SOME PROPERTIES OF MARKOV CHAINS AND MARKOV CHAIN AGGREGATES

by

G.D. Mistriotis

Submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree

of

Master of Applied Science

Department of Electrical Engineering

Faculty of Science and Engineering

University of Ottawa

Ottawa, Ontario

June 1974

© G.D. Mistriotis, Ottawa, Canada, 1974.

C O N T E N T S

ABSTRACT

The concepts of a Markov process, Markov chain and the Markovian property are reviewed and rigorously defined along the von-Mises relative frequency lines. The notions of experiment space and system are introduced which allow the derivation of probability measures, if exist, as the limits of the relative frequencies.

The properties of the stationary finite-state and discrete-time Markov chain transition matrices are feviewed. Some original results are stated like, a simple and powerful criterion for the irreducibility of transition matrices and a simple test for the accessibility between two states. These results are generalized for any non-negative matrix.

The transition matrix, common to all members of a homogeneous set of Markov chains, does not adequately describe the Markov chains collectively but, only individually. Therefore, the concept of the Markov Chain Aggregate (MCA) is introduced which is the population of the Markov chains in question. The MCA is completely described by its state-vector whose entries account for the number of the members of the MCA being at the corresponding state. The expectancy and the variance-covariance matrix are estimated in terms of its initial value and transition matrix of the MCA.

ACKNOWLEDGEMENTS

The author is greatly indebted to his thesis supervisor Professor M.Krieger for his encouragement and constructive criticism at all stages of this research and, particularly, at the painful stage of drafting the thesis.

This research has been initiated during the course of Quantitative Analysis (May 1971-April 1972) offered by the Institute for Policy Analysis of the University of Toronto. Professor A.R.Dobell, the course director at this time, encouraged the author to study the properties of transition matrices for the purpose of modelling income distribution. His advice, specific suggestions and various discussions extended with timewise generosity are gratefully appreciated.

Since April 1972, the author worked in the Planning Branch of the Treasury Board Secretariat on the, so-called, POLSIM project. Dr. C.J.Hindle, the author's director, has shown on many occasions theoretical and practical interests on Markov chains. His enthusiastic conjectures (called by himself "gut's feelings") and the discussions that followed were a continuous source of knowledge. The author wishes to thank him and express his deep appreciation for his encouragement and guidance.

Finally, grateful thanks are extended to Messrs N.Spyratos, C.B.Marriott, R.Arnott and Prof. N.Georganas for their comments of the first drafts of this thesis.

SYMBOLS AND NOTATIONS

$\cap, \cup, \subset$	Set intersection, union and inclusion, respectively.
$\in$	Set membership symbol, being read "belongs to".
$f:B \to S$	Mapping f, from the set B into the set S.
f^{-1}	Inverse mapping.
$E[a]$	Expectancy of the random scalar or vector a.
$Var[a]$	Variance of the random scalar a, or variance-covariance matrix of the random vector a.
a,v,b...	Scalars or vectors denoted by lower case letters.
A,M,N,...	Matrices denoted by upper case letters.
A',v',...	Transposed of matrix A, vector v,..etc.
$A>0, v>0$	Denote elementwise strict positivity of matrix A and the vector or scalar v.
$A\geq 0, v\geq 0$	Denotes elementwise non-negativity.
$\overline{1}$, $\overline{1}_n$	Denote the vector $(1,1,\ldots,1)'$ of appropriate dimension and of dimension n, respectively.
e_i	Denotes the unit vector $(0,\ldots,0,1,0,\ldots,0)'$ along the i-th coordinate.
I, I_n	Denote the identity matrix of appropriate order and of order n, respectively.
diag[v]	Denotes a diagonal matrix whose diagonal elements are the coordinates of the vector $v=(v_1,v_2,\ldots)'$
P^n_{ij} or p^n_{ij}	Denotes the (i,j) element of the square matrix $P^n = P\ldots P$ (n times). By definition $P^0 = I$ for any non-zero square matrix P.

$L^{(m)}$ Denotes a matrix which is superscripted by m.

$(v_1,\ldots,v_n)'$ Denotes the vector $\begin{vmatrix} v_1 \\ \vdots \\ v_n \end{vmatrix}$, i.e. it is written in its transposed form for obvious reasons of typing convenience.

CHAPTER I. INTRODUCTION

The Russian mathematician A.A. Markov (1856-1922) introduced in 1907 the mathematical concept of a Markov chain. Markov chains have been generalized by other mathematicians, especially by Kolmogorov, to the concept of a Markovian process.

Since almost the beginning of this century, the concept of a Markov process has been studied extensively and used to model a variety of phenomena in almost any field of applied science. In particular, it is a typical mathematical tool used to-day in the field of physics and engineering. The subject has met an enormous interest by engineers throughout the world, especially, during the interval from 1955 to-day.

It seems that, it is almost hopeless for one to be involved with a popular subject like this and still claim that he can produce a certain result that it is original. This statement seems stronger for the most simple type of Markov processes, namely, the Markov chains. This is, perhaps, almost true for the pure mathematician. Indeed, it seems that there is very little ground for further relevant contributions in the domain of Markov chains at the abstract mathematical level. However, this argument is not valid for engineers and other applied researchers.

An engineer, biologist, or any other type of applied researcher can be easily convinced that he can further contribute into the domain of Markov

chains by undertaking a certain case-study. By a case-study we mean a real-life problem which can be modelled by a Markov chain and demands concrete numerical results accompanied with their analysis for the purpose of understanding the system in question.

This research has been initiated by exactly the above mentioned motive. The author was faced with two particular real-life phenomena. First, the Income Distribution problem and second, the Canadian Interprovincial Migration problem. Both problems are usually modelled by some finite-state Markov chains with time-invariant probability parameters.

Abstracting from the original problems at hand, the author concentrated on two points of general interest. First, he wanted to define precisely and rigorously what a Markov chain is. Second, to study the mathematical properties of the transition matrices which describe the Markov chains. The results with respect to the first point are presented in appendix A and chapter II. While, the results with respect to the second point are presented in chapters III and IV.

I.1. About the Definition of Markov Processes.

The definition of a Markov process is given by the various involved textbooks in a variety of styles. Those that do not pretend giving a rigorous definition are excused for any logical inconsistency. However, those that

claim formal and rigorous exposition should offer a definition that is general enough to fit any practical or theoretical case at hand. Specifically, all textbooks agree on the point that a Markov process is a random process that satisfies the, so-called, Markovian property. The Markovian property, on the other hand, states that certain conditional probabilities coincide. Therefore, the basic question arises: What is conditional probability? The answer is given by all textbooks unanimously: Conditional probability of an event E (i.e. appropriate subset of a sample or state-space) conditional to a non-null event E_o is the real number

$$Pr[E|E_o] = \frac{Pr[E \cap E_o]}{Pr[E_o]}$$

Therefore, it is evident that in a Markovian system the "unconditional" probabilities Pr[A], for all events A, should be meaningful and defined.

It is common knowledge, in practice, that for certain Markov processes the "unconditional" probabilities are meaningless, while, the conditional ones are considered as being defined and meaningful. Obviously, we are at some dead end as far as the basic and essential definition of a conditional probability.

Motivated by the need to introduce the concept of the conditional probability independently of the unconditional one, the author was forced to review the whole basic structure of the probabilistic concepts as presented in appendix A. His approach is based on a combination of the classical concept of probability and the measure-theoretic one.

The author associates the concept of probability with phenomena of repetitive nature. For such phenomena there exist accumulated information and data in either actual or potential form. This information is assumed to be available and form a countable set called an experiment space. From the experiment space one can define, for each event, the sequence of relative frequencies which in their limit (if exist) become the probabilities of the events in question. These probabilities are checked to satisfy all "axioms" of the probability measures. I.e., these probabilities define certain mappings $\mu: \mathcal{F} \to [0,1]$ from a σ-algebra into the closed unit interval $[0,1]$ and also satisfy the well-known Kolmogorov axioms.

As explained in appendix A, this approach allows one to define the conditional probabilities independently of the unconditional ones. Thus, one is now able to define the concept of a Markovian process and the Markovian property on more realistic grounds.

I.2. About the Properties of Transition Matrices

In section II.2 it is explained that a Markov chain with discrete-time is completely described by its single-step transition matrix $P=(p_{ij})$. Thus, we identified the study of a certain Markov chain by the study of its transition matrix. We restricted our attention to the stationary, finite-state, discrete-time Markov chains.

In section III.1 we define the communication relation between two states of the state-space of a square non-negative matrix. This relation is proven to be an equivalence relation which, therefore, partitions the state-space into equivalence communication classes. Any class of the state-space is characterized to be either a recurrent or a transient class. Thus, by relabelling the states, i.e. by proper combined row and column elementary operation, one can derive the canonical form of a non-negative matrix and, in particular, a transition matrix.

In section III.2 we, first, state a few basic propositions related to the transition matrices, i.e. TH.1,TH.2 and TH.3. Their purpose is to be used for the proofs of TH.4 and TH.5 as well as to assist us on various other propositions later on. The proposition TH.4 states an important property of finite Markov chains. Specifically, if the order of a non-negative square matrix is $n \geq 1$ and a state j can reach another state i then, the transition $j \rightarrow i$ is possible in at most n time-steps. This proposition is suggested by Gill's paper ([23], pp.49-50) on finite-state systems. However, our proof is a modified version of Gill's one. Finally, the proposition TH.5 offers a powerful criterion for the irreducibility of non-negative square matrices.

In section III.3 we derive the canonical form of any square non-negativ matrix by using three stated propositions which are considered to be corollaries of TH.4 and TH.5. The procedure is explicit in all its technical details

and an example is given of a moderate size matrix. Furthermore, a FORTRAN program is written for this purpose and is listed in appendix B. The material intended to cover the case of transition matrices. However, an obvious generalization is made to cover any square matrix from any number field. By the end of the section we comment on this generalization and we give as an example a linear dynamic system.

Chapter IV deals, in general, with the powers of transition matrices and their limits as their exponent tends to infinity.

In section IV.1 it is stated the well-known Frobenius theorem. The following propositions TH.6 and its corollary 6.1 are in one form or another known in the literature of linear algebra and have been included for the purpose of proper material exposition.

In section IV.2 we establish that the matrix powers of any transition matrix are asymptotically periodic with period some integer $d \geq 1$. Of course, $d=1$ means convergence of the matrix powers. By corollary 8.1, a transition matrix power sequence $\{P^m\}$ converges if and only if P has not any eigenvalue of unity magnitude other than $\lambda=1$. A useful matrix sequence $A_m=(P+\ldots+P^m)/m$; $m=1,2,\ldots$, for any transition matrix P, is defined and is called the sequence of the "average" transition matrices. It is proven in TH.9 that $\{A_m\}$ always converges and also, if $\lim_{m\to\infty} P^m$ exists, $\lim_{m\to\infty} P^m=\lim_{m\to\infty} A_m$.

The "average transition matrix" can be intuitively interpreted so that, it can be considered as a substitute of the original sequence $\{P^m\}$ in order to describe the long run behavior of the Markov chain in question.

In section IV.3 we analytically evaluate the limit of the average transition matrix. Its evaluation assumes that the transition matrix is in its canonical form.

Finally, in section IV.4 we offer a numerical example of a moderate transition matrix size.

Chapters III and IV present material that is mostly either explicitly or implicitly already stated in the literature. However, two contributions of this research to the theory of Markov chains should be mentioned. These are the propositions TH.5 and TH.8.

I.3 About Markov Chain Aggregates.

Markov chains usually model real-life objects which are members of homogeneous sets. In some cases, the study of a particular Markov chain is inadequate to describe the behavior of the homogeneous Markov chain set in which it belongs. Instead, the set of all the Markov chains, members of a certain homogeneous group, should be studied collectively.

In section II.3 we introduce the concept of a Markov Chain Aggregate (MCA) which is a finite set of Markov chains. An MCA is described by its state-vector $x(t) = (x_1(t), x_2(t), \ldots)'$ whose entries $x_1(t), x_2(t), \ldots$ are non-negative integers and denote the number of the MCA members that are at the states $1, 2, \ldots$, respectively.

In chapter V we present the few findings we obtained on the MCAs. In section V.2 we present the proposition TH.10. By this proposition we evaluate the expected value and the variance-covariance matrix of the state-vector. These statistical parameters of the system are expressed in terms of the state-vector initial value and the transition matrix. One should notice that the theorem is valid for stationary or not, discrete or continuous-time type of an MCA.

Finally, in section V.3, we discuss the stationary, discrete-time MCA case and we examine the asymptotic behavior of its state-vector $x(t)$.

CHAPTER II

MARKOV CHAINS AND CHAIN AGGREGATES

CHAPTER II. MARKOV CHAINS AND CHAIN AGGREGATES.

In this chapter we intend to introduce the two concepts of the Markov Chain and Markov Chain aggregate. The Markov chain is a special case of a class of systems known as Markovian systems. The Markov chain aggregate is a finite set of Markov chains with common transition matrix and as such it is an interesting system in many practical cases.

II.1. Markovian Systems.

A process is called Markovian system if it is described by a random variable $x(t); t \in T$, where T is a time set, and for any two points in time t and $t+h$ $(h \geq 0)$ the value $x(t+h)$ of the random variable x at time $t+h$ does not depend on the entire past evolution of the system, but rather on its last value $x(t)$.

In the above statement we used terms like process, system, random variable and expressions like "the value,..,does not depend on..etc.". Although there is a common understanding of these terms and similar expressions, still there is some ground for ambiguities and misunderstandings. However, one might suggest some revision of this verbal statement; we do not think that this would improve considerably the situation. Also another might suggest the use of mathematical language. Eventually, mathematical expressions provide the ultimate accuracy.

However, any symbol and concept involved should be rigorously defined; otherwise, mathematical expressions are more confusing than verbal statements.

To our view, rigorous definition of any mathematical concept means that it should be defined as either a set or a mapping from a certain set into another. Anythingelse, that does not comply with this requirement, does not serve mathematical rigor and usually generates gaps in communication. Thus, we will attempt to define the concept of a Markovian system along the prescribed guidelines. The reason for doing this, is the fact that we were unable to refer to a certain textbook or paper which defines the Markovian system concept without leaving undefined certain terms and concepts:

In appendix A, section A.4 we defined the concept of a system being a quadruple $\Sigma=(B,S,\mathcal{F},V)$. B is a set $(\neq\emptyset)$, called the background population, whose points are the objects under study. S is another set $(\neq\emptyset)$, called state-space or sample-space. If any of its points is associated with some point $b\in B$, then it is called "state" of the point $b\in B$. The "state" is supposed to provide all the information one needs to know about the associated point of the background population. Any mapping $e:B\rightarrow S$ is called experiment or observation and its image $e(b)$, for a particular point $b\in B$, is called the outcome of experiment e for the point $b\in B$. Furthermore, V is a countable set of experiments, called experiment space of the system Σ. The experiment space $V=\{e_1,e_2,\ldots\}$ is supposed to contain a comprehensive account of our experience on the objects under study, which are, as mentioned above, the points of the

background population. Finally, $\mathcal{F}$ is a Borel-field or σ-algebra defined on S. Its members are known as Borel-sets or events. In appendix A we elaborate on the above concepts.

A system $\Sigma=(B,S,\mathcal{F},V)$ is called a process if time is involved in the structure of the background population. Specifically, Σ is a process if the background population is in the Cartesian product form $B=C\times T$ where, T is a time-set, i.e. T is the real line or any of its subsets or, more generally, any totally ordered set to which we attach time meaning, and C is any arbitrary non-empty set. For example, suppose that we are interested to study the position in the 3-dimensional space R^3 of a set of certain particles a,b,c,... and at any point in time $t\in T$. Clearly, the state-space is $S=R^3$ and the background population is $B=\{a,b,c,...\}\times T$. Given a certain particle, say a, and time t, then the tuple (a,t) is one of the "objects" under study. Furthermore, given an experiment $e:B\rightarrow R^3$ the outcome e(a,t) specifies the position of the particle a at time t in the state-space R^3.

For a process $\Sigma=(C\times T,S,\mathcal{F},V)$ in which $C=\{a\}$, i.e. C consists of only one point, we can delete the set C and write $\Sigma=(T,S,\mathcal{F},V)$ for reasons of simplicity. Of course, the set C although deleted is implied in existence within the structure of the system. We restrict our attention to processes in this form $\Sigma=(T,S,\mathcal{F},V)$ with time-set T the real line or the set of real integers. In fact, the objects of study is some single object, which is implied, in conjunction with the points in time $t\in T$. The system Σ might be stochastic or might not. In the first case probability measures are attached to each event $E\in\mathcal{F}$, while in the second case, probability measures are meaningless

for such a system as it stands.

In a similar way as in section A.7 of appendix A, from the system

$$\Sigma = (T,S,\tilde{f},V) \tag{II.1}$$

one can generate other systems in the following general way:
Consider any subsets $T' \subset T$ and $V' \subset V$ where, V' is any arbitrary countable subset of V and T' is any arbitrary non-empty subset of T, and define the quadruple

$$\Sigma' = (T',S,\tilde{f},V') \tag{II.2}$$

which, by definition, is a system and, in particular, a process.
We note that it is possible that the original system (II.1) not to be stochastic while, the derived one (II.2) might be as such. Also we note that the choice of T' and V' is, in principle, general and arbitrary. However, if the choice of T' and V' is prescribed in a certain way one can derive processes which are meaningful and useful in probability theory.

Consider the process (II.1), which is not necessarily stochastic, and choose $n \geq 1$ finite number of points $t_1 > t_2 > \ldots > t_n$ from T. Also, let us choose an equal number of non-empty events $E_1, E_2, .., E_n$ from $\tilde{f}$. These events are not necessarily distinct like the time points $t_1, .., t_n$.
Now we can define

$$T' = T'(t_1,\ldots,t_n) = \{t \in T \mid t > t_1\} \tag{II.3}$$

$$V' = V'(t_1,\ldots,t_n;E_1,\ldots,E_n) = \{e \in V \mid e(t_i) \in E_i;\ i=1,\ldots,n\} \tag{II.4}$$

Assuming that V' is countable we can define the process

$$\Sigma' = \Sigma'(t_1,\ldots,t_n;E_1,\ldots,E_n) = (T',S,\tilde{f},V') \qquad \text{(II.5)}$$

The interpretation of the process (II.5) is illustrated in figure 1. First, our attention is focused on the points in time $t > t_1$ being the points of T' as defined by (II.3). Second, we concentrate on those experiments whose outcomes at $t_1,\ldots,t_n$ are lying within the events $E_1,\ldots,E_n$, respectively. For example, the shown in figure 1 experiment e_1 is not included in V' while, e_2 and e_3 are included because they pass through $E_1,..,E_n$ at $t_1,\ldots,t_n$, respectively.

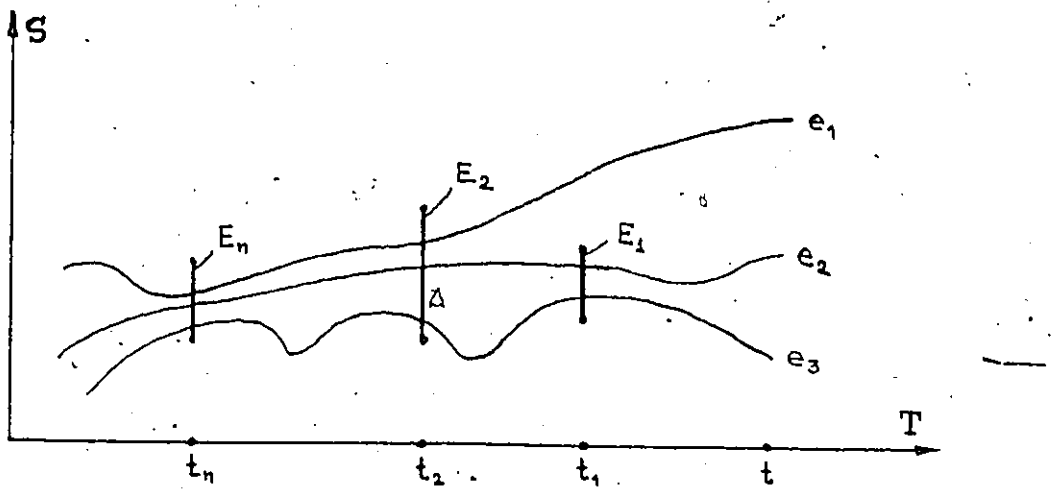


Figure 1.

