenvalues of P in (4.1)



So we must study furhter about the boundary of D_n .

N. A. Dmitriev & E. B. Dynkin [5] proved the following

Theorem 5. (N. A. Dmitriev & E. B. Dynkin) (*)
Let λ be an eigenvalue of stochastic matrix of order n . If $\arg(\lambda) \leq 2\pi/n$, then λ is in C_n .

(*) This theorem is quoted from [9] . Unfortunately the author has never seen the original paper [5] .

Theorem 5 implies that the segments $[1, \omega_n]$ and $[1, \omega_n^{n-1}]$ are parts of the boundary of D_n. The remaining part of it has never been known exactly. The following theorems will give us a clue for the study about it.

Let $\mathcal{M}_n^1$ be the set of stochastic matrices of order n such that each row of them has at most two nonzero entries, and E_n^1 be the range of eigenvalues of stochastic matrices in $\mathcal{M}_n^1$.

Theorem 6. The boundary of D_n is contained in E_n^1.

Several lemmas are necessary for the proof of this theorem.

Lemma 5. Let λ be a simple eigenvalue of a stochastic matrix P of order n, and t and u be a right and a left eigenvectors of P corresponding to λ. For arbitrary matrix B and a small number ϵ, the matrix $P + \epsilon B$ has an eigenvalue

$$(4.2) \qquad \lambda' = \lambda + \frac{uBt}{ut}\epsilon + O(\epsilon^2).$$

For the proof of this lemma, see [12].

Let δ be the smallest value of nonzero entries of P, and $\mathcal{B}(P)$ be the set of matrices B such that $P + \delta B$ is a stochastic matrix. Let F be the range of uBt/ut for $B \in \mathcal{B}(P)$.

Lemma 6. F is a closed convex polygon containing the origin.

Proof. The proof of this lemma is obvious.

Lemma 7. If p_{ij}, $p_{ik} > 0$, then $z = u_i(t_j - t_k)/ut$ and $-z$ are in F.

Proof. We consider the matrix B such that $b_{ij} = 1$, $b_{ik} = -1$ and other entries are equal to zero. Then it is clear that both B and $-B$ are in $\mathcal{B}(P)$. Hence $z = uBt/ut$ and $-z = u(-B)t/ut$ are in F.

Lemma 8. If λ is on the boundary of D_n, then the origin is on the boundary of F.

Proof. If λ is on the boundary of D_n and if the origin is in the interior of F, then there exists a number z such that z is in F but $\lambda + \epsilon z$ is out of D_n for every positive $\epsilon \leqq 1$. Since z is in F, there exists a matrix B in $\mathcal{B}(P)$ such that $z = uBt/ut$. By Lemma 5, $P + \epsilon B$ has an eigenvalue $\lambda' = \lambda + \epsilon z + O(\epsilon^2)$. Since $\lambda + \epsilon z$ is out of D_n, λ' is also out of D_n for sufficiently small ϵ. Since $P + \epsilon B$ is a stochastic matrix, this is in contradiction to the definition D_n.

Lemma 9. If λ is on the boundary of D_n, and if p_{ij}, $p_{ik} > 0$, then $z = u_i(t_j - t_k)/ut$ is on the same edge of F as the origin.

Proof. By Lemma 7, three points z, 0 and $-z$ are in F, and by Lemma 8, the origin is on the boundary of F. Since F is a convex polygon, these three points must be on the same edge of F.

Now we shall prove Theorem 6.

Proof of Theorem 6. Since eigenvalues of stochastic matrices are uniformly continuous on the entries, it is clear that D_n is closed. So for every point λ on the boundary of D_n, there exists a stochastic matrix having an eigenvalue λ. So, we are sufficient to prove that if an eigenvalue λ of P is on the boundary of D_n and if p_{ij}, p_{ik}, $p_{il} > 0$, then there exists a stochastic matrix P' with the same eigenvalue λ such that P' has the same entries as P other than three entries p'_{ij}, p'_{ik}, p'_{il} and one of the three entries is equal to zero. For the simplicity of the proof, we shall prove only the case where λ is a simple eigenvalue of P.

If $p_{ij}, p_{ik}, p_{il} > 0$, then by Lemma 9 three points

$$z^i_{jk} = u_i(t_j - t_k)/ut , \quad z^i_{kl} = u_i(t_k - t_l)/ut \tag{4.3}$$

and the origin are on the same edge of F. So, $z^i_{jk} = 0$ or $z^i_{kl} = c\, z^i_{jk}$ where c is a real number. Hence we should consider three cases, (i) $u_i = 0$, (ii) $t_j = t_k$ and (iii) $t_k - t_l = c\,(t_j - t_k)$.

(i) If $u_i = 0$, then $uQ = \lambda u$ for every Q formed by replacing the i-th row of P with a stochastic vector. Hence especially, the matrix P' formed by replacing the i-th row of P with a stochastic vector whose i-th entry is equal to 1 and others are equal to zero, has an eigenvalue λ.

(ii) When $t_j = t_k$, we consider the stochastic matrix P' formed from P by replacing p_{ij} with 0 and p_{ik} with $p_{ij} + p_{ik}$. Then $P't = Pt = \lambda t$. Hence P' has an eigenvalue λ.

(iii) Now we consider the case where $t_k - t_l = c\,(t_j - t_k)$. If $c = -1$, then we have $t_l = t_j$, and we can construct P' in the same way as in (ii). Hence we assume that $c \neq -1$. Then we can write as

$$(4.4) \qquad t_k = \frac{1}{1 + c}\,(t_l + ct_j)\ .$$

We consider the stochastic matrix P' formed from P by replacing the three entries p_{ij} , p_{ik} and p_{il} with

$$(4.5) \qquad p'_{ij} = p_{ij} + \frac{c}{1 + c}\,p_{ik}\ ,$$

$$p'_{ik} = 0 \qquad \text{and}$$

$$p'_{il} = p_{il} + \frac{1}{1 + c}\,p_{ik}\ .$$

Since

$$(4.6) \qquad p'_{ij}t_j + p'_{il}t_l = p_{ij}t_j + p_{ik}t_k + p_{il}t_l\ ,$$

we have $P't = Pt = \lambda t$. Hence P' has an eigenvalue λ . Thus in every case we can construct a stochastic matrix P' having an eigenvalue λ such that one of p'_{ij} , p'_{ik} , p'_{il} is zero. Thus we have proved the theorem.

In the proof of Theorem 6, we have dealt with the i-th row only. The same idea can be used for any pair of two rows with two nonzero entries. Namely, if $p_{ij}, p_{ik} > 0$ and $p_{hl}, p_{hm} > 0$, then one of

$$(4.7) \qquad z^i_{jk} = u_i(t_j - t_k)/\,ut$$

and

(4.8) $\qquad z^h_{lm} = u_h(t_l - t_m)/ut$

is equal to zero, or $z^h_{lm} = c\, z^i_{jk}$ $(\neq 0)$. In the former case, we can show that there exists a stochastic matrix P' with the same eigenvalue λ such that P' has the same entries as P except for four entries $p_{ij}, p_{ik}, p_{hl}, p_{hm}$ and one of the four entries is equal to zero. Hence in order to study about the boundary of D_n, we can restrict our considerations to stochastic matrices in $\mathfrak{M}^1_n$ satisfying

<u>Alignment condition</u>. For any choice of two rows of P with two nonzero entries p_{ij}, p_{ik} and p_{hl}, p_{hm}, if exist, the relation

(4.9) $\qquad u_i(t_j - t_k) = c\, u_h(t_l - t_m)$.

holds for a real number c which may depend on the choice of the rows. We denote by $\mathfrak{M}^2_n$ the set of stochastic matrices in $\mathfrak{M}^1_n$ satisfying the Alignment condition, and by E^2_n the range of eigenvalues of stochastic matrices in $\mathfrak{M}^2_n$. Then we can summarize the above discussion as in the following

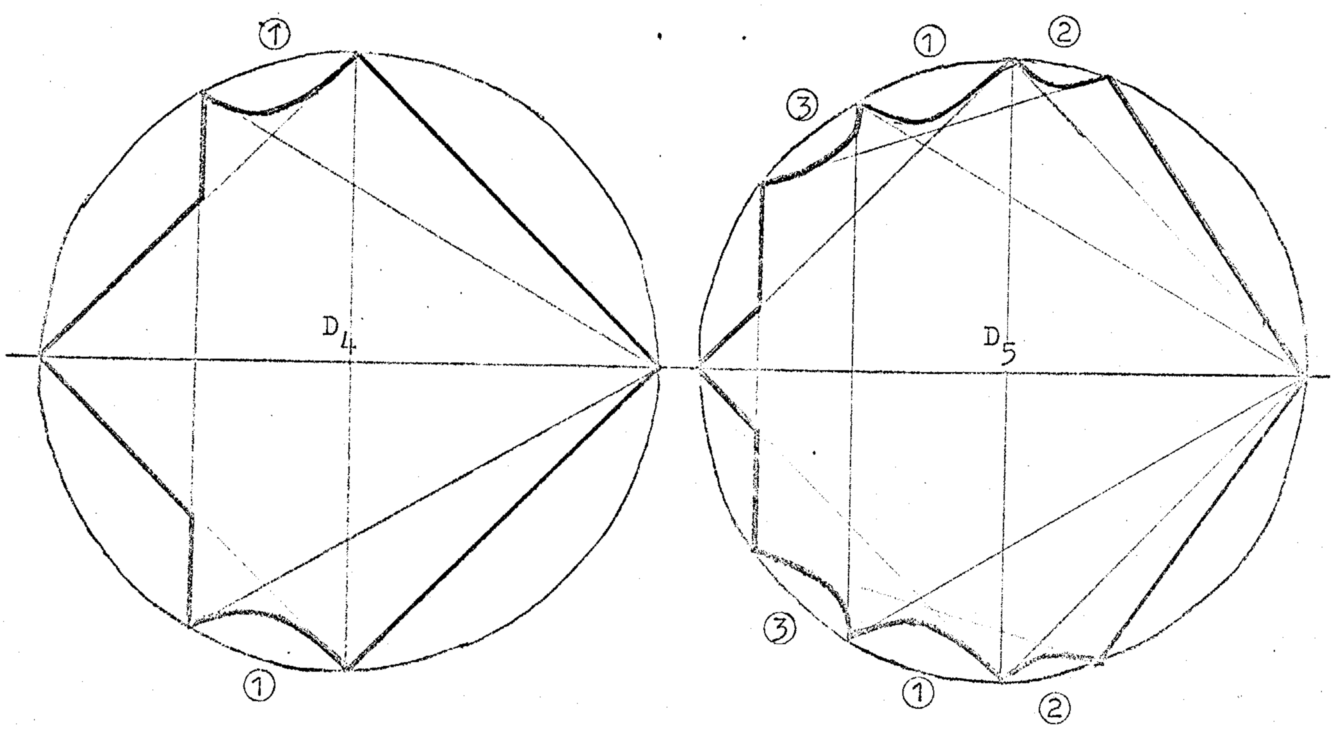
<u>Theorem 7</u>. The boundary of D_n is contained in E^2_n.

Unfortunately we know few sufficient conditions for the Alignment condition. Hence we have not yet determined the boundary of D_n exactly. The author supposes that the boundary of D_n is attained by simple stochastic matrices and conjectures that it is contained in the union of the boundary of $D_{n-1} \cup C_n$ and E_n^3 defined below.

We consider the matrix defined in (3.10) and denote it here by P_n. We denote by $P_n(i,\alpha)$ the matrix formed from P_n by replacing the (i,1)-th entry with α and the (i,i+1)-th entry with $1-\alpha$. For example, the matrix in (4.1) can be denoted by $P_4(3,\alpha)$ in this notation. We let $\mathfrak{M}_n^3$ be the set of stochastic matrices $P_n(i,\alpha)$ for $i = 1,2,\cdots,n-1$ and $0 \leq \alpha \leq 1$, and E_n^3 be the range of eigenvalues of stochastic matrices in $\mathfrak{M}_n^3$.

If the conjectur is true, D_4 and D_5 become as in Fig. 5.

<u>Fig. 5</u>. D_4 and D_5 by the conjecture



① $P_4(3,\alpha)$ ② $P_5(4,\alpha)$ ③ $P_5(3,\alpha)$

5. A Lower Bound for the Number of States of the Underlying Markov Chain

Now we return to the example in Section 1. As stated there, the curves for the observations may be considered as curves for damped oscillations, and the period is about ten years and the amplitudes become about halfs in one period. If the motions of these curves are approximated by the distribution of a Markov chain, then the transition matrix P of the chain has an eigenvalue

$$(5.1) \qquad \lambda = \left(\frac{1}{2}\right)^{\frac{1}{10}} \cdot \left\{ \cos \frac{2\pi}{10} + i \sin \frac{2\pi}{10} \right\}$$

$$= .756 + .548\, i \,.$$

From Theorem 3 we know that there is no stochastic matrix of order 3 with an eigenvalue λ. Hence it is impossible to approximate the curves in Fig. 1 by the distribution of a Markov chain with three states. We should rather consider that the process is not a Markov chain but a lumped process of a Markov chain.

We consider that there exists an underlying Markov chain with n states. The state space of it is divided into r sets by a partition $A = \{A_1, A_2, \cdots, A_r\}$. We assume that we cannot observe the state of the chain but can observe the set in which the chain is. We denote the state of the chain at time m by f_m and the observed set by g_m. The observed process $\{g_m\}$ is called as a lumped process of the chain $\{f_m\}$ or as a function of the chain.

In this example, the observable sets A_1, A_2, A_3 are Camel, Lucky Strike and Chesterfield, respectively, but we do not know the state space of the underlying chain. However using Theorem 5, we can obtain a lower bound of the number of states of the underlying Markov chain. The transition matrix of the chain must have an eigenvalue λ given by (5.1). The angle θ between the real axis and the line connecting 1 and λ, is

$$(5.2) \qquad \theta = \tan^{-1} \frac{.548}{1 - .756} = \tan^{-1} 2.25 = .367\pi .$$

The angle between the real axis and the line connecting 1 and ω_n is $(n-2)\pi/2n$. Hence $\theta \leqq (n-2)\pi/2n$ if and only if $n \geqq 8$. Thus λ cannot be in D_n for $n \leqq 7$, and we conclude that the underlying Markov chain must have at least 8 states.

However it is another problem whether we can approximate the observed process by a lumped process of a Markov chain with just eight states. Probably it is very difficult problem to estimate the transition matrix of the underlying chain from a set of observations of distributions of the observed process.

C. K. Burke & M. Rosenblatt [1] studied conditions for a function of a finite Markov chain being a Markov chain too. E. J. Gilbert [4], S. W. Dharmadhikari [2], [3], A. Heller [7] and others studied the problem of constructing the transition matrix of the underlying Markov chain from the full informations of the probability laws of the observed process. However our new problem is more and more difficult than the problems studied by the above authors. Because we can only use the information of the distribution of the observed process, and cannot use the informations about transitions.

6. A Remark on Approximations of Lumped Processes by Markov Chains

In the last section we showed a method for obtaining a lower bound of the number of states of the underlying Markov chain from observations of the distribution of the lumped process. However it is often impossible to estimate the transition matrix of the underlying Markov chain. In such a case we cannot help approximating the lumped process by a Markov chain. Hence we shall study about such approximations.

We consider the problem of approximating the observed process $\{ g_m \}$ by a Markov chain with a transition matrix $Q = (q_{\mu\nu})$. It is natural to expect that the lumped process and the approximate Markov chain have the same one-step transition probabilities. If the underlying Markov chain $\{ f_m \}$ start with the stationary distribution $\alpha = (a_i)$, the one-step transition probabilities of $\{ g_m \}$ are given by

$$
\begin{aligned}
(6.1) \qquad q_{\mu\nu} &= \Pr \{ g_{m+1} = A_\nu \mid g_m = A_\mu \} \\
&= \sum_{i \in A_\mu} \Pr \{ f_{m+1} \in A_\nu \mid f_m = i \} \cdot \Pr \{ f_m = i \} \Big/ \sum_{i \in A_\mu} \Pr \{ f_m = i \} \\
&= \sum_{i \in A_\mu} a_i \sum_{j \in A_\nu} p_{ij} \Big/ \sum_{i \in A_\mu} a_i , \\
&\qquad (\mu, \nu = 1, 2, \dots , r) .
\end{aligned}
$$

If the approximate Markov chain has the transition matrix Q defined by (6.1), then it has the same stationary distribution as the lumped process, even if the two processes do not start with their stationary distributions. (G. Guardabassi & S. Rinaldi [6] considered the problem of finding a Markov chain with the same stationary distribution as the lumped process by a topological approach. His method also uses this approximate chain.)

If the underlying Markov chain is lumpable (see [8] p.124), then the lumped process $\{g_m\}$ is a Markov chain with the same transition matrix as the approximate Markov chain, and hence if we can estimate the transition matrix of the approximate chain well, then the approximation will be fairly good. If the underlying Markov chain is weakly lumpable (see [8] p.132), then the lumped process is a Markov chain when the underlying chain start with the stationary distribution. Hence if we can estimate the transition matrix of the approximate chain well and if the underlying chain start with a distribution near to the stationary one, then the approximation is fairly good.

Thus in special cases the approximation may be good. However generally the approximation is not so good. We shall see such a situation by the point of view of the rate of convergence of the distribution to the stationary one in the next section. As stated in Section 2, the rate of convergence of a Markov chain is mainly governed by eigenvalues having the maximal absolute value among eigenvalues of the transition matrix other than 1 . So our problem is to examine the relations between eigenvalues of P and Q . This problem is very difficult, so we shall study the simplest case.

7. Approximation of a Lumped Process of a Markov Chain with Three States by a Markov Chain with Two States

We suppose that the underlying chain has three states, 1 , 2 and 3 , and that the states 2 and 3 are lumped together for the lumped process. The underlying chain has the transition matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} . \tag{7.1}$$

Then the stationary distribution $\alpha = (a_1, a_2, a_3)$ of it is given by

$$\begin{aligned} a_1 &= b_1/(b_1 + b_2 + b_3) \\ a_2 &= b_2/(b_1 + b_2 + b_3) \\ a_3 &= b_3/(b_1 + b_2 + b_3) \end{aligned} \tag{7.2}$$

where

$$\begin{aligned} b_1 &= p_{21}p_{31} + p_{21}p_{32} + p_{23}p_{31} \\ b_2 &= p_{12}p_{31} + p_{12}p_{32} + p_{13}p_{32} \\ b_3 &= p_{12}p_{23} + p_{13}p_{21} + p_{13}p_{23} \end{aligned} \tag{7.3}$$

By (6.1), the transition matrix $Q = (q_{\mu\nu})$ of the approximate Markov chain is given by

$$(7.4) \qquad \begin{aligned} q_{11} &= p_{11} \\ q_{12} &= p_{12} + p_{13} \\ q_{21} &= (a_2p_{21} + a_3p_{31})/(a_2 + a_3) \\ q_{22} &= \left[a_2(p_{22} + p_{23}) + a_3(p_{32} + p_{33})\right]/(a_2 + a_3) \end{aligned}$$

We denote the three eigenvalues of P by 1, λ_1 and λ_2, and the two eigenvalues of Q by 1 and λ'. Then λ_1 and λ_2 are two roots of the equation

$$(7.5) \qquad \lambda^2 - (p_{11} + p_{22} + p_{33} - 1)\lambda + |P| = 0 ,$$

and λ' is given by

$$(7.6) \qquad \lambda' = q_{11} - q_{21} .$$

The rate of convergence of the approximate Markov chain is governed by λ', and that of the lumped process, or the underlying Markov chain, is governed by $\max\{|\lambda_1|, |\lambda_2|\}$. So we shall consider the ratio

$$(7.7) \qquad \rho = \frac{\lambda'}{\max\{|\lambda_1|, |\lambda_2|\}} .$$

If $\rho < 1$, then the rate of convergence of the approximate Markov chain is faster than that of the lumped process. The following theorem gives a sufficient condition that $\rho < 1$.

Theorem 8. $\lambda_1 < \lambda' < \lambda_2$ if and only if

(7.8) $$(p_{21} - p_{31})\left(\frac{p_{12}}{a_2} - \frac{p_{13}}{a_3}\right) < 0 .$$

Proof. We consider a function

(7.9) $$f(\lambda) = \lambda^2 - (p_{11} + p_{22} + p_{33} - 1)\lambda + |P| .$$

Since λ_1 and λ_2 are roots of the equation $f(\lambda) = 0$, $\lambda_1 < \lambda' < \lambda_2$ if and only if $f(\lambda') < 0$. Hence the theorem is easily obtained from the relation

(7.10) $$f(\lambda') = \frac{a_2 a_3 (a_1 + a_2 + a_3)}{(a_2 + a_3)^2}(p_{21} - p_{31})\left(\frac{p_{12}}{a_2} - \frac{p_{13}}{a_3}\right).$$

In order to examine the tendency of ρ, the author calculated the density of ρ by a Monte Carlo simulation under the condition that rows of P are mutually independent random vectors and each row has the uniform distribution on the triangular defined by $p_{i1} + p_{i2} + p_{i3} = 1$ and $p_{ij} \geqq 0$ in the Eucledian space of order three.

Fig. 6 is the histogram of ρ in the 20000 experiments. It has a mode about $\rho = 1.0$ and it has a similar shape to the uniform distribution for $\rho < 0.9$ and to the exponential distribution for $\rho > 1.0$. The proportion that $\rho < 1.0$ is about 82 % and the proportion that $\rho > 1.0$ is about 18 %.

Thus in most cases the approximation causes the rate of convergence to be faster. These experiments treats the simplest case that two of three states are combined together. We can guess that if many states are combined, then in almost all cases the rate of convergence becomes much faster.

In many applications, we have failed to apply Markov chain models for time series. As discussed above, we can guess that in most cases, the cause of the failures is in the adoption of too small number of states. Hence, when we try to apply a Markov chain model to a time series, we should take care of adopting sufficiently many number of states. If the number of states is smaller than the real process, the rate of convergence of the approximate Markov chain would be greatly faster than that of the real process.

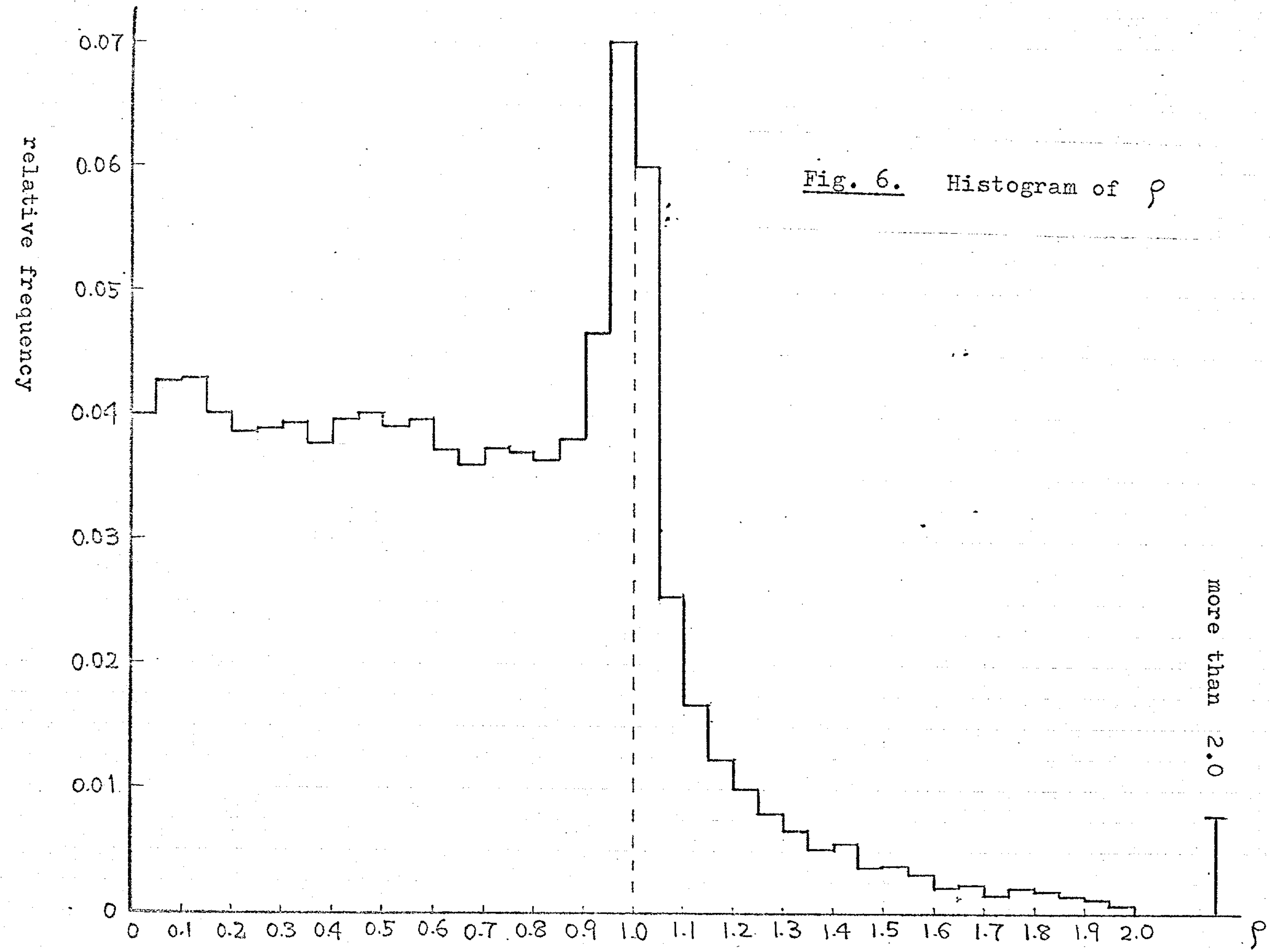


Fig. 6. Histogram of ρ

References

[1] Burke, C.J. & Rosenblatt, M., "A Markovian function of a Markov chain ", Ann. Math. Stat. 29 (1958), 1112-1122.

[2] Dharmadhikari, S.W., "Functions of finite Markov chains", Ann. Math. Stat. 34 (1963), 1022-1032.

[3] Dharmadhikari, S.W., "A characterisation of a class of functions of finite Markov chains", Ann. Math. Stat. 36 (1965), 524-528.

[4] Gilbert, E.J., "On the identifiability problem for functions of finite Markov chains", Ann. Math. Stat. 30 (1959), 688-697.

[5] Dmitriev, N.A. & Dynkin E.B., "On the characteristic roots of stochastic matrices", Dokl. Akad. Nauk SSSR 49 (1945), 159-162

[6] Guardabassi, G. & Rinaldi, S., "Two problems in Markov chains: a topological approach", Operations Research 18 (1970), 324-333.

[7] Heller, A., "On stochastic processes derived from Markov chains", Ann. Math. Stat. 36 (1965), 1286-1291.

[8] Kemeny, J.G. & Snell, J.L., Finite Markov Chains. Van Nostrand Comp., Princeton (1960).

[9] Marcus, M. & Ming, H., A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Inc., Boston, 1964.

[10] Telser, L.G., "Least-square estimates of transition probabilities", Chap. 11 of Measurement in Economics: Studies in Mathematical Economics in Memory of Yehuda Grunfeld. Stanford Unv. Press, Stanford Cal. (1963).

[11] Theil, H. & Guido, R., "A quadratic programming approach to the estimation of transition probabilities", Management Sci. 12 (1966), 714-721.

[12] Wilkinson, M.A., The Algebraic Eigenvalue Problem. Clarendon Press, Oxford, (1965).

CHAPTER 3 ON THE EFFECTS OF SMALL DEVIATIONS IN THE TRANSITION MATRIX OF A FINITE MARKOV CHAIN

1. Introduction

Let $P = \left\{ p_{ij} \; ; \; i,j = 1,2,\cdots,r \right\}$ be the transition matrix of a stationary, regular Markov chain M and $\alpha = \left\{ a_i \; ; \; i = 1,2,\cdots,r \right\}$ be its limiting vector which represents the stationary distribution. Suppose that there is another regular Markov chain M' with the transition matrix $P' = \left\{ p'_{ij} \right\}$ and the limiting vector $\alpha' = \left\{ a'_i \right\}$. If P' is close to P, we can expect that α' is also close to α. Then how close are they? If P and P' are exactly known, then we can answer the question by calculating both α and α'. However in most practical cases, we only know that P' is close to P. Such a situation arises whenever we have to infer the transition matrix of a Markov chain. In such a case, we can only get an approximate value of the transition matrix, and we want to know the range where the real limiting vector is in.

This problem is rather difficult than it appears. Because, each entry a_i of the limiting vector is written as a quotient of two determinants of matrices, and generally it is not easy to determine the range of variation of a determinant caused by small changes of its entries.

P. J. Shweitzer [2] gave an answer of this problem by giving a perturvation series expansion of α' in powers of a matrix $U = (P' - P)(I - P + A)^{-1}$ representing the difference between P and P', where A is the matrix with α in each row. However, since U includes the inverse of a matrix, we cannot easily guess the value of U when we only know that P' is close to P. J. L. Smith [3] showed that if $p_{ij} - \Delta p_{ij}^{-} \leqq p'_{ij} \leqq p_{ij} + \Delta p_{ij}^{+}$ for all i and j, then α' lies in a convex cone in the r-th order Eucledian space bounded by at most $3r$ hyperplanes. His method gives a precise information about α', but we can hardly guess the range of α' without calculating a linear programming problem.

In this paper, we obtain a simple bound for α' using special properties of the matrix $(I - P)$. This method can also be applied for other quantities, e.g., the mean values of first passage times, the variances of first passage times, taboo probabilities, and so on.

We obtain a bound for α' in Section 2, and bounds for other basic quantities in Section 4. In Sections 5 and 6, we also obtain simple bounds for the variances of the basic quantities for the Markov chain M', when p'_{ij} are mutually independent random variables.

2. A Bound for Limiting Vector α'

We consider two regular Markov chains M and M' with a common state space $S = \{ s_1, s_2, \cdots, s_r \}$. We denote their transition matrices by $P = \{ p_{ij} \}$ and $P' = \{ p'_{ij} \}$ and their limiting vectors by $\alpha = \{ a_i \}$ and $\alpha' = \{ a'_i \}$. If P' is close to P, then we can expect that α' is also close to α. We have many measures for closeness. Here we are convenient to adopt the ratios p'_{ij} / p_{ij} as a measure. Because, in dealing with a Markov chain, it is very impotant whether individual p_{ij} is equal to zero or not, and so ratios between transition probabilities are rather meaningful than differences. Hence we assume that

$$(2.1) \qquad (1+\epsilon)^{-1} p_{ij} \leqq p'_{ij} \leqq (1+\epsilon) p_{ij} \qquad (i,j = 1,2,\cdots,r \ ; \ i \neq j)$$

where ϵ is a positive constant. Before we obtain a bound for a'_k, we prepare some lemmas.

2.1 Preliminary lemmas

By the assumption of regularity, the limiting vector α of the Markov chain M is the unique positive stochastic row vector satisfying $\alpha P = \alpha$. Let $\overline{P}_k$ $(k = 1,2,\cdots,r)$ be the (k,k)-th cofactor of the matrix $(I - P)$, i.e., $\overline{P}_k$ be the determinant of the matrix formed by deleting the k-th row and the k-th column from $(I - P)$, where I is the r r identity matrix. And let $U = \sum_{k=1}^{r} \overline{P}_k$. We use the same notations with primes for corresponding quantities for the Markov chain M'.

Lemma 1. For each k $(k = 1,2,\cdots,r)$,

(2.2) $$a_k = \overline{P}_k / U .$$

Proof. This is an easy consequence of the Cramer's rule for equations $\alpha P = \alpha$ and $\sum_{i=1}^{r} a_i = 1$.

Lemma 2. For each k $(k = 1,2,\cdots,r)$, $\overline{P}_k$ can be written as a sum of products of transition probabilities with plus signs:

(2.3) $$\overline{P}_k = \sum_{J_k} p_{1j_1} p_{2j_2} \cdots p_{k-1\, j_{k-1}} p_{k+1\, j_{k+1}} \cdots p_{rj_r}$$

where the summation is taken over a set J_k of ordered $(r-1)$-tuples $(j_1, j_2, \cdots, j_{k-1}, j_{k+1}, \cdots, j_r)$, $(j_i = 1,2,\cdots, i-1, i+1, \cdots, r)$.

Proof. This is a fundamental lemma, but is a trivial corollary of Lemma 9 in Section 3. Hence we omit the proof here.

Lemma 3. If $0 \leq x' \leq x$ and $0 \leq y' \leq y$, then

(2.4) $$0 \leq x' + y' \leq x + y \quad \text{and} \quad 0 \leq x'y' \leq xy .$$

Proof. This proof is obvious.

Lemma 4. If (2.1) holds, then for each k $(k = 1,2,\cdots,r)$ we have

(2.5) $$(1+\epsilon)^{-r+1} \overline{P}_k \leq \overline{P}'_k \leq (1+\epsilon)^{r-1} \overline{P}_k$$

and

(2.6) $$(1+\epsilon)^{-r+1} (U - \overline{P}_k) \leq (U' - \overline{P}'_k) \leq (1+\epsilon)^{r-1} (U - \overline{P}_k)$$

Proof. This lemma can be easily proved by (2.1) and Lemmas 2 and 3.

2.2 A bound for a_k'

Here we propose a bound for a_k'. In Section 2.4, we will show that there exists a pair of Markov chains which nearly attain this bound. So, this bound is the best one of those which use the information on a_k only. In the corollaries below, $O(x)$ represents a term such that $O(x)/x$ is bounded in a neighbourhood of the origin.

<u>Theorem 1</u>. If (2.1) holds, then for each k $(k=1,2,\cdots,r)$ we have

$$\frac{a_k}{a_k+(1+\epsilon)^{2r-2}(1-a_k)} \leq a_k' \leq \frac{a_k}{a_k+(1+\epsilon)^{-2r+2}(1-a_k)} \quad (2.7)$$

<u>Proof</u>. By Lemmas 1 and 4,

$$a_k' = \frac{\bar{P}_k'}{U'} = \frac{\bar{P}_k'}{\bar{P}_k' + (U' - \bar{P}_k')} \quad (2.8)$$

$$\leq \frac{(1+\epsilon)^{r-1}\bar{P}_k}{(1+\epsilon)^{r-1}\bar{P}_k + (1+\epsilon)^{-r+1}(U-\bar{P}_k)}$$

$$= \frac{a_k}{a_k+(1+\epsilon)^{-2r+2}(1-a_k)} .$$

This proves a half of the theorem, and the other half can be proved similarly.

<u>Corollary 1</u>. If (2.1) holds for sufficiently small ϵ, then we have

$$|a_k' - a_k| \leq 2(r-1)(1-a_k)a_k\epsilon + O(\epsilon^2) . \quad (2.9)$$

This method for obtaining a bound for a_k' can be easily adopted for the case where P' differs from P only in one row, or the case where the range of possible values of p_{ij}' is different in each row, i.e., the value of ϵ in (2.1) is different in each i. Here we shall show results for the former case, i.e., P' differs from P only in the h-th row. Theorem 2 will be proved in the same way as Theorem 1, using Lemma 5 instead of Lemma 4. We assume that

$$(2.10) \qquad (1+\epsilon)^{-1} p_{hj} \leqq p_{hj}' \leqq (1+\epsilon) p_{hj} \qquad (j = 1,2,\cdots,h-1,h+1,\cdots,r)$$

and

$$(2.11) \qquad p_{ij}' = p_{ij} \qquad (i,j = 1,2,\cdots,r\ ;\ i \neq h)\ .$$

<u>Lemma 5</u>. If (2.10) and (2.11) hold, then we have

$$(2.12) \qquad \bar{P}_h' = \bar{P}_h \text{ and } (1+\epsilon)^{-1}(U - \bar{P}_h) \leqq (U' - \bar{P}_h') \leqq (1+\epsilon)(U - \bar{P}_h),$$

and

$$(2.13) \qquad (1+\epsilon)^{-1}\bar{P}_k \leqq \bar{P}_k' \leqq (1+\epsilon)\bar{P}_k$$

and

$$(2.14) \qquad (1+\epsilon)^{-1}(U - \bar{P}_k - \bar{P}_h) \leqq (U' - \bar{P}_k' - \bar{P}_h') \leqq (1+\epsilon)(U - \bar{P}_k - \bar{P}_h)$$

for each k $(\neq h)$.

Theorem 2. If (2.10) and (2.11) hold, then we have

$$\frac{a_h}{a_h + (1+\epsilon)(1-a_h)} \leq a'_h \leq \frac{a_h}{a_h + (1+\epsilon)^{-1}(1-a_h)}, \tag{2.15}$$

and for each k $(k = 1,2,\cdots,h-1,h+1,\cdots,r)$ we have

$$\frac{a_k}{a_k + (1+\epsilon)a_h + (1+\epsilon)^2(1-a_k-a_h)} \leq a'_k \leq \tag{2.16}$$

$$\leq \frac{a_k}{a_k + (1+\epsilon)^{-1}a_h + (1+\epsilon)^{-2}(1-a_k-a_h)} .$$

Corollary 2. If (2.10) and (2.11) hold, and if ϵ is sufficiently small, then we have

$$|a'_h - a_h| \leq (1-a_h)a_h\epsilon + O(\epsilon^2), \tag{2.17}$$

and for each k $(\neq h)$ we have

$$|a'_k - a_k| \leq (2-2a_k-a_h)a_k\epsilon + O(\epsilon^2). \tag{2.18}$$

2.3 Taylor's expansion of a'_k

In order to obtain a tool for an intimate study of the difference between a'_k and a_k, we shall expand a'_k about a_k in powers of $(p'_{ij} - p_{ij})$. We first prepare some notations and lemmas.