Definition: We call a process $\Sigma=(T,S,\tilde{f},V)$ a Markovian system if and only if for every $t_1 > \ldots > t_n$ finite number of points in time and an equal number of non-empty events $E_1,\ldots,E_n$

(i) The quadruple $\Sigma' = (T',S,\tilde{f},V')$ as defined by equations (II.3),(II.4) and (II.5) is a stochastic system.

(ii) Σ satisfies the Markovian property, i.e.

$$\Pr\{E,t|E_1,t_1;..;E_n,t_n\} = \Pr\{E,t|E_1,t_1\} \quad \text{for all } E\in\tilde{f} \text{ and } t>t_1 \qquad \text{(II.6)}$$

We notice that, for a Markovian system, the so-called Markovian property is well-stated by identity (II.6). Specifically, by condition (i) the generated systems by the choice of $t_1,..,t_n;E_1,..,E_n$ and $t_1;E_1$, respectively, are stochastic. Hence, the probabilities involved in (II.6) are conditional probabilities, as defined in appendix A.7, and they are meaningful with respect to two conceptually different stochastic systems, both generated from the original process Σ. The Markovian property states that their conditional probabilities coincide for all $E\epsilon\tilde{\mathcal{F}}$ and $t>t_1$.

Traditionally, the Markovian systems are categorized in three independent ways: 1) as stationaty and non-stationary, 2) discrete and continuous state-space, 3) discrete and continuous time.

The continuous time and continuous state-space Markovian systems have been studied extensively (see [1],[4],[6],[13]). Basically, the conditional probabilities associated with these processes satisfy the well-known Kolmogorov partial differential equations.

The discrete state-space, i.e. $S=\{1,2,...\}$, either discrete or continuous time and either stationary or not Markovian systems are known as Markov Chains. This definition is due to Prof. Chung[5]. Many other authors define Markov chains in more restrictive terms. Markov chains have been extensively studied since 1907, when first were intrgduced by A.Markov. In this study we discuss certain aspects of the Markov chains. Specifically, our investigation in limited to the discrete-time Markov chains. Occasionally, however, we involve the continuous-time ones whenever a certain property is valid for both types.

II.2. Markov Chains and their Transition Matrices

Consider $\Sigma=(T,S,\tilde{\mathcal{F}},V)$ to be a Markov chain with state-space $S=\{1,2,\ldots\}$ either a finite or a countable set. The power set of S is chosen to be the Borel-field $\tilde{\mathcal{F}}$ on S and therefore, any subset of S will be considered as an event. The background population time-set T is considered to be either the real line (continuous case) or the set of real integers (the discrete-time case).

One should observe that any event $E\in\tilde{\mathcal{F}}$ is an, at most countable, union of the singletons $\{i\}$; $i\in E$ which are pairwise disjoint events. Consequently, if any probability measure function p is defined on $\tilde{\mathcal{F}}$ then, for any event $E\subset S$ its probability measure is the, at most countable, sum $p(E)=\sum_{i\in E}p(\{i\})$. For this reason the probabilities of the basic events $\{1\},\{2\},\ldots$ are sufficient to completely define the probability measure function in question. Throughout this investigation $p(\{i\})$ will be denoted by p_i while, if there are written any additional subscripts, superscripts or function arguments they will specify and denote other conditions.

Let us now pick some $j\in S$ fixed state and $t_o\in T$ fixed point in time. If we define $T'=\{t\in T \mid t\geq t_o\}$ and $V'=\{e\in V \mid e(t_o)=j\}$ then, $\Sigma'_j=(T',S,\tilde{\mathcal{F}},V')$ is a stochastic system, because Σ is a Markov chain, and induces probabilities $p(i,t\,|\,j,t_o)$ for all $i\in S$ and $t\geq t_o$. We denote these probabilities by $p_{ij}(t,t_o)$ and we interprete them as the probabilities that the system is at state $i\in S$ at time $t\geq t_o$, conditional

to that it was at state $j \in S$ at time t_o.

For any two points in time t_o and $t \geq t_o$ we can define the matrix

$$P(t,t_o) \equiv \begin{bmatrix} p_{11}(t,t_o) & p_{12}(t,t_o) & \dots \\ p_{21}(t,t_o) & p_{22}(t,t_o) & \dots \\ \dots & \dots & \dots \end{bmatrix} = \left(p_{ij}(t,t_o) \right)_{i,j \in S} \qquad \text{(II.8)}$$

which is called "the transition matrix from time t_o to time $t \geq t_o$".

Any transition matrix $P(t,t_o)$ has the following properties:

1) $P(t,t_o) \geq 0$ for all $t,t_o \in S$ $(t \geq t_o)$ (II.9a)

2) Its column sums is unity for all columns $j \in S$, i.e.

$$\sum_{j \in S} p_{ij}(t,t_o) = 1 \quad \text{for all } j \in S \text{ and } t \geq t_o$$

which in matrix form can be written

$$\bar{1}' P(t,t_o) = \bar{1}' \quad \text{for all } t \geq t_o \qquad \text{(II.9b)}$$

where, $\bar{1} = (1,1,1,\dots)'$ is a vector with all its entries equal to unity.

3) $P(t_o,t_o) = I = (\delta_{ij})$ (II.9c)

where, I is the identity matrix.

4) For any $t_3 \geq t_2 \geq t_1$ points in time

$$P(t_3,t_1) = P(t_3,t_2)\, P(t_2,t_1) \qquad \text{(II.9d)}$$

We emphasize the fact that the above properties of transition matrices are valid for any Markov chain i.e., with finitely or countably many states, stationary or not, and continuous or discrete time (see Chung [5], and Kemeny [12], pp. 33-35).

Professor K.L.Chung ([6] and [5]) has, almost exhaustively, studied the continuous-time Markov chains. One of his basic and important theorems is that a transition matrix $P(t,t_o)$ is continuous as function of t if and only if it is a measurable (elementwise) function of t. Therefore, except for practically unusual cases, one should expect the matrix $P(t,t_o)$ to be continuous in t, in which case, it is proven that $P(t,t_o)$ is differentiable and satisfies the backward and forward well-known Kolmogorov differential equations

$$\frac{d}{dt} P(t,t_o) = Q(t)\ P(t,t_o) \quad \text{and} \quad \frac{d}{dt_o} P(t,t_o) = P(t,t_o)\ Q(t)$$

where, $Q(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} (P(t+\Delta t,t)-P(t,t))$.

We do not intend to elaborate further on the continuous-time Markov chains. Instead, we will investigate the discrete-time Markov chains, i.e. $T = \{ t \mid \text{real and integer} \}$, for which (II.9d) yields,

$$P(t_o+k+1,t_o) = P(t_o+k+1,t_o+k)\ P(t_o+k;t_o) \quad \text{for any integer reals } t_o,\ k \geq 0 \ (I$$

By induction identity (II.10) yields,

$$P(t_o+k+1,t_o) = P(t_o+k+1,t_o+k)\ P(t_o+k,t_o+k-1)\ldots P(t_o+1,t_o) =$$

$$= \prod_{m=k}^{o} P(t_o+m+1,t_o+m) \qquad (II.$$

Thus, by the above identity (II.11) a multi-step transition matrix can be expressed in terms of the single-step ones.

The non-stationary discrete-time Markov chains have met very little interest in the literature. This is, perhaps, due to their property to change characteristics, i.e. their transition matrix, in any arbitrary way and in any number of time-steps.

A Markov chain is, by definition, called stationary if

$$P(t,t_o) = P(t-t_o,0) \quad \text{for all} \quad t \geq t_o$$

Hence, we may denote the k-step transition matrix by

$$P^{(k)} = P(t_o+k,t_o) = P(k,0) \quad \text{for all } t_o \in T \text{ and } k \geq 0$$

and in particular, for the single-step transition matrix we delete the superscript k=1 and we denote

$$P = P^{(1)} = P(1,0) = P(t_o+1,t_o) \quad \text{for any } t_o \in T.$$

In this case, identity (II.11) yields

$$P^{(k)} = P(t_o+k,t_o) = P.P...P = P^k \quad \text{for all } t_o \in T \text{ and } k \geq 0 \qquad \text{(II.12}$$

i.e. the superscript $k \geq 0$ becomes an exponent.

The remaining of this study will be focused on the stationary discrete-time Markov chains. These systems are completely described by their (single-step) transition matrix $P = (p_{ij})$. Any other transition matrix of the system for any number of time-steps $k \geq 0$ can be obtained by exponentiation of P, by virtue of identity (II.12).

Our approach is one that is understated in the literature of Markov

chains and not fully exploited. Mainly, our approach consists of the analysis of the transition matrices via the analysis of their spectrum. We are not, even to that effect, general enough because we only discuss the finite-state stationary case in which the spectrum of a transition matrix coincides with the set of its eigenvalues. The position of the eigenvalues of a transition matrix in the complex plane play a determinant role on the asymptotic behavior of a Markov chain.

In chapters III and IV when we refer to Markov chains or transition matrices we mean stationary, finite-state and discrete-time ones.

II.3. Markov Chain Aggregates.

Consider a finite set $\Sigma_1, \Sigma_2, \ldots, \Sigma_N$, for some $N \geq 1$, of Markov chains with common state-space $S = \{1,2,\ldots,n\}$ and common transition matrix $P(t,t_o)$ for all points in time (continuous or discrete) $t \geq t_o$. We will call this set $\Delta = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}$ a "Markov Chain Aggregate" or abbreviated, a MCA. The term "aggregate" is chosen as synonymous to a set.

The reason of the introduction of the MCA lies on the fact that, in most real-world situations, Markovian systems, and in particular Markov chains, are members of homogeneous groups. All the members of such groups appear, or assumed, to have the same chances of transition from state to state. For example, the Canadian Interprovincial Migration process can be seen as a MCA with its members as all the families residing in Canada and their state-space being the set of the ten provinces. It is clear, that the study of the "chances", i.e. probabilities, of transition from a province to another for a patricular family is not adequate for the study of the over-all inerprovincial migration problem. Thus, the set of all families, i.e. the MCA of the Canadian families, must be studied as one entity.

Another example could be an industrial plant with similar pieces of machinery. The state-space might be $S = \{$"disposable","fails to operate", "0-1,000 hours of operation","1,000-2,000 hrs",...,"10,000 & over hrs"$\}$, or any similar set with, perhaps, different structure of the accumulated hours of operation brackets. Evidently, the "good" or "bad" state of a

particular machine has little to do with the overall productivity of the plant, especially, if the plant uses a large number of them. However, if at a particular point in time we have complete account as to how many machines are in each state, this would constitute a useful information in order to assess the productivity of the plant. But this is, what we call, the state-vector of the MCA as defined below:

Definition: For a finite Markov Chain Aggregate $\Delta = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}$ for some $N \geq 1$ integer, we call state-vector at $t \in T$ the n-tuple $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))'$ whose entries $x_i(t)$; $i=1,\ldots,n$, at $t \in T$, are non-negative integers indicating how many members of Δ are in state i and at time t.

Obviously, $$\sum_{i=1}^{n} x_i(t) = N \quad \text{for all } t \in T \tag{II.13}$$ since, the MCA is a constant population set of Markov chains.

The problem associated with a certain Markon chain aggregate can be outlined as follows: At some initial point in time $t_o \in T$ the state of each member of an MCA is given. Thus, the state-vector $x(t_o)$ is assumed known, i.e. we know how many members of the MCA are in each state at $t_o \in T$. Our purpose, then, is to study and assess the value of the state-vector $x(t)$ of the MCA at any future point in time $t \geq t_o$. Specifically, since Markov chains are stochastic processes, assessment of the state-vector $x(t)$ must be understood as the evaluation of the statistical characteristics of this vector. Chapter V is devoted to this study. Our findings are at their "first attempt" stage and therefore far from being exhausting and comprehensive on the subject matter.

CHAPTER III. STRUCTURAL PROPERTIES OF TRANSITION MATRICES.

The study of transition matrices must be identified with the study of the Markov chains they characterize. In this and the next one chapter by Markov chain and transition matrix we mean stationary, discrete-time and finite-state Markov chain and transition matrix, respectively.

The first two sections of this chapter deal with basic definitions and propositions. Although the exposition was intended to be restricted to transition matrices, it has been generalized to non-negative matrices. The reason for this generalization will become clear at the last section of the chapter, where we discuss the partitioning of a finite state system, characterized by a square matrix, in recurrent and transient sub-systems.

III.1. Definitions.

D.III.1. A non-negative square matrix P of order $n \geq 1$ is called "transition matrix" if its column sums are all equal to unity, i.e.

$$\sum_{i=1}^{n} p_{ij} = 1 \quad \text{for all} \quad j = 1,\ldots,n \tag{III.1}$$

If we denote by $p_1,\ldots,p_n$ the columns and $q_1',\ldots,q_n'$ the rows of the transition matrix P, then (III.1) can take the matrix notation forms:

$$\bar{1}' P = \bar{1}' \tag{III.2}$$

$$\sum_{i=1}^{n} q_i = \bar{1} \tag{III.2a}$$

and $$\bar{1}' p_i = 1 \qquad \text{for all} \quad i = 1,\ldots,n \tag{III.2b}$$

D.III.2. For any square non-negative matrix P of order $n \geq 1$, we call any integer $i \in \{1,\ldots,n\}$ "state i". Furthermore, state i will be "accessible or reachable" from state j if there exists some non-negative integer $t \geq 0$ such that $p_{ij}^{t} > 0$. The relation of accessibility will be denoted by $j \rightarrow i$.

D.III.3. States i and j, for a given non-negative square matrix P, are said to communicate if they are accessible from each other. This will be denoted by $i \leftrightarrow j$.

The communication relation, as defined above between two states of a non-negative matrix P, is an equivalence relation because,

(a) $i \leftrightarrow i$ since $p_{ii}^{0} = 1 > 0$, i.e. $\leftrightarrow$ is reflexive,

(b) $i \leftrightarrow j$ implies, by definition, $j \leftrightarrow i$, i.e. $\leftrightarrow$ is symmetric,

(c) $i \leftrightarrow j$ and $j \leftrightarrow k$ implies, $i \leftrightarrow k$, i.e. $\leftrightarrow$ is transitive.

Proof of transitivity: Since i reaches j and j reaches k there exist non-negative integers u and v such that, $p_{ji}^{v} > 0$ and $p_{kj}^{u} > 0$.

Thus, $p_{ki}^{u+v} = (P^{u+v})_{ki} = (P^{u}\ P^{v})_{ki} = \sum_{a=1}^{n} p_{ka}^{u} p_{ai}^{v} \geq p_{kj}^{u}\ p_{ji}^{v} > 0$ and i reaches k. Similarly, k reaches i as a result of $k \rightarrow j$ and $j \rightarrow i$. Therefore, $k \leftrightarrow i$. Q.E.D.

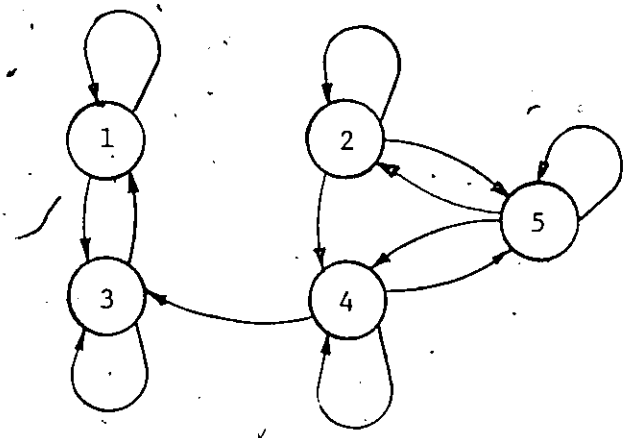
D.III.4. We will call the set of states $S = \{1,\ldots,n\}$ of a non-negative matrix P "the state-space of P". Since the communication relation is an equivalence relation on S, it can partition S into equivalence classes $S_1,\ldots,S_h$ which are disjoint subsets of S and their union is S.

We call $S_1,\ldots,S_h$ communication classes or simply classes.

Example: Let,

$$P = \begin{bmatrix} x & 0 & x & 0 & 0 \\ 0 & x & 0 & 0 & x \\ x & 0 & x & x & 0 \\ 0 & x & 0 & x & x \\ 0 & x & 0 & x & x \end{bmatrix} \begin{matrix} (1) \\ (2) \\ (3) \\ (4) \\ (5) \end{matrix}$$

(1) (2) (3) (4) (5)



where, "x" indicates non-zero entry. The shown diagram helps to visualize the two communication classes $S_1=\{1,3\}$ and $S_2=\{2,4,5\}$. We notice that any state of S_1 is reachable from any state of S_2 but, none of S_2 can be reached from any state of S_1.

Sometimes, we will say "class A reaches class B" and we will mean that any state of A reaches any state of B.

D.III.5. If the state-space $S=\{1,\ldots,n\}$ of a non-negative matrix P consists of only one communication class then, the matrix P is called "irreducible or indecomposable".

If a particular class is not accessible by any other class and also cannot reach any other class then, by proper relabelling of the states, i.e. by combined elementary row and column transformations, the non-negative matrix can be written in the form $P=\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$ where P_1 and P_2 are square non-negative matrices. Most of the times we rule out such cases, because, it leads to two independent non-negative matrices with no interaction between them.

D.III.6. If a certain class $S_m \subset S$ cannot reach any other class will be called "recurrent class" and its states "recurrent states".

If, on the other hand, $S_m \subset S$ can reach, at least, another class, it will be called "transient class" and its states "transient states".

An irreducible non-negative matrix, obviously, consists of recurrent states forming one recurrent class. However, if the matrix is not irreducible we always assume that there exists at least one transient class reaching some of the recurrent classes. The opposite case provides to the non-negative matrix in question trivially easy structure.

D.III.7. A non-negative matrix P is called "regular" or "ergodic" if for some positive integer t, $P^t > 0$.

Clearly enough, a regular matrix is irreducible while, the converse is not always true. For example, the Hadamard matrix $H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is irreducible but not regular.

The Canonical Form: Let P be a non-negative matrix and $S_1, \ldots, S_m$ $(m \geq 1)$ be its recurrent classes with $r_1, \ldots, r_m$ being the number of states in each of them, respectively. Also let $S_{m+1}, \ldots, S_{m+k}$ transient classes with $r_{m+1}, \ldots, r_{m+k}$ number of states, respectively.

By proper relabelling, i.e. by combined row and column elementary transformations, one can put first the r_1 states of the class S_1, followed by the r_2 states of the class S_2 etc., until all classes are accomodated.

Evidently, the matrix will take the form

$$P = \left[\begin{array}{ccc:ccc} P_1 & & 0 & L_{1,m+1} & & L_{1,m+k} \\ & \ddots & & & & \\ 0 & & P_m & L_{m,m+1} & & L_{m,m+k} \\ \hdashline 0 & \cdots & 0 & P_{m+1} & & L_{m+1,m+k} \\ \vdots & & \vdots & & \ddots & \\ 0 & \cdots & 0 & & & P_{m+k} \end{array}\right]$$

where, the diagonal matrices $P_1,\ldots,P_{m+k}$ are square and irreducible, while, for any $g = m+1,\ldots,m+k$ the matrices $L_{1,g},\ldots,L_{g-1,g}$ cannot all of them be zero.

This form will be called the "canonical form" of the matrix P. In section III.3 a procedure is derived for arranging a non-negative matrix in its canonical form.

III.2. General and Basic Propositions.

In this section are stated a few basic propositions which will be useful later on, especially in section III.3. Most of them are included for purpose of consistency and completeness. However, there are two important propositions in this section, namely, TH.4 and TH.5. The theorem TH.4 was suggested by Gill's paper ([23], pp. 49-50) on finite-state systems. On the other hand, theorem TH.5 provides a powerful criterion for the irreducibility of matrices.

TH.1. If P and Q are transition matrices of the same order, so is their product PQ.

Proof: Let $P' = [p_1,\ldots,p_n]$ and $Q = [q_1,\ldots,q_n]$ partitioned by their rows and columns, respectively.

For all $i,j = 1,\ldots,n$ $(PQ)_{ij} = p_i' q_j \geq 0$. Thus, PQ is non-negative.

Furthermore, $\sum_{i=1}^{n}(PQ)_{ij} = (\sum_{i=1}^{n} p_i') q_j = \bar{1}' q_j = 1$ for all $j=1,\ldots,n$.

That is all column sums of the matrix $PQ \geq 0$ are equal to unity. Q.E.D.

Corollary 1.1: If P is a transition matrix, so is any finite non-negative power P^m ; $m = 1,\ldots,k,..$

Proof: Trivial by induction theorem of the natural numbers.

TH.2. If $P_1,\ldots,P_k$ $(k \geq 2)$ are transition matrices of the same order, so is any convex combination of them.

Proof: Let $P = u_1P_1 + \dots + u_kP_k$ where, $\sum_{i=1}^{k} u_i = 1$ and $u_i \geq 0$; $i = 1,\dots,k$. Obviously, $P \geq 0$. While, $\bar{1}' P = \sum_{i=1}^{k} u_i \bar{1}' P_i = \sum_{i=1}^{k} u_i \bar{1}' = \bar{1}'$. Q.E.D.

TH.3. If P and Q are transition matrices of the same order the following inequalities hold for any $i,j = 1,\dots,n$:

$$\min_{g,h}(p_{gh}) \leq \min_{h}(p_{ih}) \leq (PQ)_{ij} \leq \max_{h}(p_{ih}) \leq \max_{g,h}(p_{gh})$$

Proof: $(PQ)_{ij} = \sum_{m=1}^{n} p_{im}\, q_{mj} \geq \sum_{m=1}^{n} (\min_{h}(p_{ih}))\, q_{mj} = \min_{h}(p_{ih})$; $i,j = 1,\dots,n$.

Similarly, $(PQ)_{ij} \leq \max_{h}(p_{ih})$ for any $i,j = 1,\dots,n$.

The other inequalities are trivially easy to prove. Q.E.D.

The interpretation of TH.3 is that if a transition matrix P is post-multiplied by any other transition matrix then each row of the product matrix is, elementwise, bounded by the minimal and maximal elements of the corresponding original matrix row. The proposition can be revised for pre-multiplication of P by another transition matrix. In this case, the property of boundness is valid for the columns of P.

Corollary 3.1. If P is a transition matrix then, for any $m \geq k \geq 0$

$$\min_{i,j}(p_{ij}^{m}) \geq \min_{i,j}(p_{ij}^{k}) \quad \text{and} \quad \max_{i,j}(p_{ij}^{m}) \leq \max_{i,j}(p_{ij}^{k}).$$

Corollary 3.2. If P is a transition matrix and for some integer $k > 0$ $P^k > 0$ then, $P^m > 0$ for all integers $m \geq k$.

The proofs of these two corollaries will be omitted as being simple and direct implications of TH.3.

TH.4. For any square non-negative matrix of order $n>1$, if state i is reachable from state j then, transition is possible in at most $k \leq n$ steps. That is, $p_{ij}^{k} > 0$ for some $k \leq n$.

Proof: Since i is reachable from j, there exists integer $k \geq 0$, such that, $p_{ij}^{k} > 0$. Let k be the smallest non-negative integer possessing this property, i.e. $p_{ij}^{k} > 0$. If $k \leq 1 \leq n$ the proposition holds. If $k > 1$ then,

$$p_{ij}^{k} = \sum_{g=1}^{n} p_{ig}\, p_{gj}^{k-1} > 0$$

Since the sum of non-negative terms is strictly positive, there exists state g_1 such that the corresponding term is strictly positive, i.e. $p_{ij}^{k} \geq p_{i,g_1}\, p_{g_1,j}^{k-1} > 0$. Using similar arguments, one can find states $g_2,\ldots,g_{k-1}$ such that,

$$p_{ij}^{k} \geq p_{i,g_1}\, p_{g_1,g_2} \cdots p_{g_{k-1},j} > 0.$$

The fact that k is the smallest integer such that $p_{ij}^{k} > 0$ implies that $g_1, g_2, \ldots, g_{k-1}$ are all distinct. For otherwise, if we assume that $g_u = g_v$ for some $u > v \geq k-1$ then,

$$p_{ij}^{k-(v-u)} \geq p_{i,g_1} \cdots p_{g_{u-1},g_u}\, p_{g_v,g_{v-1}} \cdots p_{g_{k-1},j} > 0,$$

which contradicts the original assumption that k is the smallest integer for which $p_{ij}^{k} > 0$.

Therefore, since $g_1,\ldots,g_{k-1}$ are distinct and not i or j then, $k \leq n-1$ (if $i=j$) and $k \leq n$ (if $i \neq j$). Q.E.D.

TH.5. The necessary and sufficient condition for a non-negative square matrix of order $n>1$ to be irreducible is that,

$$P + P^2 + \ldots + P^n > 0.$$

Proof: Necessity- Since P is irreducible any state i can be reached from any state j. Thus, by TH.4 for all $i,j = 1,\ldots,n$, $p_{ij}^k > 0$ for some $k \leq n$. Therefore, $(P + P^2 + \ldots + P^n)_{ij} \geq p_{ij}^k > 0$ for all $i,j = 1,\ldots,n$.

Sufficiency- If $(P + P^2 + \ldots + P^n)_{ij} > 0$ for all states $i,j = 1,\ldots,n$ then, at least one of the matrices $P, P^2, \ldots, P^n$ have entry (i,j) strictly positive. This implies that any state j reaches any other state i. Therefore, all states communicate among themselves and the non-negative matrix P is irreducible. Q.E.D.

The above proposition can be found in Gantmacher (ref. [10]). However, Gantmacher's proof is entirely different from the one offered above. One advantage of our approach is the fact that the above result can be easily generalized to matrices from any field. The same applies to all structural properties we will present in the following section.

III.3. Canonical Form of Square Non-negative Matrices.

In section III.1 the canonical form of a non-negative square matrix was introduced. The procedure to derive the canonical form requires that the communication classes have been established. Therefore, we should first show some way of establishing the communication equivalence classes.

Let P be a square non-negative matrix in its canonical form

$$P = \left[\begin{array}{ccc|ccc} P_1 & & 0 & L_{1,m+1} & \cdots & L_{1,m+k} \\ & \ddots & & & & \\ 0 & & P_m & L_{m,m+1} & \cdots & L_{m,m+k} \\ \hline & & & P_{m+1} & \cdots & L_{m+1,m+k} \\ & 0 & & & \ddots & \\ & & & 0 & & P_{m+k} \end{array}\right]$$

and define $W = P + \ldots + P^n$. Clearly, this matrix will be in the form

$$W = \left[\begin{array}{ccc|ccc} P_1^* & & 0 & L_{1,m+1}^* & \cdots & L_{1,m+k}^* \\ & \ddots & & & & \\ 0 & & P_m^* & L_{m,m+1}^* & \cdots & L_{m,m+k}^* \\ \hline & & & P_{m+1}^* & \cdots & L_{m+1,m+k}^* \\ & 0 & & & \ddots & \\ & & & 0 & & P_{m+k}^* \end{array}\right]$$

where, in particular, the diagonal matrices are the sums $P_h^* = P_h + \ldots + P_h^n$ for $h = 1,..,m+k$.

The diagonal matrices $P_1^*,\ldots,P_{m+k}^*$ of the above form for W are strictly positive by the necessity part of the TH.5 (sec. III.2).

Conversely, $P_1,\ldots,P_{m+k}$ are irreducible by the sufficiency part of the TH.5 if $P^*_1,\ldots,P^*_{m+k}$ are strictly positive. These properties can be used to recover the communication classes. To this end, we will state now a few propositions, which are all corollaries of TH.4 and TH.5 of the previous section. We consider here any non-negative matrix P (not necessarily in its canonical form) and the corresponding square non-negative matrix $W = P + \ldots + P^n$ where, $n \geq 1$ is the order of P.

Proposition 1: States $i \neq j$ belong to the same communication class if and only if $W_{ij} \neq 0$ and $W_{ji} \neq 0$.

Proof: If $i \leftrightarrow j$ then clearly, by TH.5 $W_{ij} \neq 0$ and $W_{ji} \neq 0$. Conversely, if $W_{ij} \neq 0$ then, by TH.5 $j \rightarrow i$, and also $W_{ji} \neq 0$ implies $i \rightarrow j$. Thus, $i \leftrightarrow j$ if W_{ij} and W_{ji} are both non-zero. Q.E.D.

Note: If $i \leftrightarrow j$ then, $W_{ii} \neq 0$ and $W_{jj} \neq 0$.

Proposition 2: If $W_{ii} = 0$, for some state i, then this state forms by itself a communication class.

Proof: If $j \neq i$ belongs to the same class with i then, this implies that $i \rightarrow j$ in $k \leq n$ steps, and $j \rightarrow i$ in $k'' \leq n$. This, however, implies that $i \rightarrow i$ in some number of steps k ($\leq n$, by TH.4) which yields, $W_{ii} \neq 0$ contradicting the assumption $W_{ii} = 0$. Q.E.D.

The first proposition can be used to establish the fact that two states belong or not to the same class. The second proposition takes care

of a certain trivial case of one-state class. It must be noted that in order that a state i to form a class by itself it is not necessary that $W_{ii}=0$.

Proposition 3: A communication class $C=\{i,\ldots\}$ consists of recurrent states if and only if for all $i\in C$ and all $k\notin C$ $W_{ki}=0$.

Proof: If C is a recurrent class then, any $i\in C$ does not reach any $k\notin C$ which implies $W_{ki}=0$. Conversely, if $W_{ki}=0$ for all $i\in C$ and $k\notin C$ then, there is no state of any other class reaching C. Thus, C is a recurrent state class. Q.E.D.

Using Proposition 1. we can check one by one all the states if they belong into the same class with a certain state i. Therefore, the establishment of the communication classes is a direct application of the proposition 1.

For each class, if we apply the criterion of proposition 3, we can establish whether or not it is a recurrent class. Of course, if a certain class is not recurrent it is a transient one.

Since our criteria of communication are based on the zero or non-zero values of the matrix $W = P + \ldots + P^n$, we shall use binary arithmetic in order to reduce and simplify the computational complexity. The computer FORTRAN program "MPOWER", listed in Appendix B, can recover the communication classes of any square matrix. Indeed, although the analysis has been carried out on non-negative square matrices; the results are applicable to any

square matrix whose entries are from any field. This is possible by replacing all non-zero entries by 1 (unity) and operate on the resulted non-negative matrix.

The establishment of the communication classes and whether each of them is a recurrent or a transient class does not completely solve the problem. We, further, need to order them so that the canonical form will be directly induced. To this end we can proceed as follows:

If there is not any transient class then, the problem is trivially easy. Any ordering of the (recurrent) classes is valid and will produce canonical form

$$P = \text{diag}\,(P_1,\ldots,P_m) = \begin{bmatrix} P_1 & & 0 \\ & \ddots & \\ 0 & & P_m \end{bmatrix}$$

with the diagonal matrices being square irreducible.

If there exists at least one transient class we proceed as follows: We order the recurrent classes, in any arbitrary way, and consider them first. If there is only one transient class we order it last and we are finished. If there are more than one transient class then, there is at least one transient class which is not accessible by any other class. This is true because if a certain class C_1 reaches another class C_2, this implies that C_2 does not reaches C_1. For otherwise there would be communication between C_1 and C_2. Such a class we set last and we ignore it from the further procedure. Among the remaining transient classes it is, again, found one that is not accessible by the others and order it just

before the last ordered one. Ignoring it again we proceed the same way with the remaining till we order them all.

As a final result we obtain $C_1,\ldots,C_m$ recurrent classes and $C_{m+1},\ldots,C_{m+k}$ transient ones and they are ordered by their indices. By interchanging the corresponding rows and columns of the originally given matrix P, we can set the states of C_1 class first, the states of C_2 second, etc. to derive the canonical form of P.

The above procedure will be best illustrated by an example.

Example: Consider the 11×11 transition matrix

$$P = \begin{array}{c} \begin{matrix} (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) & (11) \end{matrix} \\ \begin{bmatrix} \frac{1}{2} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{2}{3} & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{3}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \\ (7) \\ (8) \\ (9) \\ (10) \\ (11) \end{matrix}$$

As stated earlier our interest is focused on the zero and non-zero entries of the matrix P. Thus, it is convenient to replace any non-zero entry of P by the symbol "1". Addition and multiplication rules will be agreed to be as follows: 0 + 1 = 1, 0 + 0 = 0, 1 + 1 = 1, 0*0 = 0, 0*1 = 0, and 1*1 = 1. The corresponding matrices P and $W = P + \ldots + P^{11}$ become

	1	2	3	4	5	6	7	8	9	10	11
1	1	0	1	0	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	0	1	1	0
3	1	0	0	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	1	0	0	0
5	0	0	0	1	1	0	0	0	0	1	0
6	0	0	0	0	0	1	0	0	0	0	1
7	0	1	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	1	0	1	0	0	0	0
10	0	0	0	0	0	0	0	1	0	1	0
11	0	0	0	0	0	0	0	0	0	0	0

P

	1	2	3	4	5	6	7	8	9	10	11
1	1	0	1	0	0	0	0	1	0	0	1
2	0	1	0	1	1	0	1	1	1	1	0
3	1	0	1	0	0	0	0	1	0	0	1
4	0	0	0	1	1	0	0	1	0	0	0
5	0	0	0	1	1	0	0	1	0	1	0
6	0	0	0	0	0	1	0	0	0	1	1
7	0	1	0	1	1	0	1	1	1	1	0
8	0	0	0	0	0	0	0	1	0	0	0
9	0	1	0	1	1	0	1	1	1	1	0
10	0	0	0	0	0	0	0	1	0	1	0
11	0	0	0	0	0	0	0	0	0	0	0

$W = P + P^2 \ldots + P^{11}$

The elements of the matrix $W=P+P^2+..+P^{11}$ are shown above and have been computed by the FORTRAN program MPOWER (see appendix B). Next we consider how to recover the communication classes working on the above matrix W and following the rules as stated above.

1-st communication class: We start with state 1. By inspection of the first column of W we see that the only candidate state, to belong to the same class with state 1, is state 3 (according to Prop.1). Indeed, state 1

communicates with state 3 because, $W_{11}=W_{13}=W_{33}=W_{31}=1$. Thus, $C_1=\{1,3\}$ form a communication class. This class C_1 is also recurrent due to Prop. 3. Specifically, the 1st and 3rd columns of W contain everywhere zeroes except at the 1st and 3rd rows.

Conclusion: $C_1=\{1,3\}$ is recurrent-state class.

2-nd communication class: Consider state 2. From column 2 we see that states 7 and 9 are the only candidate states to be associated with state 2 as members of the same class. For the same reason as before in C_1, we derive the conclusion that $C_2=\{2,7,9\}$ is a recurrent-state class.

3-rd communication class: consider state 4. From column 4 of W we see that candidate states can be only 2,5,7 and 9. Since 2,7 and 9 belong to the already established class C_2, they are excluded. The test according to the Prop. 1 for state 5 yields that, states 4 and 5 form a communication class. Furthermore, since state 4 reaches states 2,7 and 9 the requirements of the Prop. 3 are not fulfilled and this class is transient.

Conclusion: $C_3=\{4,5\}$ is a transient communication class.

4-th communication class: consider state 6. From column 6 of W we see that this state forms by itself a class and furthermore, this class is a recurrent one. Conclusion: $C_4=\{6\}$ is a recurrent-state communication class.

5-th communication class: Consider state 8. From column 8 we see that candidate states are 1,2,3,4,5,7,9, and 10. However, excluding those states which already have been considered we end up with only state 10. The test,

according to Prop. 3, fails because, $W_{8,10}=0$, and therefore, state 8 forms by itself a class $C_5=\{8\}$. The class C_5 is, obviously, transient because, $8 \longrightarrow 1,2,3,4,5,7,9,10$.

Conclusion: $C_5=\{8\}$ is a transient-state communication class.

6-th communication class: Easily enough it can be seen that $C_6=\{10\}$ is a transient-state communication class.

7-th communication class: According to Prop. 2 the only left state 11 is a transient state and form by itself the class $C_7=\{11\}$.

Among the established classes C_1, C_2 and C_4 are recurrent-state communication classes while, C_3, C_5, C_6 and C_7 are transient-state ones. Among the transient-state classes C_7 and C_3 do not reach any transient class. While, between the remaining C_5 and C_6 the class C_6 does not reach C_5. For these reasons, we order the classes as follows: $C_1, C_2, C_4, C_7, C_3, C_6, C_5$. The canonical form of the original matrix is a direct result of this ordering and is given below

	(1)	(3)	(2)	(7)	(9)	(6)	(11)	(4)	(5)	(10)	(8)
(1)	1/2	1					1/2				
(3)	1/2	0									1/4
(2)			1/3	0	1					1/3	
(7)			2/3	1/4	0						
(9)			0	3/4	0				1/3		
(6)						1	1/2				
(11)							0				
(4)								0	1/3		1/4
(5)								1	1/3	1/3	
(10)										1/3	1/4
(8)											1/4

$= P_c$

As stated earlier the canonical form of a square matrix is obtained by proper interchanges of rows and columns, simultaneously. This procedure is known as combined elementary column and row operations. In fact, the canonical form of a given square matrix is an orthogonal transformation of the original matrix, i.e.

$P_c = M^{-1} P M$ where, $M^{-1} = M'$

and M is the identity matrix of the same order as P whose columns have been interchanged properly. For the above numerical example, the orthogonal transformation matrix $M = [e_1, e_3, e_2, e_7, e_9, e_6, e_{11}, e_4, e_5, e_{10}, e_8]$.

The canonical form for transition matrices will be used in the next chapter to study the limit properties of the transition matrices. It has been already stated that the above presented procedure for setting a non-negative matrix into its canonical form, is general enough to be applied to any square matrix whose entries are from any field. The canonical form of a matrix can be proven useful for any system that is characterized by a square matrix. For example, the linear dynamical systems

$\dot{x} = A x + B u$

can be reduced to a set of smaller order dynamical systems if the square matrix A is reducible. To show this, consider M the orthogonal matrix that reduces A into its canonical form. Thus, by the transformation

$x = M y$ or, $y = M^{-1} x = M' x$

the system becomes

$$\dot{y} = (M' A M)\, y + M'Bu = A_c y + B_1 u$$

where, A_c is the canonical form of A.

Just for illustration purpose assume that, $A_c = \begin{bmatrix} P_1 & L \\ 0 & P_2 \end{bmatrix}$ where, P_1, P_2 are square matrices of orders n_1, n_2, respectively. If we partition $y' = (y_1', y_2')$, where, y_1 is the vector containing the lst n_1 components of y, and y_2 the remaining n_2 ones then, the system becomes

$$\dot{y}_1 = P_1 y_1 + L\, y_2 + B_1 u$$

$$\dot{y}_2 = P_2 y_2 + B_2 u$$

The derived form has the technical advantage that the original system is reduced to two dynamical sub-systems both of smaller order. Furthermore, the second sub-system is independent of the first. On the other hand, the state-vector of the second can be considered as another input vector to the first sub-system.

CHAPTER IV

LIMIT PROPERTIES OF TRANSITION MATRICES

CHAPTER IV. LIMIT PROPERTIES OF TRANSITION MATRICES

The purpose of this chapter is to study the behavior of the powers of transition matrices P^m as the exponent m tends to infinity. We recall that any power P^m of a transition matrix ($m \geq 0$) consists of the transition probabilities between any two states in m time-steps for a stationary discrete-time Markov chain. The behavior of the powers P^m for any integer $m \geq 0$ is directly connected with the structure of P, i.e. its decomposition into equivalence communication classes, and also the position of its eigenvalues in the complex plane.

Section IV.1 considers the necessary theoretical background by presenting the properties of the transition matrix eigenvalues.

Section IV.2 presents the problem of the limits for the powers of the transition matrices. The existence of these limits is considered and, whenever these limits do not exist, a substitute limit is proposed, called "average transition matrix", which always exists.

Section IV.3 deals with the technical aspects of the computation of these limits. Finally, in section IV.4 a numerical example is treated in order to illustrate the findings and exposition of the previous sections.

IV.1. Eigenvalues of the Transition Matrices.

The discussion starts with the statement of the well-known Frobenius theorem. The theorem's proof is omitted but one can find a simple version in Gantmacher[10], vol.2, pp. 53-65.

Frobenius Theorem: An irreducible non-negative square matrix of order $n \geq 2$ always has some positive eigenvalue $r > 0$ that is a simple root of its characteristic equation. The magnitudes of all the other eigenvalues do not exceed r. To this "maximal" eigenvalue there corresponds an eigenvector with strictly positive coordinates.

Moreover, if the matrix has $m \geq 1$ eigenvalues $\lambda_1, \ldots, \lambda_m$ of magnitude r, then these numbers are all distinct and are roots of the equation $\lambda^m - r^m = 0$. Furthermore, the whole spectrum of the matrix eigenvalues, regarded as a system of points in the complex plane, goes over into itself under a rotation of the plane by the angle $2\pi/m$. If $m > 1$ then the matrix can be put, by means of combined elementary column and row operations, into the following "cyclic" form where there are square blocks along the main diagonal:

$$\begin{bmatrix} 0 & A_{12} & 0 & \ldots & 0 \\ 0 & 0 & A_{23} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & A_{m-1,m} \\ A_{m1} & 0 & 0 & \ldots & 0 \end{bmatrix}$$

In the proof given by Gantmacher it is, incidentally, proven that the adjoint matrix of $(r\,I - A)$ is strictly positive, where A is the originally given non-negative square matrix and r its "maximal" real eigenvalue. This result will be used in the following Th.6.