By Lemma 2, we can write $\overline{P}_k$ and U as sums of products of transition probabilities with plus signs. We denote the sum of terms in the right hand side of (2.3) containing p_{ij} by $\overline{P}_k^{ij}$ and define that $U^{ij} = \sum_{k=1}^{r} \overline{P}_k^{ij}$. Namely,

$$\bar{P}_k^{ij} = p_{ij} \frac{\partial}{\partial p_{ij}} \bar{P}_k \qquad (i,j,k = 1,2,\cdots,r \; ; \; i \neq j) \tag{2.19}$$

and

$$U^{ij} = p_{ij} \frac{\partial}{\partial p_{ij}} U \qquad (i,j,k = 1,2,\cdots,r \; ; \; i \neq j) \; , \tag{2.20}$$

where we differentiate $\bar{P}_k$ and U regarding them as functions of p_{ij}'s $(i,j = 1,2,\cdots,r \; ; \; i \neq j)$ and we do not consider that p_{ii}'s $(i = 1,2,\cdots,r)$ are variables for them. The following Lemmas 6 and 7 will be easily proved by Lemma 2 and the definitions.

Lemma 6. $\bar{P}_k^{ij}$ $(i,j,k = 1,2,\cdots,r \; ; \; i \neq j \; , \; i \neq k)$ can be written as a sum of products of transition probabilities with plus signs, and $\bar{P}_k^{kj}$ $(k,j = 1,2,\cdots,r \; ; \; k \neq j)$ is equal to zero.

Lemma 7. For each i,k such that $i \neq k$, we have

$$\bar{P}_k = \sum_{\substack{j=1 \\ j \neq i}}^{r} \bar{P}_k^{ij} \; , \tag{2.21}$$

and for each k we have

$$U - \bar{P}_k = \sum_{\substack{j=1 \\ j \neq k}}^{r} U^{kj} \; . \tag{2.22}$$

We define ϵ_{ij} by

$$p'_{ij} = p_{ij}(1 + \epsilon_{ij}) \qquad (i,j = 1,2,\cdots,r \; ; \; i \neq j) \; , \tag{2.23}$$

and assume that $|\epsilon_{ij}| \leq \epsilon$ for sufficiently small ϵ . Then by the Taylor's formula we obtain

Theorem 3. For each k $(k = 1,2,\cdots,r)$ we have

$$a_k' = a_k + a_k \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} \left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U} \right) \epsilon_{ij} + O(\epsilon^2) . \tag{2.24}$$

Proof. Applying Lemma 1 for the Markov chain M', we have

$$a_k' = \frac{\overline{P}_k'}{U'} . \tag{2.25}$$

By the Taylor's formula, we can expand it as

$$a_k' = \left.\frac{\overline{P}_k'}{U'}\right|_{p_{ij}' = p_{ij}} + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} \left\{ \frac{\partial}{\partial p_{ij}'} \left.\frac{\overline{P}_k'}{U'}\right|_{p_{ij}' = p_{ij}} \right\} (p_{ij}' - p_{ij}) + R_2 \tag{2.26}$$

where R_2 represents the residual term and it may be replaced with $O(\epsilon^2)$. By (2.19), (2.20) and (2.23),

$$\begin{aligned} &\left\{ \frac{\partial}{\partial p_{ij}'} \left.\frac{\overline{P}_k'}{U'}\right|_{p_{ij}' = p_{ij}} \right\} (p_{ij}' - p_{ij}) \\ &= \left\{ \frac{1}{U'} \left.\frac{\partial \overline{P}_k'}{\partial p_{ij}'}\right|_{p_{ij}' = p_{ij}} \right\} p_{ij}\epsilon_{ij} - \left\{ \frac{\overline{P}_k'}{U'^2} \left.\frac{\partial U'}{\partial p_{ij}'}\right|_{p_{ij}' = p_{ij}} \right\} p_{ij}\epsilon_{ij} \\ &= \frac{1}{U} \overline{P}_k^{ij} \epsilon_{ij} - \frac{\overline{P}_k}{U^2} U^{ij} \epsilon_{ij} \\ &= \frac{\overline{P}_k}{U} \left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U} \right) \epsilon_{ij} . \end{aligned} \tag{2.27}$$

Hence (2.26) becomes to

$$a' = \frac{\overline{P}_k}{U} + \frac{\overline{P}_k}{U} \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} \left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U} \right) \epsilon_{ij} + O(\epsilon^2) , \tag{2.28}$$

and using Lemma 1 again, we obtain (2.24).

2.4 Examples

We have obtained a bound for a'_k. It uses the information of a_k only, and the following Example 1 shows that the bound is nearly attained with a Markov chain which has many negligible a_i's. Hence it is expected that the bound will be greatly improved by using all of the information of the limiting vector. Unfortunately the author has never been able to get such a bound. He has thought that the following Example 2 shows some informations for studying such a bound.

Example 1. This example shows that the bound (2.9) is the best one of those which use the information of a_k only. Let

$$(2.29) \qquad P = \begin{pmatrix} p_{11} & p_{12} & & & \\ p_{21} & p_{22} & p_{23} & 0 & \\ p_{31} & & p_{33} & p_{34} & \\ p_{41} & & & p_{44} & p_{45} \\ p_{51} & 0 & & & p_{55} \end{pmatrix}.$$

Then $\overline{P}_k$'s are given by

$$(2.30) \qquad \begin{aligned} \overline{P}_1 &= (p_{21} + p_{23})(p_{31} + p_{34})(p_{41} + p_{45})\, p_{51} \\ \overline{P}_2 &= p_{12} \times (p_{31} + p_{34})(p_{41} + p_{45})\, p_{51} \\ \overline{P}_3 &= p_{12} \times p_{23} \times (p_{41} + p_{45})\, p_{51} \\ \overline{P}_4 &= p_{12} \times p_{23} \times p_{34} \times p_{51} \\ \overline{P}_5 &= p_{12} \times p_{23} \times p_{34} \times p_{45} \end{aligned}$$

Hence by Theorem 3 we have

$$(2.31)\qquad a_5' - a_5 = a_5\Big\{ \qquad + \epsilon_{12}(1 - U^{12}/U)$$
$$- \epsilon_{21}U^{21}/U + \epsilon_{23}(1 - U^{23}/U)$$
$$- \epsilon_{31}U^{31}/U + \epsilon_{34}(1 - U^{34}/U)$$
$$- \epsilon_{41}U^{41}/U + \epsilon_{45}(1 - U^{45}/U)$$
$$- \epsilon_{51}U^{51}/U + O(\epsilon^2)\Big\}.$$

Now we assume that $p_{21} = p_{31} = p_{41} = \delta$, $p_{12} = p_{23} = p_{34} = p_{45} = \delta^2$ and $p_{51} = \delta^5$ for sufficiently small δ. Then

$\overline{P}_1 = \delta^8(1+\delta)^3$, $\overline{P}_2 = \delta^9(1+\delta)^2$, $\overline{P}_3 = \delta^{10}(1+\delta)$, $\overline{P}_4 = \delta^{11}$, $\overline{P}_5 = \delta^8$.

So, we have approximately

$$(2.32)\qquad U^{21} \approx \overline{P}_1 , \quad U^{31} \approx \overline{P}_1 , \quad U^{41} \approx \overline{P}_1 , \quad U^{51} \approx \overline{P}_1 ,$$
$$U^{12} \approx \overline{P}_5 , \quad U^{23} \approx \overline{P}_5 , \quad U^{34} \approx \overline{P}_5 , \quad U^{45} \approx \overline{P}_5$$
$$\text{and} \quad U \approx \overline{P}_1 + \overline{P}_5 .$$

Thus

$$(2.33)\qquad a_5' - a_5$$
$$\approx a_5\Big\{(\epsilon_{12} + \epsilon_{23} + \epsilon_{34} + \epsilon_{45}) - (\epsilon_{21} + \epsilon_{31} + \epsilon_{41} + \epsilon_{51})\Big\}\overline{P}_1/U ,$$

and if $\epsilon_{12} = \epsilon_{23} = \epsilon_{34} = \epsilon_{45} = \epsilon$ and $\epsilon_{21} = \epsilon_{31} = \epsilon_{41} = \epsilon_{51} = -\epsilon$, then it becomes

(2.34) $$a_5' - a_5 \approx 8(1 - a_5)a_5 \epsilon \ .$$

This coincides with the bound in (2.9) for $r = 5$.

Example 2. We consider the extreme case where all p_{ij} $(i,j = 1,2,\cdots,r\ ;\ i \neq j)$ are equal to p (some constant). Then the following theorem can be proved by direct calculations.

Theorem 4. If $p_{ij} = p$ for all i,j such that $i \neq j$, then we have

(2.35) $$\overline{P}_k = r^{r-2}\, p^{r-1} \ .$$

(2.36) $$\overline{P}_k^{ij} = r^{r-3}\, p^{r-1} \quad \text{if } i \neq k \text{ and } j \neq k\ ,\ j \neq i\ .$$
$$= 2\, r^{r-3}\, p^{r-1} \quad \text{if } i \neq k \text{ and } j = k\ .$$

(2.37) $$U = r^{r-1}\, p^{r-1} \ .$$

(2.38) $$U^{ij} = r^{r-2}\, p^{r-1} \quad \text{if } j \neq i\ .$$

Then by Theorem 3, we have

(2.39) $$a_k' - a_k = a_k \frac{1}{r} \Big(\sum_{i \neq k} \epsilon_{ik} - \sum_{j \neq k} \epsilon_{kj} \Big) + O(\epsilon^2)\ .$$

Hence for this special Markov chain, we obtain a bound

(2.40) $$\left| a_k' - a_k \right| \leq 2\,(1 - a_k)\, a_k \epsilon + O(\epsilon^2)\ ,$$

since $a_k = 1/r$. In this case the multiplier $(r - 1)$ in (2.9) vanishes.

3. Fundamental Lemmas

In the preceding section, we have seen that we can derive a simple bound for a_k' from Lemma 2. This lemma treats a matrix of the form defined by (3.1) and (3.2) below. In this section we study some properties of such a matrix.

We consider a matrix $X = \{ x_{ij} ; i,j = 1,2, \cdots ,n \}$ with components

$$x_{ij} = -y_{ij} \qquad (i,j = 1,2, \cdots ,n ; i \neq j) \tag{3.1}$$

and

$$x_{ii} = \sum_{k=1}^{n} y_{ik} \qquad (i = 1,2, \cdots ,n) , \tag{3.2}$$

where y_{ij} $(i,j = 1,2, \cdots ,n)$ are some constants or variables. The following Lemmas 8,9,10 and 11 show fundamental properties of such a matrix.

<u>Lemma 8</u>. Let X be the matrix defined by (3.1) and (3.2). Then its determinant is written as a sum of products of y_{ij}'s with plus signs:

$$|X| = \sum_{J} y_{1j_1} y_{2j_2} \cdots y_{nj_n} \tag{3.3}$$

where the summation is taken over a set J of ordered n-tuples $(j_1, j_2, \cdots , j_n)$.

Proof. Since

$$(3.4)\qquad |X| = \begin{vmatrix} y_{11}+\cdots+y_{1n} & -y_{12} & \cdots & -y_{1\,n-1} & -y_{1n} \\ -y_{21} & y_{21}+\cdots+y_{2n} & \cdots & -y_{2\,n-1} & -y_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ -y_{n-1\,1} & -y_{n-1\,2} & \cdots & y_{n-1\,1}+\cdots+y_{n-1\,n} & -y_{n-1\,n} \\ -y_{n1} & -y_{n2} & \cdots & -y_{n\,n-1} & y_{n1}+\cdots+y_{nn} \end{vmatrix},$$

we can write it as a sum of products of y_{ij}'s with plus or minus signs:

$$(3.5)\qquad X = \sum_J \delta(j_1,j_2,\cdots,j_n)\, y_{1j_1}y_{2j_2}\cdots y_{nj_n} ,$$

where $\delta(j_1,j_2,\cdots,j_n) = +1$ or $= -1$, and the summation is taken over a set J of ordered n-tuples $(j_1,j_2,\cdots,j_n)$. We first note that each product in the right hand side of (3.5) contains at least one y_{ii} $(i = 1,2,\cdots,n)$, in other words, each n-tuple in J has at least one element j_i such that $j_i = i$. In the determinant of (3.4), we add all the columns from the first to the (n-1)-th, to the last column. Then the last column becomes to $(y_{11},y_{22},\cdots,y_{nn})^T$, where the superscript T represents the transpose of the vector. Hence each product in the right hand side of (3.5) must contain some y_{ii}.

We prove the lemma using the mathematical induction on n . If $n = 2$, then $|X| = y_{11}y_{22} + y_{11}y_{21} + y_{12}y_{22}$, and the lemma is true in this case. We assume that the lemma is true for $n-1$. Namely, we consider an $(n-1)\times(n-1)$ matrix $X' = \{x'_{ij}\}$ with components

$$(3.6)\qquad x'_{ij} = -y'_{ij} \qquad (i,j = 1,2,\cdots,n-1\ ;\ i\neq j)$$

and

$$(3.7) \qquad x'_{ii} = \sum_{j=1}^{n-1} y'_{ij} \qquad (i = 1,2,\cdots,n-1)$$

where y'_{ij} $(i,j = 1,2,\cdots,n-1)$ are some constants or variables, and we assume that its determinant is written as a sum of products of y'_{ij}'s with plus signs:

$$(3.8) \qquad |X'| = \sum_{J'} y'_{1j_1} y'_{2j_2} \cdots y'_{n-1\, j_{n-1}} \ .$$

We consider the terms (i.e., the products with signs) containing y_{nn} in the right hand side of (3.5). Since y_{nn} is only in the (n,n)-th entry of the determinant of (3.4), the sum of them is written as the product of y_{nn} and the determinant of the matrix X^{nn} formed by deleting the n-th row and the n-th column from the matrix X . If we set

$$(3.9) \qquad y'_{ij} = y_{ij} \qquad (i,j = 1,2,\cdots,n-1 \ ; \ i \neq j)$$

and

$$(3.10) \qquad y'_{ii} = y_{ii} + y_{in} \qquad (i = 1,2,\cdots,n-1)\ ,$$

then X^{nn} becomes of the same form as X' . Hence by the representation (3.8), $|X^{nn}|$ is also written as a sum of products of y_{ij}'s with plus signs. Thus we can conclude that the products containing y_{nn} in the right hand side of (3.5) have plus signs.

Similarly, we can prove that products containing y_{ii} $(i = 1,2,\cdots,n-1)$ in the right hand side of (3.5) have also plus signs. Since each product in the right hand side of (3.5) contains some y_{ii} , all of them have plus signs, and this proves the lemma.

Lemma 9. Let X be the matrix defined by (3.1) and (3.2). Then the (i,j)-th cofactor $X(i,j)$ of it is written as a sum of products of y'_{kl}'s $(k \neq i)$ with plus signs:

$$(3.11) \qquad X(i,j) = \sum_{J_{ij}} y_{1j_1} y_{2j_2} \cdots y_{i-1\,j_{i-1}} y_{i+1\,j_{i+1}} \cdots y_{nj_n}$$

where the summation is taken over a set J_{ij} of ordered $(n-1)$-tuples $(j_1, j_2, \cdots, j_{i-1}, j_{i+1}, \cdots, j_n)$.

Proof. We prove the lemma for two typical cofactors $X(n,n)$ and $X(n,n-1)$. In similar ways we can prove it for other cofactors.

$X(n,n)$ is the determinant of the matrix X^{nn} defined in the preceding proof. So the proof for $X(n,n)$ is essentially contained in the preceding proof.

By the definition, we have

$$(3.12) \quad X(n,n-1) = (-1)^{2n-1} \begin{array}{r|ccc|c|} & & & & n-1 \\ & y_{11}+\cdots+y_{1n} & \cdots & -y_{n\,n-2} & -y_{1n} \\ & \vdots & & \vdots & \vdots \\ & -y_{n-2\,1} & \cdots & y_{n-2\,1}+\cdots+y_{n-2\,n} & -y_{n-2\,n} \\ & -y_{n-1\,1} & \cdots & -y_{n-1\,n-2} & -y_{n-1\,n} \\ n) & - - - & - - - & - - - & - - - \end{array}$$

In the determinant of (3.12), we add all the columns from the first to the $(n-2)$-th, to the last column, and then multiply the last column by -1. Then the last column becomes as

$$(3.13)\qquad \begin{pmatrix} -y_{11}-y_{1\,n-1} \\ \vdots \\ -y_{n-2\,n-2}-y_{n-2\,n-1} \\ y_{n-1\,1}+\cdots+y_{n-1\,n-2}+y_{n-1\,n} \end{pmatrix}.$$

Hence if we set

$$(3.14)\qquad \begin{aligned} y'_{ij} &= y_{ij} \qquad (i=1,2,\cdots,n-1,\ \ j=1,2,\cdots,n-2\ ;\ i\neq j) \\ y'_{i\,n-1} &= y_{ii}+y_{i\,n-1} \qquad (i=1,2,\cdots,n-2) \\ \text{and}\quad y'_{ii} &= y_{in} \qquad (i=1,2,\cdots,n-1)\ , \end{aligned}$$

then $X(n,n-1)$ coincides with the determinant of the matrix X' defined by (3.6) and (3.7). By Lemma 8, $|X'|$ is written as (3.8), and by the relations in (3.14) we can conclude that the cofactor $X(n,n-1)$ can be written as in (3.11).

<u>Lemma 10</u>. Let J_{ij} $(i,j=1,2,\cdots,n)$ be the set of (n-1)-tuples defined in Lemma 9. Then if $i\neq j$,

$$(3.15)\qquad J_{ii}\supset J_{ij}\ ,$$

or equivalently, $X(i,i)-X(i,j)$ can be written as a sum of products of y_{kl}'s $(k\neq i)$ with plus signs:

$$(3.16)\qquad X(i,i)-X(i,j)=\sum_{J_{ii}-J_{ij}} y_{1j_1}y_{2j_2}\cdots y_{i-1\,j_{i-1}}y_{i+1\,j_{i+1}}\cdots y_{nj_n}\ .$$

Proof. $X(i,i) - X(i,j)$ ca be considered as the determinant of the matrix formed by replacing the i-th row of X with a vector which has $+1$ in the i-th entry, -1 in the j-th entry and 0's in other entries. This matrix has a special form of X with $y_{ij} = 1$ and $y_{ik} = 0$ for $k = 1, 2, \cdots, j-1, j+1, \cdots, n$. Thus this lemma can be proved directly from Lemma 1.

Lemma 11. Let X be the matrix defined by (3.1) and (3.2), and define that $y_{i0} = 1 - (y_{i1} + y_{i2} + \cdots + y_{in})$ for given i $(i = 1, 2, \cdots, n)$. Then $X(i,i) - |X|$ is written as a sum of products of y_{i0} and y_{kj}'s $(k, j = 1, 2, \cdots, n)$ with plus signs:

$$X(i,i) - |X| = \sum_{J^*} y_{1j_1} y_{2j_2} \cdots y_{nj_n}, \tag{3.17}$$

where the summation is taken over a set J^* of ordered n-tuples $(j_1, j_2, \cdots, j_n)$.

Proof. By expanding $|X|$ in cofactors by the i-th row, we have

$$|X| = (y_{i1} + y_{i2} + \cdots + y_{in})\, X(i,i) - \sum_{k \neq i} y_{ik}\, X(i,k). \tag{3.18}$$

It follows that

$$X(i,i) - |X| = y_{i0}\, X(i,i) + \sum_{k \neq i} y_{ik}\, X(i,k). \tag{3.19}$$

Hence by Lemma 9, $X(i,i) - |X|$ is written as (3.17).

4. Bounds for Some Basic Quantities for Finite Markov Chains

In Section 2.1, we saw that each element a_k of the limiting vector of a regular Markov chain M can be written as a quotient of sums of products of transition probabilities with plus signs. This fact enables us to obtain a simple bound for a'_k of other Markov chain M' in Section 2.2 The same idea can be adopted for other quantities in finite Markov chains. In this section, we show that many quantities in finite Markov chains can be represented as quotients of sums of products of transition probabilities with plus signs or as sums of such quotiests, and using the representations we obtain simple bounds for the quantities.

We will use the following notations.

$s_1, s_2, \cdots, s_r$ states of a Markov chain

P transition matrix of the chain

$M_i[\mathbf{f}]$ the mean value of a random variable $\mathbf{f}$ when the chain is started from state s_i

$\mathbf{Var}_i[\mathbf{f}]$ the variance of $\mathbf{f}$ when the chain is started from state s_i

$G = \{g_{ij}\}$ matrix with entries g_{ij}

$\gamma = \{g_i\}$ column vector with entries g_i

ξ column vector with all entries 1

I identity matrix

G_{sq} matrix whose (i,j)-th entry is g_{ij}^2

G_{dg} diagonal matrix whose i-th diagonal entry is g_{ii}

4.1 Some basic quantities for absorbing Markov chains

We consider an absorbing Markov chain M with states $s_1, s_2, \cdots, s_r$. We let $T = \{s_1, s_2, \cdots, s_s\}$ be the set of transient states, and $\tilde{T} = \{s_{s+1}, s_{s+2}, \cdots, s_r\}$ be the set of absorbing states. Then the transition matrix $P = \{p_{ij}\}$ of the Markov chain has the form

$$
(4.1) \qquad P = \begin{array}{cc} \begin{array}{cc} s & r-s \end{array} & \\ \left(\begin{array}{c|c} Q & R \\ \hline O & I \end{array}\right) & \begin{array}{c} s \\ r-s \end{array} \end{array} .
$$

We shall deal with the following quantities for the chain.

n_j number of times in state s_j before absorbing

t number of steps taken before absorption

m total number of transient states entered before absorption

b_{ij} probability starting in state s_i that the chain is absorbed in state s_j

h_{ij} probability starting in state s_i that the chain is ever in state s_j

$N = \{n_{ij}\} = \{M_i[n_j]\}$ $\quad (s_i, s_j \in T)$

$N_2 = \{Var_i[n_j]\}$ $\quad (s_i, s_j \in T)$

$\tau = \{M_i[t]\}$ $\quad (s_i \in T)$

$\tau_2 = \{Var_i[t]\}$ $\quad (s_i \in T)$

$\mu = \{M_i[m]\}$ $\quad (s_i \in T)$

$B = \{b_{ij}\}$ $\quad (s_i \in T,\ s_j \in \tilde{T})$

$H = \{h_{ij}\}$ $\quad (s_i, s_j \in T)$

We can prove that each entry of above matrices or vectors except τ_2, can be represented as a quotient of two sums of products of transition probabilities with plus signs, or as a sum of such quotients. We denote the matrix $(I - Q)$ by $\overline{Q}$ and the (i,j)-th cofactor of it by $\overline{Q}(i,j)$. Since row sums of the matrix P are 1, we can write

$$(4.2) \quad \overline{Q} = \begin{pmatrix} p_{12}+\cdots+p_{1r} & -p_{12} & \cdots & -p_{1s} \\ -p_{21} & p_{21}+p_{23}+\cdots+p_{2r} & \cdots & -p_{2s} \\ \vdots & \vdots & & \vdots \\ -p_{s1} & -p_{s2} & \cdots & p_{s1}+\cdots+p_{s\,s-1}+p_{s\,s+1}+\cdots+p_{sr} \end{pmatrix}.$$

If we set

$$(4.3) \qquad y_{ij} = p_{ij} \qquad (i,j = 1,2,\cdots,s\ ;\ i \neq j)$$

and

$$(4.4) \qquad y_{ii} = \sum_{j=s+1}^{r} p_{ij} \qquad (i = 1,2,\cdots,s),$$

then $\overline{Q}$ is of the same form as X defined by (3.1) and (3.2). Hence by Lemmas 8, 9, 10 and 11, $|\overline{Q}|$, $\overline{Q}(i,j)$, $\overline{Q}(i,i) - \overline{Q}(i,j)$ and $\overline{Q}(i,i) - |\overline{Q}|$ are written as sums of products of transition probabilities with plus signs. We shall show that quantities defined above can be represented in terms of p_{ij}, $\overline{Q}$ and $\overline{Q}(i,j)$. We use relations proved in [1].

Since $N = (I - Q)^{-1}$,

$$(4.5) \qquad M_i[n_j] = n_{ij} = \overline{Q}(j,i)/|\overline{Q}|.$$

Since $N_2 = N(2N_{dg} - I) - N_{sq}$,

(4.6) $$\mathrm{Var}_i[\mathbf{n}_j] = 2n_{ij}n_{jj} - n_{ij} - n_{ij}^2$$
$$= \frac{\overline{Q}(j,i)}{|\overline{Q}|^2}\left\{(\overline{Q}(j,j) - |\overline{Q}|) + (\overline{Q}(j,j) - \overline{Q}(j,i))\right\}.$$

Since $\tau = N\xi$,

(4.7) $$M_i[\mathbf{t}] = \sum_{k=1}^{s} n_{ik} = \frac{1}{|\overline{Q}|}\sum_{k=1}^{s}\overline{Q}(k,i) .$$

Since $\tau_2 = (2N - I)\tau - \tau_{sq}$,

(4.8) $$\mathrm{Var}_i[\mathbf{t}] = 2\sum_{k=1}^{s}\sum_{j=1}^{s} n_{ij}n_{jk} - \sum_{k=1}^{s} n_{ik} - \left\{\sum_{k=1}^{s} n_{ik}\right\}^2 .$$

This quantity cannot be written as aquotient of two sums of products of transition probabilities with plus signs (see Example 3 below).

Since $\mu = (NN_{dg}^{-1})\xi$,

(4.9) $$M_i[\mathbf{m}] = \sum_{k=1}^{s} n_{ik}/n_{kk} = \sum_{k=1}^{s}\overline{Q}(k,i)/\overline{Q}(k,k) .$$

Since $B = NR$,

(4.10) $$b_{ij} = \sum_{k=1}^{s} n_{ik}p_{kj} = \frac{1}{|\overline{Q}|}\sum_{k=1}^{s} p_{kj}\,\overline{Q}(k,i) .$$

Since $H = (N - I)N_{dg}^{-1}$,

(4.11) $$h_{ii} = (n_{ii} - 1)/n_{ii} = (\overline{Q}(i,i) - |\overline{Q}|)/\overline{Q}(i,i) ,$$

and for $j \neq i$

(4.12) $$h_{ij} = n_{ij}/n_{jj} = \overline{Q}(j,i)/\overline{Q}(j,j) .$$

Thus we have shown that each quantity considered above can be written as a quotient of sums of products with plus signs except for τ_2. However the following example shows that τ_2 cannot be represented in such a form.

Example 3. We consider an absorbing Markov chain with the transition matrix

$$(4.13)\qquad P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & 0 \\ 0 & p_{22} & 0 & p_{24} \\ 0 & 0 & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A tedious calculation shows that in this case

$$(4.14)\qquad \mathrm{Var}_1[\,t\,] = \Big\{ p_{11}p_{24}^2p_{34}^2 + (p_{12}+p_{13})(p_{12}p_{22}p_{34}^2 + p_{13}p_{24}^2p_{33}) + p_{12}p_{13}(p_{24}-p_{34})^2 \Big\} \Big/ (p_{12}+p_{13})^2p_{24}^2p_{34}^2 ,$$

and the numerator of the right hand side of (4.14) cannot written as a sum of products of p_{ij}'s with plus signs. Hence we cannot adopt our method, because Lemma 3 in Section 2 cannot be used. Only the second order moment $M_i[\,t^2]$ can be written as desired form. Since $\{M_i[\,t^2]\} = (2N - I)\tau$,

$$(4.15)\qquad M_i[\,t^2\,] = 2\sum_{\substack{k=1\\k\neq i}}^{s}\sum_{j=1}^{s} n_{ik}\,n_{kj} + \sum_{j=1}^{s}(2n_{ii}-1)\,n_{ij}$$

$$= \frac{1}{|\overline{Q}|^2}\Big[2\sum_{\substack{k=1\\k\neq i}}^{s}\sum_{j=1}^{s}\overline{Q}(k,i)\,\overline{Q}(j,k) + \sum_{j=1}^{s}\Big(2\,\overline{Q}(i,i) - |\overline{Q}|\Big)\,\overline{Q}(j,i)\Big].$$

4.2 Bounds for basic quantities for an absorbing Markov chain M'

We consider two absorbing Markov chains M and M'. We use the notations defined in Section 3.1 for quantities of the chain M, and use the same notations with primes for corresponding quantities of the chain M'. We shall obtain simple bounds for basic quantities for the chain M' under the assumption that

(4.16) $$(1+\epsilon)^{-1} p_{ij} \leqq p'_{ij} \leqq (1+\epsilon) p_{ij}$$
$$(i = 1,2,\cdots,s\ ,\ j = 1,2,\cdots,r\ ;\ i \neq j)\ .$$

For (4.23) and (4.27), we need a further assumption that

(4.17) $$(1+\epsilon)^{-1} p_{ii} \leqq p'_{ii} \leqq (1+\epsilon) p_{ii} \qquad (i = 1,2,\cdots,s)\ .$$

Our main tools are in the following

<u>Lemma 12</u>. Under the assumption (4.16), we have

(4.18) $$(1+\epsilon)^{-s} |\overline{Q}| \leqq |\overline{Q}'| \leqq (1+\epsilon)^{s} |\overline{Q}|\ ,$$

(4.19) $$(1+\epsilon)^{-s+1} \overline{Q}(i,j) \leqq \overline{Q}'(i,j) \leqq (1+\epsilon)^{s-1} \overline{Q}(i,j)\ ,$$

(4.20) $$(1+\epsilon)^{-s+1} (\overline{Q}(i,i) - \overline{Q}(i,j)) \leqq \overline{Q}'(i,i) - \overline{Q}'(i,j)$$
$$\leqq (1+\epsilon)^{s-1} (\overline{Q}(i,i) - \overline{Q}(i,j))\ .$$

Under the assumptions (4.16) and (4.17), we have

(4.21) $$(1+\epsilon)^{-s} (\overline{Q}(i,i) - |\overline{Q}|) \leqq \overline{Q}'(i,i) - |\overline{Q}'|$$
$$\leqq (1+\epsilon)^{s} (\overline{Q}(i,i) - |\overline{Q}|)\ .$$

Proof. These results are obvious from Lemmas 8 , 9 , 10 and 11 .

From (4.5)$\sim$(4.12), we can obtain the bounds using Lemma 12. Here we show the upper bounds only, because lower bounds can be obtained by replacing $(1+\epsilon)$ in the upper bounds with $(1+\epsilon)^{-1}$.

(4.22) $$M_i'[n_j'] \leqq (1+\epsilon)^{2s-1} M_i[n_j]$$ under (4.16)

(4.23) $$\mathrm{Var}_i'[n_j'] \leqq (1+\epsilon)^{4s-1} \mathrm{Var}_i[n_j]$$ under (4.16) and (4.17)

(4.24) $$M_i'[t'] \leqq (1+\epsilon)^{2s-1} M_i[t]$$ under (4.16)

(4.25) $$M_i'[m'] \leqq (1+\epsilon)^{2s-2} M_i[m]$$ under (4.16)

(4.26) $$b_{ij}' \leqq (1+\epsilon)^{2s} b_{ij}$$ under (4.16)

(4.27) $$h_{ii}' \leqq (1+\epsilon)^{2s-1} h_{ii} .$$ under (4.16) and (4.17)

(4.28) $$h_{ij}' \leqq (1+\epsilon)^{2s-2} h_{ij} \qquad (i \neq j)$$ under (4.16)

Some of these bounds can be improved using the relations

(4.29) $$\overline{Q}(i,i) = \overline{Q}(i,j) + (\overline{Q}(i,i) - \overline{Q}(i,j))$$

and

(4.30) $$|\overline{Q}| = \sum_{\substack{k=1 \\ k\neq i}}^{s} p_{ik} (\overline{Q}(i,i) - \overline{Q}(i,k)) + (\sum_{k=s+1}^{r} p_{ik}) \overline{Q}(i,i) .$$

For example, (4.28) can be improved as

$$(4.31)\qquad h'_{ij} \leqq h_{ij} \Big/ \left\{ h_{ij} + (1+\epsilon)^{-2s+2}(1 - h_{ij}) \right\},$$

and (4.21) for $i \neq j$ can be improved as

$$(4.32)\qquad n'_{ij} \leqq (1+\epsilon)\overline{Q}(i,i) \Big/ \left\{ \Big(\sum_{k=s+1}^{r} p_{ik} \Big) \overline{Q}(i,i) + (1+\epsilon)^{-2s+2} \sum_{\substack{k=1 \\ k\neq i}}^{s} p_{ik} \big(\overline{Q}(i,i) - \overline{Q}(i,k) \big) \right\}.$$

4.3 Some basic quantities for regular Markov chains

We let $s_1, s_2, \cdots, s_r$ be the states of a regular Markov chain M and $P = \{p_{ij}\}$ be its transition matrix. We consider the following quantities.

$\alpha = \{a_i\}$ limiting vector (stationary distribution of the chain)

A matrix with each row α

$Z = \{z_{ij}\} = (I - P + A)^{-1}$ fundamental matrix

$M = \{m_{ij}\}$ matrix of mean number of steps required to reach s_j for the first time, starting in s_i

$W = \{w_{ij}\}$ matrix of variances for the number of steps required to reach s_j, starting in s_i

We can apply our method for obtaining simple bounds for a_i, z_{ii}, m_{ij} and w_{ii}. However off-diagonal entries of Z and W cannot be represented as quotients of sums of products of transition probabilities with plus signs or as sums of such quotients. We can easily guess this fact. Off-diagonal entries of Z have both possibilities of taking positive values and negative values. Hence clearly they cannot be represented as quotients of positive terms. Off-diagonal entries of W are essentially the same as entries of τ_2 in Section 4.1.

Now we prove two lemmas. In this section we use the notations defined in Section 2.1, too. Besides we denote by $\overline{P}_k(i,j)$ $(i,j,k=1,2,\cdots r\ ;\ i\neq k\ ,\ j\neq k)$ the (i,j)-th cofactor of the matrix formed by deleting the k-th row and the k-th column from the matrix $(I-P)$.

Lemma 13. $\overline{P}_k(i,j)$ can be written as a sum of products of transition probabilities with plus signs:

$$\overline{P}_k(i,j) = \sum_{J_{k,ij}} \left(\prod_{\substack{l=1 \\ l\neq k,i}}^{r} p_{lj_l} \right), \tag{4.33}$$

where the summation is taken over a set $J_{k,ij}$ of ordered (r-2)-tuples.

Proof. The matrix formed by deleting the k-th row and the k-th column from $(I-P)$ is of the same form as X defined by (3.1) and (3.2) with obvious modifications of suffixes. Hence this lemma is a trivial corollary of Lemma 10.

Lemma 14. For each k,

$$z_{kk} = \frac{\overline{P}_k}{U} + \frac{1}{U^2} \sum_{\substack{j=1 \\ j\neq k}}^{r} \sum_{\substack{i=1 \\ i\neq k}}^{r} \overline{P}_j\, \overline{P}_k(i,j)\ . \tag{4.34}$$

<u>Proof</u>. We first show that $|I-P+A| = U$. If we add all the columns from the first to the (r-1)-th of the matrix $(I-P+A)$ to the last column, then the last column becomes to ξ which is the column vector with all entries 1. Next for each $i = 1,2,\cdots,r-1$, we subtract $a_i\xi$ from the i-th column. Then the matrix becomes to one formed from $(I-P)$ by replacing the last column with ξ. Since this matrix has the same determinant as that of $(I-P+A)$, if we expand the determinant of the matrix by the last column, we have

(4.35) $$|I-P+A| = \sum_{k=1}^{r} \overline{P}(k,r) ,$$

where $\overline{P}(k,r)$ is the (k,r)-th cofactor of the matrix $(I-P)$. It can be easily proved that $\overline{P}(k,r) = \overline{P}_k$. Hence by the definition of U in Section 2, (4.35) becomes

(4.36) $$|I-P+A| = U .$$

Since $Z = (I-P+A)^{-1}$,

(4.37) $$z_{kk} =$$

$$= \frac{1}{U}\begin{array}{c|cccccc|}
 & & & & k & & \\
 & p_{12}+\cdots+p_{1r}+a_1 & -p_{12}+a_2 & \cdots & & \cdots & -p_{1r}+a_r \\
 & -p_{21}+a_1 & p_{21}+\cdots+p_{2r}+a_2 & \cdots & & \cdots & -p_{2r}+a_r \\
 & \vdots & \vdots & & & & \vdots \\
k) & & & & & & \\
 & \vdots & \vdots & & & & \vdots \\
 & -p_{r1}+a_1 & -p_{r2}+a_2 & \cdots & & \cdots & p_{r1}+\cdots+p_{r\,r-1}+a_r
\end{array}$$

Each column of the determinant in (4.37) is a sum of two column vectors, one of which is of the form $a_i\xi$. Hence the determinant can be written as a sum of 2^{r-1} determinants, each of which is formed from $(I-P)$ by deleting the k-th row and the k-th column and replacing some columns with vectors of the form $a_i\xi$. Of such determinants those which have two or more columns of the form $a_i\xi$, are equal to zero. Hence the determinant in (4.37) can be written as a sum of r determinants $\overline{P}_k$ and $F_k(j)$ $(j=1,2,\cdots,k-1,k+1,\cdots,r)$, where $F_k(j)$ is the determinant of a matrix formed from $(I-P)$ by deleting the k-th row and the k-th column and replacing the j-th column with $a_j\xi$. If we expand $F_k(j)$ by the j-th column, then

$$(4.38) \qquad F_k(j) = \sum_{\substack{i=1 \\ i\neq k}}^{r} a_j\,\overline{P}_k(i,j) .$$

Hence we have

$$(4.39) \qquad z_{kk} = \frac{1}{U}\left[\overline{P}_k + \sum_{\substack{j=1 \\ j\neq k}}^{r}\sum_{\substack{i=1 \\ i\neq k}}^{r} a_j\,\overline{P}_k(i,j)\right] ,$$

and using Lemma 1, we can obtain the expression (4.34).

We can represent a_k, z_{kk}, m_{jk} and w_{kk} as quotients of sums of products of transition probabilities with plus signs or as sums of such quotients as follows.

By Lemma 1,

$$(4.40) \qquad a_k = \frac{\overline{P}_k}{U} .$$

By Lemma 14,

$$z_{kk} = \frac{\overline{P}_k}{U} + \frac{1}{U^2} \sum_{\substack{j=1 \\ j \neq k}}^{r} \sum_{\substack{i=1 \\ i \neq k}}^{r} \overline{P}_j \, \overline{P}_k(i,j) \; . \tag{4.41}$$

Since $m_{kk} = 1/a_k$,

$$m_{kk} = \frac{U}{\overline{P}_k} \; . \tag{4.42}$$

By changing the notations in (4.7),

$$m_{jk} = \frac{1}{\overline{P}_k} \sum_{\substack{i=1 \\ i \neq k}}^{r} \overline{P}_k(i,j) \quad , (j \neq k) \; . \tag{4.43}$$

Since $w_{kk} = \dfrac{2z_{kk}}{a_k^2} - \dfrac{1}{a_k}$,

$$w_{kk} = \frac{U}{\overline{P}_k} + \frac{2}{\overline{P}_k^{\,2}} \sum_{\substack{j=1 \\ j=k}}^{r} \sum_{\substack{i=1 \\ i=k}}^{r} \overline{P}_j \, \overline{P}_k(i,j) \; . \tag{4.44}$$

4.4 Bounds for basic quantities for a regular Markov chain M'

We consider two regular Markov chains M and M'. We use the notations defined in Section 4.3 for quantities of the chain M , and use the same notations with primes for corresponding quantities of the chain M'. We shall obtain simple bounds for basic quantities for the chain M' under the assumption (2.1).

Our main tools are Lemma 4 and the following

Lemma 15. If (2.1) holds, we have

$$(1+\epsilon)^{-r+2}\,\bar{P}_k(i,j) \leqq \bar{P}'_k(i,j) \leqq (1+\epsilon)^{r-2}\,\bar{P}_k(i,j) \;. \tag{4.45}$$

Proof. This is a direct corollary of Lemma 13.

From (4.40)~(4.44), we can obtain the bounds for the quantities using Lemmas 4 and 15 as follows. Here we show only the upper bounds under the assumption (2.1), because the lower bounds can be obtained by replacing $(1+\epsilon)$ in the upper bounds with $(1+\epsilon)^{-1}$.

$$a'_k \leqq (1+\epsilon)^{2r-2}\,a_k \tag{4.46}$$

$$z'_{kk} \leqq (1+\epsilon)^{4r-5}\,z_{kk} \tag{4.47}$$

$$m'_{kk} \leqq (1+\epsilon)^{2r-2}\,m_{kk} \tag{4.48}$$

$$m'_{jk} \leqq (1+\epsilon)^{2r-3}\,m_{jk} \qquad (j \neq k) \tag{4.49}$$

$$w'_{kk} \leqq (1+\epsilon)^{4r-5}\,w_{kk} \tag{4.50}$$

Some of these bounds can be improved as Theorem 1 in Section 2.2. However we omit detailed discussions here.

5. A Bound for the Variance of a'_k

In Section 2.2, we got a simple bound for a'_k. When we may consider that p'_{ij}'s are mutually independent random variables, we can also obtain a simple bound for the variance of a'_k. Our bound is the best one of those which use the information on a_k only, because it is nearly attained by the same Markov chains as in Example 1.

We consider the situation where p'_{ij} $(i,j=1,2,\cdots,r\ ;\ i\neq j)$ are mutually independent random variables distributed about p_{ij}. We assume that p'_{ij} satisfies the following three conditions.

(5.1) $$E\left\{\epsilon_{ij}\right\} = 0, \qquad \text{i.e.,} \quad E\left[p'_{ij}\right] = p_{ij}$$

(5.2) $$\mathrm{Var}\left\{\epsilon_{ij}\right\} = \sigma_{ij}^2 \leqq \sigma^2, \qquad \text{i.e.,} \quad \mathrm{Var}\left[p'_{ij}\right] = p_{ij}^2\sigma_{ij}^2 \leqq \sigma^2 p_{ij}^2$$

(5.3) $$E\left\{\left|\epsilon_{ij}^3\right|\right\} = O(\sigma^3), \qquad \text{i.e.,} \quad E\left[\left|p'_{ij} - p_{ij}\right|^3\right] = O(\sigma^3)$$

<u>Theorem 5</u>. If p'_{ij} $(i,j=1,2,\cdots,r\ ;\ i\neq j)$ are mutually independent random variables satisfying the three conditions (5.1), (5.2) and (5.3), then

(5.4) $$E\left\{a'_k\right\} = a_k + O(\sigma^2)$$

and

(5.5) $$\mathrm{Var}\left\{a'_k\right\} = a_k^2 \sum_{i=1}^{r}\sum_{\substack{j=1\\ j\neq i}}^{r}\left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U}\right)^2\sigma_{ij}^2 + O(\sigma^3).$$

<u>Proof</u>. (5.4) is an obvious consequence of Theorem3 with a slight change in the residual term, and taking the expectation of the square of (2.24), (5.5) can be easily proved.

Theorem 6. If p'_{ij} $(i,j = 1,2,\cdots,r\ ;\ i \neq j)$ are mutually independent random variables and if the three conditions (5.1), (5.2) and (5.3) are satisfied, then

$$\text{Var}\left\{ a'_k \right\} \leqq 2(r-1)\, a_k^2 (1-a_k)^2 \sigma^2 + O(\sigma^3). \tag{5.6}$$

Proof. By the preceding theorem, we have

$$\text{Var}\left\{ a'_k \right\} \leqq \sigma^2 a_k^2 \sum_{i=1}^{r} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{r} \left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U} \right)^2 \right\} + O(\sigma^3). \tag{5.7}$$

We can evaluate the sums in the braces in (5.7) as follows. If $i \neq k$, then by Lemma 7, we have

$$\begin{aligned}
\sum_{j\neq i} \left(\frac{\overline{P}_k^{ij}}{\overline{P}_k} - \frac{U^{ij}}{U} \right)^2 &= \frac{1}{(\overline{P}_k U)^2} \sum_{j\neq i} \left\{ \overline{P}_k^{ij} (U - \overline{P}_k) - \overline{P}_k (U^{ij} - \overline{P}_k^{ij}) \right\}^2 \\
&\leqq \frac{1}{(\overline{P}_k U)^2} \left[\sum_{j\neq i} \left\{ \overline{P}^{ij} (U - \overline{P}_k) \right\}^2 + \sum_{j\neq i} \left\{ \overline{P}_k (U^{ij} - \overline{P}_k^{ij}) \right\}^2 \right] \\
&\leqq \frac{1}{(\overline{P}_k U)^2} \left[\left\{ \sum_{j\neq i} \overline{P}_k^{ij} (U - \overline{P}_k) \right\}^2 + \left\{ \sum_{j\neq i} \overline{P}_k (U^{ij} - \overline{P}_k^{ij}) \right\}^2 \right] \\
&= \frac{1}{(\overline{P}_k U)^2} \left[\left\{ \overline{P}_k (U - \overline{P}_k) \right\}^2 + \left\{ \overline{P}_k (U - \overline{P}_i - \overline{P}_k) \right\}^2 \right] \\
&= (1-a_k)^2 + (1-a_i-a_k)^2 \\
&= 2(1-a_k)^2 - 2(1-a_k)a_i + a_i^2 .
\end{aligned} \tag{5.8}$$

If $i = k$, then

$$(5.9) \qquad \sum_{j \neq k} \left(- \frac{U^{kj}}{U} \right)^2 \leq \frac{1}{U^2} \left(\sum_{j \neq k} U^{kj} \right)^2 = (1 - a_k)^2 .$$

Hence (5.7) becomes

$$(5.10) \qquad \mathrm{Var}\left\{ a_k' \right\} \leq \sigma^2 a_k^2 \left[\sum_{\substack{i=1 \\ i \neq k}}^{r} \left\{ 2(1-a_k)^2 - 2(1-a_k)a_i + a_i^2 \right\} + (1-a_k)^2 \right] + O(\sigma^3)$$

$$\leq \sigma^2 a_k^2 \left[(2r-1)(1-a_k)^2 - 2(1-a_k) \sum_{i \neq k} a_i + \left(\sum_{i \neq k} a_i \right)^2 \right] + O(\sigma^3)$$

$$= 2(r-1)\sigma^2 a_k^2 (1-a_k)^2 + O(\sigma^3) .$$

For the case where P' differs from P in one row, we have the following theorem. Its proof is essentially contained in the preceding proof.

Theorem 7. If p'_{hj} $(j = 1,2, \cdots ,h-1,h+1, \cdots ,r)$ are mutually independent random variables and $p'_{ij} = p_{ij}$ $(i,j = 1,2, \cdots ,r \; ; \; i \neq h)$, and if the conditions (5.1), (5.2) and (5.3) are satisfied, then when $k \neq h$,

$$(5.11) \qquad \mathrm{Var}\left\{ a_k' \right\} \leq a_k^2 \left\{ (1-a_k)^2 + (1-a_h-a_k)^2 \right\} \sigma^2 + O(\sigma^3) ,$$

and when $k = h$

$$(5.12) \qquad \mathrm{Var}\left\{ a_h' \right\} \leq a_h^2 (1-a_h)^2 \sigma^2 + O(\sigma^3) .$$

Example 4. The bound in Theorem 6 is the best one of those which use the information of a_k only. In order to show this fact, we can use the Markov chain in Example 1 again. By the approximation (2.33), we have

$$\text{(5.13)} \qquad \operatorname{Var}\{a_5'\} = E\{(a_5' - a_5)^2\}$$

$$\approx a_5^2 (1-a_5)^2 \{ \sigma_{12}^2 + \sigma_{23}^2 + \sigma_{34}^2 + \sigma_{45}^2 + \sigma_{21}^2 + \sigma_{31}^2 + \sigma_{41}^2 + \sigma_{51}^2 \}.$$

If $\sigma_{12}^2 = \sigma_{23}^2 = \cdots = \sigma_{51}^2 = \sigma^2$, then it follows that

$$\text{(5.14)} \qquad \operatorname{Var}\{a_5'\} \approx 8\, a_5^2 (1-a_5)^2 \sigma^2 ,$$

and this coincides with the bound in Theorem 6 for $r = 5$.

Example 5. We again consider the extreme case where all p_{ij} $(i,j = 1,2,\cdots,r\ ;\ i \neq j)$ are equal to p (some constant). In this case, by Theorems 4 and 5, we have

$$\text{(5.15)} \qquad \operatorname{Var}\{a_k'\} = a_k^2 \frac{1}{r^2} \Big\{ \sum_{i \neq k} \sigma_{ik}^2 + \sum_{j \neq k} \sigma_{kj}^2 \Big\} + O(\sigma^3) .$$

If $\sigma_{ik}^2 = \sigma_{kj}^2 = \sigma^2$ for all i $(\neq k)$ and j $(\neq k)$, then it follows that

$$\text{(5.16)} \qquad \operatorname{Var}\{a_k'\} = 2(r-1) \frac{1}{r^4} \sigma^2 + O(\sigma^3) .$$

6. Bounds for the Variances of Basic Quantities

The method for obtaining a simple bound for the variance of a'_k can be applied to some basic quantities for the Markov chain M' defined in Sections 4.1 and 4.3. We assume that p'_{ij} $(i,j = 1,2,\cdots,r\ ;\ i = j)$ are mutually independent random variables satisfying the three conditions (5.1), (5.2) and (5.3). Then the means of the quantities coincide with the corresponding quantities for the chain M. Here we shall show simple bounds for the variances of the quantities. Since the procedures for obtaining the bounds are essentially the same as in Section 5, we omit the proofs for them.

6.1 Bounds for the variances of basic quantities for an absorbing Markov chain M'

$$\mathrm{Var}\left\{ M'_i[n'_j] \right\} \leq (2s-1)\,\sigma^2 \left\{ M_i[n_j] \right\}^2 \tag{6.1}$$

$$\mathrm{Var}\left\{ M'_i[t'] \right\} \leq (2s-1)\,\sigma^2 \left\{ M_i[t] \right\}^2 \tag{6.2}$$

$$\mathrm{Var}\left\{ M'_i[m'] \right\} \leq 2(s-1)\,\sigma^2 \left\{ M_i[m] \right\}^2 \tag{6.3}$$

$$\mathrm{Var}\left\{ b_{ij} \right\} \leq 2s\,\sigma^2 \left\{ b_{ij} \right\}^2 \tag{6.4}$$

$$\mathrm{Var}\left\{ h_{ij} \right\} \leq 2(s-1)\,\sigma^2 \left\{ h_{ij} \right\}^2 \qquad (i \neq j) \tag{6.5}$$

For the cases of $\mathrm{Var}'_i[n'_j]$ and h_{ii}, we cannot get bounds by this method. Because, the representations (4.6) and (4.11) contain $(\overline{Q}(j,j) - |\overline{Q}|)$ and $(\overline{Q}(i,i) - |\overline{Q}|)$. By Lemma 11, $(\overline{Q}(i,i) - |\overline{Q}|)$ contains not only p_{ij} $(j = 1,2,\cdots,r\ ;\ j \neq i)$ but also p_{ii}. Since we cannot assume that all p'_{ij} $(j = 1,2,\cdots,r)$ are mutually independent random variables, we cannot apply our method for these cases.

6.2 Bounds for the variances of basic quantities for a regular Markov chain M'

(6.6) $$\operatorname{Var}\{a_k'\} \leqq 2(r-1)\,\sigma^2\{a_k\}^2$$

(6.7) $$\operatorname{Var}\{z_{kk}'\} \leqq 8(r-1)\,\sigma^2\{z_{kk}\}^2$$

(6.8) $$\operatorname{Var}\{m_{kk}'\} \leqq 2(r-1)\,\sigma^2\{m_{kk}\}^2$$

(6.9) $$\operatorname{Var}\{m_{jk}'\} \leqq (2r-3)\,\sigma^2\{m_{jk}\}^2 \qquad (j \neq k)$$

(6.10) $$\operatorname{Var}\{w_{kk}'\} \leqq 8(r-1)\,\sigma^2\{w_{kk}\}^2$$

Acknowledgement. The author wishes to thank Proffessor H. Morimura for his guidance and encouragement.

References

[1] Kemeny, J.G. & Snell, J.L., Finite Markov Chains. Van Nostrand Comp., New York (1960).

[2] Schweitzer, P.J., "Perturbation theory and finite Markov chains", J. Appl. Prob. 5 (1968), 401-413.

[3] Smith, J.L., "Approximate stationary probability vectors of a finite Markov chain", SIAM J. Appl. Math. 20 (1971), 612-618.

CHAPTER 4 A LUMPING METHOD FOR NUMERICAL CALCULATIONS OF STATIONARY DISTRIBUTIONS OF MARKOV CHAINS

1. Introduction

The Markov chain technique is one of the most powerful tools for analyses of various stochastic models such as random walks, queueing models, inventory models, stochastic models in the reliability theory and so on. In such applications, we often have to calculate the stationary distributions of Markov chains. Numerical values of them can be obtained by ordinary numerical calculation methods for systems of equations such as the Gaussian elimination method, the Gauss-Seidel iterative method and so on. However when we deal with Markov chains with special structures, we sometimes have more suitable methods of numerical calculations.

In applications of Markov chains, we have often to treat Markov chains with considerably many states and sometimes with infinitely many states. Numerical calculations of stationary distributions of such Markov chains will require many computing efforts. For example, if we calculate the stationary distribution of a Markov chain of a certain type (see the note at the end of this section) by the Jacobi iterative method, the number of iterations required is approximately proportional to the square of the number of states. So, for Markov chains with numerous states we should develope more suitable numerical methods which skillfully use the special structures of the chains.

In this paper, a lumping method is proposed. It uses the fact that we can construct a Markov chain which has the same stationary distribution as the lumped process of of the Markov chain under consideration. Using this idea, we can split the original system of equations into several subsystems of equations, and we solve these subsystems iteratively improving the parameters.

This method can be used for any finite Markov chains, and it is especially effective for Markov chains of the following type. The state space of a Markov chain is divided into several sets in a natural way. If once the chain enters in a set, it stays in the set in a long time. So the chainrarely moves from one set to another.

Note. We consider a Markov chain with s states having the following transition matrix.

$$P = \begin{pmatrix} 1-p-q & p & 0 & 0 & \cdots & 0 & q \\ q & 1-p-q & p & 0 & \cdots & 0 & 0 \\ 0 & q & 1-p-q & p & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-p-q & p \\ p & 0 & 0 & 0 & \cdots & q & 1-p-q \end{pmatrix}$$

This matrix has eigenvalues

$$\lambda_k = (1-p-q) + (p+q)\cos\frac{2k\pi}{s} + (p-q)\,i\,\sin\frac{2k\pi}{s}$$

$$(k = 0, 1, 2, \cdots, s-1).$$

If we calculate the stationary distribution α of the chain iteratively by the formula

$$\alpha(n+1) = \alpha(n)\,P\,,$$

then the rate of convergence is mainly governed by λ_1 which has the maximal absolute value except for $\lambda_0 = 1$. Hence the number of iterations required is approximately proportional to

$$-1/\log|\lambda_1| \approx \{(1-p-q)(p+q) + 8pq\}^{-1} \cdot \left(\frac{s}{2\pi}\right)^2$$

for large s.

2. Fundamental Theorem

We consider a regular Markov chain with the state space $S = \{1, 2, \cdots, s\}$. We let its transition matrix be $P = (p_{ij})$ and its stationary distribution, or the limiting vector of P, be $\alpha = (a_i)$. Now we assume that the state space S is divided into r sets by a partition $A = \{A_1, A_2, \cdots, A_r\}$, where $A_1 \cup A_2 \cup \cdots \cup A_r = S$ and $A_k \cap A_l = \emptyset$ for $k \neq l$. Without loss of generality, we may assume that

$$(2.1) \qquad \begin{aligned} A_1 &= \{1, 2, \cdots, t_1\} \\ A_2 &= \{t_1+1, t_1+2, \cdots, t_2\} \\ &\vdots \\ A_r &= \{t_{r-1}+1, t_{r-1}+2, \cdots, s\}. \end{aligned}$$

We denote the number of states in the set A_k by $s_k = t_k - t_{k-1}$. Then the transition matrix P can be divided into r^2 submatrices.

$$(2.2) \qquad P = \begin{pmatrix} P^{11} & P^{12} & \cdots & P^{1r} \\ P^{21} & P^{22} & \cdots & P^{2r} \\ \vdots & \vdots & & \vdots \\ P^{r1} & P^{r2} & \cdots & P^{rr} \end{pmatrix}$$

where $P^{kl} = (p_{ij} \,;\, i \in A_k, \, j \in A_l)$ is an $s_k \times s_l$ matrix which represents the probabilities of transitions from states in A_k to states in A_l. We can also divide the limiting vector α into r subvectors.

(2.3) $\qquad \alpha = (\alpha^1, \alpha^2, \cdots, \alpha^r)$

where

(2.4) $\qquad \alpha^k = (a_i \ ; i \in A_k) = (a_{t_{k-1}+1} , a_{t_{k-1}+2} , \cdots , a_{t_k}).$

Now we prepare some notations. We denote by ρ^k the s_k-th order column vector with all entries 1 , and by V the $s \times r$ matrix whose k-th column is a vector with 1's in the components corresponding to states in A_k and 0's otherwise, that is,

(2.5) $$V = \begin{pmatrix} \rho^1 & & & 0 \\ & \rho^2 & & \\ & & \ddots & \\ 0 & & & \rho^r \end{pmatrix}.$$

We let $\beta = (b_1, b_2, \cdots , b_r) = \alpha V$, then we have

(2.6) $$b_k = \alpha^k \rho^k = \sum_{i \in A_k} a_i ,$$

namely, b_k is the total weight on the set A_k of the limiting vector α . We denote by $\gamma^k = (g_i^k \ ; i \in A_k)$ the restriction of α on the set A_k , that is, γ^k is the s_k-th order stochastic row vector with components defined by

(2.7) $$g_i^k = a_i / b_k \qquad (i \in A_k).$$

We let U be the $r \times s$ matrix whose k-th row is a vector with g_i^k in the component corresponding to state i $(\in A_k)$ and 0's otherwise, that is,

$$(2.8) \qquad U = \begin{pmatrix} \gamma^1 & & & 0 \\ & \gamma^2 & & \\ & & \ddots & \\ 0 & & & \gamma^r \end{pmatrix} .$$

Then we can easily prove that $\beta U = \alpha$. Finally we let $Q = (q_{kl} ; k,l = 1,2, \cdots ,r) = UPV$. Then we have

$$(2.9) \qquad q_{kl} = \gamma^k P^{kl} \rho^l = \frac{1}{b_k} \sum_{i \in A_k} \sum_{j \in A_l} a_i p_{ij} .$$

Now we prove the fundamental theorem.

<u>Theorem 1</u>. Q is a stochastic matrix and has the limiting vector β.

<u>Proof</u>. Since all row sums of U, P and V are equal to 1, the row sums of Q are also equal to 1, and since all entries of U, P and V are nonnegative, all entries of Q are also nonnegative. Hence Q is a stochastic matrix. By the equations $\beta U = \alpha$, $\alpha P = \alpha$ and $\alpha V = \beta$, we have $\beta Q = \beta$. Thus Q has the limiting vector β.

This theorem can be interpreted in terms of stochastic processes as follows. We denote the Markov chain under consideration by $\{ \xi_n \,;\, n=0,1,2,\cdots \}$. We consider the lumped process $\{ \varepsilon_n \}$ of ξ_n with the partition $A = \{ A_1, A_2, \cdots, A_r \}$, that is, we consider the process $\{ \varepsilon_n \}$ defined by the following relation:

$$(2.9) \qquad \varepsilon_n = A_k \quad \text{if and only if} \quad \xi_n \in A_k .$$

This lumped process $\{ \varepsilon_n \}$ has the stationary distribution β. On the other hand, we can consider a Markov chain on A with the transition matrix $Q = UPV$. The above theorem states that this Markov chain has also the stationary distribution β.

Hence if we know all γ's, then in order to obtain α, we may consider a Markov chain with the transition matrix Q instead of the original chain. Thus we can solve the problem for obtaining α in three steps.

1. Calculate vectors γ^k $(k=1,2,\cdots,r)$.
2. Calculate β from $Q = UPV$.
3. Obtain α by the relation $\alpha = \beta U$.

Generally we cannot calculate γ^k alone, and hence we cannot obtain α by adopting this procedure directly. However this procedure suggests us an iterative method for numerical calculations of stationary distributions of finite Markov chains.

3. A Lumping Method

In the last section we saw that if we know γ's , then we can obtain β and α from the stochastic matrix Q . Similarly we can show that if we know β and γ's other than γ^k , then we can obtain γ^k from an $s_k \times s_k$ stochastic matrix Q^k. In order to show this fact, we shall consider a partition $B_k = \{ t_{k-1}+1 , t_{k-1}+2 , \cdots$ $\cdots , t_k , S-A_k \}$, i.e., B_k consists of states in A_k and the set $S-A_k$ of all other states. We can prove an anaogous theorem to Theorem 1 for this partition B_k . In this case matrices corresponding to V and U are

$$(3.1)\qquad V^k = \begin{array}{cl} \begin{array}{cc} \quad A_k \quad & S-A_k \end{array} & \\ \left(\begin{array}{ccc|c} 0 & \cdots & 0 & 1\\ \vdots & & \vdots & \vdots \\ 0&\cdots&0&1\\ \hline 1 & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & 1 & 0 \\ \hline 0&\cdots&0&1\\ \vdots&&\vdots&\vdots\\ 0&\cdots&0&1 \end{array}\right) & \begin{array}{l} \\ A_1\cup\cdots\cup A_{k-1}\\ \\ \\ A_k \\ \\ \\ A_{k+1}\cup\cdots\cup A_r \\ \\ \end{array}\end{array}$$

and

$$(3.2) \qquad U^k = \begin{array}{c} \qquad A_1 \cup \cdots \cup A_{k-1} \qquad\qquad A_k \qquad\qquad A_{k+1} \cup \cdots \cup A_r \\ \left(\begin{array}{ccc|ccc|ccc} 0 & \cdots & 0 & 1 & & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & & 1 & 0 & \cdots & 0 \\ \hline u_1^k & \cdots & u_{t_{k-1}}^k & 0 & \cdots & 0 & u_{t_k+1}^k & \cdots & u_s^k \end{array}\right) \begin{array}{l} \\ A_k \\ \\ S-A_k \end{array} \end{array}$$

where

$$(3.3) \qquad u_i^k = \frac{a_i}{1-b_k} \qquad (i \in S - A_k)$$

$$= \frac{b^l g_i^l}{1-b_k} \qquad (i \in A_l \ ; \ l \neq k).$$

Then Q^k becomes as

$$(3.4) \qquad Q^k = \begin{pmatrix} P^{kk} & \zeta^k \\ \phi^k & v^k \end{pmatrix}$$

where P^{kk} is the submatrix of P corresponding to $A_k \times A_k$, $\zeta^k = (r_i^k \ ; \ i \in A_k)$ is an s_k-th order column vector whose component r_i^k is given by

$$(3.5) \qquad r_i^k = \sum_{j \in S-A_k} p_{ij} = 1 - \sum_{j \in A_k} p_{ij} \qquad (i \in A_k),$$

$\phi^k = (q_j^k \ ; \ j \in A_k)$ is an s_k-th order row vector whose component q_j^k is given by

$$(3.6) \qquad q_j^k = \sum_{i \in S-A_k} u_i^k \, p_{ij}$$

$$= \frac{1}{1 - b_k} \sum_{\substack{l=1 \\ l \neq k}}^{r} \sum_{i \in A_l} b_l \, g_i^l \, p_{ij} \qquad (j \in A_k) ,$$

and finally

$$(3.7) \qquad v^k = 1 - \sum_{j \in A_k} q_j^k .$$

Then by the analogous theorem to Theorem 1, Q^k has the limiting vector

$$(3.8) \qquad \eta^k = (\alpha^k, 1 - b_k) = (a_{t_{k-1}+1} , \cdots , a_{t_k} , 1 - b_k) .$$

Hence we can get γ^k by restricting η^k on the set A^k .

Thus we have seen that we can calculate each one of β and γ 's from the knowledge of others. This fact suggests us an iterative method for the numerical calculation of α .

A LUMPING METHOD

We start with appropriate initial values $\gamma^1(0), \cdots, \gamma^r(0)$, and proceed along the following iteration scheme:

0a. Calculate $Q(n+1)$ from $\gamma^1(n), \gamma^2(n), \cdots, \gamma^r(n)$.

b. Calculate $\beta(n+1)$ from $Q(n+1)$.

1a. Calculate $Q^1(n+1)$ from $\beta(n+1), \gamma^2(n), \cdots, \gamma^r(n)$.

b. Calculate $\gamma^1(n+1)$ from $Q^1(n+1)$.

$\vdots$

ka. Calculate $Q^k(n+1)$ from $\beta(n+1), \gamma^1(n+1), \cdots, \gamma^{k-1}(n+1), \gamma^{k+1}(n), \cdots, \gamma^r(n)$.

b. Calculate $\gamma^k(n+1)$ from $Q^k(n+1)$.

$\vdots$

ra. Calculate $Q^r(n+1)$ from $\beta(n+1), \gamma^1(n+1), \cdots, \gamma^{r-1}(n+1)$.

b. Calculate $\gamma^r(n+1)$ from $Q^r(n+1)$, and return to 0a replacing n with $n+1$.

We may use any convergence criterion. A criterion is

$$\left|b_k(n+1) - b_k(n)\right| < \varepsilon \quad \text{and} \quad \left|g_i^k(n+1) - g_i^k(n)\right| < \varepsilon \tag{3.9}$$

for every $k = 1, 2, \cdots, r$ and $i \in A_k$.

4. Some Remarks for the Lumping Method

The lumping method defined in the last section proposes the iterative calculation of and 's along the iteration sheme. But it does not provide the method for obtaining and 's from the stochastic matrices Q's . Hence we have many variations of the lumping method. In this section we state some remarks which will be useful in practices. The following theorem gives a base for modifications of the lumping method.

Theorem 2. Let $Q = (q_{ij})$ be a $t \times t$ regular stochastic matrix and $\xi = (x_i)$ be its limiting vector. For positive constants $c_1, c_2, \cdots, c_t$, define a matrix $\tilde{Q} = (\tilde{q}_{ij})$ by

$$(4.1) \qquad \tilde{q}_{ij} = c_i \, q_{ij} \qquad (i,j = 1,2,\cdots,t \ ; \ i \neq j)$$

and

$$(4.2) \qquad \tilde{q}_{ii} = c_i \, q_{ii} + 1 - c_i \qquad (i = 1,2,\cdots,t),$$

and define the vector $\tilde{\xi} = (\tilde{x}_i)$ by

$$(4.3) \qquad \tilde{x}_i = c \, x_i / c_i \qquad (i = 1,2,\cdots,t)$$

where

$$(4.4) \qquad c = (\sum_{i=1}^{t} x_i / c_i)^{-1}.$$

Then $\tilde{\xi} \tilde{Q} = \tilde{\xi}$, and if $\tilde{Q}$ is a stochastic matrix, then $\tilde{\xi}$ is its limiting vector.

Proof. This theorem can be easily proved by a direct calculation of $\tilde{\xi}\tilde{Q}$.

Remark 1. Instead of Q^k , we may use the stochastic matrix

$$(4.5) \qquad \tilde{Q}^k = \begin{pmatrix} P^{kk} & \zeta^k \\ \tilde{\phi}^k & \tilde{v}^k \end{pmatrix} .$$

where $\tilde{\phi}^k = (1 - b_k)\phi^k$ and $\tilde{v}^k = (1 - b_k) v^k + b_k$. By Theorem 2 we can easily show that Q^k and $\tilde{Q}^k$ bring the same γ^k . This modification simplifies the calculations of the elements in the last row of Q^k , namely, the element $\tilde{q}_j^k$ of $\tilde{\phi}^k$ is given by

$$(4.6) \qquad \tilde{q}_j^k = \sum_{\substack{l=1 \\ l \neq k}}^{r} \sum_{i \in A_l} b_l \, g_i^l \, p_{ij} \qquad (j \in A_k),$$

and need not be divided by $(1 - b_k)$.

Remark 2. Instead of Q^k , we may use the stochastic matrix

$$(4.7) \qquad \bar{Q}^k = \begin{pmatrix} P^{kk} & \zeta^k \\ \bar{\phi}^k & \bar{v}^k \end{pmatrix}$$

where $\bar{\phi}^k = c\,\phi^k$ and $\bar{v}^k = c\, v^k + 1 - c$ for a constant $c \leq 1/(1 - \tilde{v}^k)$. We can also show that Q^k and $\bar{Q}^k$ bring the same γ^k by Theorem 2. This modification is effective if we use an iterative method for obtaining η^k from Q^k and if q_j^k 's are so small that the rate of convergence is very slow.

Remark 3. Instead of Q^k, we may use the stochastic matrix $\bar{\bar{Q}}^k$ obtained from $\bar{Q}^k$ by multiplying off-diagonal elements with some positive constant and modifying diagonal elements so that the row sums are 1. This modification does not change the value of γ^k too. This modification is effective if we use an iterative method for obtaining η^k from Q^k and if the rate of convergence is very slow. This is one of the ordinary acceleration techniques.

Remark 4. Suppose that the chain cannot reach any state in the set A_k from each state out of A_k without passing through the state j in A_k. In such a case, we can calculate γ^k only from P^{kk} and need not know the values of β and other γ's. Because, the last row of Q^k has only two non-zero entries q_j^k and $v^k = 1 - q_j^k$. Hence we can modify Q^k by Remark 2 so that $q_j^k = 1$ and all other entries of the last row are zero. This modified $\bar{Q}^k$ can be determined by P^{kk} only, and hence we can obtain γ^k without knowledges of β and other γ's. Thus if once we calculate γ^k, then we need not calculate γ^k in the successive iterations.

This situation can also be interpreted as follows. We consider the partition $A' = \{1, 2, \cdots, t_{k-1}, t_k + 1, \cdots, s, A_k\}$. As pointed at the end of Section 2, if we know γ^k, then we are sufficient to calculate the stationary distribution β' of Q' corresponding to the partition A'. This means that the original problem can be reduced to the problem of finding the stationary distribution of a Markov chain with the transition matrix Q'. Thus if we find a set A_k with the property mentioned above, then we can decrease the number of states and save plenty of computing time.

<u>Remark 5</u>. When we calculate $\beta(n+1)$ and $\eta^k(n+1)$ by an iterative method, we need not to get so accurate values since $Q(n+1)$ and $Q^k(n+1)$ are only approximate values of Q and Q^k. Hence we need not make to many iterations for obtaining the values of $\beta(n+1)$ and $\eta^k(n+1)$. In Example 3 (stated in the next section) the numbers of iterations in each step are 6 or 8.

<u>Remark 6</u>. When we calculate $\beta(n+1)$ and $\eta^k(n+1)$ by an iterative method, we may use $\beta(n)$ and $\eta^k(n)$ as initial values. However generally it is more effective to use the values of β and η^k calculated from the latest approximate values of β and γ's as initial values.

5. Examples

In this section we deal with two simple examples and one application to a practical problem. Example 1 is an example of the usage of the modification stated in Remark 4. Example 2 is an example of the ordinary use of the lumping method with the Gauss-Seidel iterative method. Example 3 is an application of the method to an analysis of the effect of the speed class sequencing in the air traffic control. In the example the numbers of multiplications and divisions required for obtaining the stationary distribution of a Markov chain are compared for two method, the lumping method and the ordinary Gauss-Seidel iterative method.

<u>Example 1</u>. We consider a Markov chain with the transition matrix

$$
(5.1) \qquad P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\
\left[\begin{array}{ccc|ccc|ccc}
 & .2 & .2 & .4 & & & .2 & & \\
.2 & & .2 & & & & .6 & & \\
 & .3 & .1 & .2 & & & .4 & & \\
\hline
.3 & & & & .3 & .4 & & & \\
.2 & & & .3 & .4 & .1 & & & \\
 & & & .2 & .6 & & .2 & & \\
\hline
 & & .5 & & & & & .2 & .3 \\
 & & & .2 & & & & .2 & .6 \\
 & & & .3 & & & .1 & .2 & .4
\end{array}\right]
\end{array}
$$

where entries with 0's are represented by blanks. The chain can enter the set $A_4 = (4,5,6)$ only through the state 4, and can enter the set $A_7 = (7,8,9)$ only through the state 7. So when we calculate the stationary distribution α, it is effective to use the lumping method with the partition $A = \{1,2,3,A_4,A_7\}$. As stated in Remark 4, the weight distribution γ^4 and γ^7 of α in the sets A_4 and A_7 can be calculated easily.

We shall calculate γ^4. We first form the matrix Q^4.

$$(5.2) \qquad Q^4 = \begin{pmatrix} 0 & .3 & .4 & .3 \\ .3 & .4 & .1 & .2 \\ .2 & .6 & 0 & .2 \\ 1. & 0 & 0 & 0 \end{pmatrix}$$

The limiting vector $\eta^4 = (y_4^4, y_5^4, y_6^4, y^4)$ of Q^4 can be obtained by the Cramer's rule. So y_j^4 $(j = 4,5,6)$ is proportional to h_j^4 which is the determinant of the matrix formed from $(I - Q^4)$ by deleting the row and the column corresponding to the state j. Hence the entry g_j^4 of γ^4 corresponding to j is given by

$$(5.3) \qquad g_j^4 = h_j^4 / (h_4^4 + h_5^4 + h_6^4) \qquad (j = 4,5,6).$$

Here we have

$$(5.4) \qquad h_4^4 = .54 \ , \quad h_5^4 = .54 \ , \quad h_6^4 = .27 \ .$$

It follows that

(5.5) $\gamma^4 = (.4 , .4 , .2)$.

Similarly we have

(5.6) $\gamma^7 = (.4 , .2 , .4)$.

Then the problem is reduced to the problem of calculating the stationary distribution of a Markov chain with the transition matrix

(5.7)
$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ A_4 \\ A_7 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & A_4 & A_7 \end{array} \\ \begin{pmatrix} 0 & .2 & .2 & .4 & .2 \\ .2 & 0 & .2 & 0 & .6 \\ 0 & .3 & .1 & .2 & .4 \\ .2 & 0 & 0 & .76 & .04 \\ 0 & 0 & .2 & .16 & .64 \end{pmatrix} \end{array} .$$

This chain has the stationary distribution

(5.8) $\beta = (.1 , .05 , .1 , .45 , .3)$.

Hence we obtain the stationary distribution α of the original chain as

(5.9) $\alpha = (.10 , .05 , .10 , .18 , .18 , .09 , .12 , .06 , .12)$.

Example 2. We shall calculate the stationary distribution α of a Markov chain with the transition matrix

$$(5.10) \qquad P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \left[\begin{array}{ccc|ccc|ccc} .8 & .1 & & & & & & & .1 \\ .1 & .7 & .2 & & & & & & \\ & .1 & .6 & .3 & & & & & \\ \hline & & .1 & .5 & .4 & & & & \\ & & & .1 & .4 & .5 & & & \\ & & & & .1 & .3 & .6 & & \\ \hline & & & & & .1 & .2 & .7 & \\ & & & & & & .1 & .1 & .8 \\ .9 & & & & & & & .1 & 0 \end{array}\right] \end{array}$$

We will use the lumping method with the partition $A = \{(1,2,3), (4,5,6), (7,8,9)\}$ and calculate $\beta(n+1)$ and $\eta^k(n+1)$ by the Gauss-Seidel iterative method and adopt the modifications stated in Remarks 5 and 6. We start with the initial values

$$(5.11) \qquad \gamma^1(0) = (1/3,\ 1/3,\ 1/3)\ , \quad \gamma^2(0) = (1/3,\ 1/3,\ 1/3)$$
$$\text{and} \quad \gamma^3(0) = (1/3,\ 1/3,\ 1/3).$$

Then the matrix $Q(1)$ becomes

$$(5.12) \qquad Q(1) = \begin{pmatrix} .866667 & .100000 & .033333 \\ .033333 & .766667 & .200000 \\ .300000 & .033333 & .666667 \end{pmatrix} .$$

In order to get the value of $\beta(1)$, we use the Gauss-Seidel iterative method for the system of linear equations

$$(5.13) \qquad (x_1, x_2, x_3) = (0, 0, 1) + (x_1, x_2, x_3)\begin{pmatrix} .866667 & .100000 & -1 \\ .033333 & .766667 & -1 \\ .300000 & .033333 & 0 \end{pmatrix}$$

We start with the initial value

$$(5.14) \qquad x(0) = (.333333 \quad .333333 \quad .333334)$$

Then the successive values of the vector x are as follows

$$(5.15) \qquad \begin{aligned} x(1) &= (\ .400000 \quad .306667 \quad .293333\ ) \\ x(2) &= (\ .444889 \quad .289378 \quad .265733\ ) \\ x(3) &= (\ .474937 \quad .278209 \quad .246854\ ) \\ x(4) &= (\ .494942 \quad .271016 \quad .234042\ ) \\ x(5) &= (\ .508197 \quad .266400 \quad .225403\ ) \\ x(6) &= (\ .516939 \quad .263447 \quad .219614\ ) \end{aligned}$$

As stated in Remark 5, we may regard this $x(6)$ as the value of $\beta(1)$.