The Frobenius theorem establishes the existence of a maximal positive eigenvalue but does not assess its magnitude. The following theorem Th.6 gives some bounds for this Frobenius eigenvalue and will be used later on.

TH.6. Let P be an irreducible non-negative square matrix of order $n>1$ and $S_j = \sum_{i=1}^{n} p_{ij}$; $j = 1,\ldots,n$ be its column sums. If r is the "maximal" real eigaenvalue of P then,

either, $r = \min_j(S_j)$ if $\min_j(S_j) = \max_j(S_j)$

or, $\min_j(S_j) < r < \max_j(S_j)$.

Proof: According to the concluding remark of the Frobenius theorem statement, the adjoint matrix of $(rI - P)$ is strictly positive. That is,

$B = (b_{ij}) = \mathrm{adj}(rI-P) > 0.$

Also, since r is an eigenvalue of P, the matrix $(rI - P)$ is singular.

Therefore, $(rI - P)\, B = (0)$

or, $r\, b_{ij} - \sum_{k=1}^{n} p_{ik} b_{kj} = 0$ for all $i,j = 1,\ldots,n.$

For any fixed j we sum over $i = 1,\ldots,n$ to derive

$$r\,(\sum_{i=1}^{n} b_{ij}) = \sum_{k=1}^{n}(\sum_{i=1}^{n} p_{ik})\, b_{kj} = \sum_{k=1}^{n} S_k\, b_{kj} \qquad \text{(IV.1)}$$

but since $b_{ij} > 0$ for all $i,j = 1,\ldots,n$ we can define

$u_k = b_{kj}/(\sum_{i=1}^{n} b_{ij}) > 0$; $k = 1,\ldots,n$, for which $\sum_{k=1}^{n} u_k = 1$.

Therefore, equation (IV.1) becomes, $r = \sum_{k=1}^{n} u_k S_k$ (IV.2)

If $\min_j(S_j) = \max_j(S_j)$ then, $S = S_1 = \ldots = S_n$ and (IV.2) yields

$r = (\sum_{k=1}^{n} u_k) S = 1 \cdot S = S = \min_j(S_j) = \max_j(S_j)$.

If $\min_j(S_j) < \max_j(S_j)$ then, there exist states u and v such that,

$S_v > \min_j(S_j)$ and $S_u < \max_j(S_j)$. Thus,

$$r = \sum_{\substack{k=1 \\ k \neq v}}^{n} u_k S_k + u_v S_v \geq (1-u_v) \min_j(S_j) + u_v S_v > \min_j(S_j).$$

Similarly, we can prove $r < \max_j(S_j)$. Q.E.D.

In the case of transition matrices this theorem produces more concrete results as stated in the following proposition.

Corollary 6.1: An irreducible transition matrix has a simple maximal eigenvalue $r = 1$.

Proof: Since all column sums of a transition matrix are equal to unity then, the maximal Frobenius eigenvalue, by the above TH.6, is $r = 1$. This eigenvalue is simple as required by the Frobenius theorem and all other eigenvalues of the transition matrix are located within the unit circle of the complex plane. Q.E.D.

The above propositions considered the eigenvalues of irreducible non-negative matrices. Since any transition matrix is not necessarily irreducible one cannot apply directly the above propositions. However, the canonical form includes irreducible matrices along its diagonal. Furthermore, the eigenvalues of the transition matrix are the eigenvalues of the irreducible diagonal blocks and this enables us to use the above propositions.

Consider a transition matrix P in its canonical form

$$P = \begin{bmatrix} P_1 & & 0 & L_{1,m+1} & \cdots & L_{1,m+k} \\ & \ddots & & & & \\ & & P_m & L_{m,m+1} & \cdots & L_{m,m+k} \\ & & & P_{m+1} & & L_{m+1,m+k} \\ & 0 & & & \ddots & \\ & & & 0 & & P_{m+k} \end{bmatrix}$$

where, the diagonal matrices are square and irreducible.

The diagonal matrices $P_1,\ldots,P_m$, corresponding to the recurrent classes, are irreducible transition matrices. Therefore, by corollary 6.1 each of them has a simple maximal eigenvalue $r = 1$.

The diagonal matrices $P_{m+1},\ldots,P_{m+k}$, corresponding to the transient classes, are irreducible and they have maximal eigenvalue $r<1$. This is true, because non of them is a transition matrix and in each of them at least one column sum is strictly less than unity. Thus, by virtue of TH.6 $r<1$.

If we consider the transient classes block

$$Q_o = \begin{bmatrix} P_{m+1} & \cdots & L_{m+1,m+k} \\ & \ddots & \\ 0 & & P_{m+k} \end{bmatrix}$$

and since $\det(\lambda I - Q_o) = \prod_{i=1}^{k} \det(\lambda I - P_{m+i})$ (see [22], page 124). the eigenvalues of Q_o are the the eigenvalues of $P_{m+1}, \ldots, P_{m+k}$. Therefore, all eigenvalues of Q_o are strictly within the unit circle of the complex plane, i.e. $|\lambda| < 1$ for all eigenvalues of Q_o.

The above result, about the eigenvalues of Q_o, will be used in the next section IV.2. Actually in this chapter we will use the following abbreviated canonical form of transition matrices

$$P = \left[\begin{array}{ccc:c} P_1 & \cdots & 0 & L_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & P_m & L_m \\ \hdashline 0 & \cdots & 0 & Q_o \end{array}\right]$$

in which all transient classes (if any) are lumped together into one group. We emphasize the above stated property: If λ is any eigenvalue of Q_o then, $|\lambda| < 1$.

IV.2. Powers of Transition Matrices and their Limits.

The first proposition to be proven considers the convergence of the powers of a non-negative matrix. The necessary and sufficient condition for these powers to converge is stated in terms of the position of its eigenvalues in the complex plane.

TH.7. For a non-negative square matrix P, its limit $\lim_{m\to\infty} P^m = 0$ if and only if all of its eigenvalues are strictly within the unit circle of the complex plane.

Proof: Consider a non-negative square matrix P whose all eigenvalues are in magnitude strictly less than unity, i.e. $|\lambda|<1$ for all eigenvalues of P. Let J be its Jordan canonical form, obtained by the similarity transformation

$$J = M P M^{-1} \quad \text{or,} \quad P = M^{-1} J M$$

where, M is the proper similarity transformation matrix and $J = \text{diag}(J_1,\ldots,J_s)$ with $J_1,\ldots,J_s$ being the appropriate Jordan blocks of orders $n_1,\ldots,n_s$, respectively.

The m-th power of P, for $m \geq 0$, becomes

$J^m = M P^m M^{-1}$ and similarly, $P^m = M^{-1} J^m M$

Therefore, $\lim_{m\to\infty} P^m$ exists and is zero if and only if $\lim_{m\to\infty} J^m$ exists and is zero. But, $J^m = \text{diag}(J_1^m,\ldots,J_s^m)$ thus, if we consider any Jordan block J_i of order $n_i \geq 1$ and let λ_i be its associated eigenvalue of

P, for which by assumption $|\lambda_i|<1$, we derive the following conclusions:

If $n_i=1$ then, $J_i=[\lambda_i]$ which yields, $J_i^m=[\lambda_i^m]$ and $\lim_{m\to\infty} J_i^m = \lim_{m\to\infty} \lambda_i^m = 0$ because, $|\lambda_i|<1$.

If $n_i>1$ then,

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots \\ 0 & \lambda_i & 1 & \cdots \\ \cdot & \cdot & \cdots & \cdots \\ 0 & 0 & & \lambda_i \end{bmatrix}$$

which yields, for any $m>n_i$

$$J_i^m = \begin{bmatrix} \binom{m}{0}\lambda_i^m & \binom{m}{1}\lambda_i^{m-1} & \binom{m}{2}\lambda_i^{m-2} & \cdots \\ 0 & \binom{m}{0}\lambda_i^m & \binom{m}{1}\lambda_i^{m-1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

That is, the entries of the matrix J_i^m are either zero or in the form

$a_m = \binom{m}{k}\lambda_i^{m-k}$ for some $k = 0,1,\ldots,n_i-1$.

If $k = 0$ then, $a_m = \lambda_i^m$ which converges to zero as $m \to \infty$.

If $k \geq 1$ then, $\lim_{m\to\infty}\left|\frac{a_{m+1}}{a_m}\right| = \lim_{m\to\infty} \frac{(m+1)}{(m+k-1)}|\lambda_i| = |\lambda_i| < 1$. Therefore, a_m converges to zero, again, as $m \to \infty$.

This concludes the sufficiency part of the theorem.

P, for which by assumption $|\lambda_i|<1$, we derive the following conclusions:

If $n_i=1$ then, $J_i=[\lambda_i]$ which yields, $J_i^m=[\lambda_i^m]$ and $\lim_{m\to\infty} J_i^m = \lim_{m\to\infty} \lambda_i^m = 0$ because, $|\lambda_i|<1$.

If $n_i>1$ then,

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots \\ 0 & \lambda_i & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \lambda_i \end{bmatrix}$$

which yields, for any $m>n_i$

$$J_i^m = \begin{bmatrix} \binom{m}{0}\lambda_i^m & \binom{m}{1}\lambda_i^{m-1} & \binom{m}{2}\lambda_i^{m-2} & \cdots \\ 0 & \binom{m}{0}\lambda_i^m & \binom{m}{1}\lambda_i^{m-1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

That is, the entries of the matrix J_i^m are either zero or in the form

$a_m = \binom{m}{k}\lambda_i^{m-k}$ for some $k = 0,1,\ldots,n_i-1$.

If $k = 0$ then, $a_m = \lambda_i^m$ which converges to zero as $m \to \infty$.

If $k \geq 1$ then, $\lim_{m\to\infty}\left|\frac{a_{m+1}}{a_m}\right| = \lim_{m\to\infty} \frac{(m+1)}{(m+k-1)}|\lambda_i| = |\lambda_i|<1$. Therefore, a_m converges to zero, again, as $m \to \infty$.

This concludes the sufficiency part of the theorem.

Assume now that $\lim_{m\to\infty} P^m = 0$. The theorem's claim is that if λ is an eigenvalue of P then $|\lambda|<1$.

Since λ is an eigenvalue of P, there exists an eigenvecter $v \neq 0$ such that, $P\,v = \lambda v$, which yields, $P^m\,v = \lambda^m\,v$.

But, $\lim_{m\to\infty}(P^m v) = (\lim_{m\to\infty} P^m)\,v = 0 \cdot v = 0$

Therefore, $\lim_{m\to\infty}(\lambda^m)\,v = 0$, which implies $\lim_{m\to\infty}(\lambda^m) = 0$, because $v \neq 0$.

Hence, $|\lambda|<1$, for otherwise, $\lim_{m\to\infty}(\lambda^m) \neq 0$. Q.E.D.

An important property of the transition matrices is the asymptotic periodicity of their powers. This periodicity in its trivial form of period unity means convergence. The next two propositions deal with this matter.

TH.8. For any transition matrix P, its powers P^m ; $m \geq 1$ can be written in the form of two matrix powers sum

$$P^m = S^m + C^m$$

where, S and C matrices are such that, the power sequence $\{C^m\}$ converges to zero as $m \to \infty$ and the power sequence $\{S^m\}$ is periodic with period some integer $d \geq 1$ (i.e. $S^m = S^{m+d}$; $m = 1,2,..$). Furthermore, the period $d = 1$ if and only if P has no eigenvalues on the unit circle, other than $\lambda = 1$.

Proof: Consider any transition matrix P in its canonical form

$$P = \left[\begin{array}{ccc|c} P_1 & & & L_1 \\ & \ddots & & \vdots \\ & & P_k & L_k \\ \hline & & & Q_o \end{array}\right] \quad \text{with powers,} \quad P^m = \left[\begin{array}{ccc|c} P_1^m & & & L_1^{(m)} \\ & \ddots & & \vdots \\ & & P_k^m & L_k^{(m)} \\ \hline & & & Q_o^m \end{array}\right]$$

The eigenvalues of P are the eigenvalues of $P_1,\ldots,P_k$ and Q_o. Each P_i $(i = 1,\ldots,k)$ is an irreducible transition matrix with $d_i \geq 1$, eigenvalues located on the unit circle and $\delta_i \geq 0$ eigenvalues located strictly within the unit circle. Thus, P has $\lambda_1,\ldots,\lambda_N$, where $N = \sum_{i=1}^{k} d_i \geq k$, eigenvalues on the unit circle. As stated before, all eigenvalues of Q_o are located strictly within the unit circle.

By Frobenius theorem any unit magnitude λ_j $(j = 1,\ldots,N)$ is cyclic with period d = least common multiple$(d_1,\ldots,d_k) \geq 1$, i.e. $\lambda_j^{m+d} = \lambda_j^m$ for any $m=1,2,\ldots$

Let J be the Jordan canonical form of P, i.e. $P = M\,J\,M^{-1}$ and $J = M^{-1}P\,M$. The powers of P and J take the form $P^m = M\,J^m\,M^{-1}$ and $J^m = M^{-1}\,P^m\,M$; $m = 1,2,\ldots$

Since P^m is a transition matrix for all $m = 1,2,\ldots$ then P^m is bounded (elementwise). Therefore, J^m must be bounded (elementwise). But this implies that any Jordan block associated with some eigenvalue of magnitude unity must be of order one. For otherwise, any such Jordan block, of order 2 or more, would have terms, in its superdiagonal of its matrix powers, in the form

$$a_m = \binom{m}{1} \lambda^{m-1} = m\,\lambda^{m-1} \qquad \text{with } |\lambda| = 1$$

which becomes unbounded as m tends to infinity.

Thus, J can be put in the form

$J = \text{diag}([\lambda_1],\ldots,[\lambda_N],J_{N+1},\ldots)$ while, its powers become

$J^m = \text{diag}([\lambda_1^m],\ldots,[\lambda_N^m],J_{N+1}^m,\ldots) = U^m + V^m \quad ;m = 1,2,3,\ldots$

where, $U = \text{diag}([\lambda_1],\ldots,[\lambda_N],0,\ldots)$ is, obviously, periodic with period d, i.e. $U^{m+d} = U^m$ for all m=1,2,..., and $V = \text{diag}(0,\ldots,0,J_{N+1},..)$ for which $\lim_{m\to\infty} V^m = 0$ as is shown in the proof of TH.7.

Finally,

$$P^m = M\,J^m\,M^{-1} = M\,(\,U^m + V^m\,)\,M^{-1} = M\,U^m\,M^{-1} + M\,V^m\,M^{-1} =$$

$$= S^m + C^m \quad ;m = 1,2,\ldots.$$

where, $S = M\,U\,M^{-1}$ periodic matrix with period $d \geq 1$, i.e.

$$S^{m+d} = M\,U^d\,M^{-1} = S^m \quad ;m = 1,2,3,\ldots.$$

and $C = M\,V\,M^{-1}$ such that, $\lim_{m\to\infty} C^m = \lim_{m\to\infty}(M\,V^m\,M^{-1}) = M(\lim_{m\to\infty} V^m)\,M^{-1} = 0.$

Moreover, d = 1 if and only if $d_1 = \ldots = d_k = 1$ which means that all eigenvalues of P on the unit circle are equal to the positive unity. Q.E.D.

Corollary 8.1. For a transition matrix P, the limit $\lim(P^m)$ exists if and only if there is no eigenvalue on the unit circle other than $\lambda = 1$.

Proof: This proposition is a direct result of TH.8 since $P^m = S^m + C^m$;m=1,2,.. converge if and only if the period of S is $d = 1$. In this case, $S^m = S$

for all $m = 1,2,\ldots$ and $\lim_{m\to\infty} P^m = S + \lim_{m\to\infty} C^m = S$. Q.E.D.

The Corollary 8.1 was our target. It establishes the sufficient and necessary condition for powers of a transition matrix to converge. There are, indeed, transition matrices whose powers do not converge. For example, the matrix powers of the following two matrices

$$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix} \quad \text{with } \lambda_1 = 1 \text{ and } \lambda_2 = -1$$

$$\begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix} \quad \text{with } \lambda_1 = 1 \text{ and } \lambda_2, \lambda_3 = -\tfrac{1}{2} - j\tfrac{\sqrt{3}}{2}$$

do not converge. While, the following matrix

$$\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{with } \lambda_1 = 1,\ \lambda_2, \lambda_3 = -\tfrac{1}{4} - j\tfrac{\sqrt{7}}{4},\ \lambda_4 = 0.$$

has powers that converge, despite its pseudo-cyclic appearance.

A useful matrix sequence, associated closely to the matrix powers of a transition matrix, is the well-known "average transition matrix"

$$A_m = \frac{P + \ldots + P^m}{m} \quad ;m = 1,2,\ldots.$$

where, P is a given transition matrix.

For a fixed $m \geq 1$ the matrix A_m is the arithmetic average of all transition matrices $P,\ldots,P^m$. For large enough m the matrix A_m has an obvious interpretation. For example, suppose that $(A_m)_{ij} \approx 1/5$, for

some entry (i,j) of the matrix A_m. This means that a Markov chain, with transition matrix P and starting from state j, visits state i on the average once every five time-steps. Intuitively speaking, we should expect that A_m always converges as $m \to \infty$ and; indeed, it does.

As we will prove in the following TH.9 if $\{P^m\}$ converges then, it converges to the same limit as $\{A_m\}$. For this reason and also for the interpretation we can attach to A_m, for large enough m, we argue that the limit of A_m, as $m \to \infty$, can serve as a substitute of the; occasionally, non-existing limit of the matrix sequence $\{P^m\}$. Of course, by the term substitute we are referring to the entity $\lim_{m\to\infty} P^m$ as it is intuitively interpreted, rather than its precise mathematical meaning.

TH.9. For any transition matrix P and any integer $m \geq 1$ the "average transition matrix" $A_m = (P + \ldots + P^m)/m$ has the following properties:

(a) It is a transition matrix.

(b) It has always a limit as $m \to \infty$.

(c) If $A = \lim_{m\to\infty} A_m$ then, $P^k A = A P^k = A$ for any finite $k=1,2,\ldots$

(d) If $\lim_{m\to\infty} P^m$ exists then $\lim_{m\to\infty} P^m = \lim_{m\to\infty} A_m = A$.

Proof: (a) Since $A_m = \sum_{i=1}^{m} u_i P^i$ with $u_i = 1/m > 0$ and $\sum_{i=1}^{m} u_i = 1$ then, A_m is a convex combination of the transition matrices $P,\ldots,P^m$ and by TH.2 is a transition matrix.

(b) Kemeny [12] proves the existence of this limit using the existence of converging subsequences in a compact space. Gantmacher [10] (pp.88-90) proves it using analytical properties of the characteristic polynomial of P.

(c) Let $A_m = \frac{P + \ldots + P^m}{m}$ for some integer $m \geq 1$. For any integer $k \geq 1$

$$P^k A_m = \frac{P^{k+1} + \ldots + P^{k+m}}{m} = \frac{(P + \ldots + P^m)}{m} P^k = A_m P^k$$

and in the limit as $m \to \infty$ $\quad P^k A = A P^k$ for any $k = 1,2,\ldots$

Also $P^k A_m = \frac{P^{k+1} + \ldots + P^{k+m}}{m} = \frac{P+\ldots+P^k+\ldots+P^{k+m}}{k+m} \cdot \frac{m+k}{m} - \frac{k}{m} A_k =$

$$= A_{k+m}\left(\frac{m+k}{m}\right) - \frac{k}{m} A_k$$

which as $m \to \infty$ and for fixed $k \geq 1$ yields, $P^k A = A$.

(d) Let $Q = \lim P^m$. By definition of this limit, given any $\epsilon > 0$ there exists positive integer M such that,

$\| P^m - Q \| < \frac{1}{2}\epsilon$ for all $m > M$.

If we define $M' = 1 + \max\left(\frac{2}{\epsilon}\sum_{k=1}^{M}\|P^k - Q\| , M\right)$ then, for any $m > M'$

$$\| A_m - Q \| = \left\| \frac{1}{m}\sum_{k=1}^{m}(P^k - Q)\right\| \leq \frac{1}{m}\sum_{k=1}^{m}\|P^k - Q\| =$$

$$= \frac{1}{m}\sum_{k=1}^{M}\|P^k - Q\| - \frac{1}{m}\sum_{k=M+1}^{m}\|P^k - Q\|$$

$$< \epsilon/2 + \frac{1}{m}(m-M)\frac{\epsilon}{2} < \epsilon$$

therefore, $\lim_{m\to\infty} A_m = Q = \lim_{m\to\infty} P^m$. Q.E.D.