Next we calculate $\gamma^1(1)$ from $Q^1(1)$. $Q^1(1)$ becomes

$$(5.16) \qquad Q^1(1) = \begin{pmatrix} .8 & .1 & .0 & .1 \\ .1 & .7 & .2 & .0 \\ .0 & .1 & .6 & .3 \\ .136389 & .0 & .018179 & .845432 \end{pmatrix}$$

In order to get the value of $\eta^1(1)$, we use the Gauss-Seidel iterative method for the system of linear equations

$$(5.17) \qquad (y_1, y_2, y_3, y_4) = (0, 0, 0, 1) + (y_1, y_2, y_3, y_4) \begin{pmatrix} .8 & .1 & .0 & -1 \\ .1 & .7 & .2 & -1 \\ .0 & .1 & .6 & -1 \\ .136389 & .0 & .018179 & 0 \end{pmatrix}$$

We may use the initial value $y(0) \doteq (\ .111111 \quad .111111 \quad .111111 \quad .666667\ )$. However in this case we adopt the initial value $y(0) = (\ .172313 \quad .172313 \quad .172313 \quad .483061\ )$ for the sake of the modification stated in Remark 6. Then the successive values of y are

$$(5.18) \qquad \begin{aligned} y(1) &= (\ .220965 \quad .159947 \quad .144159 \quad .474929\ ) \\ y(2) &= (\ .257542 \quad .152133 \quad .125556 \quad .464769\ ) \\ y(3) &= (\ .284636 \quad .147513 \quad .113286 \quad .454565\ ) \\ y(4) &= (\ .304458 \quad .145034 \quad .105243 \quad .445265\ ) \end{aligned}$$

By the modification stated in Remark 5, we regard this value $y(4)$ as the value of $\eta^1(1)$. Then the vector $\gamma^1(1)$ is given by

$$(5.19) \qquad \gamma^1(1) = (\ .548835 \quad .261447 \quad .189718\ ).$$

We note that in order to get β and γ's by the iterative method, we need not calculate the last columns of the matrices Q's . In practices, we may directly form the matrices in the systems of equations such as in (5.13) and (5.17) . In this case, we next form the system of equations

$$(5.20) \qquad (y_1, y_2, y_3, y_4) = (0, 0, 0, 1) + (y_1, y_2, y_3, y_4)\begin{pmatrix} .5 & .4 & .0 & -1 \\ .1 & .4 & .5 & -1 \\ .0 & .1 & .3 & -1 \\ .039945 & .0 & .009939 & 0 \end{pmatrix}$$

Starting with the initial value $y(0) = (.087816 \quad .087816 \quad .087816 \quad .736552)$, we obtain the value $\eta^2(1) = y(4) = (.077432 \quad .063168 \quad .057446 \quad .801954)$. Hence we have $\gamma^2(1) = (.390980 \quad .318956 \quad .290064)$. In like manner we have $\gamma^3(1) = (.325518 \quad .295372 \quad .379111)$.

Using these values for γ's , we can form the system of equations for $\beta(2)$,

$$(5.21) \qquad (x_1, x_2, x_3) = (0, 0, 1) + (x_1, x_2, x_3)\begin{pmatrix} .888201 & .056915 & -1 \\ .039098 & .786864 & -1 \\ .341200 & .032552 & 0 \end{pmatrix}$$

and starting with the initial value $x(0) = \beta(1)$, we have $\beta(2) = x(6) = (.605926 \quad .205617 \quad .188457)$.

We can proceed in like manner.

Example 3. The author used the lumping method for obtaining the stationary distribution of a Markov chain with the transition diagram in Figures 8 and 9 in Chapter 5. The state space of the chain is divided into six classes in a natural way. Hence it is expected that the lumping method is effective than ordinary methods. He first modified the problem by adopting the modification stated in Remark 4 to the sets in which states are combined with arcs with probability one in the transition diagram. Then he calculated the stationary distribution using the lumping method with the partition with six classes.

He also examined the numbers of multiplications and divisions required for the computations of the modified problem in two methods, the lumping method and the Gauss-Seidel iterative method. We can see from Table 1 below that the lumping method is especially effective for the cases in which the number of the states is large and the rate of convergence in the Gauss-Seidel method is slow.

Table 1. Comparison of the numbers of multiplications and divisions between the lumping method and the Gauss-Seidel method

p	q	r	k	the lumping method	the Gauss-Seidel method
1/3	1/3	1/3	3	1466	1556
			7	3938	8190
.5	.3	.2	3	1754	2972
			7	3938	8612
.7	.29	.01	3	1754	more than 4592
			7	5666	more than 10712

CHAPTER 5 A SEQUENCING MODEL WITH AN APPLICATION TO THE SPEED CLASS SEQUENCING IN THE AIR TRAFFIC CONTROL

1. Introduction

Somewhat to relieve the air traffic congestion at a terminal, sequencing methods of landing aircrafts by their speeds have been proposed and analyzed with computer simulations by many authors in private papers. In this paper we analyze theoretically how the landing capacity can be increased by adopting the speed class sequencing under some idealized conditions.

We set up a model in Section 3, while in Section 2 the speed class sequencing in the air traffic control is briefly explained and also the plan of the model is shown from the point of view of the speed class sequencing. The model is so simple that it can also be applied to many other queueing problems of the following type: There are two (or more) types of customers. A server serves customers one by one, and needs some transfer times besides substantial service times when the types of consecutive customers are different with each other. Then how can we decrease the mean service time (or increase the capacity) in a busy period by reordering (sequencing) the customers ?

Problems of this type could be analyzed using the theory of multi-queues or the theory of priority queues. However they are very difficult and only solved in the simplest cases. Though our model is very idealized, it will have many applications for the sake of the simplicity. For example, it may be applied to some queueing models in computer systems or to some job-shop scheduling models.

2. The Speed Class Sequencing

We shall start with a variant of the model used by A. Blumstein [1] for the analysis of the landing capacity. We consider a single runway which is used for landings only. Aircrafts pass through an imaginary gate in space and follow a common glide path to the runway. To schedule the passing times through the gate, two conditions for security must be satisfied:

(a) When an aircraft passes through the gate, the distance from the next aircraft should be greater than d (some fixed distance).

(b) There can be only one aircraft on the runway.

Let us define that the service time of an aircraft is the time interval between the passing times through the gate of it and of the preceding one. Then the service time of an aircraft depends not only (on) its own speed but also on the speed of the preceding one. We wish to schedule the passing times through the gate so that each service time is as short as possible. When should the aircraft pass the gate in order to have the shortest service time under the above conditions?

Let the length of the glide path and the runway be m and l respectively. (See Fig. 1.) Suppose that an aircraft with speed v_1 follows one with speed v_0. (We assume that the speed of each aircraft is constant in the considering area about the terminal. Modifications to the real situation would be obvious.) Let t_0 be the scheduled passing time of the preceding aircraft through the gate. Then it goes out the runway at $t_0 + (m + l)/v_0$. If the following aircraft passes the gate at t_1, it reaches the runway at $t_1 + m/v_1$. Hence by the condition (a), t_1 must be greater than $t_0 + d/v_1$, and by the condition (b) $t_1 + m/v_1$ must be greater than $t_0 + (m + l)/v_0$. For the shortest service time of the following aircraft, we should schedule its passing times through the gate as

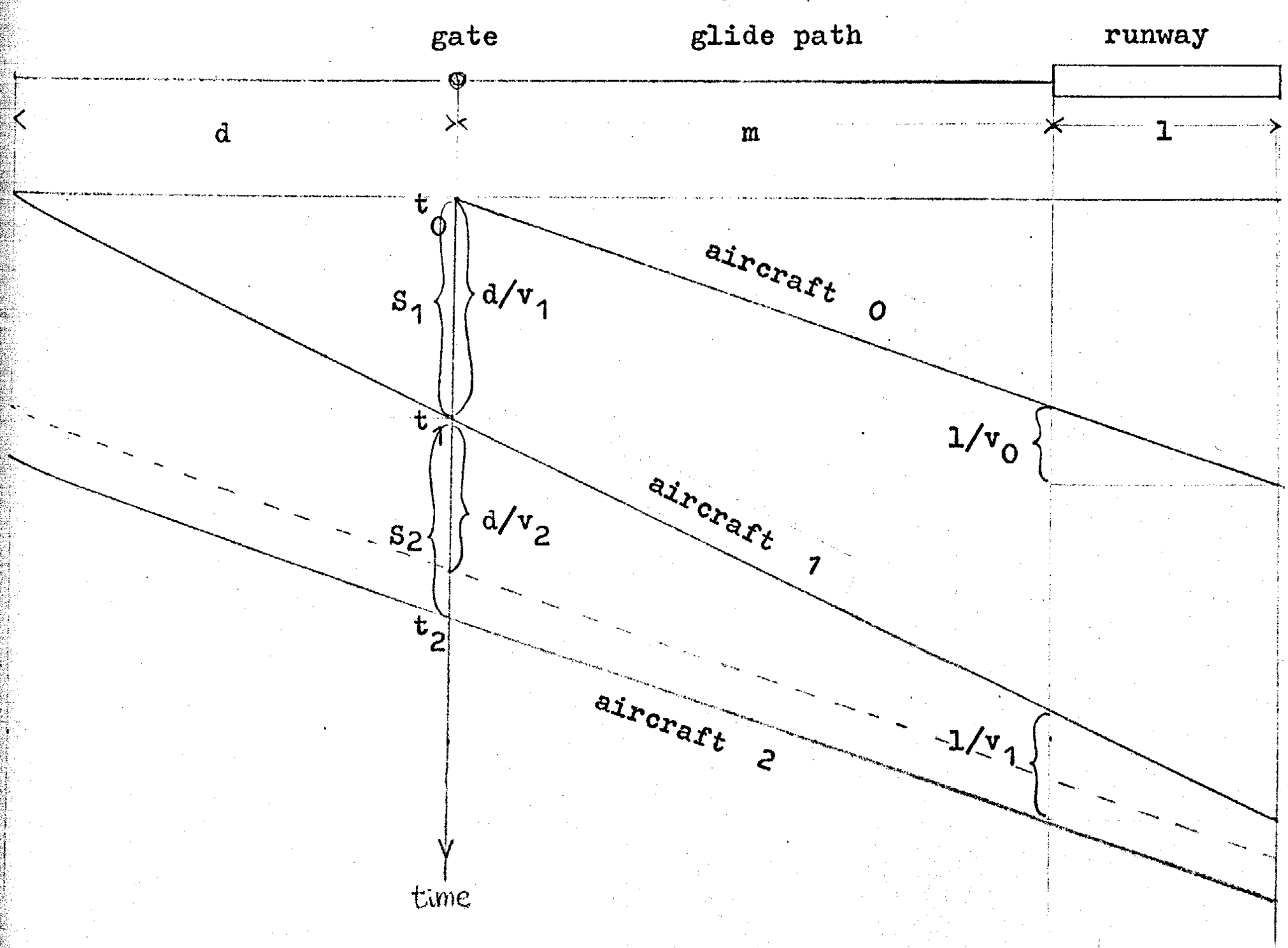
(1) $$t_1 = t_0 + \max \left(d/v_1 \, , \, (m + l)/v_0 - m/v_1 \right) ,$$

and then the shortest service time of it is given by

(2) $$S_1 = t_1 - t_0 = \max \left(d/v_1 \, , \, (m + l)/v_0 - m/v_1 \right) .$$

Since ordinarily $d > l$ for the sake of the unaccuracy of the estimation of the position, in the speed class sequencing it is preferable to schedule the passing time of an aircraft so that it follows another aircraft with the same or slower speed, or conversely speaking, so that it is followed by an aircraft with the same or faster speed.

Fig. 1



We can also regard the service time S_1 in (2) as the sum of the essential service time d/v_1 and the transfer time $S_1 - d/v_1$. Then non zero transfer time is needed only when the following aircraft is slower than the preceding one. From this point of view, our aid is to decrease the rate of transfer time or to decrease the rate of changes from a fast aircraft to a slow one by sequencing. For this purpose we should schedule them so that as many aircrafts as possible with the same speed land continuously. However if many aircrafts with the same speed land continuously, there is an aircraft with faster or slower speed having a long wait, and this is inadmissible in the practical situations. Hence we should seek an optimal sequencing policy in the set of policies satisfying some restrictions which assre that no aircraft has a long wait. This will be expressed in the four kinds of restrictions in our model in Section 3.

Though the speeds of landing aircrafts may be different with each other, ordinarily they can be predicted by the types of the aircrafts only, and in the speed class sequencing, we have to schedule the passing times through the gate by the information on the types of aircrafts. Hence we classify aircrafts by their types, and in the simplest case, we consider two types of aircrafts (e.g., jet planes and turbo-prop planes).

We will concern the case in which there are always sufficiently many aircrafts waiting for landing. It does not matter if few aircrafts are waiting. This altitude enable us to ignore the laws concerning the arrivals such as inter-arrival times.

3. The Model with Two Types of Customers

Now we shall set up our model. We consider a servicing system with a single server. Following three assumptions are made in the model.

A_1 : There are two types of customers and we distinguish them by type 0 or type 1 .

A_2 : Each customer has his own service time, and when the types of two consecutive customers are different with each other, the server needs some transfer time besides the own service times of the customers. Furthermore we assume that the own service time S_i $(i=0,1)$ of a customer of type i is constant, and that the transfer time T_i $(i=0,1)$ from type $(1-i)$ to type i is also constant. (This assumption can be replaced with a loose one. See Remark 1 at the end of this paper.)

A_3 : There are always sufficiently many customers waiting for service. (This assumption is introduced in order to simplify the model so that the laws for the input, such as the assumption of a renewal input and the distribution of the inter-arrival times, need not be specified. Hence it can be replaced with a loose one or with one defined more definitely. See Remark 2.)

By the above assumptions, the arriving customers can be represented by a sequence $(\epsilon_1, \epsilon_2, \epsilon_3, \cdots)$ where ϵ_n is equal to 0 or 1 according to the type of the n-th customer. We will use the term "the service time" of a customer as the sum of his own service time and the transfer time preceding it. Then our problem is to find a policy which minimize the mean service time of the first N customers. In order to state the problem more definitely, some preparations are needed.

We consider the first N customers whose types are represented by a sequence $(\epsilon_1, \epsilon_2, \cdots, \epsilon_N)$. A "sequencing policy" or simply "policy" is a permutation $\begin{pmatrix} 1 & 2 & \cdots & N \\ m_1 & m_2 & \cdots & m_N \end{pmatrix}$. Customers are reordered by a policy and the n-th customer gets the m_n-th position in the new sequence. The types of the customers in the new sequence are represented by $(\epsilon_1', \epsilon_2', \cdots, \epsilon_N')$ where $\epsilon_{m_n}' = \epsilon_n$. Then the service time of the m-th customer of the new sequence is given by

$$(3) \qquad S(m) = \begin{cases} S_{\epsilon_m'} & \text{if} \quad \epsilon_m' = \epsilon_{m-1}' , \\ S_{\epsilon_m'} + T_{\epsilon_m'} & \text{if} \quad \epsilon_m' = \epsilon_{m-1}' . \end{cases}$$

Hence the (arithmetic) mean service time of the N customers is given by

$$(4) \qquad \overline{S}_N = \frac{1}{N}\left[\sum_{m=1}^{N} S(m)\right] = \frac{1}{N}\left[N_0 S_0 + N_1 S_1 + M_0 T_0 + M_1 T_1\right]$$

$$= p_N S_0 + q_N S_1 + \alpha_N T_0 + \beta_N T_1$$

where N_i $(i = 0,1)$ is the number of m's such that $1 \leqq m \leqq N$ and $\epsilon_m' = i$, and M_i $(i = 0,1)$ is the number of m's such that $2 \leqq m \leqq N$ and $\epsilon_m' = i$, $\epsilon_{m-1}' = 1 - i$. We will call $p_N = N_0/N$ as the ratio of customers of type 0 (in the first N customers) and $q_N = N_1/N$ as the ratio of customers of type 1. And we will call $\alpha_N = M_0/N$ as the ratio of 1-0 changes (in the first N customers) and $\beta_N = M_1/N$ as the ratio of 0-1 changes.

In order to minimize the mean service time $\overline{S}_N$, if $T_0 > T_1$ we might reorder the input sequence so that all customers of type 0 are served first and then customers of type 1 are served. When N is large, under such a policy a customer of type 1 has a long wait. However in most of servicing systems, such a situation is inadequate. Especially for aircrafts, waiting time is very expensive, and such a situation arises excessive injustices. Hence we should restrict our policy so that no customer has a long wait, and find an optimal policy under the restriction.

Here we shall study the following four ranks of restrictions. (For other restrictions, see Remark 3.)

R_0 : $m_n = n$ (i.e., the first-come first-served policy).

R_1 : $m_n < m_{n+k+i}$ for all n and i (> 0).

R_2 : $|m_n - n| \leq k$ for all n.

R_3 : $m_n \leq n + k$ for all n.

In the above restrictions, k expresses the allowance limit of exchanges, and when $k=0$, R_1, R_2 and R_3 coincide with R_0. Let P_i $(i=0,1,2,3)$ be the set of all policies satisfying the restriction R_i. Then clearly $P_i \subset P_j$ if $i < j$.

Now we can state our problem definitely. For each input sequence, we say that a policy is optimal in P_i if it minimize the mean service time $\overline{S}_N$.

<u>Problem</u>. Find an optimal policy in P_i $(i=1,2,3)$.

The answer of this problem is given in the next section.

4. An Optimal Policy

Now we seek an optimal policy in each P_i . We can prove that the following procedure makes an optimal policy. In the procedure, we decide the servicing order one by one. So it is convenient to use the terms concerning the queue in the servicing system. We assume that customers in the queue are in the order of arrivals.

Procedure I : Serve the first customer in the queue of the same type as one being served, unless it is out of accordance with the restriction R_i . If it is out of accordance with R_i or if there is no customer in the queue of the same type as one being served, then serve the customer at the top of the queue.

In order to determine a policy by this procedure, we must choose a customer to be served first. Two special cases are considered.

Policy I_0 : First serve the first customer of type 0, and then adopt Procedure I.

Policy I_1 : First serve the first customer of type 1, and then adopt Procedure I.

Policy I_0 or Policy I_1 may not be in P_i , since the first service may violate the restriction R_i. But at least one of them is in P_i, and we can prove that either Policy I_0 or Policy I_1 is optimal in P_i. We will conveniently call the optimal one of them as Policy I.

Before proving the optimality, let us study an example adopting Policy I_0 to a sequence under the restriction R_1. (See Fig. 2.) In this case Procedure I can be rephrased as follows.

Rephrased Form of Policy I for R_1 :

If the server is servicing a customer of type 0 and if the customer at the top of the queue is of type 1 , then before servicing the top of the queue, let the server serve customers of type 0 in the first $(k+1)$ customers of the queue. And then let him serve customers of type 1 untill a customer of type 0 comes to the top of the queue. In the contrary case, do the same way exchanging the roles of 0 and 1 .

Fig. 2. An example of Policy I_0 with the restriction R_1 for $k=3$

input sequence 0 1 0 1 1 1 0 0 1 0 1 0 1 0 1 0

reordered sequence 0 0 1 1 1 1 1 0 0 0 0 0 1 1 1 1

1 0 1 0 0 1 0 1 1 $M_0 = 9$ $M_1 = 10$

1 0 0 0 0 0 1 1 1 $M_0 = 2$ $M_1 = 3$

Hence the mean service time $\bar{S}$ is decreased

from $\bar{S} = \frac{12}{25} S_0 + \frac{13}{25} S_1 + \frac{9}{25} T_0 + \frac{10}{25} T_1$ (under FCFS)

to $\bar{S} = \frac{12}{25} S_0 + \frac{13}{25} S_1 + \frac{2}{25} T_0 + \frac{3}{25} T_1$ (under Policy I_0)

The proof of the optimality of Policy I

Here we prove the optimality of Policy I for the restriction R_1. We can prove the optimality for the other restrictions in the same line. We prove it in two steps.

Lemma 1. There exists an optimal policy in the class P_1^* ($\subset P_1$) of policies satisfying that $m_n < m_{n'}$ for each pair n and n' such that $n < n'$ and $\epsilon_n = \epsilon_{n'}$. (In other words, we are sufficient to seek an optimal policy in the class P_1^* of policies under which the first-come first-served rule is adopted for the customers of the same type.)

Proof. Let the input sequence be $(\epsilon_1, \epsilon_2, \epsilon_3, \cdots, \epsilon_N)$, and under a policy in P_1 the n-th customer get the m_n-th position in the reordered sequence. Suppose that for some n and n' ($n < n'$), $\epsilon_n = \epsilon_{n'}$, $m_n = x$, $m_{n'} = y$ and $x > y$. By the definition of R_1

(5) $$x < m_{n+k+i}$$

for all positive i. Then, since $y < x$,

(6) $$y < m_{n+k+i}$$

for all positive i, and since $n < n'$,

(7) $$x < m_{n'+k+i}$$

for all positive i. Hence there is another policy in P_1 which reorders the sequence in the same manner as the original policy except that $m_n = y$

and $m_{n'} = x$. Since $\epsilon_n = \epsilon_{n'}$, the ratios of 1-0 changes and 0-1 changes are the same in both policies, and this shows that an optimal policy is in the class P_1^* .

<u>Theorem 2</u>. Either Policy I_0 or Policy I_1 is optimal in P_1 .

<u>Proof</u>. Let us consider an optimal policy in P_1^* and name it as Policy II . Suppose that $(m-1)$ customers have been served under Policy II and that $\epsilon'_{m-1} = 0$ (i.e., a customer of type 0 has just been served). If the n_i-th customer $(i = 0,1)$ in the original sequence is the first customer of type i in the queue, then under Policy II , either the n_0-th or the n_1-th customer is served next. Three cases can be considered. In order to satisfy the restriction R_1 ,

(i) the n_0-th customer must be served next.

(ii) the n_1-th customer must be served next.

(iii) both the n_0-th and the n_1-th customers may be served next.

In the first two cases we have no alternative, and we can consider that Policy II adopts Procedure I in this step. In the third case, if $m_{n_0} = m$ (i.e., the n_0-th customer is served next), Policy II adopts Procedure I in this step. If $m_{n_1} = m$, we can prove that there exists another optimal policy which adopts Procedure I in this step.

Suppose that we are in the third case and that $m_{n_1} = m$ under Policy II . Let $M_0^{II}(j,j')$ be the number of 1-0 changes from the j-th customer to the j'-th customer in the reordered sequence. Then the number of 1-0 changes in the N customers is given by

(8) $$M_0^{II} = M_0^{II}(1,N) = M_0^{II}(1,m-1) + M_0^{II}(m-1,m) + M_0^{II}(m,N)$$
$$= M_0^{II}(1,m-1) + 0 + M_0^{II}(m,N) \ .$$

We define Policy III as follows. Under Policy III, the server serves $(m-1)$ customers in the same manner as Policy II, and then serves the n_0-th and the n_1-th customers in this order. And he serves the remaining $(N-m-1)$ customers in the queue in the same order as Policy II except for the n_0-th customer. Then Policy III adopts Procedure I at the m-th service, and we can easily prove that Policy III is in P_1^*. The number of 1-0 changes under Policy III is given by

(9) $$M_0^{III} = M_0^{III}(1,N) = M_0^{III}(1,m-1) + M_0^{III}(m-1,m+1) + M_0^{III}(m+1,N)$$
$$= M_0^{III}(1,m-1) + 0 + M_0^{III}(m+1,N) \ .$$

Now we compare $M_0^{II}(m,N)$ with $M_0^{III}(m+1,N)$. The reordered segment from the $(m+1)$-th costomer to the N-th one under Policy III is the same as the reordered segment from the m-th customer to the N-th one under Policy II except that a customer of type 0 has been gone out. Hence

(10) $$M_0^{II}(m,N) \geqq M_0^{III}(m+1,N) \ .$$

Since $M_0^{II}(1,m-1) = M_0^{III}(1,m-1)$, we have

(11) $$M_0^{II} \geqq M_0^{III}$$

by (8), (9) and (10).

For the number of 0-1 changes, we can also prove that $M_1^{II} \geqq M_1^{III}$ in a similar way. Hence we have

(12) $$\bar{S}_N^{II} \geqq \bar{S}_N^{III} .$$

It follows that Policy III is also optimal, since Policy II is optimal.

Using the above result, by the mathematical induction on m, we can prove that there exists an optimal policy in P_1^* which adopts Procedure I in each step. Such an optimal policy is either Policy I_0 or Policy I_1, and this proves the theorem.

5. The Limiting Mean Service Time under an Optimal Policy

Now we shall investigate how we can decrease the mean service time by adopting the sequencing. For the purpose, we are convenient to assume that the input sequence is a Bernoulli sequence with probabilities p of type 0 and q of type 1 $(p+q=1)$. Namely we assume that $\epsilon_1, \epsilon_2, \cdots, \epsilon_N$ are mutually independent random variables with the common distribution

$$P(\epsilon_i = 0) = p \quad \text{and} \quad P(\epsilon_i = 1) = q .$$

Then by the law of large numbers, the ratios p_N and q_N of customers of type 0 and of type 1 converges to p and q with probability one as $N \to \infty$. By the renewal theorem, we can prove that the ratio α_N of 1-0 changes under Policy I converges to a limit α with probability one as $N \to \infty$. The ratio β_N of 0-1 changes under Policy I also converges to the same limit α with probability one as $N \to \infty$, since always $|M_0 - M_1| \leqq 1$. Hence we can write the limiting mean service time as

$$\bar{S} = pS_0 + qS_1 + \alpha(T_0 + T_1) \quad . \tag{13}$$

Table 1 shows the representation of the limit α for each restriction. In the remaining of this section, we shall prove it. In the subsequent discussions, we sometimes discuss as if N is infinity and omit the word "limiting".

Table 1. The limiting ratio α of 1-0 changes under an optimal policy with the restriction R_i

restriction	α
R_0	pq
R_1	$pq/(1 + 2kpq)$
R_2	$pq/(1 + 2pq\ E(Z_k))$ (*)
R_3	$pq/(1 + k)$

(*) Z_k is a random variable having the distribution

$$\Pr\left\{Z_k = t\right\} = \binom{t-1}{k-1}\left(p^k q^{t-k} + p^{t-k} q^k\right) \quad \text{if } k \leqq t \leqq 2k-1, \tag{14}$$

$$= 0 \qquad \text{otherwise.}$$

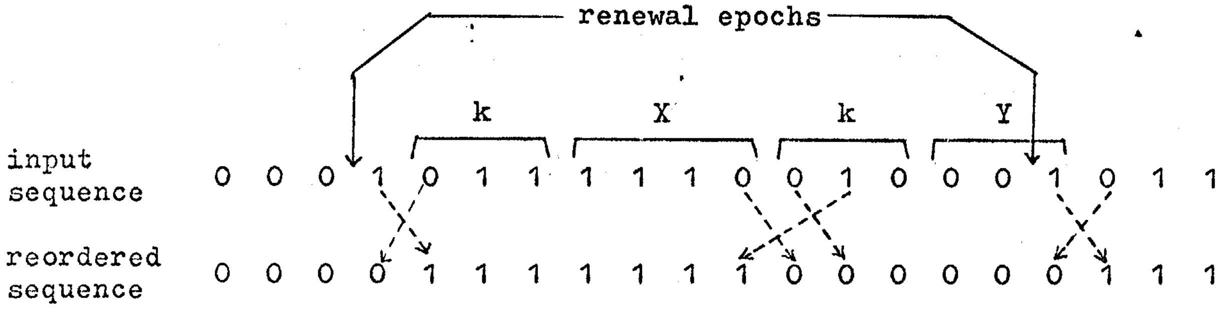
Table 2. $E(Z_k)$ for small k

k	$E(Z_k)$
0	0
1	1
2	$2(1 + pq)$
3	$3(1 + pq + 2p^2q^2)$
4	$4(1 + pq + 6p^2q^2 + 5p^3q^3)$

(i) The case under the restriction R_1

In this case Procedure I can be rephrased as in Section 4. Hence, considering the input sequence as a discrete time stochastic process, we may take the time as a renewal epoch when a customer of type 1 comes to the top of the queue during a continuous service of customers of type 0 (see Fig. 3).

Fig. 3. Renewal epochs under Policy I in P_1



Then the length of a cycle from a renewal epoch to the next renewal epoch is given by

(15) $$L = 2k + X + Y$$

where X and Y are random variables representing the number of customers required in a Bernoulli sequence up to and including the first arrival of a customer of type 0 and of type 1 respectively. Since X and Y have geometric distributions, the mean cycle length is given by

(16) $$E(L) = 2k + \frac{1}{p} + \frac{1}{q} = \frac{1}{pq}(1 + 2kpq).$$

In each cycle, just one 1-0 change occurs. Hence by the renewal theorem, the limiting ratio α of 1-0 changes exists with probability one and is given by

(17) $$\alpha = 1/E(L) = pq/(1 + 2kpq) .$$

(ii) The case under the restriction R_2

We may take similar renewal epochs as in (i) (see Fig. 4). In this case a cycle consists of four mutually independent random variables W_1, X, W_2 and Y. Customers of type 0 in W_1 are served before the customer of type 1 who is the last one in preceding Y, and customers of type 1 in W_2 are served before the customer of type 0 who is the last one in X. X and Y have the same distributions as in (i). W_i $(i=1,2)$ is the number of customers required in a Bernoulli sequence untill at least k customers of one type arrive. Then we can easily show that W_i has the distribution given by (14). Hence it follows that the mean cycle length

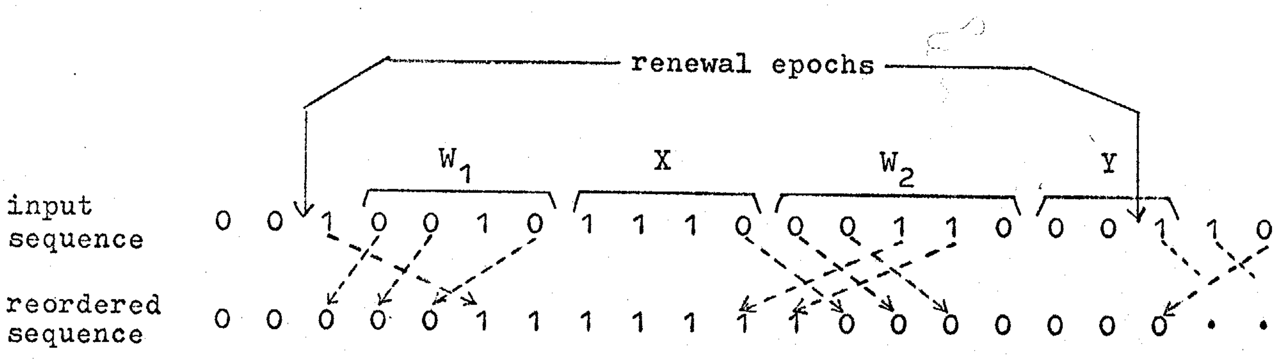
$$(18) \qquad E(L) = E(W_1) + E(X) + E(W_2) + E(Y)$$

$$= 2\,E(Z_k) + \frac{1}{p} + \frac{1}{q} ,$$

and that the limiting ratio of 1-0 changes

$$(19) \qquad \alpha = pq/(1 + 2pq\,E(Z_k)) .$$

Fig. 4. Renewal epochs under Policy I in P_2



(iii) The case under the restriction R_3

We use similar renewal epochs as in (i), too (see Fig. 5). In this case a cycle consists of four mutually independent random variables U, X, V and Y. Customers of type 0 in U are served before the customer of type 1 who is the last one in preceding Y, and customers of type 1 in V are served before the customer of type 0 who is the last one in X. X and Y have the same distributions as in (i). U is the number of customers required in a Bernoulli sequence up to and including the k-th arrival of type 0, and V is a similar one for type 1. Then U and V have negative binomial distributions. Hence it follows that the mean cycle length

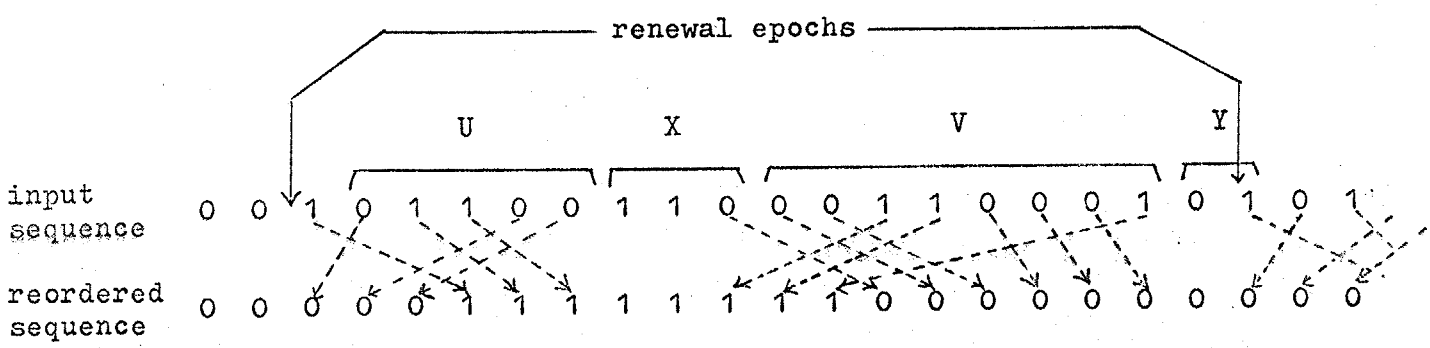
$$E(L) = E(U) + E(X) + E(V) + E(Y) \tag{20}$$

$$= \frac{k}{p} + \frac{1}{p} + \frac{k}{q} + \frac{1}{q} = \frac{k+1}{pq},$$

and that the limiting ratio of 1-0 changes

$$\alpha = pq/(1+k)\ . \tag{21}$$

Fig. 5. Renewal epochs under Policy I in P_3



6. Comparison of Policy I in the Three Restrictions

We have considered three restrictions R_1, R_2 and R_3, and obtained an optimal policy in each case. However the optimality in here reflects only the capacity of the system. In general, ameasure for goodness of a policy may contain more factors. In the speed class sequencing, the measure might have factors such as the capacity of the runway, the simplicity of the control, the mean number of aircrafts which go ahead of a certain aircraft, and so on. Hence we should decide the measure taking care of the pattern of the landing aircrafts, the ability of the control system, the security in the sequencing, feelings of people for the delays caused by the sequencing, and so on.

If the measure is determined, we might be to seek an optimal policy with respect to the measure. However, such approach is sometimes difficult, and there is another approach in which we have been engaged. In the approach, we choose several policies and study their properties. And in the application, we adopt one of them which is best of all with respect to the measure. Hence we should study properties of Policy I for each restrictions.

Here we consider only the mean number μ of aircrafts which go ahead of a certain aircraft. μ can be also considered as the mean delay $E[m_n - n]^+$ caused by the sequencing, or as a half of the mean absolute departure $E|m_n - n|$ of m_n from n, since $E(m_n) = n$. Table 3 shows the results of the calculations of μ, and Fig. 6 shows a numerical example for $p = 0.7$ and $q = 0.3$. There may be an optimal policy in P_i which has smaller μ

than Policy I . For example, if the input sequence is $(0, \cdots, 0, 1, 0, \cdots, 0)$, then for each restriction R_i , $\mu = k/N$ under Policy I , but $\mu = 0$ under the first-come first-served policy which is also optimal for the sequence. Hence we might be to seek an optimal policy in P_i which minimize μ. Unfortunately, we have not yet find such an optimal policy. However under such a policy, probably we cannot decide the servicing order one by one, and so the control is more complicated than Policy I .