Corollary 9.1. If A is the limit of the average transition matrix A_m; m=1,2,.. of a transition matrix P then, each column of A is an eigenvector of P corresponding to a $\lambda = 1$ eigenvalue of P.

Moreover, if P is irreducible then, the columns of A are identical.

Proof: By TH.9(c) we have $P A = A$ which implies that the columns of A are eigenvectors of P for the eigenvalue $\lambda = 1$. If, however, P is irreducible such eigenvector is unique, since $\lambda = 1$ is a simple eigenvalue. Therefore, in this case, the columns of A are identical to this eigenvector. Q.E.D.

IV.3. Analytical Evaluation of the Limits.

Consider a transition matrix in its canonical form

$$P = \begin{bmatrix} P_1 & & & L_1 \\ & \ddots & & \vdots \\ & & P_k & L_k \\ & & & Q_o \end{bmatrix} \qquad \text{(IV.3)}$$

where, $P_1,\ldots,P_k$ are transition matrices corresponding to the recurrent classes and Q_o is a square non-negative matrix block corresponding to all transient classes considered as one group.

The m-th power of P, for any integer $m \geq 1$, is in the form

$$P^m = \begin{bmatrix} P_1^m & & & L_1^{(m)} \\ & \ddots & & \vdots \\ & & P_k^m & L_k^{(m)} \\ & & & Q_o^{(m)} \end{bmatrix} \qquad ;m = 1,2,\ldots \qquad \text{(IV.4)}$$

where, the diagonal matrices are the m-th powers of the matrices $P_1,\ldots,P_k,Q_o$ and $L_1^{(m)},\ldots,L_k^{(m)}$, with superscript m, denote the corresponding blocks in the last column.

If we form the average transition matrix of P for any integer $m = 1,2,\ldots$ we derive

$$A_m = \begin{bmatrix} A_{1,m} & & & B_{1,m} \\ & \ddots & & \vdots \\ & & A_{k,m} & B_{k,m} \\ & & & B_{o,m} \end{bmatrix} \qquad \text{(IV.5)}$$

where,
$$A_m = \frac{1}{m}\sum_{i=1}^{m} P^i \qquad (IV.6)$$

$$A_{u,m} = \frac{1}{m}\sum_{i=1}^{m} P_u^i \qquad ;u = 1,\ldots,k \qquad (IV.7)$$

$$B_{o,m} = \frac{1}{m}\sum_{i=1}^{m} Q_o^i \qquad (IV.8)$$

$$B_{u,m} = \frac{1}{m}\sum_{i=1}^{m_1} L_u^{(i)} \qquad ;u = 1,\ldots,k \qquad (IV.9)$$

By TH.9(b) the limit of A_m exists as $m \to \infty$. Let us denote this limit by

$$A = \left[\begin{array}{ccc|c} A_1 & & & B_1 \\ & \ddots & & \vdots \\ & & A_k & B_k \\ \hline & & & B_o \end{array}\right] \qquad (IV.10)$$

First, we claim that $\boxed{B_o = 0.}$ (IV.11)

To prove (IV.11) we must take into account that Q_o has eigenvalues strictly within the unit circle of the complex plane. This was explained at the end of the section IV.1. Therefore, by TH.7 $Q = \lim_{m\to\infty} Q_o^m = 0$. Thus, $B_o = \lim_{m\to\infty}(\frac{1}{m}\sum_{i=1}^{m} Q_o^i)$ can be proven to be $Q = 0$ by arguments identical to the ones used in the proof of TH.9(d).

Second, since $P_1,\ldots,P_k$ are irreducible transition matrices, by virtue of Corollary 9.1 the matrices $A_1,\ldots,A_k$ are in the form

$$\boxed{A_i = [v_i, v_i, \ldots, v_i]} \qquad ;i = 1,\ldots,k \qquad (IV.12)$$

That is, all columns of the square matrices A_i ;i=1,...,k are identical to the unique eigenvector v_i of the corresponding matrix P_i. We should note that the eigenvector v_i is unique in the sense that, it is an eigenvector of the matrix P_i, for $\lambda = 1$, with sum of its coordinates unity.

Third and final, from the relation $A = AP$ we derive,

$B_i = A_i L_i + B_i Q_o$;i = 1,...,k, which yields

$$\boxed{B_i = A_i L_i (I-Q_o)^{-1} \quad ;i=1,\ldots,k} \qquad \text{(IV.1}$$

One should notice that $I-Q_o$ is not singular matrix because $\lambda = 1$ is not an eigenvalue of Q_o. Furthermore, $(I-Q_o)^{-1} \geq 0$ since Q_o has eigenvalues strictly less than unity (see [2], pp.438-439).

The matrices B_i;i=1,...,k possess an interesting property. For any i=1,...,k let us define $Z = L_i(I-Q_o)^{-1}$ so that, (IV.13) becomes $B_i = A_i Z$. On the other hand, by (IV.12) we can write

$$A_i = [v_i,\ldots,v_i] = \begin{bmatrix} v_{1i}(1,1,\ldots,1) \\ v_{2i}(1,1,\ldots,1) \\ \vdots \end{bmatrix} = \begin{bmatrix} v_{1i}\bar{1}' \\ v_{2i}\bar{1}' \\ \vdots \end{bmatrix}$$

Thus, the u-th row, j-th column entry of B_i is

$$(B_i)_{uj} = v_{ui}\left(\sum_m z_{mj}\right) = v_{ui}c_j$$

where, the factor c_j does not depend on the row index u.

Therefore, the matrix B_i can be written in the form

$$B_i = \left[c_{1i}v_i,\ c_{2i}v_i, \dots \right] \qquad ;i=1,\dots,k \tag{IV.14}$$

i.e. the columns of B_i are all multiples of the $\lambda = 1$ eigenvector of P_i.

Equation (IV.14) can be used directly, in some cases, for the computation of B_i. For example, if there is only one recurrent class then the coefficients $c_{11}=c_{21}=\dots=1$ and the matrix B_1 takes the form

$$B_1 = A_1 L_1 (I-Q_o)^{-1} = \left[v_1, v_1, \dots \right] \tag{IV.15}$$

The above results have some peculiar implications. For example, consider the transition matrix

$$P = \begin{bmatrix} P_1 & L_1 \\ 0 & Q_o \end{bmatrix}$$

where, P_1 is an irreducible transition matrix.
Also consider the limit of the average transition matrix

$$A = \begin{bmatrix} A_1 & B_1 \\ 0 & 0 \end{bmatrix}$$

The original choice of P consists of the choices for P_1, L_1, and Q_o. The average transition matrix should be expected to depend on at least two of these matrices P_1, L_1 and Q_o. However, the truth of the matter is that it depends only on P_1 and is independent of the choice of the matrices L_1 and Q_o. Specifically,

$$A = \begin{bmatrix} v_1 & v_1 & \dots \\ 0 & 0 & \dots \end{bmatrix} \quad \text{where, } v_1 \text{ is the } \lambda = 1 \text{ eigenvector of } P_1.$$

IV.4. A Numerical Example.

Consider a 15-state Markov Chain with transition matrix

$$P = \begin{array}{c} \begin{matrix} (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) & (11) & (12) & (13) & (14) & (15) \end{matrix} \\ \left[\begin{array}{ccc|ccc|cccccc|ccc} .4 & .0 & .5 & & & & & & & & & & .0 & .0 & .0 \\ .3 & .6 & .5 & & O & & & & O & & & & .0 & .6 & .0 \\ .3 & .4 & .0 & & & & & & & & & & .0 & .0 & .0 \\ \hline & & & .0 & 1. & 1. & & & & & & & & & \\ & O & & .5 & .0 & .0 & & & O & & & & & O & \\ & & & .5 & .0 & .0 & & & & & & & & & \\ \hline & & & & & & .0 & .0 & .0 & .0 & .0 & .6 & .0 & .1 & .0 \\ & & & & & & .0 & .0 & .0 & .0 & .4 & .0 & .0 & .0 & .0 \\ & O & & & O & & .3 & .5 & .0 & .0 & .0 & .0 & .0 & .0 & .2 \\ & & & & & & .2 & .5 & .0 & .0 & .0 & .0 & .0 & .1 & .0 \\ & & & & & & .5 & .0 & .0 & .0 & .0 & .0 & .0 & .0 & .1 \\ & & & & & & .0 & .0 & 1. & 1. & 1. & .0 & .0 & .0 & .1 \\ \hline & & & & & & & & & & & & .4 & .0 & .6 \\ & O & & & O & & & & O & & & & .6 & .0 & .0 \\ & & & & & & & & & & & & .0 & .2 & .0 \end{array}\right] \end{array} \begin{matrix} (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \\ (7) \\ (8) \\ (9) \\ (10) \\ (11) \\ (12) \\ (13) \\ (14) \\ (15) \end{matrix}$$

recurrent states — transient states

or in its abbreviated partitioned form,

$$\left[\begin{array}{ccc|c} P_1 & 0 & 0 & L_1 \\ 0 & P_2 & 0 & L_2 \\ 0 & 0 & P_3 & L_3 \\ \hline 0 & 0 & 0 & Q_o \end{array}\right]$$

The limit $A = \lim A_m = \lim(\frac{1}{m}\sum_{i=1}^{m} P^i)$ will be in the form

$$A = \left[\begin{array}{ccc:c} A_1 & & & B_1 \\ & A_2 & & B_2 \\ & & A_3 & B_3 \\ \hdashline & & & B_o \end{array}\right]$$

Matrix B_o: By equation (IV.11) $B_o = 0$.

Matrix A_1: The eigenvector of P_1, for $\lambda = 1$, is $v_1 = (20,45,24)'$. Normalizing it so that its entries sum up to unity yields,

$v_1 = (20/89 \quad 45/89 \quad 24/89)'$

Therefore, by equation (IV.12)

$$A_1 = [v_1 \quad v_1 \quad v_1] = \begin{bmatrix} 20/89 & 20/89 & 20/89 \\ 45/89 & 45/89 & 45/89 \\ 24/89 & 24/89 & 24/89 \end{bmatrix}$$

Matrix A_2: The matrix P_2 is cyclic with period $d=2$, because it is in the form $P_2 = \begin{vmatrix} 0 & C_1 \\ C_2 & 0 \end{vmatrix}$ where, the diagonal blocks are square matrices. By Frobenius theorem there are two eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = -1$, on the unit circle. The other eigenvalue must, by Frobenius theorem again, be equal to zero. For if $\lambda_3 \neq 0$ is eigenvalue of P_2 then, $-\lambda_3$ must as well be an eigenvalue of P_2, which is not possible.

The eigenvector for $\lambda = 1$ is found to be $v_2 = (\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{4})'$.

Therefore,

$$A_2 = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{vmatrix}$$

Matrix A_3: The matrix $P_3 = \begin{vmatrix} 0 & 0 & C_1 \\ C_2 & 0 & 0 \\ 0 & C_3 & 0 \end{vmatrix}$ has symmetric eigenvalues by means of rotation by $2\pi/3$. Therefore, it has three simple eigenvalues on the unit circle and the remaining three are zero because P_3 is a singular matrix (it has three identical columns).

The normalized eigenvector of P_3 for $\lambda = 1$ is found to be

$v_3 = (1/5 \quad 2/15 \quad 19/150 \quad 16/150 \quad 1/10 \quad 1/3)'$

and $A_3 = [v_3, v_3, v_3, v_3, v_3, v_3]$ (6×6 matrix)

Matrix $(I-Q_o)^{-1}$: It is found to be $\frac{5}{66}\begin{vmatrix} 25 & 3 & 15 \\ 15 & 15 & 9 \\ 3 & 3 & 15 \end{vmatrix}$ and by (IV.13) we derive

Matrix $B_1 = A_1 L_1 (I-Q_o)^{-1} = \frac{3}{22}[5v_1, 5v_1, 3v_1]$

Matrix $B_2 = A_2 L_2 (I-Q_o)^{-1} = 0$ since $L_2 = 0$

Matrix $B_3 = A_3 L_3 (I-Q_o)^{-1} = \frac{1}{22}[7v_3, 7v_3, 13v_3]$.

Since A is a transition matrix, the columns of B_1, B_2, B_3 must sum to unity and, indeed, they do so.

In the above numerical example, the recurrent class blocks P_1, P_2 and P_3 have periods 1,2, and 3, respectively. We recall that the period of an irreducible transition matrix coincides with the number of its unit magnitude eigenvalues. Thus, the power sequence $\{P^m\}$ is asymptotically periodic with period $d=6$ and, of course, does not converge as $m \to \infty$.

IV.5. Concluding Remarks.

We have identified the study of transition matrix powers with the study of transition probabilities between any two states and in any finite number of time-steps. We have proven that the powers $P^m; m=1,2,..$ of a transition matrix P are asymptotically periodic, i.e. they consist of a periodic component and another one which converges to zero as $m \to \infty$. In the case where the period is unity then, the powers converge.

It has been also shown that the average transition matrix $A_m = \frac{1}{m} \sum_{k=1}^{m} P^k \quad ; m=1,2,\ldots$ always converges as $m \to \infty$ and we derived analytical formulas for the computation of this limit. If the original transition matrix P has converging power sequence then, the two limits coincide. On the other hand, if the original matrix has not converging power series then, in the long run, i.e. for number of time-steps large enough, the average transition matrices can serve as a substitute of the oscillating powers of the original matrix.

By the numerical example of the section IV.4, it has been illustrated that the analytical formulas, for the evaluation of $A = \lim \frac{1}{m} \sum_{i=1}^{m} P^i$, are computationally simple. Furthermore, by computing A one by-passes the problem of the existence of $\lim P^m$ which, if exists, is A.

The dominant mathematical concept which is used in this chapter is the spectrum of a transition matrix. The position of the eigenvalues of a transition matrix in the complex plane determines completely its asymptotic behavior.

CHAPTER V

MARKOV CHAIN AGGREGATES

CHAPTER V. MARKOV CHAIN AGGREGATES

V.1. Introduction.

In the previous chapters III and IV some probabilistic properties of the stationary discrete-time finite-state Markov chains were presented. The study focused on the investigation of the transition matrices which completely describe the evolution of such systems in time. Here we will consider a finite Markov Chain Aggregate $\Delta=\{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}$ which, as defined in section II.3, is a finite set of finite-state Markov chains having common transition matrix for all its members. An MCA is completely described by a non-negative integer vector, called state-vector of the MCA, whose coordinates denote the number of the MCA members in the corresponding state.

The MCA, as a mathematical model, can describe numerous real-life processes in the fields of engineering, economics, biology, genetics etc. The main analytical problem one has to study is that, given the state-vector $x_o = x(t_o)$ at some initial time t_o, the value of the state-vector $x(t)$ should be predicted at any point in time $t \geq t_o$. Of course, since an MCA is a stochastic system the prediction of its state-vector will be obtained by means of its expectancy and its dispersion around this expectancy, measured by its variance-covariance matrix. These two statistical parameters are shown in section V.2 to be dependent on the initial state-vector and the transition matrix from the initial time t_o to the final time $t \geq t_o$ in question.

In particular, when an MCA is stationary then there exists some fixed transition matrix P such that, $P(t,t_o) = P^{(t-t_o)}$ for any $t \geq t_o$. In such cases, there are some interesting, from the practical point of view, limit properties of the MCA statistical parameters, which are presented in the section V.3.

V.2. Expectancy and Variance-Covariance Matrix of MCA.

Consider a finite Markov Chain Aggregate of either discrete or continuous-time and with finite state-space $S = \{1,\ldots,n\}$. Let $N \geq 1$ be the MCA's number of members. Furthermore, let

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t))' \quad ; t \geq t_o$$

denote its state-vector at time $t \geq t_o$, where, t_o is some given initial point in time.

The following proposition refers to the evaluation of the statistical parameters of the state-vector $x(t)$ in terms of the initial state-vector and the transition matrix $P(t,t_o)$.

TH.10. Let $\Delta = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_N\}$ be an MCA with given initial state-vector $x_o = x(t_o)$ and transition matrix $P(t,t_o)$; $t \geq t_o$.
The expected value and variance-covariance matrix of the state-vector $x(t)$, at any point in time $t \geq t_o$, are given by

$$E[x(t) \mid x_o] = P(t,t_o)\, x_o \tag{V.1}$$

$$Var[x(t) \mid x_o] = diag[P(t,t_o)\, x_o] - P(t,t_o)\, diag[x_o]\, P(t,t_o)' \tag{V.2}$$

Note: Given a vector $v = (v_1,\ldots,v_n)'$ then, by the notation $diag[v]$ we mean the diagonal matrix with the corresponding coordinates of v located along its main diagonal.

Proof: Before we proceede we will introduce some notations for reasons of technical convenience.

Let us denote the given initial state-vector by

$x_o = (a_1,\ldots,a_j,\ldots,a_n)'$

the final state-vector at $t \geq t_o$ by

$x = (x_1,\ldots,x_i,\ldots,x_n)'$

and finally, the transition matrix by $P(t,t_o) = P = (p_{ij})$

If we pick a particular state $j \in S$ then, the a_j members of the MCA being in state j at time t_o will be distributed among the states $i=1,\ldots,n$ according to a multinomial distribution with probabilities $p_{1j},\ldots,p_{ij},\ldots,p_{nj}$, i.e. the entries of the j-th column of the transition matrix $P = (p_{ij})$. Let the non-negative integer vector

$$y_j = (y_{1j},\ldots,y_{ij},\ldots,y_{nj})' \quad ;j \in S \tag{V.3}$$

denote this distribution, i.e. y_{ij} denotes the number of the MCA members which are at state i at time $t \geq t_o$ and whose initial state was j.

Obviously, $$\sum_{i=1}^{n} y_{ij} = a_j \quad ;j = 1,\ldots,n \tag{V.4}$$

By its definition, the state-vector $x = x(t)$ is given by the vector sum

$$x = x(t) = \sum_{j=1}^{n} y_j \tag{V.5}$$

First, we notice that the vectors y_j ;j=1,...,n are pairwise stochastically indepedent. Also, as stated above, these vectors are multinomially distributed with parameters a_j and $p_{1j},\ldots,p_{nj}$. Their expected values and variance-covariance matrices are given by Sverdrup 21 ,vol. 2,pp. 41-42, to be for any j=1,...,n

$$E[y_j] = a_j(p_{1j},\ldots,p_{nj})' = a_j p_j \tag{V.6}$$

$$Var[y_j] = E[(y_j - a_j p_j)(y_j - a_j p_j)'] = a_j \operatorname{diag}[p_j] - a_j p_j p_j' \tag{V.7}$$

where, a_j (scalar) is the j-th coordinate of the initial state-vector and $p_j=(p_{1j},\ldots,p_{nj})'$, considered as a column vector, is the j-th column of the transition matrix $P = P(t,t_o)$.

From identity (V.5) and equation (V.6) we derive

$$E[x] = \sum_{j=1}^{n} E[y_j] = \sum_{j=1}^{n} a_j p_j$$

or, in matrix form

$$E[x] = P\,(a_1,\ldots,a_n)' = P\,x_o$$

Thus, the equation (V.1) is proven to be valid. This result could be obtained by simpler means and can be considered as well known. However, the variance-covariance matrix equation, as stated by (V.2), can be considered as a contribution of the present investigation.

Using identity (V.5) and equation (V.1) we derive

$$\mathrm{Var}[x] = E[(x - E[x])(x - E[x])'] =$$

$$= E[(\sum_{j=1}^{n} y_j - \sum_{j=1}^{n} a_j p_j)(\sum_{j=1}^{n} y_j - \sum_{j=1}^{n} a_j p_j)'] =$$

$$= E[\sum_{j=1}^{n}\sum_{k=1}^{n} (y_j - a_j p_j)(y_k - a_k p_k)'] =$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} E[(y_j - a_j p_j)(y_k - a_k p_k)']$$

and by stochastic indepedence of the vectors $y_j; j=1,\ldots,n$

$\mathrm{Var}[x]$ becomes

$$\mathrm{Var}[x] = \sum_{j=1}^{n} E[(y_j - a_j p_j)(y_j - a_j p_j)'] =$$

$$= \sum_{j=1}^{n} \mathrm{Var}[y_j] =$$

$$= \sum_{j=1}^{n} a_j \, \mathrm{diag}[p_j] - \sum_{j=1}^{n} a_j \, p_j p_j'$$

the first term in the above expression can take the form

$$\sum_{j=1}^{n} a_j \, \mathrm{diag}[p_j] = \sum_{j=1}^{n} \mathrm{diag}[a_j p_j] = \mathrm{diag}[\sum_{j=1}^{n} a_j p_j] = \mathrm{diag}[P\, x_o]$$

Also we note that the j-th column of P can be given by the matrix expression $p_j = P\, e_j$ where, e_j is the unit vector along its j-th coordinate, i.e. $e_j = (0,..,1,..,0)'$.