Table 3. The mean number μ of aircrafts which go ahead of a certain aircraft under Policy I

restriction	μ
R_1	$kpq\{1+(k-1)pq\}/(1+2kpq)$
R_2	$pq\{(k+3/2)E(Z_k)-k\}/(1+2kpqE(Z_k))$
R_3	$2kpq/(k+1)+k/2$

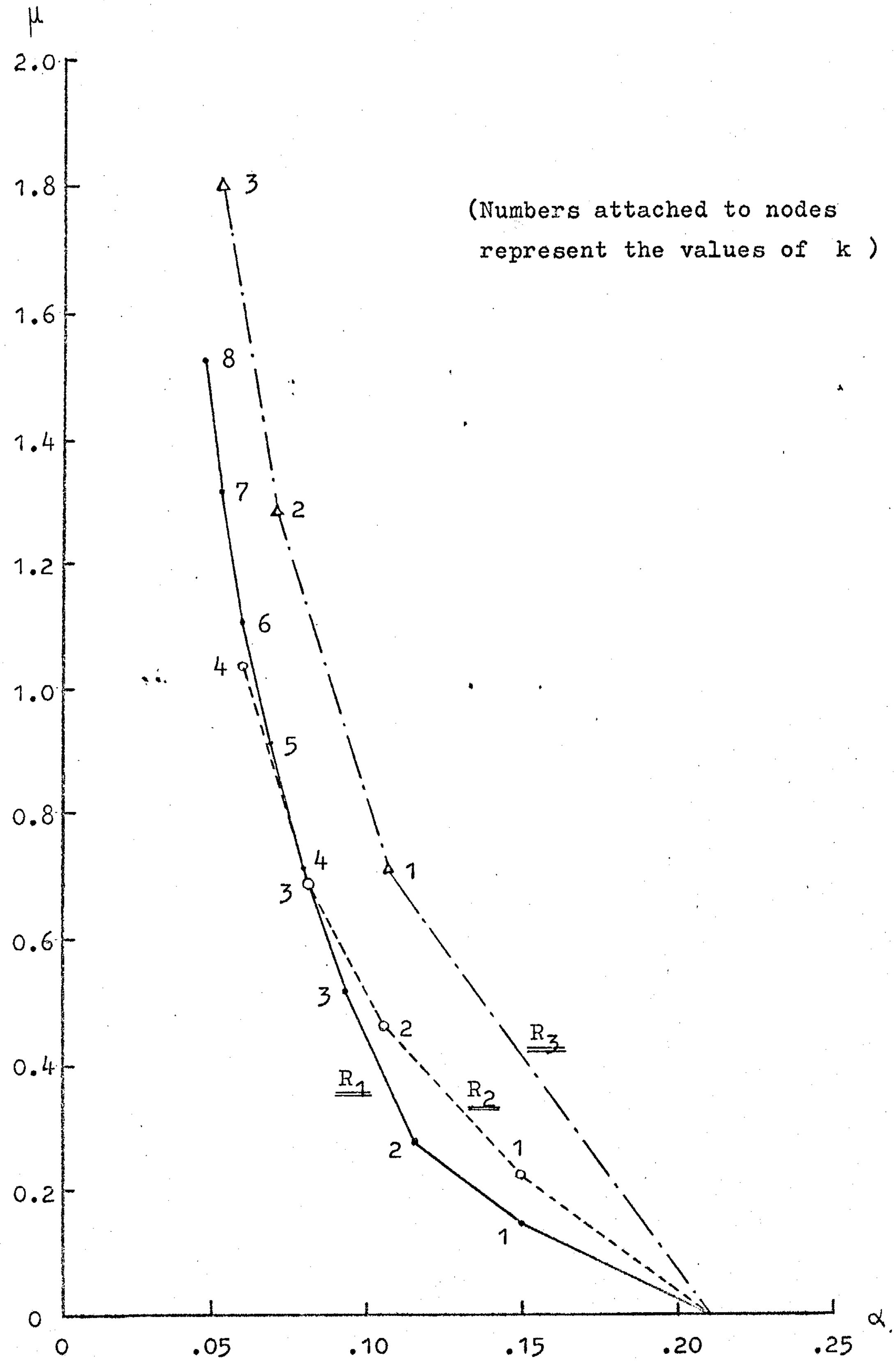
In order to prove the results in Table 3, we return to Section 5. For the case under the restriction R_1 , we consider a cycle in Fig. 3. Aircrafts of type 1, which arrive after an aircraft of type 0 and whichare served before it, are in the third interval denoted by k . If the i-th aircraft in the interval is of type 0 , the mean number of aircrafts of type 1 which go ahead of it is $(k-i)q$. And for the last aircraft in the second interval denoted by X , the mean number of aircrafts of type 1 which go ahead of it is kq . These are the only aircrafts of type 0 for which $m_n > n$. We can obtain similar results for aircrafts of type 1 for which $m_n > n$. Hence the mean number μ of aircrafts which go ahead of a certain aircraft is given by

$$\text{(22)} \qquad \mu = \frac{pq}{1 + 2kpq}\left[\sum_{i=1}^{k}(k-i)pq + kq + \sum_{i=1}^{k}(k-i)qp + kp\right]$$

$$= \frac{kpq}{1 + 2kpq}\left[1 + (k-1)pq\right] .$$

This proves the result in Table 3 for R_1. For the restrictions R_2 and R_3 we only notice that it is convenient to calculate rather $E\left[n - m_n\right]^+$ than $E\left[m_n - n\right]^+$, and omit the proofs.

Fig. 6. The mean number μ of aircrafts which go ahead of a certain aircraft and the ratio α of 0-1 changes under Policy I ($p = 0.7$, $q = 0.3$)



7. An Example with Three Types of Customers

If the number of types of customers are more than two, we can not adopt the renewal theoretical techniques used in the preceding sections. In such a case, we should use a Markov chain technique in order to calculate the limiting ratios of changes of the type of service. As an example we consider a servicing system with three types of customers under the restriction R_1 (defined in Section 3). In this case we can hardly find an optimal policy, because, if we try to get an optimal policy, the selction of the next served customer depends not only on the type of the customer being served, but also on the whole sequence. This shows an essential difference between the cases of two types of customers and of three or more types of customers. Hence inthis case we are content with a policy adopting Procedure I . We study with

Policy I' : First serve the first customer, and then adopt Procedure I .

Since our aid is to obtain the limiting ratios of changes of the type of service, we can take an infinite sequence $(\epsilon_1, \epsilon_2, \epsilon_3, \cdots)$ as the input sequence. We assume that it is a sequence of mutually independent random variables ϵ_n $(n = 1, 2, 3, \cdots)$ with a common distribution

$$\Pr(\epsilon_n = a) = p, \quad \Pr(\epsilon_n = b) = q \quad \text{and} \quad \Pr(\epsilon_n = c) = r \qquad (p + q + r = 1). \tag{22}$$

In order to calculate the limiting ratios of changes of the type of service, we must define a Markov chain. However it is preferable to study an example of services under Policy I' before we define the Markov chain.

Let us consider an example for $k = 3$ in Fig. 7. Customer 1 is served first. Then Customer 2 comes to the top of the queue, but he is of different type with Customer 1. Hence we seek customers of type a among Customers $3 \sim 5$ who are the only customers being able to be served before Customer 2 under the restriction R_1. We find that Customer 4 is of type a. Then we decide that Customer 4 is served next and Customer 2 is served after him. Thus the server serves these customers in the order ; 1 , 4 , 2 . When Customer 2 is served, Customer 3 is at the top of the queue. Since he is of different type with Customer 2, we seek customers of type b among Customers $4 \sim 6$ who are the only customers being able to be served before Customer 3 under the restriction R_1. (Customer 4 has already been served and he is not in the queue. But in order to clarify the range for search, it is convenient to treat as if he were in the queue.) However we can find no customer of type b , and so we decide that Customer 3 is served next. When Customer 3 is served, Customer 5 is at the top of the queue. Since he is of the same type as Customer 3, he is served immediately after Customer 3. The same situations arise for Customers 6 and 7. When Customer 7 is served, Customer 8 is at the top of the queue. Since he is of different type with Customer 7, we seek customers of type c among Customers $9 \sim 11$. Since Customer 9 is of type c , we decide that he is served immediately after Customer 7, and then Customer 8 is served. ....

Fig. 7. An example of transitions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
input sequence	a	b	c	a	c	c	c	a	c	b	b	c	a	b	b	a	c	c	a	a
	*	*	*	*	*	∅	∅	0	1	2	3	*	*	∅	∅	0	1	2	3	*
states	∅	0	1	2	3	*	∅	∅	∅	0	1	2	3	*	*	∅	∅	∅	∅	∅
	∅	∅	0	1	2	3	*	*	*	*	*	0	1	2	3	*	*	*	*	∅
reordered sequence	a	a	b	c	c	c	c	c	a	a	b	b	b	b	c	c	c	a	a	a
	1	4	2	3	5	6	7	9	8	13	10	11	14	15	12	17	18	16	19	20

Now we define a Markov chain which represents the transitions of the type of service. There are many such Markov chains, but the one defined below has a fairly small number of states.

Since we are interested in the limiting ratios of changes of the type of service, we need not to know exactly when a change of the type of service occurs, but need only to know the fact that it occurs at some time. Hence we define the Markov chain so that it can indicate only the occurrences of changes of the type of service. A change of the type of service occurs if and only if a customer is not served immediately after a customer of the same type. And in such a case, he comes to the top of the queue when a customer of different type is served. Hence we can consider that these three contents represent the same situation. It is convenient to explain the definition of the Markov chain by following the transitions in the example in Fig. 7.

We consider that the time of the Markov chain corresponds to the customer's number. The state of the Markov chain at time n $(n = 1,2,3,\cdots)$ is denoted by a triple $(s_a(n),\ s_b(n),\ s_c(n))$, and each element $s_i(n)$ $(i = a,b,c)$ of it varies over a set $\{\emptyset, 0, 1, 2, \cdots, k, *\}$. In Figures 7 and 8, for the convenience of printing, we write the state in the form of a column vector.

In the example in Fig. 7, the Markov chain starts with the state $(*, \emptyset, \emptyset)$. Generally, $\emptyset$ in $s_i(n)$ $(i = a,b,c\ ;\ n = 1,2,3,\cdots)$ represents such a situation that Customer n is not of type i and that even if he is of type i, he will not be served immediately after a customer of type i. And $*$ in $s_i(n)$ represents such a situation that if Customer n is of type i, then he will be served immediately after a customer of type i. In this case Customer 1 is the first customer and there is no customer who is served before him. However it is convenient to consider that there is a customer (say Customer 0) of the same type as he and that he is served immediately after the customer. Hence here we set $s_a(1) = *$.

Customer 2 is of type b but, since $s_b(1) = \emptyset$, he will not be served immediately after a customer of type b. We represent this situation by setting $s_b(2) = 0$. Generally 0 in $s_i(n)$ $(i = a,b,c\ ;\ n = 1,2,3,\cdots)$ represents such a situation that Customer n is of type i and he will not be served immediately after a customer of type i. Hence if $s_i(n) = 0$, then Customer n will come to the top of the queue when a customer of different type is served. Returning to the example, $s_c(2)$ remains at $\emptyset$ and $s_a(2)$ also remains at $*$, because if Customer 2 is of type a, he will be served immediately after Customer 1. Hence the state of the Markov chain at time 2 becomes to $(*, 0, \emptyset)$.

When a customer of different type with one being served comes to the top of the queue, we must seek customers of the same type as one being served. In this case we must seek customers of type a among Customers $3 \sim 5$. In order to indicate these customers among whom we seek customers of type a , we set $s_b(3) = 1$, $s_b(4) = 2$ and $s_b(5) = 3$. And we can indicate that we seek customers <u>of type a</u> by setting $s_a(3) = s_a(4) = s_a(5) = *$ (see the meaning of $*$ stated above). Generally if $s_i(n) = 0$ $(i = a,b,c\ ;\ n = 1,2,3, \cdots)$, then in order to indicate the range for search, we set $s_i(n+1) = 1$, $s_i(n+2) = 2$, $\cdots$, $s_i(n+k) = k$. Of course, in the case, customers of type i among Customers $(n+1) \sim (n+k)$ are served successively after Customer n. Hence besides the above meaning, ν in $s_i(n+\nu)$ $(\nu = 1,2, \cdots ,k)$ represents the same situation as $*$, too.

For the state at time 3, we have already shown that $s_a(3) = *$ and $s_b(3) = 1$. Since Customer 3 is of type c and $s_c(2) = \emptyset$, $s_c(3)$ becomes to 0 by the same reason as Customer 2. Hence the state at time 3 is $(* , 1 , 0)$.

When Customer 3 comes to the top of the queue, we seek customers of type b among Customers $4 \sim 6$. In order to indicate these customers, we set $s_c(4) = 1$, $s_c(5) = 2$ and $s_c(6) = 3$. And in order to indicate that we seek customers <u>of type b</u> , we set $s_b(6) = *$ since $s_b(4)$ and $s_b(5)$ have already been set to 2 and 3 respectively. Thus the state of the Markov chain from time 1 to time 5 become as in Fig. 7.

For the state at time 6, $s_b(6)$ and $s_c(6)$ have been set to $*$ and 3 respectively. The state at time 5 in which $s_a(5) = *$ and $s_b(5) = 3$ implies that customers of type a among Customers $3 \sim 5$ are served before Customer 2 and that Customer 6 can not be served before Customer 2. Hence even if Customer 6 is of type a , he will not be served immediately after a customer of type a . In this case Customer 6 is not of type a and so $s_a(6)$ becomes to $\emptyset$. Thus the state at time 6 is $(\emptyset, *, 3)$. Generally if $s_i(n-1) = *$ $(i = a,b,c \ ; \ n = 2,3,4, \cdots)$ and if $s_j(n-1) = k$ for some j $(= a,b,c$ and $\neq i)$, then $s_i(n) = 0$ when Customer n is of type i and $s_i(n) = \emptyset$ when Customer n is not of type i .

At time 7, $s_b(7)$ becomes $\emptyset$ by the same reason as above stated, and $s_a(7)$ remains at $\emptyset$ since Customer 7 is not of type a . Since $s_c(6) = 3$ and Customer 7 is of type c , he will be served immediately after a customer of type c . So $s_c(7)$ becomes to $*$. Thus the state at time 7 is $(\emptyset, \emptyset, *)$. Generally if $s_i(n-1) = k$ $(i = a,b,c \ ; \ n = 2,3,4, \cdots)$, then $s_i(n)$ becomes to $*$.

At time 8, since Customer 8 is of type a , $s_a(8) = 0$ and $s_b(8)$ and $s_c(8)$ remain at $\emptyset$ and $*$ respectively. Hence the state at time 8 becomes to $(0, \emptyset, *)$. We can procede in like manner.

We can tabulate the rules of the transition from $s_i(n-1)$ to $s_i(n)$ $(i = a,b,c \ ; \ n = 2,3,4, \cdots)$ as in Table 4. In table 4, ϵ_n represents the type of Customer n .

Table 4. The rules of the transition from $s_i(n-1)$ to $s_i(n)$ $(i=a,b,c\ ;\ n=2,3,4,\cdots)$

$s_i(n-1) \rightarrow s_i(n)$

1. $\emptyset \longrightarrow \emptyset$ if $\epsilon_n \neq i$ — ϵ_n : the type of Customer n

   $\longrightarrow 0$ if $\epsilon_n = i$

2. $0 \longrightarrow 1$

   $1 \longrightarrow 2$

   $\cdots\cdots$

   $k-1 \longrightarrow k$

3. $k \longrightarrow *$

4. $* \longrightarrow *$ if $s_j(n-1) \neq k$ for every j $(=a,b,c$ and $\neq i)$

   $\longrightarrow \emptyset$ if $s_j(n-1) = k$ for some j $(\neq i)$ and if $\epsilon_n \neq i$

   $\longrightarrow 0$ if $s_j(n-1) = k$ for some j $(\neq i)$ and if $\epsilon_n = i$

This Markov chain can indicate the occurrence of a change of the type of service. If $s_i(n) = 0$ $(n=1,2,3,\cdots)$ for some i $(=a,b,c)$, then Customer n will not be served immediately after a customer of the same type. Hence we can know the changes of the type of service by the appearances of 0. In order to know them, it is convenient to see the transitions of $*$. Because, each 0 is follwed by an $*$ in the same element of the state just after k epochs, and from one such $*$ to the next such $*$, $*$'s fill the same elements of each state as the former $*$.

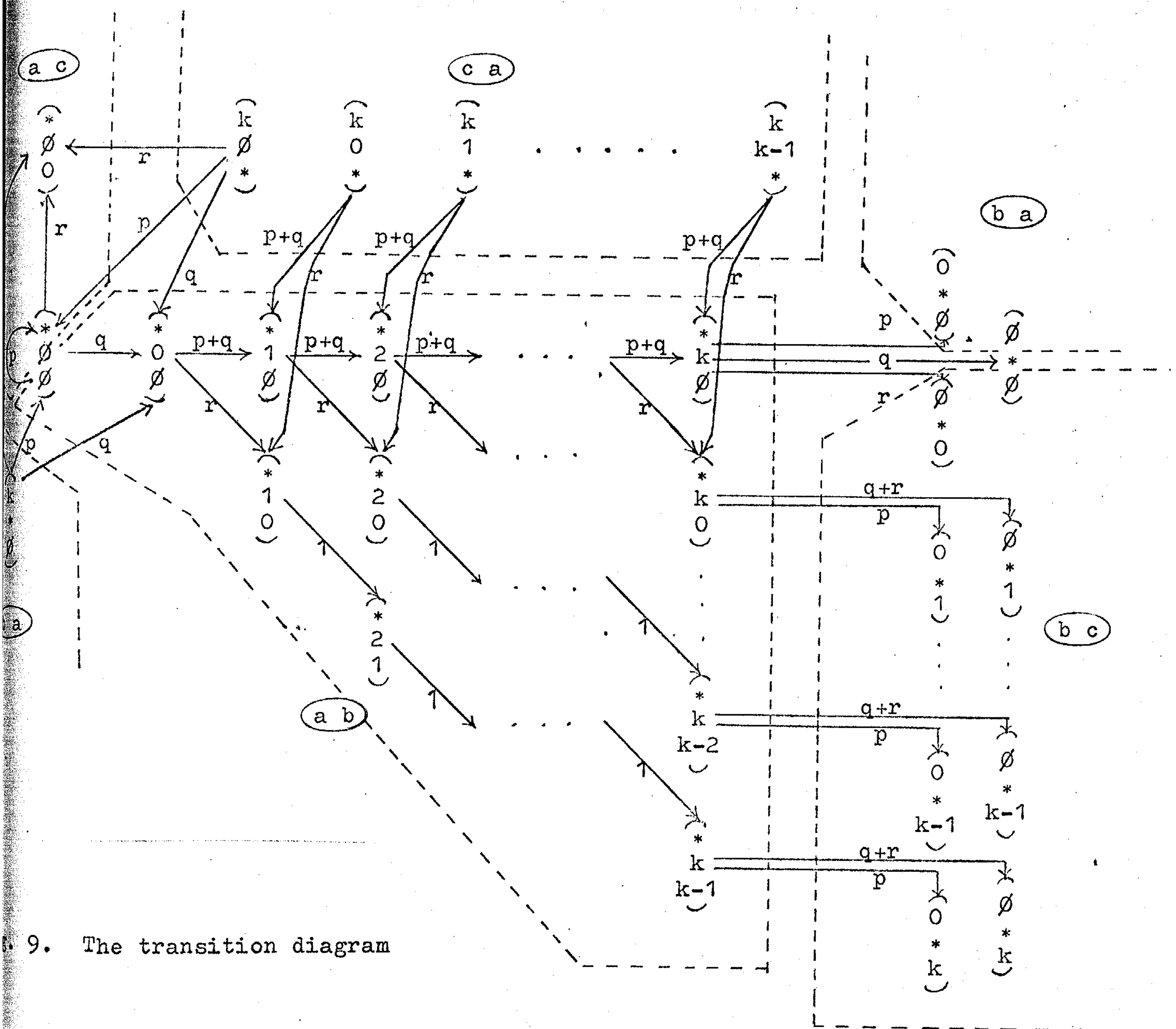
The set of possible states (s_a, s_b, s_c) $(s_a, s_b, s_c = \emptyset, 0, 1, 2, \cdots, k, *)$ of the Markov chain can be divided in to six classes (a b), (a c), (b a), (b c), (c a) and (c b). For the purpose, we introduce an order $\prec$ into the set of the symbols as follows ;

$$\emptyset \prec 0 \prec 1 \prec 2 \prec \cdots\cdots \prec k \prec * .$$

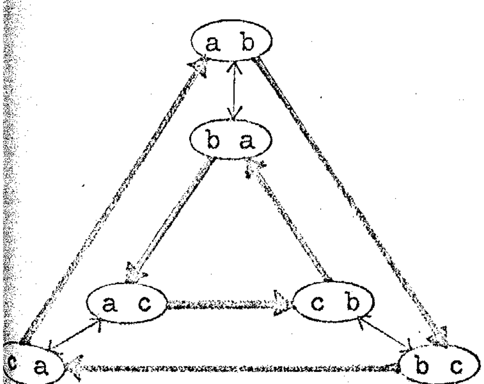
The class (i j) $(i, j = a, b, c \; ; \; i \neq j)$ consists of states (s_a, s_b, s_c) for which s_i is most dominant and s_j is secondarily dominant of the three in the order $\prec$. For example, $(*, \emptyset, 2)$ belongs to (a c) and $(2, 3, *)$ belongs to (c b). By this rule, $(*, \emptyset, \emptyset)$ can not belong to any class. However from this state, the Markov chain can only go to either $(*, 0, \emptyset)$ or $(*, \emptyset, 0)$, except for staying in the state. Hence for the sake of convenience we consider that the half of it (strictly speaking, $q/(q+r)$ of it) belongs to (a b) and the remaining half of it (strictly speaking, $r/(q+r)$ of it) belongs to (a c). (This is justified by the property of a lumped Markov chain, but here we omit the discussion about it.) For the states $(\emptyset, *, \emptyset)$ and $(\emptyset, \emptyset, *)$, we can also treat in like manner.

Then the transition diagram of the Markov chain can be written as Figures 8 and 9. Fig. 8 shows a part of it about the class (a b), and Fig. 9 shows the relations between classes. A transition from a state in (i j) to a state in (j l) $(i, j, l = a, b, c \; ; \; i \neq j$ and $j \neq l)$ implies a transition of $*$ from s_i to s_j, or a change of the type of service from type i to type j. Hence in order to obtain the limiting ratios of changes of the type of service, we may try to obtain the rate of transitions from one class to another.

Fig. 8. The detailed subdiagram of transitions about the class (a b)



9. The transition diagram



Computational results. The author calculated the limiting ratios of changes of the type of service using a computer. At that time, he used the idea of a lumped Markov chain and shortened the computing times (on this subject, see [2]). As a measure of the utility of the sequencing, we use the ratio γ of the mean service time under Policy I' to that under the first-come first-served policy (FCFS). γ is also the reciprocal of the ratio of the capacity under Policy I' to that under the first-come first-served policy.

Fig. 10 shows the value of the ratio γ for various triples (p, q, r) when $k=4$, $S_a = S_b = S_c = T_{ac} = 2$, $T_{ab} = T_{bc} = 1$ and $T_{ba} = T_{ca} = T_{cb} = 0$, where S_i $(i = a,b,c)$ is the service time of a customer of type i, and T_{ij} $(i,j = a,b,c\ ;\ i \neq j)$ is the transfer time from type i to type j. It shows that in the case the sequencing gives an increase in the capacity of about 10 per cent. Fig. 11 shows the change of the value of the ratio γ for small k. The line for $(.5, .0, .5)$ is the lower bound of γ and γ can not be under the line. The upper bound is the line $\gamma = 1$.

Fig. 10. The rate γ of the mean service time under Policy I' to that under the first-come first-served policy for $k = 4$

$$\gamma = \frac{\bar{S} \text{ under Policy I'}}{\bar{S} \text{ under FCFS}}$$

$S_a = S_b = S_c = T_{ac} = 2$

$T_{ab} = T_{bc} = 1$

$T_{ba} = T_{ca} = T_{cb} = 0$

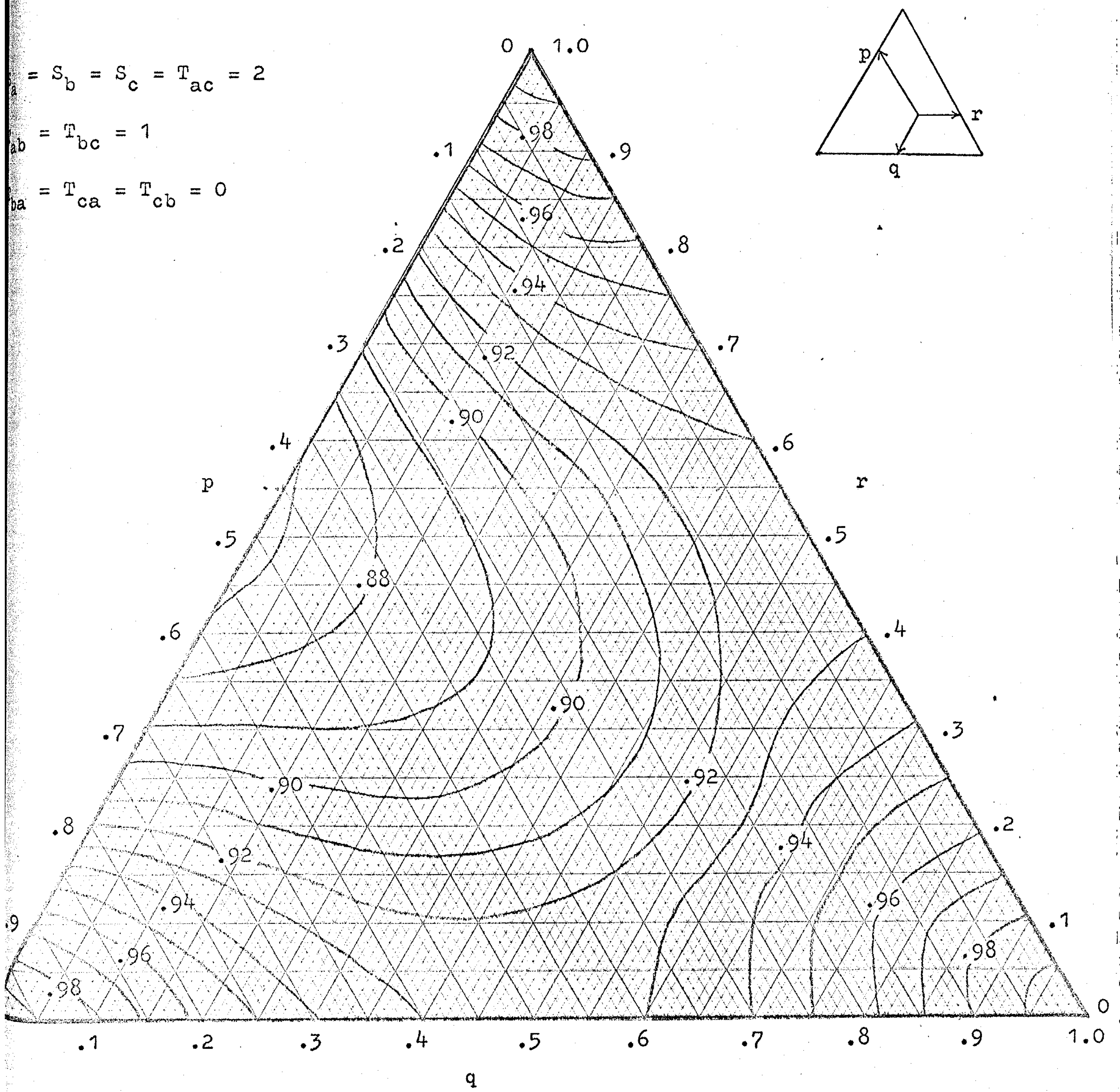


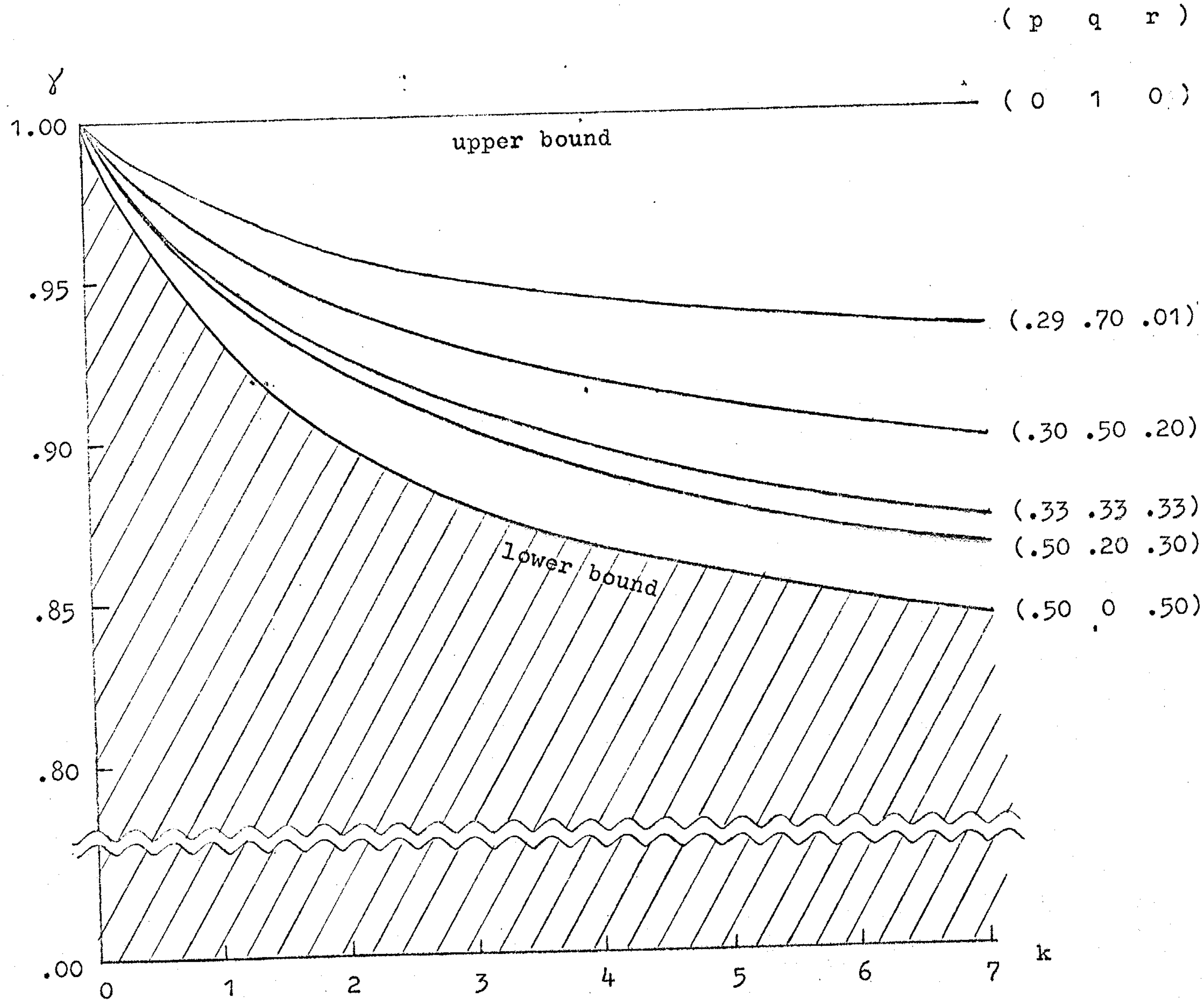
Fig. 11. The change of the ratio γ for small k

$$\gamma = \frac{\bar{S} \text{ under Policy I'}}{\bar{S} \text{ under FCFS}}$$

$S_a = S_b = S_c = T_{ac} = 2$

$T_{ab} = T_{bc} = 1$

$T_{ba} = T_{ca} = T_{cb} = 0$



8. Remarks

Here we gather some remarks on generalizations of the model. Remarks 1 and 2 are concerning to the fundamental assumptions, and Remark 3 is concerning to the restrictions of policies. Some remarks on a Markov chain method for a system with four or more types of customers are stated in Remark 4 .

Remark 1. In the assumption A_2 (in Section 3), we need not any restriction for the own service times of customers in the discussions of the optimality. Transfer times from one type to another may be (possibly dependent) random variables with a common expectaion if we loose the definition of the optimality to minimize the expectation $E(\overline{S})$ of the mean service time.

In order to obtain the limiting mean service time, we need some conditions for the renewal theorem. For example, we may require that the own service times of type 0 , the own service times of type 1 , the transfer times from type 1 to type 0 and the transfer times from type 0 to type 1 are mutually independent random variables having common distributions with finite expectations respectively.

Remark 2. The assumption A_3 (in Section 3) needs only the assurance of the adoptability of any policy in P_i . Especially, for the restrictions R_1 and R_2 , it is sufficient to assure that k or more customers are always waiting in the queue.

Remark 3. We can consider other restrictions than R_0, R_1, R_2 and R_3, such as

$$R_4 : \quad m_n \geq n - k \qquad \text{for all } n.$$

This assumption R_4 would not be adequate for the speed class sequencing in the air traffic control. However we would be able to use this assumption for a servicing system with a finite waiting room.

We can also generalize restrictions R_i $(i = 1,2,3,4)$ so that k is different with types of customers.

R_1' : $\quad m_n < m_{n+k_0+i}$ for all n such that $\epsilon_n = 0$ and for all i (>0).

$\quad\quad m_n < m_{n+k_1+i}$ for all n such that $\epsilon_n = 1$ and for all i (>0).

R_2' : $\quad |m_n - n| \leq k_0$ for all n such that $\epsilon_n = 0$.

$\quad\quad |m_n - n| \leq k_1$ for all n such that $\epsilon_n = 1$.

R_3' : $\quad m_n \leq n + k_0$ for all n such that $\epsilon_n = 0$.

$\quad\quad m_n \leq n + k_1$ for all n such that $\epsilon_n = 1$.

R_4' : $\quad m_n \geq n - k_0$ for all n such that $\epsilon_n = 0$.

$\quad\quad m_n \geq n - k_1$ for all n such that $\epsilon_n = 1$.

The limiting ratio α of 1-0 changes under an optimal policy with restriction R_i' can be obtained similarly. Their results are as follows.

restriction	α
R_1'	$pq/(1 + (k_0 + k_1)\, pq)$
R_2'	$pq/(1 + pq\, E(Z_{k_0k_1} + Z_{k_1k_0}))$
R_3' , R_4'	$pq/((k_0 + 1)p + (k_1 + 1)q)$

where $Z_{kk'}$ is a random variable having the following distribution. If $k < k'$,

$$\Pr\{Z_{kk'} = t\} = \begin{cases} \binom{t-1}{k-1} p^k q^{t-k} & (k \leq t < k') \\ \binom{t-1}{k-1} p^k q^{t-k} + \binom{t-1}{k'-1} p^{t-k'} q^{k'} & (k' \leq t \leq k + k' - 1) \\ 0 & (\text{otherwise}) \end{cases},$$

and if $k > k'$,

$$\Pr\left\{ Z_{kk'} = t \right\} = \begin{cases} \binom{t-1}{k'-1} p^{t-k'} q^{k'} & (k' \leqq t < k) \\ \binom{t-1}{k-1} p^{k} q^{t-k} + \binom{t-1}{k'-1} p^{t-k'} q^{k'} & (k \leqq t \leqq k + k' - 1) \\ 0 & (\text{otherwise}) \end{cases}$$

Remark 4. The example in Section 7 can be generalized for systems with four or more types of customers. The limiting ratios of changes between servicing types can be analyzed similarly using Markov chains. For a system with four types of customers, states of a Markov chain are represented in quadruples of states of individual types. These states of individual types and transition rules between them are the same as in Section 7. Then the transition diagram for the Markov chain has twelve parts and becomes more complicated.

ACKNOWLEDGEMENT

The author wishes to thank Mr. A. Kato for his suggestion of this problem and Professor H. Morimura for his advice and encouragement.

References

[1] Blumstein, A., "The landing capacity of a runway," Operations Reserch 7 (1959), 752-763.

[2] Takahashi, Y., "An iterative method for the numerical calculation of the limiting vector of a Markov chain," (in preparation).

APPENDIX 1

KŌDAI MATH. SEM. REP.
21 (1969), 426—447

MARKOV CHAINS WITH RANDOM TRANSITION MATRICES

BY YUKIO TAKAHASHI

Introduction.

Let P^t $(t=1, 2, 3, \cdots)$ be the transition matrix from epoch $t-1$ to t of a Markov chain with a finite state space S, and $\alpha^{(n)}$ $(n=0, 1, 2, \cdots)$ be the probability distribution at n. Then we have

$$(*)\qquad \alpha^{(n)}=\alpha^{(0)}P^1\cdots P^n.$$

Now we assume that $\alpha^{(0)}$ and P^t are mutually independent random variables and that $\alpha^{(n)}$ is defined by (*). Then $\{\alpha^{(n)}\}$ is a Markov process on the space of probability distributions on S. ($\alpha^{(n)}$ represents the probability distribution at n, starting with the initial distribution $\alpha^{(0)}$ and following to the *random* transition matrices P^t.) Such a process will be called a "*Markov chain with random transition matrices*" (*M.C. with R.T.M.*).

The author intended to generalize ordinary Markov chains as briefly mentioned above, by the following reason.

Markov chains have been applied in many fields, and one of their applications is in the analysis or prediction of market shares. Many authors have worked with so-called Markov brand-switching models, in which $\alpha^{(n)}$ represents the market shares at epoch n and P^t represents the transition matrix from epoch $t-1$ to t. In many cases they assume that these Markov chains (they consider that $\alpha^{(n)}$ is the distribution of a Markov chain at step n) are stationary. However some other authors have given warnings of the failure of stationarity and of other defects of these models (see e.g. A. S. C. Ehrenberg [1]). The author thinks that one of the causes of the warnings is in the assumption that $\alpha^{(0)}$ and P^t are *a priori* given (known) and hence have no stochastic fluctuation. The transition matrix P^t reflects the choices of purchasers, and so it is essentially stochastic. Hence it seems to be natural to consider that P^t and $\alpha^{(n)}$ are random variables. The most simple stochastic model for market shares is the model using our M.C. with R.T.M.

In this paper, properties of M.C. with R.T.M., in particular moments of $\alpha^{(n)}$ and conditions for the convergence (in law) of $\alpha^{(n)}$, are given. And we classify stationary and irreducible M.C. with R.T.M. into three groups; ergodic chains, aperiodic and non-ergodic chains, and periodic chains. Finally we prove some ergodic theorems.

Received June 17, 1969.

144

1. Markov chains with random transition matrices.

Let $S=\{1, 2, \cdots, s\}$ be a finite state space, $\mathcal{A}$ be the set of all probability measures on S, and $\mathcal{P}$ be the set of all stochastic matrices with index sets S and S. We may be consider $\mathcal{A}$ and $\mathcal{P}$ as subsets of s- and s^2-dimensional Euclidean spaces respectively;

$$\mathcal{A}=\left\{\alpha=(\alpha_1, \cdots, \alpha_s) \,\middle|\, \alpha_i \geqq 0 \quad i=1, \cdots, s, \quad \sum_{i=1}^{s} \alpha_i=1\right\}, \tag{1. 1}$$

$$\mathcal{P}=\left\{P=\begin{pmatrix} p_{11}\cdots p_{1s} \\ \vdots \quad\quad \vdots \\ p_{s1}\cdots p_{ss}\end{pmatrix} \,\middle|\, p_{ij}\geqq 0 \quad i, j=1, \cdots, s, \quad \sum_{j=1}^{s} p_{ij}=1\right\}. \tag{1. 2}$$

Therefore we can define random variables which take values on $\mathcal{A}$ or $\mathcal{P}$.