Hence, $\sum_{j=1}^{n} a_j \, p_j p_j' = \sum_{j=1}^{n} a_j \, P e_j e_j' P' = P \, (\sum_{j=1}^{n} a_j e_j e_j') \, P'$

but, $a_j e_j e_j'$ is an $n \times n$ matrix being everywhere zero except at its diagonal entry (j,j) where it assumes value a_j. Therefore,

$$\sum_{j=1}^{n} a_j e_j e_j' = \mathrm{diag}[x_o]$$

thus, $\mathrm{Var}[x] = \mathrm{diag}[P\, x_o] - P \, \mathrm{diag}[x_o] \, P'$ Q.E.D.

V.3. The Stationary Discrete-Time MCA.

In the special case where, the members of the MCA are stationary discrete-time Markov chains then, the transition matrix $P(t,t_o)$ depends only on the difference $t-t_o \geq 0$ where, the integers t_o and t are the initial and final time, respectively. For reasons of convenience, therefore, we can set $t_o=0$ while, the final time $t=0,1,2,\ldots$ Furthermore, the transition matrix $P(t,0)=P^t$ for any $t=0,1,\ldots$ where, P is the single time-step transition matrix (see section II.2). Thus, equations (V.1) and (V.2) become

$$\overline{x(t)} = E\left[x(t) \mid x(t_o)=x_o\right] = P^t x_o \qquad ;t = 0,1,2,\ldots \tag{V.10}$$

$$Var\left[x(t) \mid x(t_o)=x_o\right] = diag\left[P^t x_o\right] - P^t\, diag\left[x_o\right] P'^t \qquad ;t=0,1,\ldots \tag{V.11}$$

By theorem TH.8 (chapter IV) the power sequence $\{P^t\}$ is asymptotically periodic with period $d \geq 1$ determined by the structure of P and its unity magnitude eigenvalues. That is, the matrix P^t, for any integer $t \geq 1$, can be written in the form

$$P^t = C^t + S^t \qquad ;t = 1,2,\ldots \tag{V.12}$$

where; $C^{t+d} = C^t$ for some integer $d \geq 1$ and all $t=1,2,\ldots$

while; $S^t \longrightarrow [0]$ as $t \to \infty$.

By substitution of (V.12) into the equations (V.10) and (V.11), we derive that expected value and variance-covariance matrix of the MCA state-vector $x(t)$ are asymptotically periodic, like the $\{P^t\}$ sequence, with the same period $d \geq 1$. Of course, period $d=1$ means convergence.

For the further investigation of the asymptotic behavior of the MCA state-vector, we first introduce some notations

Let P be in its canonical form

$$P = \begin{bmatrix} P_1 & & & L_1 \\ & \ddots & & \vdots \\ & & P_k & L_k \\ & & & P_o \end{bmatrix} \qquad (V.13)$$

where, $P_1,\ldots,P_k$ are irreducible transition matrices of orders $n_1,\ldots,n_k$, respectively. and correspond to the recurrent classes with $d_1,\ldots,d_k \geq 1$ number of eigenvalues of unit magnitude, respectively. We recall that $P_1,\ldots,P_k$ are asymptotically periodic with periods $d_1,\ldots,d_k$, respectively. Also, the matrix P_o of order $n_o \geq 0$ corresponds to the set of the transient states.

The powers of P, for any integer $t \geq 1$, are in the form

$$P^t = \begin{bmatrix} P_1^t & & & L_1^{(t)} \\ & \ddots & & \vdots \\ & & P_k^t & L_k^{(t)} \\ & & & P_o^t \end{bmatrix} \qquad (V.14)$$

We partition the state-vector $x(t)$ in the following way

$$x(t)=\begin{bmatrix} x^1(t) \\ \vdots \\ x^k(t) \\ x^o(t) \end{bmatrix} \qquad ;t = 0,1,2,... \tag{V.15}$$

where, $x^1(t),...,x^k(t),x^o(t)$ are column vectors, of dimensions $n_1,...,n_k,n_o$, respectively, and each of them contains the coordinates of the states of $x(t)$ which belong to the communication class $1,...,k,0$, respectively.

In particular, we denote this partition for the initial state-vector by

$$x(0) = x_o = \begin{bmatrix} y_1 \\ \vdots \\ y_k \\ y_o \end{bmatrix} \tag{V.16}$$

Finally, we introduce the following notation for the diagonal matrix formed by a given vector $v' = (v_1,...,v_n)$

$$\hat{v} = \text{diag}(v) = \begin{bmatrix} v_1 & & 0 \\ & v_2 & \\ & & \ddots \\ 0 & & v_n \end{bmatrix}$$

By substitution of (V.14), (V.15) and (V.16) into the equations (V.10) and (V.11) we derive

$$\overline{x(t)} = E\left[x(t)\,\middle|\,x(t_o)=x_o\right] = \begin{bmatrix} P_1^t y_1 + L_1^{(t)} y_o \\ \vdots \\ P_k^t y_k + L_k^{(t)} y_o \\ P_o^t y_o \end{bmatrix} \tag{V.17}$$

$$\mathrm{Var}\left[x(t)\,\middle|\,x(t_o)=x_o\right] = \widehat{\overline{x(t)}} - \begin{bmatrix} P_1^t \hat{y}_1 P_1'^t & & \\ & \ddots & \\ & & P_k^t \hat{y}_k P_k'^t \\ & & & P_o^t \hat{y}_o P_o'^t \end{bmatrix}$$

$$- \begin{bmatrix} L_1^{(t)} \hat{y}_o L_1'^{(t)} & \cdots & L_1^{(t)} \hat{y}_o L_k'^{(t)} & L_1^{(t)} \hat{y}_o P_o'^t \\ \vdots & & \vdots & \vdots \\ L_k^{(t)} \hat{y}_o L_1'^{(t)} & \cdots & L_k^{(t)} \hat{y}_o L_k'^{(t)} & L_k^{(t)} \hat{y}_o P_o'^t \\ P_o^t \hat{y}_o L_1'^{(t)} & \cdots & P_o^t \hat{y}_o L_k'^{(t)} & 0 \end{bmatrix} \tag{V.18}$$

The limits, as $t \to \infty$, of the above involved matrix blocks do not always exist. Therefore, we cannot consider, in general, equations (V.17) and (V.18) in their limit form. However, since $P_o^t \to 0$ as $t \to \infty$, for large enough time t the above equations take the simpler forms

$$\overline{x(t)} = E\left[x(t)\middle|x(t_o)=x_o\right] = \begin{bmatrix} P_1^t y_1 + L_1^{(t)} y_o \\ \vdots \\ P_k^t y_k + L_k^{(t)} y_o \\ 0 \end{bmatrix} \quad \text{(for large } t\text{)} \qquad \text{(V.19)}$$

$$\mathrm{Var}\left[x(t)\middle|x(t_o)=x_o\right] = \hat{x}(t) - \begin{bmatrix} P_1^t y_1 P_1'^t & & & 0 \\ & \ddots & & \vdots \\ & & P_k^t y_k P_k'^t & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$- \begin{bmatrix} L_1^{(t)} y_o L_1^{(t)}, & \cdots & L_1^{(t)} y_o L_k^{(t)}, & 0 \\ \vdots & & \vdots & \vdots \\ L_k^{(t)} y_o L_1^{(t)}, & \cdots & L_k^{(t)} y_o L_k^{(t)}, & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \quad \text{(for large } t\text{)} \qquad \text{(V.20)}$$

From the above equations we can derive a few useful conclusions. If we assume that $y_o=0$, i.e. no member of the MCA was initially on any transient state, then two states m_1 and m_2 belonging to different communication classes are stochastically independent, i.e.

$$E\left[\left(x_{m_1}(t) - E\left[x_{m_1}(t)\right]\right)\left(x_{m_2}(t) - E\left[x_{m_2}(t)\right]\right)\right] = 0 \quad \text{for all } t = 1,2,\ldots$$

This result can be derived from (V.18) if we set $y_o = 0$. Therefore, the correlation between any two states from different communication classes is induced by the existence of some members of the MCA in some transient

states. Otherwise, such correlation is zero.

Consider now the case where the limit $\lim_{t\to\infty} P^t$ exists. Let us denote it by

$$\lim_{t\to\infty} P^t = \begin{bmatrix} A_1 & & & B_1 \\ & \ddots & & \vdots \\ & & A_k & B_k \\ & & 0 & 0 \end{bmatrix}$$

We recall that $A_i = [v_i,\ldots,v_i]$;i = 1,...,k, where, v_i is the unique eigenvector of P_i;i = 1,...,k corresponding to its eigenvalue $\lambda=1$ (equation (IV.12)).

Define the scalars $N_i = \bar{1}'y_i$;i=1,...,k, which are the total number of the MCA members in the recurrent classes i=1,...,k, respectively. It can be easily verified that

$$A_i y_i = N_i v_i \quad ;i = 1,\ldots,k$$

and $$A_i \hat{y}_i A_i' = N_i v_i v_i' \quad ;i = 1,\ldots,k$$

Therefore, the limits as t tends to infinity take the form

$$\overline{x(\infty)} = \lim_{t\to\infty} \overline{x(t)} = \begin{bmatrix} N_1 v_1 \\ \vdots \\ N_k v_k \\ 0 \end{bmatrix} + \begin{bmatrix} B_1 \\ \\ B_k \\ 0 \end{bmatrix} y_o \qquad \text{(V.22}$$

and

$$\mathrm{Var}\left[x(\infty)\,\middle|\,x_o\right] = \begin{bmatrix} N_1(\widehat{v}_1 - v_1 v_1') & & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots\cdots & N_k(\widehat{v}_k - v_k v_k') & 0 \\ 0 & \cdots\cdots & 0 & 0 \end{bmatrix}$$

$$+\begin{vmatrix} \widehat{(B_1 y_o)} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \widehat{(B_k y_o)} & 0 \\ 0 & & 0 & 0 \end{vmatrix} - \begin{vmatrix} B_1 \widehat{y}_o B_1' & \cdots & B_1 \widehat{y}_o B_k' & 0 \\ \vdots & & \vdots & \vdots \\ B_k \widehat{y}_o B_1' & \cdots & B_k \widehat{y}_o B_k' & 0 \\ 0 & \cdots & 0 & 0 \end{vmatrix} \qquad (V.23)$$

The above equations can be interpreted in an easy way. For example, consider the m-th state taken from the n-th recurrent class ($1 \le n \le k$). Let $v_n' = (v_{n1},\ldots,v_{nm},\ldots)$ be the $\lambda=1$ eigenvector of the corresponding irreducible transition matrix P_n, and $x_m^n(\infty)$ be the number of the MCA members in this state as t tends to infinity (see notation (V.15)). Clearly, $x_m^n(\infty)$ is a random variable with expectancy

$$E\left[x_m^n(\infty)\right] = N_n\, v_{nm} + \sum_i b_i\, y_{oi}$$

where, summation over $i=1,\ldots,n_o$ is taken with respect to all coordinates of the transient initial state-vector y_o, and the coefficients $b_1,\ldots,b_{n_o}$ are taken from the appropriate row of the matrix $\begin{bmatrix} B_1 \\ \vdots \\ B_k \\ 0 \end{bmatrix}$.

Its variance from equation (V.23) is computed to be

$$\mathrm{Var}\left[x_m^n(\infty)\,\middle|\,x_o\right] = N_n(v_{nm} - v_{nm}^2) + \sum_i (b_i y_{oi} - b_i^2 y_{oi}) =$$

$$= N_n\, v_{nm}(1 - v_{nm}) + \sum_i b_i y_{oi}(1 - b_i)$$

Therefore, the transient initial state-vector y_o "contributes", in general, to the expectancy as well as the dispersion around this expectancy, i.e. the variance, of any recurrent state. This contribution is additive, and non-negative as can be easily checked.

Example: Consider the case of one recurrent and one transient class of an MCA with transition matrix

$$P = \begin{vmatrix} P_1 & L_1 \\ 0 & Q_o \end{vmatrix}$$

Furthermore, assume that P_1 has only the $\lambda = 1$ eigenvalue of magnitude unity. Thus, $\lim_{t\to\infty} P^t$ exists.

Let $x_o = (y_{11}, y_{12}, \ldots, y_{o1}, y_{o2}, \ldots)'$ be the initial state-vector of the MCA, where, $y_{11}, y_{12}, \ldots$ correspond to the recurrent states and $y_{o1}, y_{o2}, \ldots$ correspond to the transient ones.

Denote the $\lim_{t\to\infty} P^t$ by

$$\lim_{t\to\infty} P^t = \begin{vmatrix} A_1 & B_1 \\ 0 & 0 \end{vmatrix}$$

where, $A_1 = [v_1, \ldots, v_1]$ for v_1 being the $\lambda=1$ eigenvector of A_1. Also by equation (IV.15) B_1 is in the form $B_1 = [v_1, \ldots, v_1]$ (not necessarily with the same number of columns as A_i has).

If we define $N = \sum_i y_{1i} + \sum_i y_{oi}$, i.e. N is the total number of the MCA members, we derive from equations (V.22) and (V.23)

$$E\left[x(\infty)\middle| x_o\right] = \begin{bmatrix} N v_1 \\ \hline 0 \end{bmatrix}$$

$$Var\left[x(\infty)\middle| x_o\right] = \left|\begin{array}{c|c} N(v_1 - v_1 v_1') & 0 \\ \hline 0 & 0 \end{array}\right|$$

Therefore, as time t tends to infinity the state-vector x(t) is expected to become indepedent of the initial state-vector in terms of its two statistical parameters, namely, its expectancy and its variance-covariance matrix.

Numerical application: Let it be

$$P = \left|\begin{array}{ccc|c} 1/2 & 1/3 & 1/2 & \\ 1/2 & 2/3 & 1/4 & \\ \hline 0 & 0 & 1/4 & \end{array}\right| = \left|\begin{array}{c|c} P_1 & L_1 \\ \hline 0 & Q_o \end{array}\right|$$

which yields, $v_1 = \begin{vmatrix} 2/5 \\ 3/5 \end{vmatrix}$

Therefore, if N denotes the number of the MCA members it is found

$$E\left[x(\infty)\right] = N \begin{vmatrix} 2/5 \\ 3/5 \\ 0 \end{vmatrix} \quad \text{while,} \quad Var\left[x(\infty)\right] = \frac{6}{25} N \begin{vmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{vmatrix}$$

1) We can claim with certainty that no member of the MCA will eventually be in state 3.

2) We expect, eventually, 40% of the members of the MCA to be in state 1 and 60% in state 2. These expectations are accompanied by a standard deviation $\sqrt{6N}/5$ for both states 1 and 2.

In order to visualize the practical meaning of these results we will establish some confidence intervals for the state variable $x_1(\infty)$.

By Tshebysheff's lemma (see Uspensky, pp.182) if u is a non-negative random variable with expected value $\bar{u}$ then,

$$\Pr\left\{u \leq t^2\bar{u}\right\} > 1 - \frac{1}{t^2} \qquad \text{for any } t \geq 1.$$

Setting $u = \left(x_1(\infty) - E[x_1(\infty)]\right)^2 = (x_1(\infty) - 2N/5)^2$ and $t = 2$ we derive

$$\Pr\left\{|x_1(\infty) - 2N/5| \leq 2\sqrt{6N}/5\right\} > .75$$

which means that with probability more than 75% $x_1(\infty)$ will be within the interval $(2(N-\sqrt{6N})/5\ ,\ 2(N-\sqrt{6N})/5)$. The shown table presents at the confidence level, at least, 75% the lower and upper bound of $x_1(\infty)$ for various values of N. In a similar way one can treat the state variable $x_2(\infty)$.

N	Lower	Upper
6	0	4.8
24	4.8	14.4
600	216	264
60,000	23,760	24,240

CHAPTER VI. CONCLUSIONS

In chapter II we defined the concepts of Markov system, Markov chain and Markov chain aggregate. The term "Markov Chain Aggregate" is introduced for the first time and we hope it is descriptive enough to survive the test of time.

In chapters III and IV we studied the canonical form of transition matrices (actually, non-negative) and the asymptotic behavior of the transition matrix powers. The spectrum of a transition matrix, consisting of its eigenvalues, plays a determinant role on the asymptotic behavior of the transition matrix powers. However, we treated the finite-state case only. It should be interesting to extend all these results to the countably infinite number of states case. As a matter of fact, some of the stated propositions can be seen, with minor modifications, that are valid for countably infinite number of states.

In chapter V we just touched the subject of Markov Chain Aggregates. We believe that the MCA can model many real-life systems. Thus, it is a useful mathematical tool for practical purposes and applications.

APPENDIX A

PROBABILISTIC CONCEPTS DEDUCED BY EXPERIMENT SPACE

A.1. Motivation

A.2. Background Population, State-Space and Experiment

A.3. Borel-Field, Events

A.4. Experiment Space, System

A.5. Probability Measure Functions, Stochastic System

A.6. Random Variable and Random Process

A.7. Conditional Probability

A.1. Motivation

Probability theory mainly uses two approaches. The first, called the relative frequency approach, is based on the idea that the probability of an event is the relative frequency of this event in its limit as it was introduced by von Mises [16] in 1915. The second, called the axiomatic approach, is based on the abstract notion of probability as introduced by Kolmogorov [13] in 1933. Kolmogorov called probability-measure a real-valued function, defined on a σ-algebra (or, Borel-field), assuming values in the closed interval $[0,1]$, and satisfying a certain set of axioms. This approach is called axiomatic because the existence of probability-measures is postulated. Its axioms do not make any inference as to how probabilities are obtained and computed but assumes them to be known.

There is no logical inconsistency between the two approaches. To-day both are valid and the general impression is that if one has to estimate probability measures he has to use the von Mises definition. However, if probability measures are, or can be assumed to be, known then the most appropriate approach is the axiomatic one. The attractiveness of the axiomatic approach lies in the fact that, with a few axioms, one can cover in a general form all mathematical properties of what one is supposed to call probability. On the other hand, the weakness of this approach

lies in the fact that Kolmogorov axioms are incomplete as he himself has stated (see 13, pp.1-4). Therefore, working with an incomplete system of axioms we are bound to eventually commit logical inconsistencies. The literature on probability theory contains certain types of logical inconsistencies and we will indicate some of them.

Many authors to-day define probability space to be the triple $(S, \mathcal{F}, \mu)$ where, S is either known as sample space or state-space; $\mathcal{F}$ is a Borel-field or σ-algebra defined on S, and μ is a probability measure function defined on $\mathcal{F}$. If the triple $(S, \mathcal{F}, \mu)$, as outlined above, is meant to represent something of the surrounding world, even in an abstract and general form, two things are implied:

(a) There must exist some entity or object to which S, $\mathcal{F}$ and μ are referred.

(b) For any measurable set (or event) $E \in \mathcal{F}$ its probability measure must have the usual meaning that one intuitively attaches to probabilities. In other words, the probability must be a number that quantifies a phenomenon of randomness and chance.

The probability space in the form of the triple $(S, \mathcal{F}, \mu)$ is an abstract structure and as such does not explicitly involve the notion of chance and randomness. There is nothing wrong with this approach provided that one retains logical con-

or

sistency. To be specific, the probability space is described by two sets S and $\mathcal{F}$, and a mapping μ. Its definition is rigorous for the precise reason that it involves only sets and mappings. Any additional concept must be defined along these lines, i.e. it must be defined as either a set or a mapping. Otherwise, one commits logical inconsistency. This type of inconsistency is not unusual in the literature of probability. For example, most of the textbooks on probability theory or applications start with sets, σ-algebras, probability measures etc. Suddenly, their style changes and concepts like outcome, experiment, random variable or function etc. are introduced on an entirely intuitive basis or are not defined at all and are left as self-evident concepts.

In this discussion it is intended to illustrate that probabilistic concepts can be rigorously defined. Our method will be the axiomatic one. However, since Kolmogorov's axioms were not meant to induce concepts like stochastic systems, random variables and processes etc. we will review the whole conceptual structure related to the notion of probability.