Given a sequence of probability spaces $(\Omega^t, \mathcal{F}^t, \Pr^t)$ $(t=0, 1, 2, \cdots)$, a vector valued random variable

$$\boldsymbol{\alpha}^{(0)}(\omega^0)=(\boldsymbol{\alpha}_i^{(0)}(\omega^0) \quad i=1, 2, \cdots, s) \tag{1. 3}$$

on Ω^0 with values in $\mathcal{A}$, and matrix valued random variables

$$\boldsymbol{P}^t(\omega^t)=(\boldsymbol{p}_{ij}^t(\omega^t) \quad i, j=1, 2, \cdots, s) \tag{1. 4}$$

on Ω^t $(t=1, 2, 3, \cdots)$ with values in $\mathcal{P}$, then we define the product probability space $(\Omega, \mathcal{F}, \Pr)$ by

$$(\Omega, \mathcal{F}, \Pr)=\prod_{t=0}^{\infty}(\Omega^t, \mathcal{F}^t, \Pr^t), \tag{1. 5}$$

and random variables $\boldsymbol{\alpha}^{(0)}(\omega)$ and $\boldsymbol{P}^t(\omega)$ on Ω by

$$\boldsymbol{\alpha}^{(0)}(\omega)=\boldsymbol{\alpha}^{(0)}(\omega^0) \quad \text{and} \quad \boldsymbol{P}^t(\omega)=\boldsymbol{P}^t(\omega^t) \qquad t=1, 2, 3, \cdots \tag{1. 6}$$

respectively, where

$$\Omega \ni \omega=(\omega^0, \omega^1, \omega^2, \cdots). \tag{1. 7}$$

DEFINITION. A *Markov chain with random transition matrices* (*M.C. with R.T.M.*) $\{\boldsymbol{\alpha}^{(n)}(\omega)\}$ is a Markov process on Ω with values in $\mathcal{A}$ of the form

$$\boldsymbol{\alpha}^{(n)}(\omega)=\boldsymbol{\alpha}^{(0)}(\omega)\boldsymbol{P}^1(\omega)\cdots\boldsymbol{P}^n(\omega). \tag{1. 8}$$

In the following sections we mainly study the stationary case where all probability spaces $(\Omega^t, \mathcal{F}^t, \Pr^t)$ $(t=1, 2, 3, \cdots)$ but $(\Omega^0, \mathcal{F}^0, \Pr^0)$ are identical and random variables $\boldsymbol{P}^t(\omega^t)$ have a common distribution. We will refer to such a chain as a *stationary M.C. with R.T.M.*

We denote the n-step transition probability matrix by

$$\boldsymbol{P}^{(n)}(\omega)=(\boldsymbol{p}_{ij}^{(n)}(\omega))=\boldsymbol{P}^1(\omega)\cdots\boldsymbol{P}^n(\omega). \tag{1. 9}$$

145

Clearly $\{P^{(n)}(\omega)\}$ is a Markov process and so we sometimes call it also a M.C. with R.T.M.

In the following sections, ω will be omitted when no confusion arises.

2. Moments.

We first calculate the moments of $\boldsymbol{a}^{(n)}(\omega)$ and of $P^{(n)}(\omega)$. We prepare some notations.

Let S_k $(k=1, 2, 3, \cdots)$ be the set of all ordered k-tuples $(i_1, \cdots, i_k)$ of states in S (e.g. when $S=\{1, 2\}$, $S_1=\{(1), (2)\}$; $S_2=\{(1, 1), (1, 2), (2, 1), (2, 2)\}$, $\cdots$). The k-th moment of $\boldsymbol{a}^{(n)}(\omega)$ is denoted by the row vector

$$\xi_k^{(n)}=(\xi_k^{(n)}(i_1, \cdots, i_k),\ (i_1, \cdots, i_k)\in S_k) \tag{2. 1}$$

where

$$\xi_k^{(n)}(i_1, \cdots, i_k)=E\{\boldsymbol{a}_{i_1}^{(n)}(\omega)\cdots\boldsymbol{a}_{i_k}^{(n)}(\omega)\}, \tag{2. 2}$$

and the k-th moment matrix of $P^t(\omega)$ is denoted by

$$\Sigma_k^t=(\sigma_k^t(i_1, \cdots, i_k; j_1, \cdots, j_k),\ (i_1, \cdots, i_k; j_1, \cdots, j_k)\in S_k\times S_k) \tag{2. 3}$$

where

$$\sigma_k^t(i_1, \cdots, i_k; j_1, \cdots, j_k)=E\{p_{i_1j_1}^t(\omega)\cdots p_{i_kj_k}^t(\omega)\}. \tag{2. 4}$$

In the stationary case Σ_k^t will be abbreviated to Σ_k. The same notations as the moments of $P^t(\omega)$ will be used for the moments of $P^{(n)}(\omega)$ with brackets on their shoulders.

We note that $\xi_k^{(n)}$ may be considered as a probability measure on S_k and that Σ_k^t and $\Sigma_k^{(n)}$ are stochastic matrices. In fact

$$\begin{aligned}\sum_{(i_1,\cdots,i_k)\in S_k}\xi_k^{(n)}(i_1, \cdots, i_k)&=\sum_{(i_1,\cdots,i_k)\in S_k}E\{\boldsymbol{a}_{i_1}^{(n)}\cdots\boldsymbol{a}_{i_k}^{(n)}\}\\ &=E\left\{\left(\sum_{i_1=1}^{s}\boldsymbol{a}_{i_1}^{(n)}\right)\cdots\left(\sum_{i_k=1}^{s}\boldsymbol{a}_{i_k}^{(n)}\right)\right\}=1\end{aligned} \tag{2. 5}$$

which proves the first statement, and a similar calculation leads to the second. Furthermore, we may prove the following

THEOREM 1. *For a M.C. with R.T.M. we have*

$$\Sigma_k^{(n)}=\Sigma_k^1\cdots\Sigma_k^n \quad \textit{and} \quad \xi_k^{(n)}=\xi_k^{(0)}\Sigma_k^{(n)}=\xi_k^{(0)}\Sigma_k^1\cdots\Sigma_k^n. \tag{2. 6}$$

In particular, for a stationary chain we have

$$\Sigma_k^{(n)}=(\Sigma_k)^n \quad \text{and} \quad \xi_k^{(n)}=\xi_k^{(0)}(\Sigma_k)^n. \tag{2. 7}$$

Proof. We shall show that $\xi_k^{(n)}=\xi_k^{(0)}\Sigma_k^{(n)}$. By the independence of variables we have

$$
\begin{aligned}
\xi_k^{(n)}(i_1, \cdots, i_k) &= E\left\{\left(\sum_{l_1=1}^{s} \alpha_{l_1}^{(0)} p_{l_1 i_1}^{(n)}\right) \cdots \left(\sum_{l_k=1}^{s} \alpha_{l_k}^{(0)} p_{l_k i_k}^{(n)}\right)\right\} \\
&= \sum_{(l_1, \cdots, l_k) \in S_k} E\{\alpha_{l_1}^{(0)} \cdots \alpha_{l_k}^{(0)} \cdot p_{l_1 i_1}^{(n)} \cdots p_{l_k i_k}^{(n)}\} \\
&= \sum_{(l_1, \cdots, l_k) \in S_k} E\{\alpha_{l_1}^{(0)} \cdots \alpha_{l_k}^{(0)}\} E\{p_{l_1 i_1}^{(n)} \cdots p_{l_k i_k}^{(n)}\} \qquad (2.8) \\
&= \sum_{(l_1, \cdots, l_k) \in S_k} \xi_k^{(0)}(l_1, \cdots, l_k) \cdot \sigma_k^{(n)}(l_1, \cdots, l_k; i_1, \cdots, i_k).
\end{aligned}
$$

The first relation in (2. 6) can be proved by a similar calculation. Q.E.D.

Now we consider the convergence problem of $\boldsymbol{\alpha}^{(n)}(\omega)$. We need a lemma for the convergence of a sequence of random variables (e.g. see Feller [3] p. 244).

LEMMA 2. *Let $\{X^n\}$ be a uniformly bounded sequence of random variables in r-dimensional Euclidean space and μ_k^n be the k-th moment vector of X^n. X^n converges in law to a limit X if, and only if, for each k, μ_k^n converges to a limit vector μ_k. In this case μ_k is the k-th moment vector of X.*

Applying this lemma to our chains, we obtain the following theorem.

THEOREM 3. *A M.C. with R.T.M. converges in law if, and only if, $\xi_k^{(n)} = \xi_k^{(0)} \Sigma_k^1 \cdots \Sigma_k^n$ converges for each k.*

3. Classification of stationary M.C. with R.T.M. (I)—Periodicity.

The theory of ordinary Markov chains suggests us that stationary M.C. with R.T.M. could be classified by similar ideas. For convenience we consider $P^{(n)}$-process instead of $\boldsymbol{\alpha}^{(n)}$-process. Theorem 1 shows that for the expectation of $P^{(n)}$ we can consider the ordinary Markov chain with transition matrix Σ_1. It seems to be natural to classify stationary M.C. with R.T.M. by the properties of M.C. Σ_1 (we refer to an ordinary stationary Markov chain by its transition matrix and write as M.C. Σ_1). We might define that a stationary M.C. with R.T.M. is *irreducible* (*reducible*) if M.C. Σ_1 is irreducible (reducible), and that a stationary and irreducible M.C. with R.T.M. is aperiodic (periodic) if M.C. Σ_1 is aperiodic (periodic). This definition of an irreducible chain is adequate in the sense that for each pair i, j ($\in S$) there is an n such that

$$
\Pr\{p_{ij}^{(n)}(\omega) > 0\} > 0. \qquad (3.1)
$$

However Example 1 below shows that the above definition of an aperiodic chain does not seem to be adequate.

In the following sections (except in Theorem 8) we will consider stationary and irreducible M.C. with R.T.M. only, and sometimes the words "irreducible" and "stationary" will be omitted.

EXAMPLE 1. Let $s=3$ and the distribution of P^t be

$$\Pr\{\boldsymbol{P}^t=P_1\}=\Pr\{\boldsymbol{P}^t=P_2\}=\Pr\{\boldsymbol{P}^t=P_3\}=\frac{1}{3} \tag{3.2}$$

where

$$P_1=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix},\quad P_2=\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\quad\text{and}\quad P_3=\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}. \tag{3.3}$$

Then we have

$$\Sigma_1=\begin{pmatrix}1/3&1/3&1/3\\1/3&1/3&1/3\\1/3&1/3&1/3\end{pmatrix}. \tag{3.4}$$

Hence M.C. Σ_1 is aperiodic.

Now let

$$P_4=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\quad P_5=\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}\quad\text{and}\quad P_6=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}. \tag{3.5}$$

Since

$$\begin{aligned}
&P_1P_1=P_2P_2=P_3P_3=P_4, &&P_4P_1=P_5P_2=P_6P_3=P_1,\\
&P_1P_2=P_2P_3=P_3P_1=P_5, &&P_4P_2=P_5P_3=P_6P_1=P_2,\\
&P_1P_3=P_2P_1=P_3P_2=P_6, &&P_4P_3=P_5P_1=P_6P_2=P_3,
\end{aligned} \tag{3.6}$$

we have

$$\Pr\{\boldsymbol{P}^{(n)}=P_1\}=\Pr\{\boldsymbol{P}^{(n)}=P_2\}=\Pr\{\boldsymbol{P}^{(n)}=P_3\}=\frac{1}{3} \tag{3.7}$$

if n is odd, and

$$\Pr\{\boldsymbol{P}^{(n)}=P_4\}=\Pr\{\boldsymbol{P}^{(n)}=P_5\}=\Pr\{\boldsymbol{P}^{(n)}=P_6\}=\frac{1}{3} \tag{3.8}$$

if n is even. Hence we would rather say that this M.C. with R.T.M. has "period two".

Thus we must make a new definition of the period of a stationary M.C. with R.T.M. By Lemma 2, for the convergence of $\boldsymbol{P}^{(n)}$ we are enough to examine the convergence of its moments only. We shall show that there is an integer $r\geqq 1$ such that $\Sigma_k^{(nr+m)}=(\Sigma_k)^{nr+m}$ converges as $n\to\infty$ for each k and m $(m=0,1,2,\cdots,r-1)$. (Convergence of a sequence of matrices means element-wise convergence.)

By the theory of ordinary Markov chains, each state $(i_1, \cdots, i_k) \in S_k$ has its own period with respect to M.C. Σ_k. We define that $(j_1, \cdots, j_h) \in S_h$ and $(i_1, \cdots, i_k) \in S_k$ be *equivalent* if both consist of the same states in S. (For example, $(1, 1, 2)$ is equivalent to $(2, 1, 2, 2)$.) Let $(j_1, \cdots, j_h)$ be equivalent to $(i_1, \cdots, i_k)$. By the very definition we have

$$\sigma_k^{(n)}(i_1, \cdots, i_k; i_1, \cdots, i_k) = E\{p_{i_1 i_1}^{(n)} \cdots p_{i_k i_k}^{(n)}\} \tag{3.9}$$

and

$$\sigma_h^{(n)}(j_1, \cdots, j_h; j_1, \cdots, j_h) = E\{p_{j_1 j_1}^{(n)} \cdots p_{j_h j_h}^{(n)}\}. \tag{3.10}$$

So the equivalence of $(i_1, \cdots, i_k)$ and $(j_1, \cdots, j_k)$ implies that the values in both braces in the right sides of (3.9) and (3.10) vanish simultaneously. Therefore $\sigma_k^{(n)}(i_1, \cdots, i_k; i_1, \cdots, i_k) = 0$ if, and only if, $\sigma_h^{(n)}(j_1, \cdots, j_h; j_1, \cdots, j_h) = 0$. Thus equivalent states have the same period if they are periodic. And it is easily shown that if a state in S_k is transient (with respect to M.C. Σ_k), then each state, which is equivalent to it, is also transient. Therefore equivalent states have a common period.

Let r be the least common multiple of the periods of states in S_s (s is the number of states in S). Then $\Sigma_s^{(nr+m)} = (\Sigma_s)^{nr+m}$ converges as $n \to \infty$ for each m ($m = 0, 1, 2, \cdots, r-1$) and the limit matrices are different for different m's. We note that each state in S_k has an equivalent state in S_s. Hence r is also a common multiple of the periods of states in S_k and $\Sigma_k^{(nr+m)} = (\Sigma_k)^{nr+m}$ converges as $n \to \infty$ for each m ($m = 0, 1, 2, \cdots, r-1$). Therefore applying Lemma 2 to r requences $\{\boldsymbol{P}^{(nr+m)}\}$ ($m = 0, 1, 2, \cdots, r-1$) we obtain the following

THEOREM 4. *For a stationary and irreducible M.C. with R.T.M. there exists unique integer $r \geqq 1$ such that r sequences $\{\boldsymbol{P}^{(nr+m)}\}$ $(m = 0, 1, 2, \cdots, r-1)$ converge in law as $n \to \infty$ and that their limit distributions are different from each other.*

DEFINITION. The *period* of a stationary and irreducible M.C. with R.T.M. is the number whose existence is assured in Theorem 4.

The discussion preceding Theorem 4 shows that the period of a M.C. with R.T.M. is the least common multiple of the periods of states in S_s. Turning to Example 1, the states $(1, 1, 1)$, $(2, 2, 2)$ and $(3, 3, 3)$ have period one with respect to M.C. Σ_3 and other states in S_3 have period two. Hence the chain has period two. Thus the new definition of the period seems to be adequate.

4. Classification of stationary M.C. with R.T.M. (II)—Ergodicity.

In the last section we classified stationary and irreducible M.C. with R.T.M. by their periods. However, there is another and more essential classification; ergodic chains or non-ergodic chains. We shall start with two examples.

EXAMPLE 2. Let $s = 2$ and the distribution of $\boldsymbol{P}^t$ be

149

$$\Pr\{\boldsymbol{P}^t=P_1\}=\Pr\{\boldsymbol{P}^t=P_2\}=\frac{1}{2} \tag{4.1}$$

where

$$P_1=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \quad \text{and} \quad P_2=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}. \tag{4.2}$$

Then

$$\Sigma_1=\begin{pmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{pmatrix}, \tag{4.3}$$

and it is easily shown that this chain is aperiodic. Since

$$P_1P_1=P_2P_2=P_1 \quad \text{and} \quad P_1P_2=P_2P_1=P_2, \tag{4.4}$$

we have

$$\Pr\{\boldsymbol{P}^{(n)}=P_1\}=\Pr\{\boldsymbol{P}^{(n)}=P_2\}=\frac{1}{2} \tag{4.5}$$

for each n. Hence if $\boldsymbol{\alpha}^{(0)}=(p, 1-p)$ with probability one,

$$\Pr\{\boldsymbol{\alpha}^{(n)}=(p, 1-p)\}=\Pr\{\boldsymbol{\alpha}^{(n)}=(1-p, p)\}=\frac{1}{2}. \tag{4.6}$$

Thus the distribution of $\boldsymbol{\alpha}^{(n)}$ is independent of n, and the limit distribution is given also by (4.6). We note that the limit distribution depends on $\boldsymbol{\alpha}^{(0)}$.

EXAMPLE 3. Let S and the distribution of $\boldsymbol{P}^t$ be as in Example 2, but this time we put

$$P_1=\begin{pmatrix}1/2 & 1/2\\ 1 & 0\end{pmatrix} \quad \text{and} \quad P_2=\begin{pmatrix}1/2 & 1/2\\ 0 & 1\end{pmatrix}. \tag{4.7}$$

Then Σ_1 is given by (4.3) and the chain is aperiodic too. But the distribution of $\boldsymbol{\alpha}^{(n)}$ is not so simple as the preceding example. Direct calculation shows that

$$\Pr\left\{\boldsymbol{P}^{(n)}=\begin{pmatrix}p_{n,k} & 1-p_{n,k}\\ q_{n,k} & 1-q_{n,k}\end{pmatrix}\right\}=\Pr\left\{\boldsymbol{P}^{(n)}=\begin{pmatrix}1-p_{n,k} & p_{n,k}\\ 1-q_{n,k} & q_{n,k}\end{pmatrix}\right\}=\frac{1}{2^n} \tag{4.8}$$

where

$$p_{n,k}=\frac{2k-1}{2^n} \quad \text{and} \quad q_{n,k}=\frac{k}{2^{n-1}} \qquad (k=1, 2, \cdots, 2^{n-1}). \tag{4.9}$$

Therefore $\boldsymbol{P}^{(n)}$ converges in law to a random matrix

150

$$\begin{pmatrix} q & 1-q \\ q & 1-q \end{pmatrix} \tag{4.10}$$

where q is a random variable following to the uniform distribution on the unit interval $[0, 1]$. Hence even if $\boldsymbol{\alpha}^{(0)}=(p, 1-p)$ with probability one, $\boldsymbol{\alpha}^{(n)}$ converges in law to

$$(q, \quad 1-q) \tag{4.11}$$

which does not depend on $\boldsymbol{\alpha}^{(0)}$.

Two chains in above have the same Σ_1, but their behaviors are quite different. Therefore we have to distinguish aperiodic chains into two types.

DEFINITION. A stationary and irreducible M.C. with R.T.M. is *ergodic* if it is aperiodic and its limit distribution does not depend on the initial variable $\boldsymbol{\alpha}^{(0)}$.

It is easily shown that a chain is ergodic if, and only if, $\boldsymbol{P}^{(n)}$ converges in law to a random variable

$$\boldsymbol{Q}=\begin{pmatrix} q_1 \cdots q_s \\ \vdots \quad\quad \vdots \\ q_1 \cdots q_s \end{pmatrix} \tag{4.12}$$

which has the same row vectors. In section 6 we shall obtain some necessary and sufficient conditions for ergodicity.

5. Dual processes.

We shall define the "dual process" which plays an important role in ergodic theorems, and introduce some notations.

We have denoted the n-step transition probability matrix by

$$\boldsymbol{P}^{(n)}(\omega)=(p_{ij}^{(n)}(\omega))=\boldsymbol{P}^1(\omega)\cdots\boldsymbol{P}^n(\omega), \qquad \omega\in\Omega. \tag{5.1}$$

If the order of multiplications in (5.1) is reversed, the value of the matrix differs from (5.1), so we denote it by

$$\hat{\boldsymbol{P}}^{(n)}(\omega)=(\hat{p}_{ij}^{(n)}(\omega))=\boldsymbol{P}^n(\omega)\cdots\boldsymbol{P}^1(\omega), \qquad \omega\in\Omega. \tag{5.2}$$

Clearly $\hat{\boldsymbol{P}}^{(n)}$ is a Markov process, and we will call it the *dual process* (of a chain $\boldsymbol{\alpha}^{(n)}$ or $\boldsymbol{P}^{(n)}$). Similarly, we denote the n-step transition probability matrix from epoch m $(m=1, 2, \cdots)$ by

$${}^m\boldsymbol{P}^{(n)}(\omega)=({}^m p_{ij}^{(n)}(\omega))=\boldsymbol{P}^{m+1}(\omega)\cdots\boldsymbol{P}^{m+n}(\omega), \qquad \omega\in\Omega, \tag{5.3}$$

and its dual by

$${}^m\hat{\boldsymbol{P}}^{(n)}(\omega)=({}^m\hat{p}_{ij}^{(n)}(\omega))=\boldsymbol{P}^{m+n}(\omega)\cdots\boldsymbol{P}^{m+1}(\omega), \qquad \omega\in\Omega,$$ (5.4)

For a stationary M.C. with R.T.M., random variables $P^t(\omega)$ are mutually independent and have a common distribution. Hence the distributions of $P^{(n)}$, $\hat{P}^{(n)}$, ${}^mP^{(n)}$ and ${}^m\hat{P}^{(n)}$ coincide with each other. So, for any s^2-dimensional Borel set B

(5. 5) $$\Pr\{P^{(n)}\in B\}=\Pr\{\hat{P}^{(n)}\in B\}=\Pr\{{}^mP^{(n)}\in B\}=\Pr\{{}^m\hat{P}^{(n)}\in B\}.$$

We denote the maximum and the minimum of the j-th column of $P^{(n)}[\hat{P}^{(n)}]$ by

(5. 6) $$M_j^{(n)}(\omega)=\operatorname*{Max}_{1\leqq i\leqq s} p_{ij}^{(n)}(\omega) \qquad \left[\hat{M}_j^{(n)}(\omega)=\operatorname*{Max}_{1\leqq i\leqq s} \hat{p}_{ij}^{(n)}(\omega)\right]$$

and

(5. 7) $$m_j^{(n)}(\omega)=\operatorname*{Min}_{1\leqq i\leqq s} p_{ij}^{(n)}(\omega) \qquad \left[\hat{m}_j^{(n)}(\omega)=\operatorname*{Min}_{1\leqq i\leqq s} \hat{p}^{(n)}(\omega)\right]$$

respectively. Since

(5. 8) $$\hat{p}_{ij}^{(n+1)}=\sum_{k=1}^{s} p_{ik}^{n+1}\hat{p}_{kj}^{(n)},$$

we have

(5. 9) $$\hat{p}_{ij}^{(n+1)}\leqq \hat{M}_j^{(n)}\sum_{k=1}^{s} p_{ik}^{n+1}=\hat{M}_j^{(n)}$$

and similarly

(5. 10) $$\hat{p}_{ij}^{(n+1)}\geqq \hat{m}_j^{(n)}.$$

Hence we obtain the following relation:

(5. 11) $$0\leqq \hat{m}_j^{(1)}\leqq \hat{m}_j^{(2)}\leqq\cdots\leqq m_j^{(n)}\leqq\cdots\leqq \hat{M}_j^{(n)}\leqq\cdots\leqq \hat{M}_j^{(2)}\leqq \hat{M}_j^{(1)}\leqq 1.$$

Since $\{\hat{M}_j^{(n)}(\omega)\}$ and $\{\hat{m}_j^{(n)}(\omega)\}$ are bounded monotone sequences, there exist their limit variables $\hat{M}_j(\omega)$ and $\hat{m}_j(\omega)$ for each j:

(5. 12) $$\lim_{n\to\infty} \hat{m}_j^{(n)}(\omega)=\hat{m}_j(\omega)\leqq \hat{M}_j(\omega)=\lim_{n\to\infty} \hat{M}_j^{(n)}(\omega).$$

Using these $\hat{M}_j$ we define the matrix valued random variable $\hat{M}$ with the same row vectors by

(5. 13) $$\hat{M}(\omega)=\begin{pmatrix}\hat{M}_1(\omega)\cdots\hat{M}_s(\omega)\\ \vdots \\ \hat{M}_1(\omega)\cdots\hat{M}_s(\omega)\end{pmatrix}.$$

6. Ergodic theorems.

In this section we shall obtain some ergodic theorems. Theorem 5 shows the fundamental relation between the ergodicity and the convergence of the dual process, and Theorem 6 gives us a good criterion for the ergodicity. Theorem 9 also gives a necessary and sufficient condition for the ergodicity, but it might be useful to study non-ergodic chains. Theorem 8 treats the non-stationary case and states that under some mild assumptions the effect of the initial variable vanishes in the long run.

THEOREM 5. *For a stationary and irreducible M.C. with R.T.M., the following four statements are equivalent:*

(a) *The chain is ergodic.*

(b) $\boldsymbol{P}^{(n)}(\omega)$ $[\hat{\boldsymbol{P}}^{(n)}(\omega)]$ *converges in low to* $\hat{\boldsymbol{M}}(\omega)$.

(c) $\hat{\boldsymbol{P}}^{(n)}(\omega)$ *converges in probability to* $\hat{\boldsymbol{M}}(\omega)$.

(d) $\hat{\boldsymbol{P}}^{(n)}(\omega)$ *converges with probability one to* $\hat{\boldsymbol{M}}(\omega)$.

(*We note that the expression in* (b) *is justified by the relation* (5. 5).)

Proof. As stated in section 5, (a) is equivalent to

(a′) $\boldsymbol{P}^{(n)}(\omega)$ $[\hat{\boldsymbol{P}}^{(n)}(\omega)]$ converges in law to a random variable

$$Q(\omega')=\begin{pmatrix} q_1(\omega') \cdots q_s(\omega') \\ \vdots \qquad \vdots \\ q_1(\omega') \cdots q_s(\omega') \end{pmatrix} \qquad (\omega'\in\Omega') \tag{6. 1}$$

which has the same row vectors, where $(\Omega', \mathcal{F}', \Pr')$ is a certain probability space.

By the well known theorems for convergences of random variables, it is clear that (d) implies (c) and that (c) implies (b). Also it is obvious that (b) implies (a′). So we need only to show that (c) implies (d) and that (a′) implies (c).

Suppose that (c) is satisfied. Then there is a subsequence $\{\hat{\boldsymbol{P}}^{(n_k)}\}$ which converges with probability one to $\hat{\boldsymbol{M}}$, i.e., for every i, j

$$\hat{p}_{ij}^{(n_k)}\to\hat{M}_j \qquad (k\to\infty) \quad \text{w.p. 1.} \tag{6. 2}$$

Hence

$$\hat{m}_j^{(n_k)}\to\hat{M}_j \qquad (k\to\infty) \quad \text{w.p. 1.} \tag{6. 3}$$

By the monotonicity of $\{\hat{m}_j^{(n)}\}$, (6. 3) implies that

$$\hat{m}_j^{(n)}\to\hat{M}_j \qquad (n\to\infty) \quad \text{w.p. 1.} \tag{6. 4}$$

Since for every i, j

$$\hat{m}_j^{(n)}\leqq\hat{p}_{ij}^{(n)}\leqq\hat{M}_j^{(n)}, \tag{6. 5}$$

we have

$$\hat{p}_{ij}^{(n)}\to\hat{M}_j \qquad (n\to\infty) \quad \text{w.p. 1.} \tag{6. 6}$$

which is the same as (d).

Now we show that (a′) implies (c). Let I be an interval of continuity for the distribution of $\boldsymbol{Q}$ (i.e., I is open and its boundary has probability zero. See Feller [3] p. 242.), then (a′) implies that

153

$$\Pr\{\hat{P}^{(n)}\in I\}\to\Pr'\{Q\in I\}\qquad(n\to\infty). \tag{6.7}$$

We can choose a finite set of points $\{a_\nu\}$ $(\nu=0,1,\cdots,u)$ such that each a_ν is a point of continuity for each marginal distribution of q_j and that

$$a_0<0,\quad a_u>1\quad\text{and}\quad 0<a_\nu-a_{\nu-1}<\delta\qquad(\nu=1,\cdots,u) \tag{6.8}$$

for arbitrary given positive number δ. Let

$$\left.\begin{aligned} I(\nu_1,\cdots,\nu_s)=&(a_{\nu_1-1},a_{\nu_1})\times\cdots\times(a_{\nu_s-1},a_{\nu_s})\\ &\cdots\cdots\\ &\times(a_{\nu_1-1},a_{\nu_1})\times\cdots\times(a_{\nu_s-1},a_{\nu_s})\end{aligned}\right\}\ s\text{ times}. \tag{6.9}$$

Then $I(\nu_1,\cdots,\nu_s)$ $(\nu_j=1,\cdots,u)$ are intervals of continuity of the distribution of Q and they are mutually exclusive. From (6. 7) we have

$$\Pr\{\hat{P}^{(n)}\in I(\nu_1,\cdots,\nu_s)\}\to\Pr'\{Q\in I(\nu_1,\cdots,\nu_s)\}. \tag{6.10}$$

Summing up them with respect to $(\nu_1,\cdots,\nu_s)$, we obtain that

$$\begin{aligned}\Pr\Big\{\hat{P}^{(n)}\in\bigcup_{(\nu_1,\cdots,\nu_s)}I(\nu_1,\cdots,\nu_s)\Big\}&=\sum_{(\nu_1,\cdots,\nu_s)}\Pr\{\hat{P}^{(n)}\in I(\nu_1,\cdots,\nu_s)\}\\ &\to\sum_{(\nu_1,\cdots,\nu_s)}\Pr'\{Q\in I(\nu_1,\cdots,\nu_s)\}.\end{aligned} \tag{6.11}$$

We first show that the right side of (6. 11) is equal to one. Let

$$\begin{aligned} I^*(\nu_1,\cdots,\nu_s)=&(a_{\nu_1-1},a_{\nu_1})\times\cdots\times(a_{\nu_s-1},a_{\nu_s})\\ &\left.\begin{aligned}&\times(a_0,a_u)\times\cdots\times(a_0,a_u)\\ &\cdots\cdots\\ &\times(a_0,a_u)\times\cdots\times(a_0,a_u)\end{aligned}\right\}\ s-1\text{ times}\end{aligned} \tag{6.12}$$

and

$$\left.\begin{aligned} I^*=&(a_0,a_u)\times\cdots\times(a_0,a_u)\\ &\cdots\cdots\\ &\times(a_0,a_u)\times\cdots\times(a_0,a_u)\end{aligned}\right\}\ s\text{ times}. \tag{6.13}$$

Since Q has the same row vectors, we have

$$\begin{aligned}&\Pr'\{Q\in I(\nu_1,\cdots,\nu_s)\}\\ &=\Pr'\{a_{\nu_1-1}<q_1<a_\nu,\cdots,a_{\nu_s-1}<q_s<a_{\nu_s}\}\\ &=\Pr'\{Q\in I^*(\nu_1,\cdots,\nu_s)\}.\end{aligned} \tag{6.14}$$

Hence we obtain the desired result as follows:

$$\begin{aligned}
&\sum_{(\nu_1,\cdots,\nu_s)} \Pr'\{\boldsymbol{Q}\in I(\nu_1,\cdots,\nu_s)\}\\
&=\sum_{(\nu_1,\cdots,\nu_s)} \Pr'\{\boldsymbol{Q}\in I^*(\nu_1,\cdots,\nu_s)\}\\
&=\Pr'\left\{\boldsymbol{Q}\in \bigcup_{(\nu_1,\cdots,\nu_s)} I^*(\nu_1,\cdots,\nu_s)\right\}\\
&=\Pr'\{\boldsymbol{Q}\in I^*\}\\
&=\Pr'\{0\leqq \boldsymbol{q}_j\leqq 1 \quad \text{for all } j\}\\
&=1,
\end{aligned} \tag{6.15}$$

that is

$$\sum_{(\nu_1,\cdots,\nu_s)} \Pr'\{\boldsymbol{Q}\in I(\nu_1,\cdots,\nu_s)\}=1. \tag{6.16}$$

Next we calculate the left side of (6. 11), then

$$\begin{aligned}
&\Pr\left\{\hat{\boldsymbol{P}}^{(n)}\in \bigcup_{(\nu_1,\cdots,\nu_s)} I(\nu_1,\cdots,\nu_s)\right\}\\
&=\Pr\left\{\bigcup_{(\nu_1,\cdots,\nu_s)} (a_{\nu_j-1}<\hat{p}_{ij}^{(n)}<a_{\nu_j} \quad \text{for all } i,j)\right\}\\
&=\Pr\left\{\bigcup_{(\nu_1,\cdots,\nu_s)} (a_{\nu_j-1}<\hat{m}_j^{(n)} \quad \text{and} \quad a_{\nu_j}>\hat{M}_j^{(n)} \quad \text{for all } j)\right\}\\
&\leqq\Pr\left\{\bigcup_{(\nu_1,\cdots,\nu_s)} (\hat{M}_j^{(n)}-\hat{m}_j^{(n)}<\delta \quad \text{for all } j)\right\}\\
&=\Pr\{\hat{M}_j^{(n)}-\hat{m}_j^{(n)}<\delta \quad \text{for all } j\}\\
&\leqq\Pr\{|\hat{p}_{ij}^{(n)}-\hat{M}_j^{(n)}|<\delta \quad \text{for all } i,j\}.
\end{aligned} \tag{6.17}$$

Combining (6. 16) and (6. 17) with (6. 11), we obtain that for any positive number δ

$$\Pr\{|\hat{p}_{ij}^{(n)}-\hat{M}_j^{(n)}|<\delta \quad \text{for all } i,j\}\to 1 \quad (n\to\infty). \tag{6.18}$$

If we denote the length of a vector P in s^2-dimensional Euclidean space by $||P||$, we have

$$\begin{aligned}
&\Pr\{||\hat{\boldsymbol{P}}^{(n)}-\hat{\boldsymbol{M}}||>\delta\}\\
&\leqq \sum_{i,j=1}^{s} \Pr\left\{(\hat{p}_{ij}^{(n)}-\hat{M}_j)^2>\frac{\delta^2}{s^2}\right\}\\
&=\sum_{i,j=1}^{s} \Pr\left\{|\hat{p}_{ij}^{(n)}-\hat{M}_j|>\frac{\delta}{s}\right\}.
\end{aligned} \tag{6.19}$$

155

For any positive number δ, each term in the last summation in (6. 19) tends to zero as $n\to\infty$, so we have proved that (a′) implies (c). Q.E.D.

THEOREM 6. *A stationary and irreducible M.C. with R.T.M. is ergodic if, and only if,*

$$\Pr\{\hat{m}_j(\omega)>0\}>0 \qquad \textit{for some } j, \tag{6. 20}$$

or equivalently if, and only if, for some j and N

$$\Pr\{\hat{p}_{ij}^{(N)}(\omega)>0 \quad \textit{for all } i\}=\Pr\{p_{ij}^{(N)}(\omega)>0 \quad \textit{for all } i\}>0. \tag{6. 21}$$

Theorem 6 is an easy corollary of Theorem 8 or of Theorem 9, but the proof of Theorem 8 is complicated while the proof of Theorem 6 is rather simpler by using the dual processes. The structures of both proofs are similar to each other, hence we shall prove Theorem 6 first and then modify it for Theorem 8. To prove these theorems we need the following lemma.

LEMMA 7. *Let $P=(p_{ij})$, $Q=(q_{ij})$ and $R=QP=(r_{ij})$ be stochastic matrices (i.e., they are elements of $\mathcal{P}$). We denote the maximum and the minimum of the j-th column of P by*

$$M_j=\operatorname*{Max}_{1\leqq i\leqq s} p_{ij} \qquad \textit{and} \qquad m_j=\operatorname*{Min}_{1\leqq i\leqq s} p_{ij} \tag{6. 22}$$

and similarly those of R by

$$M_i'=\operatorname*{Max}_{1\leqq i\leqq s} r_{ij} \qquad \textit{and} \qquad m_j'=\operatorname*{Min}_{1\leqq i\leqq s} r_{ij}. \tag{6. 23}$$

If for some j_0 there is a number $\delta>0$ such that

$$q_{ij_0}>\delta \tag{6. 24}$$

for every i, then

$$M_j'-m_j'\leqq\left(1-\frac{\delta}{2}\right)(M_j-m_j) \tag{6. 25}$$

for every j.

Proof of Lemma 7. Through this proof, j is arbitrarily fixed. When $M_j=m_j$, (6. 25) is trivial, because $m_j\leqq m_j'\leqq M_j'\leqq M_j$ as in (5. 11). Hence we may assume that $M_j>m_j$. Let J be a subset of S defined by

$$J=\left\{i\,\middle|\,p_{ij}\leqq\frac{1}{2}(M_j+m_j)\right\}. \tag{6. 26}$$

Since

$$r_{ij}=\sum_{k=1}^{s} q_{ik}p_{kj}, \tag{6. 27}$$

we have

(6. 28)

Similarly,

(6. 29)

Subtracti

(6. 30)

for if $j_0\in$

Proof

may supp

(6. 31)

for every

(6. 32)

Since

(6. 33)

for every

(6. 34)

and so w

Next

Theorem

and a pos

(6. 35)

we have for some i

$$M'_j = r_{ij} = \sum_{k\in J} q_{ik} p_{kj} + \sum_{k\notin J} q_{ik} p_{kj}$$

(6. 28)
$$\leqq M_j \sum_{k\in J} q_{ik} + \frac{1}{2}(M_j + m_j) \sum_{k\notin J} q_{ik}$$

$$= M_j - \frac{1}{2}(M_j - m_j) \sum_{k\notin J} q_{ik}.$$

Similarly, for some i' we have

(6. 29)
$$m'_j = r_{i'j} \geqq m_j + \frac{1}{2}(M_j - m_j) \sum_{k\in J} q_{i'k}.$$

Subtracting (6. 29) from (6. 28) we obtain the desired result:

(6. 30)
$$M'_j - m'_j \leqq (M_j - m_j)\left(1 - \frac{1}{2}\sum_{k\in J} q_{ik} - \frac{1}{2}\sum_{k\notin J} q_{i'k}\right) < (M_j - m_j)\left(1 - \frac{\delta}{2}\right),$$

for if $j_0 \in J$ then $\sum_{k\in J} q_{i'k} > \delta$ and if $j_0 \notin J$ then $\sum_{k\notin J} q_{ik} > \delta$. Q.E.D.

Proof of Theorem 6. First we shall prove the necessity. By Theorem 5 we may suppose that

(6. 31)
$$\hat{p}_{ij}^{(n)} \to \hat{M}_j \qquad (n\to\infty) \quad \text{w.p. 1.}$$

for every i, j and so we have

(6. 32)
$$\hat{m}_j = \hat{M}_j \qquad \text{w.p. 1.}$$

Since

(6. 33)
$$\sum_{j=1}^{s} \hat{p}_{ij}^{(n)} = 1$$

for every i and n, (6. 31) and (6. 32) implies that

(6. 34)
$$\sum_{j=1}^{s} \hat{m}_j = 1 \qquad \text{w.p. 1.}$$

and so we conclude (6. 20).

Next we shall prove the sufficiency by showing that (6. 21) implies (c) in Theorem 5. We may restate (6. 21) as follows; for some j there is an integer N and a positive number δ such that

(6. 35)
$$\Pr\{\hat{p}_{ij}^{(N)} > \delta \quad \text{for all } i\} > 0.$$

157

For these N and δ, let

(6. 36) $$A_k=\{\omega\in\Omega \mid {}^{kN}\hat{p}_{ij}^{(N)}>\delta \ \text{ for all } i\}, \qquad (k=1, 2, 3, \cdots).$$

Then A_k are mutually independent events and moreover from (5. 5) and (6. 35) we have

(6. 37) $$\sum_{k=1}^{\infty}\Pr\{A_k\}=\infty.$$

Therefore using Borel-Cantelli lemma, it follows that

(6. 38) $$\Pr\{A_k \text{ occurs infinitely often}\}=1.$$

In other words, if the random variable $K^{(n)}(\omega)$ denotes the number of occurrences within the n events $\{A_1, A_2, \cdots, A_n\}$, for each integer r

(6. 39) $$\Pr\{K^{(n)}\geqq r\}\to 1 \qquad (n\to\infty).$$

We divide the whole space into 2^n events as

(6. 40) $$\Omega=\bigcup_{(i_1,\cdots,i_n)}\left\{\bigcap_{k=1}^{n} B_{ki_k}\right\}$$

where $i_k=0$ or 1, and $B_{k0}=A_k$, $B_{k1}=A_k^c$. Then the number of zeros in $\{i_k, k=1, \cdots, n\}$ is equal to the value of $K^{(n)}$. If $\omega\in B_{k0}=A_k$ then by Lemma 7 we have

(6. 41) $$\hat{M}_j^{((k+1)N)}(\omega)-\hat{m}_j^{((k+1)N)}(\omega)\leqq\left(1-\frac{\delta}{2}\right)(\hat{M}_j^{(kN)}(\omega)-\hat{m}_j^{(kN)}(\omega)).$$

Because, we may replace P, Q and R in Lemma 7 by $\hat{\boldsymbol{P}}^{(kN)}$, ${}^{kN}\hat{\boldsymbol{P}}^{(N)}$ and $\hat{\boldsymbol{P}}^{((k+1)N)}$ respectively.

Now we use the inequality (6. 41) for $\omega\in B_{k0}$ and use the inequality

(6. 42) $$\hat{M}_j^{((k+1)N)}(\omega)-\hat{m}_j^{((k+1)N)}(\omega)\leqq\hat{M}_j^{(kN)}(\omega)-\hat{m}_j^{(kN)}(\omega)$$

for $\omega\in B_{k1}$, then for $\omega\in\{K^{(n)}\geqq r\}$ we have

(6. 43) $$\hat{M}_j^{((k+1)N)}(\omega)-\hat{m}_j^{((k+1)N)}\leqq\left(1-\frac{\delta}{2}\right)^r.$$

For each $\delta'>0$ there is an integer r such that $(1-\delta/2)^r<\delta'$, and therefore for arbitrarily small δ' we have

(6. 44) $$\Pr\{\hat{M}_j^{(nN)}-\hat{m}_j^{(nN)}<\delta'\}\geqq\Pr\{K^{(n)}\geqq r\}\to 1 \qquad (n\to\infty).$$

By the monotonicity of $\hat{M}_j^{(n)}$ and $\hat{m}_j^{(n)}$, (6. 44) implies that

(6. 45) $$\Pr\{\hat{M}_j^{(n)}-\hat{m}_j^{(n)}<\delta'\}\to 1 \qquad (n\to\infty).$$

So we have

158

$$\Pr\{|\hat{p}_{ij}^{(n)}-\hat{M}_j|<\delta'\}\to 1 \qquad (n\to\infty). \tag{6.46}$$

Repeating the discussion at the last paragraph in the proof of Theorem 5, we complete the proof.

Next we shall consider the non-stationary case where in general no limit exist. However, by generalizing Theorem 6, we can show that the effect of the initial variable vanishes in the long run.

THEOREM 8. *If there is an increasing sequence $\{n_k\}$ of integers and a positive number δ such that*

$$\sum_{k=1}^{\infty}\Pr\{{}^{n_k}p_{ij}^{(n_{k+1}-n_k)}>\delta \quad \textit{for some } j \textit{ and all } i\}=\infty, \tag{6.47}$$

then for any positive number ε

$$\Pr\{M_j^{(n)}(\omega)-m_j^{(n)}(\omega)<\varepsilon \quad \textit{for every } j\}\to 1 \qquad (n\to\infty). \tag{6.48}$$

Proof. For a non-stationary M.C. with R.T.M., the basic relation (5.5) does not hold and so we cannot use the concept of the dual process. Hence we shall define a substitutional process with which this proof can be done in parallel with the proof of Theorem 6.

Select a large integer L and define random variables $\tilde{P}_L^t(\omega)$ $(t=1,\cdots,L)$ by

$$\tilde{P}_L^t(\omega)=P^{L+1-t}(\omega), \qquad \omega\in\Omega \tag{6.49}$$

and a process $\{\tilde{P}_L^{(n)}\}$ by

$$\tilde{P}_L^{(n)}(\omega)=\tilde{P}_L^n(\omega)\cdots\tilde{P}_L^1(\omega)={}^{L-n}P^{(n)}(\omega) \qquad (n=1,\cdots,L). \tag{6.50}$$

We will use the same symbols with tilde and L, instead of hat, for corresponding variables as those in the dual process.

Let us denote the events in the braces of (6.47) by

$$C_k=\{\omega\in\Omega\,|\,{}^{L-n_{k+1}}\tilde{p}_{ij}^{(n_{k+1}-n_k)}(\omega)>\delta \quad \text{for some } j \text{ and all } i\} \tag{6.51}$$

for $n_{k+1}\leqq L$, then $\{C_k;\, n_{k+1}\leqq L\}$ are mutually independent. Hence by (6.47), Borel-Cantelli lemma assures that

$$\Pr\{C_k \text{ occurs infinitely often}\}=1 \tag{6.52}$$

or if the random variable $\tilde{K}^{(n)}(\omega)$ denotes the number of occurrences within the n events $\{C_1, C_2, \cdots, C_n\}$, then for each integer r

$$\Pr\{\tilde{K}^{(n)}\geqq r\}\to 1 \qquad (n\to\infty). \tag{6.53}$$

Next, let $k_0=\text{Max}\{k\,|\,n_{k+1}\leqq L\}$ and divide Ω as

$$\Omega=\bigcup_{(i_1,\cdots,i_n)}\left\{\bigcap_{k=1}^{k_0}\tilde{B}_{ki_k}\right\} \tag{6.54}$$

159

where $i_k=0$ or 1, and $\tilde{B}_{k0}=C_k$, $\tilde{B}_{k1}=C_k^c$. For $\omega\in C_k$, by Lemma 7,

$$\tilde{M}_{Lj}^{(L-n_k)}(\omega)-\tilde{m}_{Lj}^{(L-n_k)}(\omega)\leqq\left(1-\frac{\delta}{2}\right)(\tilde{M}_{Lj}^{(L-n_{k+1})}(\omega)-\tilde{m}_{Lj}^{(L-n_{k+1})}(\omega)). \tag{6.55}$$

Because, we may replace P, Q and R in Lemma 7 by $\tilde{P}_L^{(L-n_{k+1})}$, ${}_{L-n_{k+1}}\tilde{P}_L^{(n_{k+1}-n_k)}$ and $\tilde{P}_L^{(L-n_k)}$ respectively. Therefore, if we use the inequality (6. 55) for $\omega\in\tilde{B}_{k0}$ and use the inequality

$$\tilde{M}_{Lj}^{(L-n_k)}(\omega)-\tilde{m}_{Lj}^{(L-n_k)}(\omega)\leqq\tilde{M}_{Lj}^{(L-n_{k+1})}(\omega)-\tilde{m}_{Lj}^{(L-n_{k+1})}(\omega) \tag{6.56}$$

for $\omega\in\tilde{B}_{k1}$, then for $\omega\in\{\tilde{K}^{(k_0)}\geqq r\}$ we have

$$\begin{aligned}M_j^{(L)}(\omega)-m_j^{(L)}(\omega)&=\tilde{M}_{Lj}^{(L)}(\omega)-\tilde{m}_{Lj}^{(L)}(\omega)\\&\leqq(\tilde{M}_{Lj}^{(L-n_{k_0})}(\omega)-\tilde{m}_{Lj}^{(L-n_{k_0})}(\omega))\left(1-\frac{\delta}{2}\right)^r\\&\leqq\left(1-\frac{\delta}{2}\right)^r.\end{aligned} \tag{6.57}$$

For each $\delta'>0$, there is an integer r such that $(1-\delta/2)^r<\delta'$. Hence we have

$$\Pr\{M_j^{(L)}-m_j^{(L)}<\delta' \quad \text{for all } j\}\geqq\Pr\{\tilde{K}^{(k_0)}\geqq r\}. \tag{6.58}$$

By (6. 53), for each $\varepsilon>0$, there is an integer k' such that for $k\geqq k'$

$$\Pr\{\tilde{K}^{(k')}\geqq r\}>1-\varepsilon. \tag{6.59}$$

Therefore for any $L(\geqq n_{k'})$

$$\Pr\{M_j^{(L)}-m_j^{(L)}<\delta' \quad \text{for all } j\}>1-\varepsilon \tag{6.60}$$

which shows (6. 48).

Now we shall prove one more theorem which is useful to examine the structure of a non-ergodic, stationary and irreducible chain.

THEOREM 9. *A stationary and irreducible M.C. with R.T.M. is ergodic if, and only if, for each pair of subsets I and K of S ($I\neq\phi$, $K\neq\phi$ and $I\cap K=\phi$)*

$$\Pr\{\textit{there is an } n=n(\omega) \textit{ such that } \hat{p}_{ik}^{(n)}(\omega)=0 \textit{ for } i\in I,\ k\notin I \textit{ and for } i\in K,\ k\notin K\}<1. \tag{6.61}$$

Proof. We first prove the sufficiency in several steps.

(i) Since the number of possible pairs (I, K) is finite, (6. 61) implies that there are positive numbers δ and γ such that for each pair (I, K)

$$\Pr\{\text{there is an } n=n(\omega) \text{ such that } \hat{p}_{ik}^{(n)}(\omega)<\delta \text{ for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\}<1-\gamma. \tag{6.62}$$

For arbitr
the followi
(6. 63)
and
(6. 64)
We put
(6. 65)
and
(6. 66)
and define
(6. 67)
and
(6. 68)
(ii) F
the sequen
(6. 69)
and
(6. 70)
Then if n_{l+1}
(6. 71)
for $i\in I$, $k\notin$
(6. 72)

160

For arbitrarily fixed j, let $\{q_{ij}\}$ $(i=1, \cdots, s)$ be a set of rational numbers satisfying the following conditions:

$$0 \leqq q_{ij} \leqq 1 \qquad i=1, \cdots, s \tag{6.63}$$

and

$$\operatorname*{Max}_{1\leqq i\leqq s} q_{ij}=M>m=\operatorname*{Min}_{1\leqq i\leqq s} q_{ij}. \tag{6.64}$$

We put

$$I=\{i \mid q_{ij}=M\} \tag{6.65}$$

and

$$K=\{i \mid q_{ij}=m\} \tag{6.66}$$

and define the positive numbers d and ε by

$$d=\operatorname{Min}\left[\operatorname*{Min}_{i\notin I}(M-q_{ij}), \operatorname*{Min}_{i\notin K}(q_{ij}-m)\right] \tag{6.67}$$

and

$$\varepsilon=\frac{d\delta}{2}. \tag{6.68}$$

(ii) From now on we concentrate on the j-th column of $\hat{\boldsymbol{P}}^{(n)}(\omega)$. We define the sequence of random times $\boldsymbol{n}_t(\omega)$ inductively by

$$\boldsymbol{n}_1(\omega)=\begin{cases}\operatorname{Min}\{n \mid |\hat{p}_{ij}^{(n)}(\omega)-q_{ij}|<\varepsilon \quad \text{for all } i\},\\ \infty \quad \text{if the set in above braces is empty,}\end{cases} \tag{6.69}$$

and

$$\boldsymbol{n}_t(\omega)=\begin{cases}\operatorname{Min}\{n>\boldsymbol{n}_{t-1}(\omega) \mid |\hat{p}_{ij}^{(n)}(\omega)-q_{ij}|<\varepsilon \quad \text{for all } i\},\\ \infty \quad \text{if } \boldsymbol{n}_{t-1}(\omega)=\infty \text{ or if } \boldsymbol{n}_{t-1}(\omega)<\infty \text{ and the}\\ \qquad \text{set in above braces is empty.}\end{cases} \tag{6.70}$$

Then if $\boldsymbol{n}_{t+1}(\omega)<\infty$, we have

$${}^{\boldsymbol{n}_t}\hat{p}_{ik}^{(\boldsymbol{n}_{t+1}-\boldsymbol{n}_t)}<\frac{2\varepsilon}{d}=\delta \tag{6.71}$$

for $i\in I$, $k\notin I$ and for $i\in K$, $k\notin K$. In fact, since

$$\hat{p}_{ij}^{(\boldsymbol{n}_{t+1})}=\sum_{k=1}^{s} {}^{\boldsymbol{n}_t}\hat{p}_{ik}^{(\boldsymbol{n}_{t+1}-\boldsymbol{n}_t)}\hat{p}_{kj}^{(\boldsymbol{n}_t)}, \tag{6.72}$$

for $i\in I$ and $k_0\notin I$ we have

$$
\begin{aligned}
M-\varepsilon<\hat{p}_{ij}^{(n_{t+1})}&<\sum_{k=1}^{s}{}^{n_t}\hat{p}_{ik}^{(n_{t+1}-n_t)}(q_{kj}+\varepsilon)\\
&=\varepsilon+\sum_{k\neq k_0}q_{kj}\hat{p}_{jk}^{(n_{t+1}-n_t)}+q_{k_0j}{}^{n_t}\hat{p}_{ik_0}^{(n_{t+1}-n_t)}\\
&\leqq\varepsilon+M(1-{}^{n_t}\hat{p}_{ik_0}^{(n_{t+1}-n_t)})+q_{k_0j}{}^{n_t}\hat{p}_{ik_0}^{(n_{t+1}-n_t)}\\
&=\varepsilon+M-(M-q_{k_0j}){}^{n_t}\hat{p}_{ik_0}^{(n_{t+1}-n_t)}.
\end{aligned}
\tag{6.73}
$$

Hence for $i\in I$, $k_0\notin I$ we have

$$
{}^{n_t}\hat{p}_{ik_0}^{(n_{t+1}-n_t)}<\frac{2\varepsilon}{M-q_{k_0j}}\leqq\frac{2\varepsilon}{d}=\delta. \tag{6.74}
$$

Similarly the same relation holds for $i\in K$, $k_0\notin K$.

(iii) Next we shall show that

$$
\Pr\{n_{t+1}(\omega)<\infty\}<(1-\gamma)^t. \tag{6.75}
$$

We may divide the event $\{n_{t+1}(\omega)<\infty\}$ into disjoint events as

$$
\{n_{t+1}(\omega)<\infty\}=\bigcup_{\nu<\infty}\bigcup_{\mu<\infty}\{n_t(\omega)=\nu,\ n_{t+1}(\omega)=\nu+\mu\}. \tag{6.76}
$$

By the result obtained in (ii) we have

$$
\{n_t(\omega)=\nu,\ n_{t+1}(\omega)=\nu+\mu\}\subset\{{}^{\nu}\hat{p}_{ik}^{(\mu)}(\omega)<\delta\quad\text{for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\}, \tag{6.77}
$$

while

$$
\{n_t(\omega)=\nu,\ n_{t+1}(\omega)=\nu+\mu\}\subset\{n_t(\omega)=\nu\}. \tag{6.78}
$$

Therefore

$$
\begin{aligned}
&\{n_t(\omega)=\nu,\ n_{t+1}(\omega)=\nu+\mu\}\\
&\subset\{{}^{\nu}\hat{p}_{ik}^{(\mu)}(\omega)<\delta \text{ for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\}\cap\{n_t(\omega)=\nu\}.
\end{aligned}
\tag{6.79}
$$

The event $\{n_t=\nu\}$ depends only on the fraction $(\omega^1,\cdots,\omega^\nu)$ of $\omega=(\omega^0,\omega^1,\omega^2,\cdots)$, and the event $\{{}^{\nu}\hat{p}_{ik}^{(\mu)}<\delta$ for $i\in I$, $k\notin I$ and for $i\in K$, $k\notin K\}$ depends on the fraction $(\omega^{\nu+1},\cdots,\omega^{\nu+\mu})$. So they are independent. Hence we have

$$
\begin{aligned}
&\Pr\{n_t=\nu,\ n_{t+1}=\nu+\mu\}\\
&\leqq\Pr\{{}^{\nu}\hat{p}_{ik}^{(\mu)}<\delta \text{ for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\}\cdot\Pr\{n_t=\nu\}.
\end{aligned}
\tag{6.80}
$$

Therefore

$$
\begin{aligned}
\Pr\{\boldsymbol{n}_{t+1}<\infty\} &= \sum_{\nu<\infty}\sum_{\mu<\infty}\Pr\{\boldsymbol{n}_t=\nu,\ \boldsymbol{n}_{t+1}=\nu+\mu\} \\
&\leqq \sum_{\nu<\infty}\sum_{\mu<\infty}\Pr\{\boldsymbol{n}_t=\nu\}\Pr\{{}^{\nu}\hat{\boldsymbol{p}}_{ik}^{(\mu)}<\delta \text{ for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\} \\
&= \sum_{\nu<\infty}\Pr\{\boldsymbol{n}_t=\nu\}\cdot\Pr\{\text{there is an } \boldsymbol{n}(\omega)>\nu \text{ such that } {}^{\nu}\hat{\boldsymbol{p}}_{ik}^{(\boldsymbol{n})}<\delta \\
&\qquad\qquad \text{for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\} \\
&= \sum_{\nu<\infty}\Pr\{\boldsymbol{n}_t=\nu\}\cdot\Pr\{\text{there is an } \boldsymbol{n}(\omega)>0 \text{ such that } \hat{\boldsymbol{p}}_{ik}^{(\boldsymbol{n})}<\delta \\
&\qquad\qquad \text{for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\} \\
&< (1-\delta)\sum_{\nu<\infty}\Pr\{\boldsymbol{n}_t=\nu\}=(1-\delta)\Pr\{\boldsymbol{n}_t<\infty\}.
\end{aligned}
\tag{6.81}
$$

Using this relation repeatedly we obtain the desired result (6.75).

(iv) If $\hat{\boldsymbol{p}}_{ij}^{(n)}(\omega)$ $(i=1,\cdots,s)$ has an accumulation point in the interval $(q_{ij}-\varepsilon/2,$ $q_{ij}+\varepsilon/2)$ $(i=1,\cdots,s)$, then $\hat{\boldsymbol{p}}_{ij}^{(n)}(\omega)$ visits the interval $(q_{ij}-\varepsilon,\ q_{ij}+\varepsilon)$ $(i=1,\cdots,s)$ infinitely often. Therefore if $A(q_{ij})$ denotes the event that $\hat{\boldsymbol{p}}_{ij}^{(n)}(\omega)$ has an accumulation point in the interval $(q_{ij}-\varepsilon/2,\ q_{ij}+\varepsilon/2)$ $(i=1,\cdots,s)$, then for every t

$$A(q_{ij})\subset\{\boldsymbol{n}_t<\infty\}. \tag{6.82}$$

Hence

$$\Pr\{A(q_{ij})\}\leqq\lim_{t\to\infty}\Pr\{\boldsymbol{n}_t<\infty\}\leqq\lim_{t\to\infty}(1-\gamma)^{t-1}=0. \tag{6.83}$$

(v) Now we shall consider the case in which the chain is not ergodic. Let B be the event that $\hat{\boldsymbol{P}}^{(n)}(\omega)$ does not converge to $\hat{\boldsymbol{M}}(\omega)$. For every $\omega\in B$, there exists some j and $\{q_{ij}\}$ such that $\omega\in A(q_{ij})$. Therefore

$$B\subset\bigcup_{j=1}^{s}\bigcup_{\{q_{ij}\}}A\{q_{ij}\}. \tag{6.84}$$

Since q_{ij} are rational numbers, the number of possible $\{q_{ij}\}$ is countable. Therefore $\Pr\{B\}=0$. This completes the proof of the sufficiency.

Now we shall prove the necessity. Suppose that there exists a pair of subsets (I,K) of S $(I\neq\phi,\ K\neq\phi$ and $I\cap K=\phi)$ with which

$$
\begin{aligned}
&\Pr\{\text{there is an } \boldsymbol{n}=\boldsymbol{n}(\omega) \text{ such that } \hat{\boldsymbol{p}}_{ik}^{(\boldsymbol{n})}(\omega)=0 \\
&\qquad \text{for } i\in I,\ k\notin I \text{ and for } i\in K,\ k\notin K\}=1.
\end{aligned}
\tag{6.85}
$$

Let D $(\subset\mathcal{P})$ be the set of all $P=(p_{ij})$ such that $p_{ij}=0$ for $i\in I$, $k\notin I$ and for $i\in K$, $k\notin K$. Then the assumption is

$$\Pr\{\text{there is an } \boldsymbol{n}=\boldsymbol{n}(\omega) \text{ such that } \hat{\boldsymbol{P}}^{(\boldsymbol{n})}\in D\}=1, \tag{6.86}$$

If $\hat{P}^{(n)} \in D$ and ${}^{n}\hat{P}^{(m)} \in D$, then it is easily shown that $\hat{P}^{(n+m)} \in D$. Hence if we define that $t = \mathrm{Min}\{n \mid \hat{P}^{(n)} \in D\}$ then

$$\begin{aligned} &\{\text{there exist at least two } n\text{'s such that } \hat{P}^{(n)} \in D\} \\ &= \bigcup_{\nu<\infty} \{t=\nu, \text{ there exists an } n \text{ such that } {}^{\nu}\hat{P}^{(n)} \in D\}. \end{aligned} \tag{6.87}$$

Therefore we have

$$\begin{aligned} &\Pr\{\text{there exist at least two } n\text{'s such that } \hat{P}^{(n)} \in D\} \\ &= \sum_{\nu<\infty} \Pr\{t=\nu\} \cdot \Pr\{\text{there exists an } n \text{ such that } {}^{\nu}\hat{P}^{(n)} \in D\} \\ &= \sum_{\nu<\infty} \Pr\{t=\nu\} \cdot \Pr\{\text{there exists an } n \text{ such that } \hat{P}^{(n)} \in D\} \\ &= \sum_{\nu<\infty} \Pr\{t=\nu\} \cdot 1 \\ &= \Pr\{\text{there exists an } n \text{ such that } \hat{P}^{(n)} \in D\} = 1. \end{aligned} \tag{6.88}$$

Similarly for each m we can show that

$$\Pr\{\text{there exist at least } m \text{ epochs } (n\text{'s}) \text{ such that } \hat{P}^{(n)} \in D\} = 1. \tag{6.89}$$

Therefore for every N, there is an $n > N$ with probability one such that $\hat{P}^{(n)} \in D$. Hence by the monotonicity of $\hat{m}_j^{(n)}$ we have

$$\hat{m}_j^{(N)} \leqq \hat{m}_j^{(n)} \leqq \hat{p}_{ij}^{(n)} = 0 \tag{6.90}$$

for $n = n(\omega) > N$ with which $\hat{P}^{(n)} \in D$ and for $i \in I$, $j \notin I$ and $i \in K$, $j \notin K$. Since $I^c \cup K^c = S$, we have for each j

$$\hat{m}_j^{(N)} = 0 \tag{6.91}$$

with probability one, and as N is arbitrary, this implies that

$$\hat{m}_j = \lim_{N\to\infty} \hat{m}_j^{(N)} = 0. \tag{6.92}$$

On the other hand, for every N and i

$$\sum_{j=1}^{s} \hat{M}_j^{(N)} \geqq \sum_{j=1}^{s} \hat{p}_{ij}^{(N)} = 1. \tag{6.93}$$

Therefore for each ω at least one $\hat{M}_j = \lim_{N\to\infty} \hat{M}_j^{(N)}$ is strictly positive, and this contradicts to (d) in Theorem 5.

ACKNOWLEDGEMENT. The author wishes to thank Professor H. Morimura for the suggestion and the guidance.

References

[1] EHRENBERG, A. S. C., An appraisal of Markov brand-switching models. Journ. of Marketing Research 2 (1965), 347-362.

[2] FELLER, W., An introduction to probability theory and its applications. Vol. 1 (3rd ed.), John Wiley and Sons, Inc., New York (1968).

[3] FELLER, W., An introduction to probability theory and its applications. Vol 2, John Wiley and Sons, Inc., New York (1966).

DEPARTMENT OF APPLIED PHYSICS,
TOKYO INSTITUTE OF TECHNOLOGY.

APPENDIX 2

Journal of the
Operations Research
Society of Japan

VOLUME 14 November 1971 NUMBER 3 & 4

QUEUEING SYSTEMS WITH GATES

YUKIO TAKAHASHI

Tokyo Institute of Technology

(Received January 29, 1971 and Revised March 19, 1971)

1. Introduction

In this paper we study $M/M/r$ type queueing systems with gates. In these models, customers arrive at a system in a Poisson process at a rate λ and have exponentially distributed service times with mean $1/\mu$. A customer who arrives at the system, at first joins in the first queue at the gate, and when the gate opens he goes to the second queue at the counter. Servers serve customers in the second queue only. The gate closes immediately after all the customers in the first queue go to the second queue.

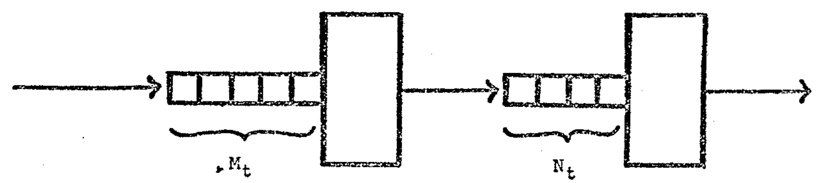


Fig. 1. A queueing system with a gate

Many types of gating rules can be considered. They may depend on the queue sizes, the total service time of the customers in the second queue, etc. But in this paper the gating rule is restricted to two types; (i) the deterministic gating rule, under which the gating intervals are constant and (ii) the exponential gating rule, under which the gating intervals are independent random variables with a common exponential distribution.

Queueing systems with the deterministic gating rule are similar to $D/M/r$ queueing systems with group arrivals and they can be analyzed with the same ideas. However queueing systems with gates generally cannot be analyzed by the ordinary techniques used for the analyses of queueing systems with group arrivals. The reason is that, in a queueing system with a gate, except for the case with the deterministic gating rule, the number of customers arriving at the system in a gating interval depends on the length of the gating interval. By the same reason, we can use queueing systems with gates as more suitable models than bulk queues for some congestion problems in a complex time sharing computer system.

We study single server queues with the exponential gating rule in Section 2, and with the deterministic gating rule in Section 3. The methods used in Section 2 are extended in Section 4 to study many server queues with the exponential gating rule.

2. Single Server Queues with the Exponential Gating Rule

2.1 Limiting Joint Distribution of Queue Sizes

Let λ, μ and η $(0<\lambda, \mu, \eta<\infty)$ be the parameters of the exponential distributions of inter-arrival times, service times and gating intervals, respectively. Denote the number of customers in the first queue at time t by M_t, and the number of customers in the second queue including ones being served at time t by N_t. Then the joint probability $Pr\{M_t=m, N_t=n\}$ $(m, n=0, 1, 2, \cdots)$ converges to a limit $p(m, n)$ as t tends to infinity, which is independent of the initial value (M_0, N_0) of the process. There are two cases; (i) $p(m, n)=0$ for all m and n, or (ii) $p(m, n)$ is the unique stationary initial distribution satisfying the system of equations

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p(m, n)=1 \tag{2.1}$$

and

$$\begin{aligned} 0=&-(\lambda+\mu+\eta)\, p(m, n)+\mu\,\{p(m, n+1)+\delta(n)\, p(m, n)\} \\ &+\lambda\,\{1-\delta(m)\}\, p(m-1, n)+\eta\delta(m)\sum_{k=0}^{n} p(k, n-k) \\ &\qquad (m, n=0, 1, 2, \cdots), \end{aligned} \tag{2.2}$$

where $\delta(0)=1$ and $\delta(x)=0$ for $x\neq 0$. As proved later, if $\lambda<\mu$ this system of equations has a unique solution, and if $\lambda\geqq\mu$ the only non-negative solution of (2.2) is $p(m, n)=0$ for all m and n, and it does not satisfy (2.1).

This system of equations can be expressed in equations of the generating function of $p(m, n)$. Denoting the generating function by

$$F(x, y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} x^m y^n\, p(m, n) \qquad (0\leqq x, y\leqq 1), \tag{2.3}$$

(2.1) becomes as $F(1, 1)=1$, and (2.2) can be written as

$$\begin{aligned} 0=&-(\lambda+\mu+\eta)\, F(x, y)+\mu\left\{\frac{1}{y}[F(x, y)-F(x, 0)]+F(x, 0)\right\} \\ &+\lambda x F(x, y)+\eta F(y, y). \end{aligned} \tag{2.4}$$

Rewriting (2.4) we obtain that

$$\begin{aligned} \{\lambda xy-(\lambda+\mu+\eta)\, y+\mu\}\, F(x, y)+\eta y F(y, y) \\ =\mu(1-y)\, F(x, 0). \end{aligned} \tag{2.5}$$

This equation includes $F(x, y)$ in three different forms; $F(x, y)$, $F(y, y)$ and $F(x, 0)$. To solve the equation, therefore, we have to eliminate two of them.

166

If we take $y=x$ and divide both sides by $(1-x)$ in (2.5), then we have a relation between $F(x, 0)$ and $F(x, x)$;

$$\mu F(x, 0)=(\mu-\lambda x)\ F(x, x). \tag{2.6}$$

If $\lambda>\mu$, for (2.6) being valid at $x=1$, $F(1, 1)$ must be zero. If $\lambda=\mu$, taking $x=1$ in (2.6), we have $F(1, 0)=0$. The monotonicity of $F(x, y)$ implies that $F(x, 0)=0$ for $0\leqq x<1$, and again from (2.6) it follows that $F(x, x)=0$ for $0\leqq x<1$. Thus if $\lambda\geqq\mu$, the only non-negative solution of (2.2) is $p(m, n)=0$ for all m and n. Hence in later discussions we assume that $\lambda<\mu$.

Combining (2.5) with (2.6) we obtain the equation

$$\{\lambda xy-(\lambda+\mu+\eta)\ y+\mu\}\ F(x, y)+\eta yF(y, y) = (1-y)(\mu-\lambda x)\ F(x, x). \tag{2.7}$$

Putting $G(x)=F(x, x)$ and substituting $y=\mu/(\lambda+\mu+\eta-\lambda x)$ in (2.7), it follows that

$$\mu\eta G\left(\frac{\mu}{\lambda+\mu+\eta-\lambda x}\right)=(\lambda+\eta-\lambda x)(\mu-\lambda x)\ G(x) \tag{2.8}$$

or

$$G(x)=g(x)\ G(f(x)) \tag{2.9}$$

where

$$f(x)=\frac{\mu}{\lambda+\mu+\eta-\lambda x}, \tag{2.10}$$

and

$$g(x)=\frac{\mu\eta}{(\lambda+\eta-\lambda x)(\mu-\lambda x)}. \tag{2.11}$$

For simplicity, we define the functions $f_n(x)$ and $g_n(x)$ recursively as follows:

$$\begin{cases} f_0(x)=x \\ f_n(x)=f(f_{n-1}(x)) \qquad (n=1, 2, 3, \cdots) \end{cases} \tag{2.12}$$

and

$$g_n(x)=g(f_n(x)) \qquad (n=0, 1, 2, \cdots). \tag{2.13}$$

We can show that

$$\lim_{n\to\infty} f_n(x)=\alpha \tag{2.14}$$

and

$$\lim_{n\to\infty} g_n(x)=1 \tag{2.15}$$

for every $0\leqq x\leqq 1$, where α is the smaller one of two positive roots of $x=f(x)$ [see Appendix A].

Repeated use of the relation (2.9) shows that

$$G(x)=g_0(x)\ g_1(x)\ \cdots\ g_n(x)\ G(f_{n+1}(x)). \tag{2.16}$$

Since the product $\prod_{n=0}^{N} g_n(x)$ uniformly converges to a function $\prod_{n=0}^{\infty} g_n(x)$ in the interval [0, 1] [see Appendix B] and $G(x)=F(x, x)$ is continuous in [0, 1], the right side of (2.16) tends to

$$G(\alpha)\ \prod_{n=0}^{\infty} g_n(x). \tag{2.17}$$

The unknown constant $G(\alpha)$ is determined by the condition $G(1)=F(1, 1)=1$, and it is

$$G(\alpha)=1/\prod_{n=0}^{\infty} g_n(1). \tag{2.18}$$

Hence we have

$$G(x)=\prod_{n=0}^{\infty}\frac{g_n(x)}{g_n(1)}. \tag{2.19}$$

Therefore using this known $G(x)$ we can obtain $F(x, y)$ from (2.7);

$$F(x, y)=\frac{-\eta yG(y)+(1-y)(\mu-\lambda x)G(x)}{\lambda xy-(\lambda+\mu+\eta)y+\mu} \tag{2.20}$$

167

for $0 \leq x,\ y \leq 1,\ y \neq f(x)$, and

$$F(x, f(x)) = \frac{1-f(x)}{f(x)} [(\mu - \lambda x)\ G'(x) - \lambda G(x)] \tag{2.21}$$

for $y=f(x)$, to be continuous at points satisfying $y=f(x)$.

We have shown that a solution of the system of equations (2.1) and (2.2) must have the generating function $F(x, y)$ given by (2.20) and (2.21). Conversely, if $\lambda < \mu$ we can show that $F(x, y)$ given by (2.20) and (2.21) is a probability generating function and it determines a probability distribution satisfying (2.2). We can also show that the probability distribution has the moments of all orders. Since it is clear that $F(x, y)$ satisfies the equation (2.4) and the condition $F(1, 1)=1$, the only fact to be proved is that $F(x, y)$ has non-negative derivatives of all orders in [0, 1], [see Feller [1], p. 221]. The proof of it is postponed to Appendix C.

Finally we shall calculate the means and the variances of the queue sizes in the limiting distributions. Since the generating function of the total number of the customers in the system is $F(x, x)=G(x)$, and that of the number of the customers in the second queue (including one being served) is $F(1, x)$, the following results can be easily obtained.

The mean and the variance of the total number of customers in the system are as follows.

$$E(M+N) = G'(1), \tag{2.22}$$

$$Var(M+N) = G''(1) + G'(1) - \{G'(1)\}^2. \tag{2.23}$$

The mean and the variance of the number of customers in the second queue including one being served are as follows.

$$E(N) = G'(1) - \frac{\lambda}{\eta}, \tag{2.24}$$

$$Var(N) = G''(1) + G'(1) - \{G'(1)\}^2 + \frac{\lambda}{\eta^2}(2\mu+\eta-\lambda) - \frac{2(\mu-\lambda)}{\eta} G'(1). \tag{2.25}$$

For the numerical values, see Figs. 4 and 5.

2.2 Limiting Sojourn Time Distribution

We shall calculate the limiting sojourn time distribution under the "first come, first served" queue discipline. Let W be the sojourn time in the system (i.e. the sum of the waiting time and the service time) of a customer who arrives at the system in the steady state at time t.

First we shall calculate the conditional distribution of W, assuming that the numbers of customers in the first queue and in the second queue including one being served just before t are m and n, respectively. We denote the first time when the gate opens after t by $t+\tau$. Then the event $\{W \leq w\}$ can be divided into two events; the first case that the server has served $k(<n)$ customers in the time interval $(t, t+\tau]$ and served more than or equal to $m+n+1-k$ customers in $(t+\tau, t+w]$, and the second case that he has served all n customers in the second queue in $(t, t+\tau]$ and served more than or equal to $m+1$ customers in $(t+\tau, t+w]$. Summing up these events for possible values of τ and k, we obtain that

$$\begin{aligned} Pr\{W \leq w \mid M_t = m,\ N_t = n\} &= \int_0^w e^{-\gamma\tau} \Big[\sum_{k=0}^{n-1} \frac{(\mu\tau)^k}{k!} e^{-\mu\tau} E_{m+n+1-k}(w-\tau) \\ &\quad + \sum_{k=n}^{\infty} \frac{(\mu\tau)^k}{k!} e^{-\mu\tau} E_{m+1}(w-\tau) \Big] d\tau \end{aligned} \tag{2.26}$$

where $E_k(x)$ is the distribution function of Erlang distribution of order k with mean k/μ. Hence the conditional probability density function of W is

$$\begin{aligned} v(w \mid m, n) = \int_0^w e^{-\gamma\tau} \Big[&\sum_{k=0}^{n-1} \frac{\mu^{m+n+1} \tau^k (w-\tau)^{m+n-k}}{k!\ (m+n-k)!} e^{-\mu w} \\ &+ \sum_{k=n}^{\infty} \frac{\mu^{m+k+1} \tau^k (w-\tau)^m}{k!\ m!} e^{-\mu w} \Big] d\tau \end{aligned} \tag{2.27}$$

and its Laplace transform is

168

$$\varphi(s|m,\ n)=\int_0^\infty e^{-sw}\ v(w|m,\ n)\ dw \tag{2.28}$$
$$=\left(\frac{\mu}{s+\mu}\right)^{m+1}\left[\left(\frac{\mu}{s+\mu}\right)^n-\frac{s}{s+\eta}\left(\frac{\mu}{s+\mu+\eta}\right)^n\right].$$

Multiplying the both sides of (2.28) by $p(m,\ n)$ and summing up for all m and n, we obtain the Laplace transform of the limiting sojourn time distribution;

$$\varphi(s)=\sum_{m=0}^\infty\ \sum_{n=0}^\infty \varphi(s|m,\ n)\ p(m,\ n) \tag{2.29}$$
$$=\frac{\mu}{s+\mu}F\left(\frac{\mu}{s+\mu},\ \frac{\mu}{s+\mu}\right)-\frac{s}{s+\eta}F\left(\frac{\mu}{s+\mu},\ \frac{\mu}{s+\mu+\eta}\right)$$
$$=G\left(1-\frac{s}{\lambda}\right).$$

The mean and the variance of the limiting sojourn time distribution can be obtained as follows.