First, we postulate the existence of the background population and the state-space. The background population is a set and its points are all the objects under study. The state-space is another set whose points represent some property, attribute, status value etc. for the points of the background population. Any mapping from the background

population into the state-space is defined to be an experiment or observation. We add to our list the concept of a σ-algebra or Borel-field defined on the state-space. Its definition is the standard mathematical one.

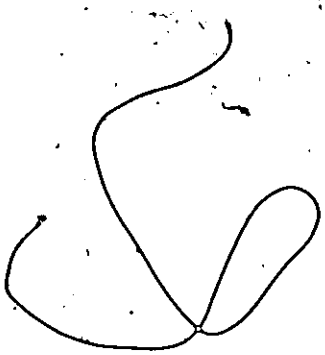
Second, the existence of a countable set of experiments, called experiment space, is postulated. The quadruple consisting of the background population, the state-space, the Borel-field and the experiment space is referred to as a system. If a given system induces probability measures, in a way that resembles relative frequency convergence, then the system is called stochastic and is represented as a quintuple consisting of the underlying system quadruple together with the class of the probability measures.

Third, for a given stochastic system we define certain mappings, induced by the system, and we call them random variables. Similarly, the concept of random process is introduced as a special case of a random variable.

Finally, as a means of application these concepts are used to define the concept of conditional probability.

From the above outline it is evident that we integrate (or at least, attempt to) the von-Mises and Kolmogorov approaches with respect to the definitions of probabilistic concepts. The benefits of such an undertaking is that rigor is used to define what is usually left to one's free discretion.

As we stated our approach is axiomatic. Incidentally however, it appeals to intuition as well. Our original objective was to think along abstract lines but, in order to review the probabilistic logical structure, we had to appeal to intuition and this was proven very fruitful.



A.2. Background Population, State-Space and Experiment

Any study has its subject which is either a concrete or an abstract object. Also it might be a single object or many objects either finite in number or infinite. For example, assume that we are interested in the machines of an industrial plant. The subject, therefore, of our attention and study will be the "population" of these machines. Another example could be a particle observed at each point in time $t \geq 0$. If we denote the particle by "p" then our attention is focussed on p in conjunction with any $t \geq 0$. Therefore, the subject of our study constitutes the set of tuples (p,t) for all $t \geq 0$. Note that the first example involves a finite number of objects, while the second, considers an uncountable number of them.

In abstract and general terms we can say that any study has as its subject the points of a non-empty set B which we call "background population".

Once a background population is given, the obvious question arises: "What do we want to do with it?"

In all cases we attach to the members of the background population characteristics, attributes, properties or status values. For example, we say machine x has accumulated y hours of operation or we say machine x produces the w type of product or both. Also we may say particle p at time t is located on position x etc.

Again in abstract terms we can say that in conjunction with the background population one has available a non-empty set S, called "State-Space" or "Sample-Space", whose points represent the characteristics, attributes etc. of the members of the background population.

We do not impose any restriction on the above described sets except that they must be non-empty. However, the above statements were not intended to be formal definitions of B and S, but merely to explain the intuitive aspects of their introduction. Formally, their existence is simply postulated as follows:

Definition 1: Let B be a non-empty set which will be called "background population" and S be another non-empty set which will be called "State-Space" or "Sample-Space".

The link between the background population and the state-space is the "experiment" or "observation". For example, if we pick a particular machine we can obtain from the industrial plant's records the information as to how many hours of operation this machine has accumulated. Thus, when we make an observation, or an experiment we associate with each member(point) of the background population a point of the state-space. For example, we say particle "p" at time "t" is in the position "x". Clearly, an experiment or observation is a mapping from B into S and as such is formally defined as follows:

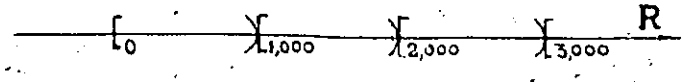
Definition 2: Given background population B and state-space S we call experiment or observation any mapping $e:B \to S$. Also for any $b \in B$ the image $e(b)$ will be called outcome of the experiment e for the point b.

In the literature, the background population is usually implied rather than explicitly stated. One can see that the concept of the experiment cannot be defined as a mapping unless the background population is explicitly introduced. We should emphasize that the non-explicit definition of the background population often leads to ambiguities. This is due to the double role played by the points of the state-space. First, as describing some property of some implied subject and second, as outcomes of particular experiments and observations.

A.3. Borel-Field, Events.

This section presents the mathematical concept of the "Borel-field" or "σ-algebra". In order to introduce it in a natural way we will proceede with the following example:

Let the background population B be the set of the machines in an industrial plant and the state-space S be the set of non-negative real numbers indicating accumulated hours in operation. In practice, one is seldom interested in the exact value of the accumulated hours of operation. Rather, one is usually interested in knowing the interval within which the accumulated hours of operation fall. In other words, if h denotes the hours of operation of a given machine we are interested whether $0 \leq h < 1,000$ or $1,000 \leq h < 2,000$ etc. Therefore, we can partition the state-space $S = \{ h \mid \text{real and } h \geq 0 \}$ into brackets as illustrated by the following figure:



If we consider a collection $\mathcal{F}$ of subsets of S consisting of the empty set and any union of the above defined brackets, this collection of subsets of S is a σ-algebra or a Borel-field on S in its most simple structure. Later we will, for completeness sake, give the mathematical definition of a Borel-field. At the moment, we emphasize the fact that we could choose 2,000 hour brackets to form a different Borel-field on S. Similarly, we

could make the number of brackets finite by taking the last bracket to be, for example, from 100,000 to infinity, or even choose brackets of unequal size, e.g. (0,255], (255,1097],..etc. In all these cases, however, we end up with a collection of subsets of S which in mathematical terms is a σ-algebra or a Borel-field on S. Thus, it is illustrated that a Borel-field defined on a state-space S is, generally, an option among alternatives. The choice of a certain such option is dictated by intuition. Intuition helps one to form the proper collection of subsets of S, called events, that are relevant to his problem at hand. Those subsets of the state-space that are excluded from being events are of no interest and it is meaningless to attach a probability measure to any of them.

For purpose of completeness, we include here the standard definition of a Borel-field.

Given a set S, a Borel-field or σ-algebra on S is a collection $\mathcal{F}$ of subsets of S having the following properties: (a) $S \in \mathcal{F}$, i.e. S itself is a member of the collection (event); (b) If E belongs to the collection so does its complement E^c, i.e. $E \in \mathcal{F}$ implies $E^c \in \mathcal{F}$; (c) if countably many $E_1, E_2, \ldots$ belong to the collection so does their union $\bigcup_i E_i$.

From the above definition one can derive the following simple conclusions: 1) The empty set belongs to the collection $\mathcal{F}$, since $S^c = \emptyset$; 2) For $E_1, E_2, \ldots \in \mathcal{F}$ their complements $E_1^c, E_2^c, \ldots \in \mathcal{F}$, thus,

their union $\bigcup_i E_i^c \in \mathcal{F}$ and its complement $\bigcap_i E_i \in \mathcal{F}$. In words, the intersection of countably many members of the collection belongs to the collection.

We summarize the above discussion into the following:

Definition 3: Given a state-space S, and $\mathcal{F}$ a Borel-field (or σ-algebra) on S, the members of $\mathcal{F}$ will be called "events" or "Borel-sets".

We have chosen above a simple bracket type of Borel-field, in order to make clear its intuitive aspects. However, there are Borel-fields very commonly used which are more complex than the above simple ones.

A.4. Experiment Space, System.

The theory of probability is a branch of applied mathematics dealing with the effects of "chance". The word chance is often used in everyday language without precise meaning. Its exact characterization has caused debate and belongs in the realm of philosophy. In order to include probability within the scope of mathematics we must proceed with the introduction of a certain set of axioms, which captures all the essential features of what we intend to call "chance". Kolmogorov did exactly this. He set up a set of axioms for the probability measures which terminated ambiguities and misuses of language. We intend to respect his consistent logical structure of axioms. Nevertheless, we will postulate not the probability measures themselves; Instead, we will move one step back and postulate the existence of an entity which induces them in a way such that Kolmogorov's axioms are satisfied. This entity contains all the information one has obtained from experience on a particular system.

We assert that lack of experience cannot yield probability measures.

For example, it is meaningless for us to talk about the probability of an inhabitant of Mars having three legs, since there is no experience or data to draw upon. On the other hand, usually the

probability that a certain face of a die, if it is cast, will turn up is meaningful. To illustrate that the meaningfulness of such probability is induced by experience and observations, consider the case of a geometrically unsymmetrical die. It is obvious that in this case we cannot attach to any of the die face a probability unless we first cast the die for at least a few hundred or thousand times and record the outcomes. However, if the die is symmetrical often we do not perform experiments in order to estimate the probabilities in question. Instead, we "assume" all faces equiprobable, and this is enough to declare the probability for each face to turn up as $1/6$. Actually, in this case we simply guess what would be the results of experiments, which as a matter of routine we should perform. Thus, our "assumption", simply, assists us to state results of some experience in its potential form rather than actual one.

The above examples did not intend to prove or completely justify our assertion. This assertion is as old as the concept of probability, and its investigation belongs to the domain of philosophy. However, if it is accepted it becomes clear that we need a certain mathematical entity which contains all the necessary information in order to be able to induce and generate probability measures. The existence of such an entity is postulated in the following definition and is called "experiment space".

Definition 4: Given a background population B, a state-space S, and a Borel-field $\mathcal{F}$ on S, we call the quadruple $\Sigma=(B,S,\mathcal{F},V)$ a system, where, $V=\{e_1,e_2,\ldots\}$ is a countable set of experiments $e_i:B\rightarrow S$;i=1,2,... and is called experiment space.

Since V consists of experiments or observations, it is indeed an entity containing information obtained from experience. The experiment space is required to contain countably many experiments. Technically, this countability offers great conveniences which will become clear in section A.5, where we consider sequences of relative frequencies. Perhaps, V can be generalized to be just any infinite set of experiments, either countable or uncountable. However, for most applications this further complexity seems unnecessary. This is due to the fact that most of the state-spaces in practice are separable spaces, i.e. they contain a countable dense subset.

For example, consider S=R the real line equipped with the Euclidean metric topology. Also let the Borel-fied $\mathcal{F}$ on R consists of all open sets, closed sets, their countable unions, and their countable intersections. Suppose now that the experiment space consists of experiments $e_1,e_2,\ldots$ whose outcomes $e_i(b)$;i=1,2,.., for any point b of the background population, are rational real numbers. Since the set of rationals is countable and also is dense in R then, countable experiments can, in principle, cover densely the real line. Furthermore, any fixed rational number

can be the outcome of countably many experiments for a given point $b \in B$. This property of the real line, to contain a countable dense subset, is possessed by many other spaces. Such spaces are called separable and we can mention: the n-dimensional Euclidean spaces ($n \geq 1$), the space of all continuous real-valued functions defined on any closed interval $[a,a'] \subset R$, the l_p ($p \geq 1$) spaces and others. Thus, from the engineering point of view the countability of the experiment space does not limit its capability for any practical purpose.

The concept of the system $\Sigma=(B,S,\mathcal{F},V)$ was introduced for two reasons. First, for reason of brevity. From now on we will refer to the "system $\Sigma=(B,S,\mathcal{F},V)$", instead of defining each of its elements $B,S,\mathcal{F},V$ separately. Second, this definition of a "system" incorporates the most primitive structure, but complete. We have the object of study which is B, its description which is S and information recorded in V. Finally, $\mathcal{F}$ expresses the list of preferred events to be considered. Actually, we even have a topology τ on the state-space S, which is the strongest topology on S contained in $\mathcal{F}$.

In the next section we will introduce the concept of the probability measures as numerical quantities which are induced by a certain type of systems, called stochastic. Probability measure, therefore, does not exist around by itself, but exists only within the context of a certain system. This thesis can

comfortably explain certain paradoxes in probability theory. For example, the Bertrand's paradox, in which different probabilities are estimated for the "same" event (see Papoulis [17], pp. 11-12), can be easily resolved. It is clear that first, one has to define the system as is, supposedly, implied by the problem's statement. It turns out that the statement of the Bertrand's paradox problem allows, by deductive logic, the construction of more than one experiment space. Thus, one can define more than one "system" which turn out to be stochastic and each of them induces a probability for the event in question. There is no reason that all these probabilities should be the same, and indeed they are not.

A.5. Probability Measure Functions, Stochastic Systems.

In this section we use the concept of a system, as defined previously, and we define relative frequencies as in von-Mises approach. Under the condition that these relative frequencies converge we derive probability measures, i.e. the limits of the relative frequencies satisfy the Kolmogorov's axioms.

Consider a system $\Sigma=(B,S,\mathcal{F},V)$. Let the characteristic functions $f_1, f_2, \ldots$ corresponding to each experiment $e_1, e_2, \ldots \in V$ be defined by

$$f_i(b,E) = \begin{cases} 1 & \text{if } e_i(b) \in E \\ 0 & \text{otherwise} \end{cases}$$

for all $b \in B$ and $E \in \mathcal{F}$.

Given an experiment $e_i \in V$, a point in the background population $b \in B$, and an event $E \in \mathcal{F}$ then, the number $f_i(b,E)$ is an indicator whether or not the outcome $e_i(b)$ falls within the event $E \in \mathcal{F}$. Furthermore, if we form the sum $\sum_{i=1}^{n} f_i(b,E)$ for any integer $n \geq 1$, this sum is a non-negative integer, not-exceeding n, and is the frequency with which the event $E \in \mathcal{F}$ occurs in the first "n" experiments of the experiment space and for the point $b \in B$.

Now we can define the relative frequencies,

$$r_n(b,E) = \frac{1}{n}\sum_{i=1}^{n} f_i(b,E) \qquad \text{for all } b \in B, \text{ and } E \in \mathcal{F}$$

Clearly, the relative frequencies are non-negative real numbers not exceeding unity. However, as n increases and tends to infinity the relative frequencies $r_n(b,E)$;for all $b \in B$, and $E \in \mathcal{F}$ do not necessarily converge. For example, there might be a trend for some of them towards $\frac{1}{2}$ for the first 1,000 experiments. Beyond these experiments the trend might switch towards $\frac{3}{4}$ for the next million experiments, and switch back to $\frac{1}{4}$ or to any other number etc. Nevertheless, the concept of probability is based on the assumption that there exist systems in which relative frequencies converge. Such systems are known as probabilistic or stochastic. In order to be indepedent from philosophical considerations we define:

Definition 5: A system $\Sigma=(B,S,\mathcal{F},V)$ is called stochastic or probabilistic if for all $b \in B$ and $E \in \mathcal{F}$ the limit of the relative frequency

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(b,E) = \lim_{n\to\infty} r_n(b,E) = \mu_b(E)$$

exists. This limit $\mu_b(E)$ will be called "probability of the event $E \in \mathcal{F}$ for the point $b \in B$".

Consider now a stochastic system $\Sigma=(B,S,\mathcal{F},V)$ with $\mu_b(E)$ the associated probabilities, as defined above, for each $b \in B$ and $E \in \mathcal{F}$. If we concentrate on a particular $b \in B$ we can define the mapping $\mu_b: \mathcal{F} \to [0,1]$ where, $\mu_b(E)$ image of μ_b for any $E \in \mathcal{F}$ is the probability for this event.

One can easily verify that the mapping μ_b is a probability measure, i.e. it satisfies the Kolmogorov's axioms:

(i) $0 \leq \mu_b(E) \leq 1$ for all $E \in \mathcal{F}$

(ii) $\mu_b(S) = 1$ and $\mu_b(\emptyset) = 0$

(iii) $\mu_b(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} b(E_i)$ for any pairwise mutually exclusive events $E_1, E_2, \ldots$ i.e. $E_i \cap E_j = \emptyset$ for $i \neq j$.

The above discussion allows us to define $M = \{\mu_b \mid b \in B\}$ the class of the probability measure functions associated with a stochastic system $\Sigma = (B, S, \mathcal{F}, V)$. If the background population is a one-point set then, M consists of only one such mapping. The same is valid when all points of the background population have common probability measure functions. This is the most common case in the literature of probability theory and applications. However, when we say that a system is stochastic, by definition, we assume that $M = \{\mu_b \mid b \in B\}$ exists and therefore, it is reasonable to call stochastic system the quintuple $\Sigma_s = (B, S, \mathcal{F}, V, M)$. We prefer to give to this quintuple a separate name like, "probability space".

In the literature, "probability space" is called a triple in the form $(S, \mathcal{F}, \mu)$ where, S is the sample-space, $\mathcal{F}$ is a Borel-field on S and μ is a probability measure function on $\mathcal{F}$. In fact, this triple is not far away from the quintuple we suggested above. In the literature, B is assumed to be a single-point set and therefore, is implied in the triple notation $(S, \mathcal{F}, \mu)$

or, equivalently, can be omitted from the quintuple Σ_s. Also, for the same reason, M consists of only one mapping, i.e. $M=\{\mu\}$ where, $\mu = \mu_b$ for the single point $b \in B$. Hence, we can replace, in this case, M by μ. However, the only completely new item in our quintuple Σ_s is the experiment space V.

To summarize the above discussion we define:

Definition 6: If $\Sigma=(B,S,\tilde{f},V)$ is a stochastic system with $M = \{\mu_b \mid b \in B\}$ the class of the probability measure functions which are, by definition, associated with the system Σ, we call the quintuple $\Sigma_s = (B,S,\tilde{f},V,M)$ a probability space.

Given a stochastic system Σ then by definition its experiment space V generates its class of probability measure functions M. One could, perhaps, argue that we do not need anymore the experiment space V. Its purpose of introduction has, perhaps, already been fulfilled. This argument has some merit but, is not completely true. The experiment space is further needed for the introduction of a few more probabilistic concepts. On the other hand, V and M can be seen as equivalent or rather "conjugate" concepts, in the sense that, one can be assumed as being generated or induced by the other. By definition of the stochastic system, given V it induces M. Conversely, given M, one could, in principle, induce an experiment space V by, for example, stochastic simulation. In the light of these arguments

both V and M can enjoy equal status in the probability space, and both are useful in probability theory. Perhaps, they are not used both at the same time, but the same is true with the the electric current and the electric charge in the theory of electricity.

A.6. Random Variable and Random Process.

Here we shall define the concept of the random variable and the random process using the previously introduced concept of probability space $\Sigma_s = (B,S,\tilde{f},V,M)$. Papoulis (pp. 85-88) defines the random variable in the following way. He associates with each experiment "e" a point in the state-space $x(e)$. Thus, he constructs a mapping "x" from the experiment space into the state-space, which he defines to be a random variable.

In our terminology, given a stochastic system $\Sigma = (B,S,\tilde{f},V)$ it induces, by definition, a probability space $\Sigma_s = (B,S,\tilde{f},V,M)$. For each point $b \in B$ of the background population B and each experiment $e \in V$ of the experiment space V we can associate a point of the state-space, say $x(b,e)$, which is the outcome of e for the point b, i.e.

$$x(b,e) = e(b) \qquad \text{for all } b \in B \text{ and } e \in V$$

Clearly, this defines a mapping from the Cartesian product $B \times V$ into the state-space S; $x: B \times V \longrightarrow S$. This mapping x is said to be a <u>random variable</u>, and, since it is a direct product of the system Σ, we can call it <u>the underlying random variable</u> of the stochastic system Σ. Furthermore, given any $b \in B$ and event $E \in \tilde{f}$, we call <u>"probability of event E for the random variable x at $b \in B$"</u> the probability measure $\mu_b(E)$, and we denote this by

$$\Pr\{x(b) \in E\} = \mu_b(E) \qquad \text{for any } b \in B \text{ and } E \in \tilde{f}.$$

Intuitively speaking, if we fix the point of interest in the background population $b \in B$ then, for each performed experiment $e \in V$ there exists a definite value for the, so-called, underlying random variable. However, a random variable is supposed to be an entity assuming values on the state-space in a chance manner. This randomness and chance is expressed in terms of the associated probability measure $\mu_b(E)$ for each event $E \in \tilde{\mathcal{F}}$.

For a certain stochastic system we have derived above a well-prescribed mapping called underlying random variable. Actually, we can generate many more random variables via measurable mappings from the state-space into any other arbitrary space.