The mean sojourn time in the system is

$$E(W)=\frac{1}{\lambda}G'(1) \qquad \left(=\frac{1}{\lambda}E(M+N)\right). \tag{2.30}$$

The variance of the sojourn time is

$$Var(W)=\frac{1}{\lambda^2}[G''(1)-\{G'(1)\}^2] \tag{2.31}$$
$$\left(=\frac{1}{\lambda^2}[Var(M+N)-E(M+N)]\right).$$

For the numerical values, see Figs. 2 and 3.

3. Single Server Queues with the Deterministic Gating Rule

3.1 Limiting Distribution of the Number of Customers in the System

Since the lengths of the gating intervals in this model are constant, we may assume without loss of generality that the gate opens at times $t=0,\ 1,\ 2,\ \cdots$. We call the time interval from $t-1$ to t, period t, and denote the number of customers in the first queue at the end of period t by M_t, and that in the second queue including one being served at that time by N_t. As M_t denotes the number of arrivals in period t, M_t $(t=1,\ 2,\ 3,\ \cdots)$ are mutually independent random variables and have the common Poisson distribution with mean λ. Let X_t denote the possible number of customers whom the server can serve in period t assuming there are sufficiently many customers in the second queue. Then M_t's and X_t's are independent of each other, and the mutually independent random variables X_t $(t=1,\ 2,\ 3,\ \cdots)$ have the common Poisson distribution with mean μ. If the initial condition is that $(M_0,\ N_0)=(m_0,\ n_0)$, then N_t's are determined by the relation

$$N_{t+1}=[N_t+M_t-X_{t+1}]^+ \qquad (t=0,\ 1,\ 2,\ \cdots), \tag{3.1}$$

where $[x]^+=x$ if $x\geqq 0$ and $=0$ if $x<0$.

The relation (3.1) has the same form as the well known equation for the waiting times in a $GI/G/1$ type queue. Hence we can easily obtain the following results [see Kingman [2]].

If $\lambda\geqq\mu$, $\lim_{t\to\infty} Pr\ \{N_t=n\}=0$ for all n, and if $\lambda<\mu$, the limiting distribution $p(n)$ of N_t exists and its probability generating function is given by

$$F(y)=\sum_{n=0}^\infty y^n\ p(n) \tag{3.2}$$
$$=\exp\left\{-\sum_{n=1}^\infty \frac{1}{n}E(1-y^{[U_n]^+})\right\},$$

where

$$U_n=\sum_{t=1}^n (M_t-X_{t+1}) \qquad (n=1,\ 2,\ 3,\ \cdots). \tag{3.3}$$

Therefore the mean and the variance of the limiting distribution of the number of customers in the second queue including ones being served at the end of a period are given by

169

$$E(N)=\sum_{n=1}^{\infty}\frac{1}{n}E([U_n]^+) \tag{3.4}$$

and

$$Var(N)=\sum_{n=1}^{\infty}\frac{1}{n}E(\{[U_n]^+\}^2). \tag{3.5}$$

For the numerical values, see Figs. 4 and 5.

3.2 Limiting Sojourn Time Distribution

Here we shall again assume the "first come, first served" queue discipline. Let W be the sojourn time of a customer in the system. The distribution of W under the condition that the customer arrives at the system at time $t-\tau$, where t is an integer and $0\leqq\tau<1$, and that $N_t=n$, is given by

$$\begin{aligned} &Pr\{W\leqq w|\tau,\ N_t=n\} \\ &\quad=\sum_{m=0}^{\infty}E_{m+n+1}(w-\tau)\frac{\{\lambda(1-\tau)\}^m}{m!}e^{-\lambda(1-\tau)} \qquad (w\geqq\tau) \end{aligned} \tag{3.6}$$

where $E_k(x)$ is the distribution function of the Erlang distribution of order k with mean k/μ. Therefore the conditional probability density function of W is

$$v(w|\tau,\ n)=\begin{cases}\displaystyle\sum_{m=0}^{\infty}\frac{\mu^{m+n+1}\{\lambda(1-\tau)\}^m(w-\tau)^{m+n}}{m!\ (m+n)!}e^{-\mu(w-\tau)-\lambda(1-\tau)} & (w\geqq\tau)\\ 0 & (w<\tau)\end{cases} \tag{3.7}$$

and its Laplace transform is

$$\begin{aligned}\varphi(s|\tau,\ n)&=\int_0^{\infty}e^{-sw}\,v(w|\tau,\ n)\,dw\\ &=\left(\frac{\mu}{s+\mu}\right)^{n+1}e^{-s\lambda(1-\tau)/(s+\mu)-s\tau}.\end{aligned} \tag{3.8}$$

Hence the Laplace transform of the limiting sojourn time distribution is

$$\begin{aligned}\varphi(s)&=\sum_{n=0}^{\infty}p(n)\int_0^1\varphi(s|\tau,\ n)\,d\tau\\ &=\frac{\mu}{s+\mu}F\left(\frac{\mu}{s+\mu}\right)\cdot\frac{s+\mu}{s(s+\mu-\lambda)}(e^{-s\lambda/(s+\mu)}-e^{-s}).\end{aligned} \tag{3.9}$$

The mean and the variance of it are given by

$$E(W)=\frac{1}{\mu}\left\{E(N)+\frac{\lambda}{2}+1\right\}+\frac{1}{2} \tag{3.10}$$

and

$$Var(W)=\frac{1}{\mu^2}\{Var(N)+E(N)+\lambda+1\}+\frac{1}{12}\left(1-\frac{\lambda}{\mu}\right)^2. \tag{3.11}$$

For the numerical values, see Figs. 2 and 3.

4. Many Server Queues with the Exponential Gating Rule

4.1 Limiting Joint Distribution of the Queue Sizes

In the case of exponential gating rule, the technique used in Section 2 can be applied to many server queues. In this section we omit the discussion concerning the existence of the limiting distribution, and restrict ourselves to discuss how to find the limiting distribution.

Let λ, μ and η be the parameters of the exponential distributions of inter-arrival times, service times and gating intervals, respectively. There are r servers, and for the existence of the proper limiting distribution we assume that $\lambda<r\mu$. We denote the number of customers in the first queue at time t by M_t, and the number of customers in the second queue including ones being served at time t by N_t. Then the limiting distribution

$$p(m,\ n)=\lim_{t\to\infty}Pr\ \{M_t=m,\ N_t=n\} \qquad (m,\ n=0,\ 1,\ 2,\ \cdots) \tag{4.1}$$

satisfies the system of equations

170

$$0=-(\lambda+\eta)\ p(m,n)+\lambda\{1-\delta(m)\}\ p(m-1,n) \tag{4.2}$$
$$+\eta\delta(m)\sum_{k=0}^{n}p(k,n-k)-\mu\left\{r-\sum_{k=0}^{r-1}(r-k)\,\delta(n-k)\right\}p(m,n)$$
$$+\mu\left\{r-\sum_{k=1}^{r-1}(r-k)\,\delta(n-k-1)\right\}p(m,n+1)$$

where $\delta(0)=1$ and $\delta(x)=0$ for $x\neq 0$.

(If $\lambda\geqq r\mu$, the only non-negative solution of (4.2) is $p(m,n)=0$ for all m and n, and if $\lambda<r\mu$, it has a unique solution satisfying the condition that $\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}p(m,n)=1$. These results can be proved by analogous methods as in Section 2. Hence the proofs are omitted here.)

Let the probability generating function of $p(m,n)$ be denoted by

$$F(x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}x^m y^n p(m,n) \qquad (0\leqq x,\ y\leqq 1). \tag{4.3}$$

Using the relation

$$\sum_{m=0}^{\infty}x^m p(m,n)=\frac{1}{n!}\left[\frac{\partial^n}{\partial y^n}F(x,y)\right]_{y=0} \tag{4.4}$$

(4.2) can be rewritten in terms of $F(x,y)$ as

$$0=-(\lambda+\eta)F(x,y)+\lambda xF(x,y)+\eta F(y,y) \tag{4.5}$$
$$-r\mu F(x,y)+\mu\sum_{k=0}^{r-1}\frac{r-k}{k!}y^k\left[\frac{\partial^k}{\partial y^k}F(x,y)\right]_{y=0}$$
$$+r\mu\frac{1}{y}[F(x,y)-F(x,0)]-\mu\sum_{k=1}^{r-1}\frac{r-k}{k!}y^{k-1}\left[\frac{\partial^k}{\partial y^k}F(x,y)\right]_{y=0}$$

To simplify the notations, we define

$$\begin{cases}F^{(0)}(x,0)=F(x,0)\\ F^{(k)}(x,0)=\left[\dfrac{\partial^k}{\partial y^k}F(x,y)\right]_{y=0} & (k=1,2,3,\cdots).\end{cases} \tag{4.6}$$

Then (4.5) can be written as

$$\{\lambda xy-(\lambda+\eta+r\mu)\,y+r\mu\}\ F(x,y)+\eta yF(y,y) \tag{4.7}$$
$$=(1-y)\,\mu\sum_{k=0}^{r-1}y^k F^{(k)}(x,0).$$

This is a generalized form of (2.5). In Section 2, three unknown functions are $F(x,y)$, $F(y,y)$ and $F(x,0)$. At first, taking $y=x$, we obtain the relation (2.6) between $F(y,y)$ and $F(x,0)$, and then substituting $y=\mu/(\lambda+\mu+\eta-\lambda x)$ we get the main functional equation (2.9). However in this case we have $r+2$ unknown functions, $F(x,y)$, $F(y,y)$, $F^{(0)}(x,0)$, $\cdots$, $F^{(r-1)}(x,0)$. In order to determine the form of $F(x,y)$, we shall first get the relations among $F^{(k)}(x,0)$'s, and write them in terms of $F^{(0)}(x,0)=F(x,0)$. Then we shall use the same technique as in Section 2 to get the functional equation of $G(x)=F(x,x)$.

Differentiate (4.7) k times ($k=1,2,\cdots,r-1$), with respect to y and put $y=0$, then

$$F^{(k)}(x,0)=\frac{1}{\mu}(\lambda+\eta+(k-1)\,\mu-\lambda x)\,F^{(k-1)}(x,0)-\frac{\eta}{\mu}a_k \tag{4.8}$$

where

$$\begin{cases}a_1=F(0,0)\\ a_{k+1}=\left[\dfrac{d^k}{dy^k}F(y,y)\right]_{y=0} & (k=2,3,\cdots,r-1).\end{cases} \tag{4.9}$$

Hence $F^{(k)}(x,0)$ is written in terms of $F(x,0)=F^{(0)}(x,0)$ such as

$$F^{(k)}(x,0)=F(x,0)\prod_{j=1}^{k-1}\frac{1}{\mu}(\lambda+\eta+j\mu-\lambda x) \tag{4.10}$$
$$-\frac{\eta}{\mu}\sum_{i=1}^{k-1}a_i\prod_{j=i}^{k-1}\frac{1}{\mu}(\lambda+\eta+j\mu-\lambda x)-\frac{\eta}{\mu}a_k.$$

Substituting the right hand side of (4.10) for $F^{(k)}(x,0)$ in (4.7), we obtain an equation including only three unknown functions, $F(x,y)$, $F(y,y)$ and $F(x,0)$ as

171

$$(4.11)\qquad \{\lambda xy-(\lambda+\eta+r\mu)\,y+r\mu\}\,F(x,y)+\eta y F(y,y)$$
$$=(1-y)\mu\,F(x,0)\left[r+\sum_{k=1}^{r-1}\frac{r-k}{k!}\,y^k\prod_{j=0}^{k-1}\frac{1}{\mu}\,(\lambda+\eta+j\mu-\lambda x)\right]$$
$$-(1-y)\,\eta\left[(r-1)\,a_1\,y+\sum_{k=2}^{r-1}\frac{r-k}{k!}\,y^k\sum_{i=1}^{k-1}a_i\prod_{j=i}^{k-1}\frac{1}{\mu}\,(\lambda+\eta+j\mu-\lambda x)\right.$$
$$\left.+\sum_{k=2}^{r-1}\frac{r-k}{k!}\,y^k\,a_k\right]$$
$$=(1-y)\mu\,F(x,0)\,B(x,y)-(1-y)\,\eta\sum_{i=1}^{r-1}a_i\,A_i(x,y),$$

where

$$(4.12)\qquad B(x,y)=r+\sum_{k=1}^{r-1}\frac{r-k}{k!}\,y^k\sum_{j=0}^{k-1}\frac{1}{\mu}\,(\lambda+\eta+j\mu-\lambda x)$$

and

$$(4.13)\qquad A_i(x,y)=\frac{r-i}{i!}\,y^i+\sum_{k=i+1}^{r-1}\frac{r-k}{k!}\,y^k\prod_{j=i}^{k-1}\frac{1}{\mu}\,(\lambda+\eta+j\mu-\lambda x)$$
$$(i=1,2,\cdots,(r-2).$$

Now we use the same technique as in Section 2. In (4.11), we take $y=x$ and divide it by $(1-x)$, then we have

$$(4.14)\qquad (r\mu-\lambda x)\,F(x,x)=\mu F(x,0)\,B(x,x)-\eta\sum_{i=1}^{r-1}a_i\,A_i(x,x).$$

Solving $F(x,0)$ from the above equation and substituting it in (4.11), we have

$$(4.15)\qquad \{\lambda xy-(\lambda+\eta+r\mu)\,y+r\mu\}\,F(x,y)+\eta y F(y,y)$$
$$=(1-y)\,(r\mu-\lambda x)\,\frac{B(x,y)}{B(x,x)}\,F(x,x)$$
$$-\eta\,(1-y)\sum_{i=1}^{r-1}a_i\left\{A_i(x,y)-\frac{B(x,y)}{B(x,x)}\,A_i(x,x)\right\}.$$

As in Section 2, we define

$$(4.16)\qquad f(x)=\frac{r\mu}{\lambda+\eta+r\mu-\lambda x}$$

and

$$(4.17)\qquad G(x)=F(x,x).$$

Taking $y=f(x)$, (4.15) will be reduced to

$$(4.18)\qquad \eta f(x)\,G(f(x))=(1-f(x))\,(r\mu-\lambda x)\,\frac{B(x,f(x))}{B(x,x)}\,G(x)$$
$$-\eta(1-f(x))\sum_{i=1}^{r-1}a_i\left\{A_i(x,f(x))-\frac{B(x,f(x))}{B(x,x)}\,A_i(x,x)\right\}.$$

It follows that

$$(4.19)\qquad G(x)=\frac{\eta f(x)}{(1-f(x))\,(r\mu-\lambda x)}\cdot\frac{B(x,x)}{B(x,f(x))}\,G(f(x))$$
$$+\frac{\eta}{r\mu-\lambda x}\sum_{i=1}^{r-1}a_i\left\{\frac{B(x,x)}{B(x,f(x))}\,A_i(x,f(x))-A_i(x,x)\right\}$$
$$=g(x)\,G(f(x))+\sum_{i=1}^{r-1}a_i\,h^i(x),$$

where

$$(4.20)\qquad g(x)=\frac{\eta f(x)}{(1-f(x))\,(r\mu-\lambda x)}\cdot\frac{B(x,x)}{B(x,f(x))}$$

and

$$(4.21)\qquad h^i(x)=\frac{\eta}{r\mu-\lambda x}\left\{\frac{B(x,x)}{B(x,f(x))}\,A_i(x,f(x))-A_i(x,x)\right\}$$
$$(i=1,2,\cdots,r-1).$$

Now we define $r+1$ sequences of functions recursively as follow:

$$(4.22)\qquad \begin{cases} f_0(x)=x \\ f_n(x)=f(f_{n-1}(x)) & (n=1,2,3,\cdots)\end{cases}$$

$$(4.23)\qquad g_n(x)=g(f_n(x))\qquad (n=0,1,2,\cdots)$$

$$(4.24)\qquad h_n^i(x)=h^i(f_n(x))\qquad (i=1,2,\cdots,r-1)\ \ (n=0,1,2,\cdots).$$

Then clearly, we have that for every $0\leqq x\leqq 1$,

$$(4.25)\qquad f_n(x) \to \alpha = \frac{1}{2\lambda}\left[\lambda+\eta+r\mu-\sqrt{(\lambda+\eta+r\mu)^2-4r\lambda\mu}\right]$$

$$(4.26)\qquad g_n(x) \to g(\alpha)=1$$

$$(4.27)\qquad h_n^i(x) \to h^i(\alpha)=0 \qquad . \ (i=1, 2, \cdots, r-1)$$

as n increases to infinity.

From (4.19), using these functions, $G(x)$ can be written as

$$(4.28)\qquad G(x)=G(\alpha)\prod_{n=0}^{\infty} g_n(x)+\sum_{i=1}^{r-1} a_i\left[h^i(x)+\sum_{n=1}^{\infty} h_n^i(x)\prod_{k=0}^{n-1} g_k(x)\right].$$

Though this expression has r unknown parameters, $a_1, a_2, \cdots, a_{r-1}$ and $G(\alpha)$, they can be determined by the simultaneous linear equations which consist of (4.9) and $G(1)=1$. Thus we have obtained $F(x, x)=G(x)$, and also $F(x, y)$ from (4.15).

The above solution is rather complicated, but for small r, we would be able to calculate the means and the variances with the help of computer.

5. Numerical Examples

Some numerical examples of single server queues with gates under the exponential or the deterministic gating rules are represented in Figs. 2~5.

Figs. 2 and 3 show the mean and the variance of the sojourn time in the system in the steady state for various values of the relative traffic intensity $\rho=\lambda/\mu$ and of the mean length $1/\eta$ of gating intervals. The graphs in solid lines represent the exponential case and those in broken lines represent the deterministic case.

Figs. 4 and 5 show the mean and the variance of the number of customers in the second queue in the steady state immediately after the gate opens. The same notations as in Figs. 2 and 3 are used here. The number of customers in the second queue immediately after the gate opens, can also be considered as the total number of customers in

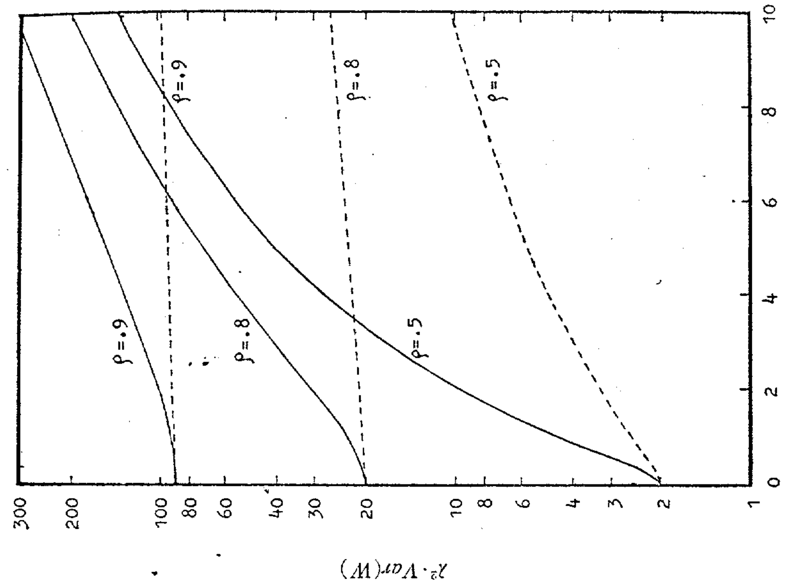


Fig. 3. The variance of the sojourn time in the system.

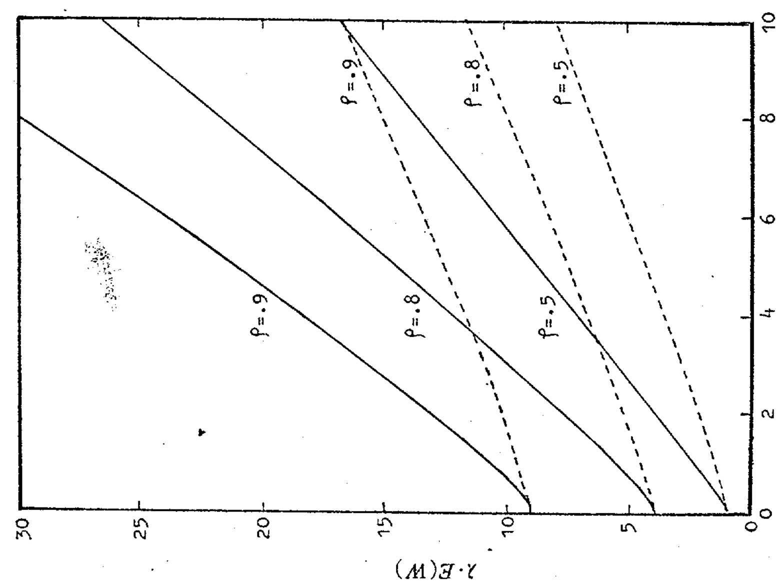


Fig. 2. The mean sojourn time in the system.

173

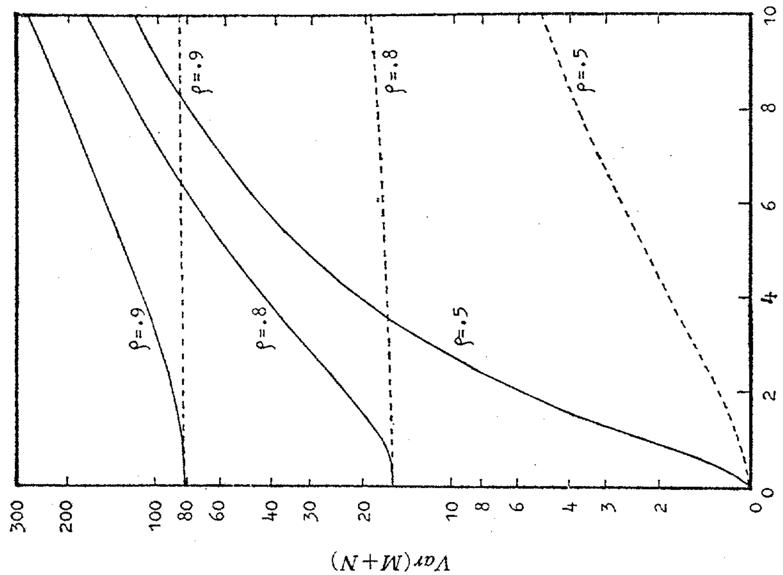


Fig. 5. The variance of the number of customers in the second queue immediately after the gate opens.

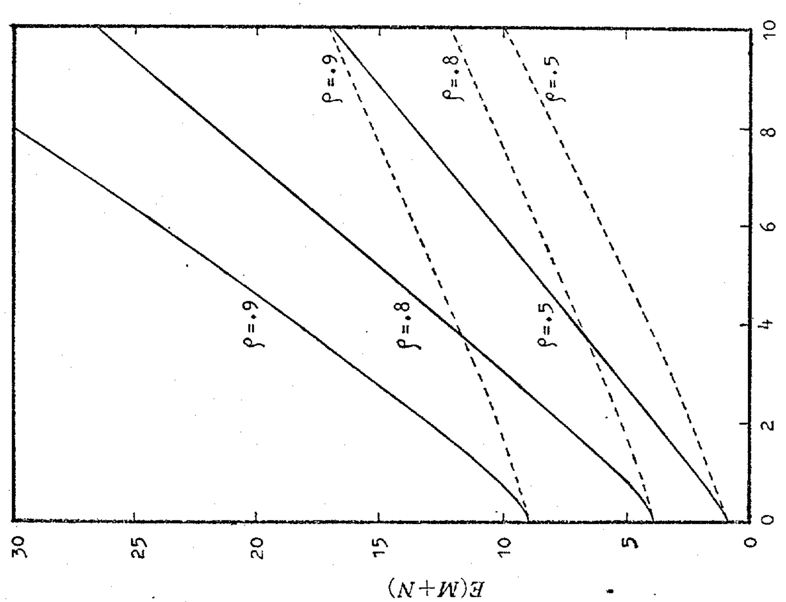


Fig. 4. The mean number of customers in the second queue immediately after the gate opens.

the system just before the gate opens, and in the case of the exponential gating rule, its distribution coincides with that for arbitrary time if the system is in the steady state. Hence the graphs in solid lines in Figs. 4 and 5 also represent the mean and the variance of the total number of customers in the system with exponential gating rule at arbitrary time in the steady state. (This is not true for the deterministic case.)

Appendices

Appendix A. Convergences of $f_n(x)$ and $g_n(x)$

Proposition A. (i) The sequence $f_n(x)$ uniformly converges to α in the interval $[0, 1]$.
(ii) The sequence $g_n(x)$ uniformly converges to 1 in $[0, 1]$.

Proof. Since α is the smaller one of two positive roots of the equation $x=f(x)$, it is given by

$$\alpha = \frac{1}{2\lambda}[\lambda+\mu+\eta-\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}]\,. \tag{A·1}$$

Clearly $0<\alpha<1$. If $x=\alpha$ then $f_n(\alpha)=\alpha$ for all n. If $x\neq\alpha$,

$$\begin{aligned}\frac{f(x)-\alpha}{x-\alpha} &= \frac{\mu-(\lambda+\mu+\eta-\lambda x)}{(x-\alpha)\,(\lambda+\mu+\eta-\lambda x)} \\ &= \frac{\alpha\lambda}{\lambda+\mu+\eta-\lambda x} \\ &= \frac{\alpha\lambda}{\mu}f(x)\,.\end{aligned} \tag{A·2}$$

Since $0<f(x)\leqq\dfrac{\mu}{\mu+\eta}$ in $[0, 1]$, we have

$$0<\frac{f(x)-\alpha}{x-\alpha}\leqq\frac{\alpha\lambda}{\mu+\eta} \qquad (<1) \tag{A·3}$$

for every $0\leqq x\leqq 1$, $x\neq\alpha$. Hence from the definition (2.12) it follows that

174

$$(\text{A}\cdot4)\qquad |f_n(x)-\alpha|=|f(f_{n-1}(x))-\alpha|$$
$$\leqq\frac{\alpha\lambda}{\mu+\eta}|f_{n-1}(x)-\alpha|$$
$$\leqq\cdots\leqq\left(\frac{\alpha\lambda}{\mu+\eta}\right)^n(x-\alpha)<\left(\frac{\alpha\lambda}{\mu+\eta}\right)^n.$$

Letting n tend to infinity, it proves (i). (ii) follows immediately from (i) and from the continuity of $g(x)$ in [0, 1].

Appendix B. Convergence of $\prod_{n=0}^{N} g_n(x)$ and Differentiability of the Limit Function $\prod_{n=0}^{\infty} g_n(x)$

Proposition B. (i) $\prod_{n=0}^{N} g_n(x)$ uniformly converges to a function $\prod_{n=0}^{\infty} g_n(x)$ in [0, 1].

(ii) The limit function $\prod_{n=0}^{\infty} g_n(x)$ has non-negative derivatives of all orders in [0. 1].

Proof. From the Taylor's formula,

$$(\text{A}\cdot5)\qquad g(x)=1+\frac{\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}}{\mu\eta}(x-\alpha)+o(x-\alpha)$$

where $o(t)$ represents a term such that $o(t)/t\to0$ as $t\to0$. From the definition (2.13) it follows that

$$(\text{A}\cdot6)\qquad \frac{\log g_{n+1}(x)}{\log g_n(x)}=\frac{\dfrac{\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}}{\mu\eta}(f_{n+1}(x)-\alpha)+o(f_{n+1}(x)-\alpha)}{\dfrac{\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}}{\mu\eta}(f_n(x)-\alpha)+o(f_n(x)-\alpha)}$$
$$=\frac{f(f_n(x))-\alpha}{f_n(x)-\alpha}+o(f_n(x)-\alpha).$$

Since $f_n(x)$ uniformly converges to α by Proposition A, (A·3) implies that

$$(\text{A}\cdot7)\qquad 0<\frac{\log g_{n+1}(x)}{\log g_n(x)}<\frac{\alpha\lambda}{\mu+\eta}<1$$

for sufficiently large n and every $0\leqq x\leqq1$. This proves (i).

In order to prove (ii), it is sufficient to prove

(a) for every $k=1, 2, 3, \cdots$, the k-th derivative of $\prod_{n=0}^{N} g_n(x)$ uniformly converges in [0, 1].

Note that $f(x)$ and $g(x)$ have positive derivatives of all orders in [0, 1], and hence $f_n(x)$ and $g_n(x)$ have also positive derivatives of all orders in [0, 1]. Since,

$$(\text{A}\cdot8)\qquad \frac{d}{dx}\left\{\prod_{n=0}^{N} g_n(x)\right\}=\left\{\prod_{n=0}^{N} g_n(x)\right\}\cdot\frac{d}{dx}\left\{\sum_{n=0}^{N}\log g_n(x)\right\},$$

differentiating both sides of (A·8) k times, we have

$$(\text{A}\cdot9)\qquad \left\{\prod_{n=0}^{N} g_n(x)\right\}^{(k+1)}=\sum_{\nu=0}^{k}\binom{k}{\nu}\left\{\prod_{n=0}^{N} g_n(x)\right\}^{(\nu)}\cdot\left[\sum_{n=0}^{N}\{\log g_n(x)\}^{(k-\nu+1)}\right]$$

where $\{h(x)\}^{(k)}$ denotes the k-th derivative of $h(x)$. Hence using the mathematical induction, (a) follows from

(b) for each $k=1, 2, 3, \cdots$, the series $\sum_{n=0}^{N}\{\log g_n(x)\}^{(k)}$ converges uniformly in [0, 1].

By the definition of $g_n(x)$, we can write as

$$(\text{A}\cdot10)\qquad \{\log g_n(x)\}^{(k)}=-\{\log(\lambda+\eta-\lambda f_n(x))\}^{(k)}-\{\log(\mu-\lambda f_n(x))\}^{(k)}$$
$$=\sum_i c_i\left[\left\{\frac{1}{\lambda+\eta-\lambda f_n(x)}\right\}^{\nu_{i0}}+\left\{\frac{1}{\mu-\lambda f_n(x)}\right\}^{\nu_{i0}}\right]\cdot\prod_{j=1}^{k}\{f_n^{(j)}(x)\}^{\nu_{ij}}$$

with positive constants c_i and $(k+1)$-tuples of integers $(\nu_{i0}, \cdots, \nu_{ik})$. Since c_i and $(\nu_{i0}, \cdots, \nu_{ik})$ are independent of n, and $f_n(x)$ uniformly converges to α, (b) follows from

(c) for each $k=1, 2, 3, \cdots$, $\sum_{n=0}^{N}\{f_n(x)\}^{(k)}$ converges uniformly in [0, 1].

175

Now we shall prove (c) by induction. First we prove that $\sum_{n=0}^{N} f_n^{(1)}(x)$ uniformly converges in [0, 1]. By the definition (2.12)

$$f_{n+1}^{(1)}(x)=\frac{d}{dx}\{f(f_n(x))\}=f^{(1)}(f_n(x))\,f_n^{(1)}(x). \tag{A·11}$$

Since $f^{(1)}(x)=\frac{\lambda}{\mu}\{f(x)\}^2$ and $0<f_n(x)\leqq\frac{\mu}{\mu+\eta}$ in [0, 1], it follows that

$$0<\frac{f_{n+1}^{(1)}(x)}{f_n^{(1)}(x)}=\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2<\frac{\lambda\mu}{(\mu+\eta)^2}\qquad(<1). \tag{A·12}$$

This proves the uniform convergence of $\sum_{n=0}^{N} f_n^{(1)}(x)$. Next we assume that for every $j=1, 2, \cdots, k-1$, the series $\sum_{n=0}^{N} f_n^{(j)}(x)$ uniformly converges in [0, 1]. Since we can write as

$$\begin{aligned}f_{n+1}^{(k)}(x)&=\{f(f_n(x))\}^{(k)}=\{f^{(1)}(f_n(x))f_n^{(1)}(x)\}^{(k-1)}\\&=f^{(1)}(f_n(x))\cdot f_n^{(k)}(x)+\sum_i c_i' f^{(\nu_{i0})}(f_n(x))\prod_{j=1}^{k-1}\{f_n^{(j)}(x)\}^{\nu_{ij}}\end{aligned} \tag{A·13}$$

with positive constants c_i' and k-tuples of integers $(\nu_{i0}, \cdots, \nu_{i\,k-1})$, we have

$$f_{n+1}^{(k)}(x)-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2 f_n^{(k)}(x)=\sum_i c_i' f^{(\nu_{i0})}(f_n(x))\prod_{j=1}^{k-1}\{f_n^{(j)}(x)\}^{\nu_{ij}}. \tag{A·14}$$

Summing up with n, we obtain that

$$\begin{aligned}f_{N+1}^{(k)}(x)&+\sum_{n=0}^{N}\left[1-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2\right]f_n^{(k)}(x)\\&=\sum_{n=0}^{N}\sum_i c_i f^{(\nu_{i0})}(f_n(x))\prod_{j=1}^{k-1}\{f_n^{(j)}(x)\}^{\nu_{ij}}.\end{aligned} \tag{A·15}$$

Since $f^{(j)}(f_n(x))$ uniformly converges to $(\lambda/\mu)^j\alpha^{j+1}$, the right hand side of (A·15) converges uniformly by the assumption. Since $f_{N+1}(x)>0$ and $(\lambda/\mu)\{f_{n+1}(x)\}^2<1$ in [0, 1], the series of positive functions $\sum_{n=0}^{N}\left[1-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2\right]f_n^{(k)}(x)$ must converge monotonically pointwise in [0, 1] by (A·15). From Proposition A, for an arbitrary positive number ε there exists an integer n_0 such that for any $n>n_0$, $|f_n(x)-\alpha|<\varepsilon$. Hence the inequalities

$$\begin{aligned}&\frac{1}{1-(\lambda/\mu)(\alpha+\varepsilon)^2}\sum_{n=n_0}^{N}\left[1-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2\right]f_n^{(k)}(x)\\&<\sum_{n=n_0}^{N} f_n^{(k)}(x)<\frac{1}{1-(\lambda/\mu)(\alpha-\varepsilon)^2}\sum_{n=n_0}^{N}\left[1-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2\right]f_n^{(k)}(x)\end{aligned} \tag{A·16}$$

hold, and we obtain that the series $\sum_{n=0}^{N} f_n^{(k)}(x)$ converges pointwise in in [0, 1] and that $f_n^{(k)}(x)$ converges to zero. Since $f_n^{(k)}(x)$ is a positive and increasing function in [0, 1], the convergence of the sequence is uniform in [0, 1]. Hence from (A·15) the convergence of $\sum_{n=0}^{N}\left[1-\frac{\lambda}{\mu}\{f_{n+1}(x)\}^2\right]f_n^{(k)}(x)$ is uniform and from (A·16) the convergence of $\sum_{n=0}^{N} f_n^{(k)}(x)$ is also uniform in [0, 1]. Thus (c) is proved and we complete the proof of (ii).

Appendix C. Absolute Monotonicity of $F(x, y)$

Proposition C. $F(x, y)$ defied by (2.20) and (2.21), has non-negative derivatives of all orders in [0, 1].

Proof. From Proposition B, $\prod_{n=0}^{\infty} g_n(x)$ has non-negative derivatives of all orders. $G(x)$ defined by (2.19) has also non-negative derivatives of all orders and by the Taylor's formula we can write as

$$G(x)=\sum_{k=0}^{\infty} a_k x^k\qquad(0\leqq x\leqq 1) \tag{A·17}$$

where a_k are non-negative coefficients. Furthermore we can write as

176

$$\text{(A·18)} \qquad \frac{x}{1-x}G(x)=\sum_{k=1}^{\infty} b_k x^k \qquad (0\leqq x<1)$$

where

$$\text{(A·19)} \qquad b_k=\sum_{i=0}^{k-1} a_i \qquad (k=1, 2, 3, \cdots).$$

Using these constants, we can rewrite $F(x, y)$ as follows:

$$\begin{aligned}\text{(A·20)} \quad F(x, y) &= \frac{f(x)}{(f(x)-y)}[-\eta y G(y)+(1-y)(\mu-\lambda x)G(x)] \\ &= \frac{\eta}{\mu}\frac{f(x)(1-y)}{f(x)-y}\left[\frac{f(x)}{1-f(x)}G(f(x))-\frac{y}{1-y}G(y)\right] \\ &= \frac{\eta}{\mu}f(x)(1-y)\sum_{k=1}^{\infty} b_k[(f(x))^{k-1}+(f(x))^{k-2}y+\cdots+y^{k-1}] \\ &= \frac{\eta}{\mu}\sum_{k=1}^{\infty} b_k(f(x))^k \\ &\quad +\frac{\eta}{\mu}f(x)y\sum_{k=1}^{\infty} a_k[(f(x))^{k-1}+(f(x))^{k-2}y+\cdots+y^{k-1}].\end{aligned}$$

Since the constants a_k and b_k are non-negative and $f(x)$ has positive derivatives of all orders in $[0, 1]$, $F(x, y)$ has also non-negative derivatives of all orders in $[0, 1]$.

References

[1] Feller, W.; *An introduction to probability theory and its applications*, Vol. 2, John Wiley & Sons, Inc., New York, 1966.
[2] Kingman, J.F.C.; "On the algebra of queues" *J. Appl. Prob.*, **3**, 285-326, 1966.

177