Consider a probability space $\Sigma_s = (B, S, \tilde{\mathcal{F}}, V, M)$ and its underlying random variable $x: B \times V \longrightarrow S$. Also, consider any measurable mapping from the state-space S into some set Q

$$f: S \longrightarrow Q$$

where, there is defined on Q a Borel-field $\tilde{\mathcal{F}}_q$. We recall that a mapping $f: S \longrightarrow Q$ is measurable if for every event $E_q \in \tilde{\mathcal{F}}_q$ the inverse image $f^{-1}(E_q) \in \tilde{\mathcal{F}}$, i.e. $f^{-1}(E_q) \in \tilde{\mathcal{F}}$ for all $E_q \in \tilde{\mathcal{F}}_q$. The composite mapping

$$f \circ x: B \times V \longrightarrow Q$$

is defined by $(f \circ x)(b,e) = f(x(b,e))$ for all $b \in \beta$ and $e \in V$. Furthermore, we can associate with each point of the background population $b \in \bar{B}$ and each event $E_q \in \tilde{\mathcal{F}}_q$ the probability measure $\mu_b(f^{-1}(E_q))$, which is well-defined because f is measurable. We denote this by the expression

$$\Pr\left\{(f\circ x)(b)\in E_q\right\} = \Pr\left\{x(b)\in f^{-1}(E_q)\right\} = \mu_b(f^{-1}(E_q)) \quad \text{for all } b\in B;\ E_q\in \tilde{f}_q.$$

Such a composite mapping $f\circ x$ is called, traditionally, random function and recently, simply random variable. The traditional term random function is more descriptive and is popular in practical fields of science. However, the term random variable ie preferred by mathematicians since there is no foundamental difference between these two concepts.

In conclusion, any measurable mapping defined on S induces a composite mapping $v = f\circ x$ from the Cartesian product $B\times V$ into the range Q of f. Furthermore, for each event $E_q\in \tilde{f}_q$ and each $b\in B$ we can associate a probability measure. We called any such mapping random variable.

To clarify the above discussion we will give a simple example.

Example: Consider a background population consisting of two dice labelled, for purpose of identification, "a" and "b", i.e. $B=\{a,b\}$. Clearly, the state-space is $S=\{1,2,3,4,5,6\}$ consisting of the six labels of the six faces for each die. The Borel-field $\tilde{f}$ is the power set of S. That is, any subset of S is considered to be an event. Let V be an experiment space that induces equal probabilities for each face of each die. That means,

$$\mu_a(\{i\}) = \mu_b(\{i\}) = 1/6 \qquad \text{for all faces } i=1,2,3,4,5,6\in S$$

Clearly, we have defined the probability space $\Sigma_s = (B,S,\tilde{f},V,M)$ where, $M=\{\mu_a,\mu_b\}$ consists of the two probability measure functions, one for each die, which incidentally are equal.

Consider now the following mapping f from S into the set $Q = \{W,L\}$ defined by

$$f(i) = \begin{cases} W & \text{if } i = 1 \text{ or } 6 \\ L & \text{if } i = 2,3,4 \text{ or } 5 \end{cases}$$

Such mapping could be defined, for example, by means of a game. That is, one wins (W) if the face of any die turns to be 1 or 6, otherwise, he loses (L). Let $\mathcal{F}_q$ the Borel-field on Q be its power set consisting of the sets $\emptyset, \{W\}, \{L\}$, Q. The inverse images of all events in Q under f are $\emptyset, \{1,6\}, \{2,3,4,5\}, S$ which are events in S. Therefore, f is a measurable mapping.

The underlying random variable x is the mapping $x: B \cdot V \rightarrow S$ such that, given one of the two dice, say "a", and an experiment $e \in V$ then, the image $x(a,e) = e(a)$, i.e. it is the outcome of this particular experiment e for the die a.

On the other hand, with any event $E \in \mathcal{F}$ we associate the probability measure denoted by

$$\Pr\{x(a) \in E\} = \mu_a(E) = \sum_{i \in E} \mu_a(\{i\}) = \sum_{i \in E} 1/6$$

However, for the random variable $f \circ x$, defined by the measurable mapping f, for each event $E_q \in \mathcal{F}_q$ we associate probability

$$\Pr\{(f \circ x)(a) \in E_q\} = \Pr\{x(a) \in f^{-1}(E_q)\}$$

Specifically,

for $E_q = \emptyset$ $\Pr\{(f \circ x)(a) \in \emptyset\} = \Pr\{x(a) \in f^{-1}(\emptyset)\} =$

$$= \Pr\{x(a) \in \emptyset\} = \mu_a(\emptyset) = 0$$

for $E_q = \{W\}$ $\Pr\{(f \circ x)(a) \in \{W\}\} = \Pr\{x(a) \in f^{-1}(\{W\})\} =$

$$= \Pr\{x(a) \in \{1,6\}\} = 1/6 + 1/6 = 2/6 = 1/3$$

for $E_q=\{L\}$ $\Pr\{(f\circ x)(a)\in\{L\}\} = \Pr\{x(a)\in f^{-1}(\{L\})\} =$

$= \Pr\{x(a)\in\{2,3,4,5\}\} = 4/6 = 2/3$

finally, for $E_q=Q=\{W,L\}$ it yields, $\Pr\{(f\circ x)(a)\in Q\} = 1$.

The associated probabilities with the events $\{W\}$ (win) and $\{L\}$ (lose) have a clear intuitive interpretation.

A special type of a random variable is the <u>random process</u>. The term process is used to specify "time" involvement in the structure of the system in question. Specifically, a system is called a process if its background population is a time set, i.e. a set whose points specify time and as such it can be the real line, any subset of the real line or, in more abstract terms, any totally ordered set. The term process is, simply, descriptive and there is no mathematical significance in the distinction between a "system", in general, and a "process". It is customary to use the expression "let x(t) be a random process", which means that, for each point in time t and an experiment e the random process assumes value $x(t)(e) = e(t)$. On the other hand, given an event E, then the probability $\Pr\{x(t)\in E\}$ is meaningful for each point in time t.

A.7. Conditional Probability.

In any contemporary textbook of probability theory, the definition of the conditional probability is given as follows:

Given a non-null event $E_o \in \tilde{f}$, i.e. of probability measure non-zero, the probability measure function μ_{E_o} on $\tilde{f}$, such that,

$$\mu_{E_o}(E) = \frac{\mu(E \cap E_o)}{\mu(E_o)} \qquad \text{for all } E \in \tilde{f},$$

is called conditional on E_o probability measure on $\tilde{f}$. Also, the probability measure $\mu_{E_o}(E)$, for any event E, is often denoted by $Pr\{E|E_o\}$, and called, probability of the event E conditional to the occurence of E_o.

This definition requires the "unconditional" probability measures $\mu(E)$; $E \in \tilde{f}$ to be meaningful and defined. However, there are systems in which the unconditional probability measures are not defined while the conditional ones are. Some of the Markovian systems are, for example, of this nature. Clearly, for these cases, the above definition cannot be used. On the other hand, for such systems we still maintain the term "conditional probability". Thus, we are either logically inconsistent or, at least, ambiguous. In order to remove this fallacy, one must either alter the definition of conditional probability or do not call conditional probabilities those for which their unconditional ones are not defined. We prefer the first alternative, i.e. redefine the concept in question in a way that does not involve the concept of unconditional probability.

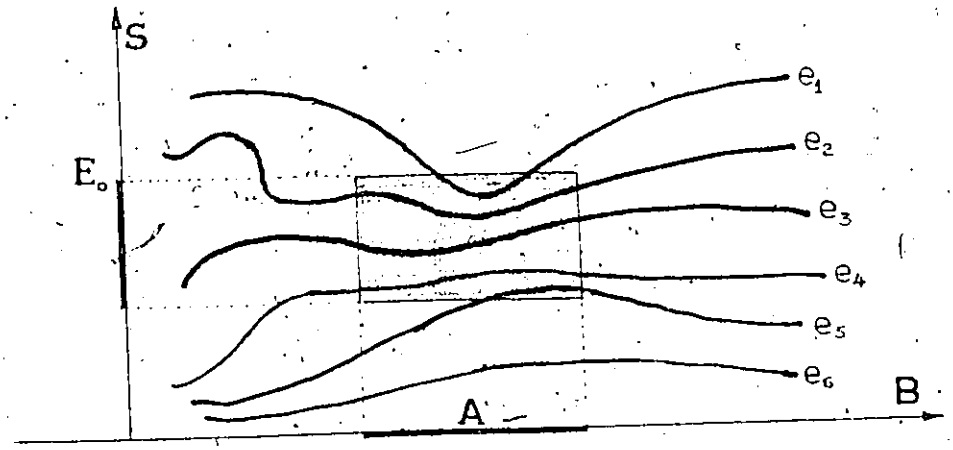
Given a system $\Sigma = (B,S,\tilde{\mathcal{F}},V)$, not necessarily stochastic, consider a non-empty subset of the background population $A \subset B$, and an event, other than the empty, $E_o \in \tilde{\mathcal{F}}$. Further, select from the experiment space V those experiments $e \in V$ whose outcomes fall in E_o for all the points of A, i.e. $e(b) \in E$ for all $b \in A \subset B$. Denote this selection by

$$V' = V'(A,E_o) = \{e \in V : e(b) \in E_o \text{ for all } b \in A \subset B\},$$

Clearly, V' is a set of experiments and if it is not finite, i.e. it is a countable set of experiments, we can form the system

$$\Sigma' = \Sigma'(A,E) = (B,S,\tilde{\mathcal{F}},V')$$

The shown figure illustrates the suggested above selection of experiments.



From $V = \{e_1,e_2,e_3,e_4,e_5,\ldots\}$ we form $V' = \{e_2,e_3,e_4,\ldots\}$ which all have outcomes in E_o for all the points of $A \subset B$.

We can briefly describe the suggested selection of experiments by saying: "experiment space V' is formed under the condition (A,E_o) on V". Similarly, the system $\Sigma' = \Sigma'(A,E_o)$ is said "to be formed by the condition (A,E_o)".

Definition 7: Given system $\Sigma = (B,S,\tilde{\mathcal{F}},V)$ if, under condition (A,E_o) for some $A \subset B$ and $E_o \in \tilde{\mathcal{F}}$, the formed system $\Sigma' = \Sigma'(A,E_o)$ is stochastic then, the induced probability measures μ_b; $b \in B$ are called probability measures conditional on (A,E_o). Also, for any event $E \in \tilde{\mathcal{F}}$ and $b \in B$ we call $\mu_b(E)$ the probability of E conditional on (A,E_o).

The above definition expresses formally what one understands as conditional probability function from the intuitive point of view. To illustrate this, consider the following:

Example: Let system $\Sigma = (T,R,\tilde{\mathcal{F}},V)$ where, the background population is the time set T and the state-space is the real line R.

Consider $A = \{t_o\}$, consisting of one point of time, and E_o some event on R. We select the experiments from V under the condition (t_o,E_o) to form the set of experiments

$$V' = V'(t_o,E_o) = \{e \in V : e(t_o) \in E_o\}$$

If we assume that V' is an infinite set of experiments, and also that the system $\Sigma' = \Sigma'(t_o,E_o) = (T,R,\tilde{\mathcal{F}},V')$ is stochastic, then it induces probability measures $\mu_t(E)$; for all $t \in T$ and $E \in \tilde{\mathcal{F}}$. The quintuple $(T,R,\tilde{\mathcal{F}},V,M)$ is a probability space, where $M = \{\mu_t : t \in T\}$ is the class of probability

measures. Let $x: T\times V' \longrightarrow R$ be the underlying random variable of this probability space for which the expression $Pr\{x(t)\in E\}$ is meaningful for any $t\in T$ and any event $E\in \tilde{\mathcal{F}}$. In order to emphasize the fact that these probabilities are meaningful with respect to the condition (t_o, E_o), we denote them by the expression

$$Pr\{x(t)\in E \mid x(t_o)\in E_o\} \quad \text{for all } t\in T \text{ and } E\in \tilde{\mathcal{F}}$$

which is the usual notation for conditional probabilities.

One notices that the above definition of conditional probability does not involve the unconditional one. Furthermore, if the unconditional probabilities exist, our definition does not require the event E_o to be non-null. To illustrate this consider the following:

Example: Consider the system $(B,S,\tilde{\mathcal{F}},V)$, such that,

$B = \{b\}$, i.e. it is one-point background population,

$S = \{1,2,3\}$ with its power set as the system's Borel-field $\tilde{\mathcal{F}}$,

and V is assumed to be given with the following property:

For every large enough integer n, among the first n^2 experiments of the experiment space V there exist $n/3$ experiments with outcome $1\in S$ and $2n/3$ experiments with outcomes $2\in S$.

Our system is clearly stochastic with probability measures

$$\mu(\{1\}) = \mu(\{2\}) = 0 \quad \text{and} \quad \mu(\{3\}) = 1$$

because for n^2 experiments the relative frequencies for events $\{1\}$ and $\{$ are $\frac{n/3}{n^2} = \frac{1}{3n}$ and $\frac{2}{3n}$, respectively, and they tend to zero as $n\to\infty$.

If we choose $E_o=\{1,2\}$ the standard definition of conditional on E_o probability cannot be used, since E_o is a null set ($\mu(E_o) = 0$). However, using the above definition, we select from V those experiments whose outcomes are either 1 or 2. From every n^2 experiments there will be selected n such experiments ($n \geq 1$). Thus, as n tends to infinity we extract from V countably many such experiments to form an experiment space V'. The derived system $(B,S,\tilde{f},V')$ is clearly stochastic with conditional probabilities

$$\mu_{E_o}(\{1\}) = 1/3, \quad \mu_{E_o}(\{2\}) = 2/3, \quad \mu_{E_o}(\{3\}) = 0 .$$

The intuitive interpretation of these probabilities is clear. For example, $\mu_{E_o}(\{1\}) = 1/3$ means that, among all experiments with outcome either 1 or 2 one out of every three has outcome 1.

Finally, we should point out that in the case where both definitions are valid, the numerical values of both conditional probabilities, derived from the standard definition and the above one, coincide.

We terminate our discussion here although we could use our approach to define a few more probabilistic concepts. For example, the concepts of joint and marginal distributions. However, our purpose was to introduce a formalized version of the relative frequency approach, and we feel that our objective has already been met. In chapter II of this study, we introduce the concept of a Markovian system. The definitions of system, stochasti system and conditional probability are used there. Without the material of this appendix we would be forced to carelessly imply certain things which, in our opinion, cannot be implied but explicitly stated and explained.

A P P E N D I X B

Appendix B: A FORTRAN program to arrange square matrices in their canonical form.

Based on the material of chapter III, it is written MPOWER, a FORTRAN program, in order to arrange a transition matrix in its canonical form. Complete listing of the program is presented here as well as the results of the example treated in chapter III.

One should note that, although the program was written for transition matrices, MPOWER can treat any square matrix as explained in chapter III.

```
      DIMENSION TR(15,15)
C
      TYPE 201
201   FORMAT(1X,'TYPE ORDER OF MATRIX  ',$)
      ACCEPT 202,N
202   FORMAT(I5)
C
      DO 5 I=1,N
      DO 5 J=1,N
5     TR(I,J)=0.
      TYPE 203
203   FORMAT(1X,'TYPE ONE-BY-ONE ALL NON-ZERO ENTRIES'/
     A  1X,'IN THE FORM: (ROW) (COLUMN) (VALUE)'/
     B  1X,'ROW=0,COLUMN=0 WILL TERMINATE INPUT'//)
C
9     ACCEPT 200,I,J,X
200   FORMAT(3I5)
      IF(I+J.EQ.0) GO TO 10
      TR(I,J)=X
      GO TO 9
10    CONTINUE
C
      CALL MPOWER(TR,N)
      STOP
      END
```

```
      SUBROUTINE MPOWER(TR,N)
C
C SUBROUTINE TO MANIPULATE NON-NEGATIVE SQUARE MATRIX 'TR' (N X N)
C
C 1ST TO COMPUTE AND PRINT THE SUM OF POWERS OF TR UP TO ITS ORDE
C
C 2ND TO DERIVE RECURRENT AND TRANSIENT CLASSES.
C
C 3RD TO SORT CLASSES AND REARRANGE STATES AND REPORT 'TR' IN ITS
C     CANONICAL FORM.
C
C
C NOTE: MATRIX P IS A COPY OF INPUT MATRIX 'TR'
C       MATRIX P IS IDENTIFIED AS HAVING ZERO AND NON-ZERO
C       ELEMENTS ONLY. NAMELY,0 OR 1.  MULTIPLICATION AND ADDITION
C       ARE CARRIED OUT AS FOLLOWS:  0+0=0, 0+1=1, 1+1=1, 0*0=0
C       0*1=0, 1*1=1.
C
      INTEGER P,Q,TEM
      DIMENSION P(15,15),Q(15,15),TEM(15),TR(15,15)
C
      DO 5 I=1,N
      DO 5 J=1,N
      K=0
      IF(ABS(TR(I,J)).GT.1.E-25) K=1
      P(I,J)=K
      Q(I,J)=0
5     CONTINUE
C
C
C
      DO 60 M=1,N
C
C COMPUTE MATRICES Q(M)=P*[I+Q(M-1)] WITH Q(0)=0
C THEREFORE, Q(1)=P, Q(2)=P+P*P, ETC.
C
C FIRST, COMPUTE (I+Q(M-1)) WHERE Q(M-1) IS STORED IN Q
C
      DO 10 I=1,N
      Q(I,I)=1
10    CONTINUE
C
C MULTIPLY P BY Q(I.E. I+Q(M-1)) AND STORE IT IN Q.
C
      DO 20 J=1,N
      DO 15 I=1,N
      TEM(I)=0
      DO 15 J1=1,N
      K=P(I,J1)*Q(J1,J)
15    TEM(I)=LAD(K,TEM(I))
      DO 18 I=1,N
18    Q(I,J)=TEM(I)
20    CONTINUE
60    CONTINUE
```

```
C
C THE MATRIX Q(N) HAS BEEN COMPUTED. NOW REPORT IT.
C
      TYPE 100,N
100   FORMAT(//1X,' MATRIX POWER SUM OF ORDER',I3)
C
      DO 30 I=1,N
30    TYPE 101,(Q(I,J),J=1,N)
101   FORMAT(1X,50I2)
      CALL CLASS(N,TR,0)
      RETURN
      END
      FUNCTION LAD(KK,LL)
      MM=KK+LL
      IF(MM) 10,10,20
10    LAD=0
      GO TO 30
20    LAD=1
30    RETURN
      END
```

```
      SUBROUTINE CLASS(N,P,Q)
      INTEGER Q(15,15),STATE(15)
      DIMENSION NSTA(15),KLASS(16),NTYPE(15),KOUNT(16),P(15,15)
     A ,QQ(15,15)
      DO 1 I=1,N
      KLASS(I)=0
1     NSTA(I)=0
C
      IC=0
      IS=0
C SWEEP ALL STATES FOR CLASSIFICATION.
      DO 50 I=1,N
      IF(KLASS(I).NE.0) GO TO 50
C STATE I IS THE FIRST IN THE IC CLASS
      IC=IC+1
      IS=IS+1
      NSTA(IS)=I
      KLASS(I)=IC
      KOUNT(IC)=1
      IF(Q(I,I).EQ.0) GO TO 50
      IP1=I+1
      IF(IP1.GT.N) GO TO 50
      DO 50 II=IP1,N
      IF(KLASS(II).NE.0) GO TO 50
      IF(Q(I,II).EQ.0.OR.Q(II,I).EQ.0) GO TO 50
      KLASS(II)=IC
      KOUNT(IC)=KOUNT(IC)+1
      IS=IS+1
      NSTA(IS)=II
50    CONTINUE
C
C ALL STATES ARE CLASSIFIED. SPECIFY NOW TYPE OF CLASSES.
      NT=1
      DO 60 IK=1,IC
      NN=NT+KOUNT(IK)
      KOUNT(IK)=NT
60    NT=NN
      KOUNT(IC+1)=NN
C
C
      NTR=0
      NRC=0
      DO 100 IK=1,IC
      I1=KOUNT(IK)
      I2=KOUNT(IK+1)-1
      DO 90 JJ=I1,I2
      J=NSTA(JJ)
      DO 90 I=1,N
      IF(Q(I,J).EQ.0) GO TO 90
      IF(KLASS(I).EQ.KLASS(J)) GO TO 90
      NTYPE(IK)=0
      NTR=NTR+1
      GO TO 100
90    CONTINUE
      NRC=NRC+1
      NTYPE(IK)=NRC
100   CONTINUE
```

```
C       IF(NTR.LE.0) GO TO 999
C
110     DO 200 IK=1,IC
        IF(NTYPE(IK).NE.0) GO TO 200
C IK CLASS IS TRANSIENT. CHECK IF GOES ONLY TO RECENT STATES.
        NTYPE(IK)=NRC+1
        I1=KOUNT(IK)
        I2=KOUNT(IK+1)-1
        DO 150 JJ=I1,I2
        J=NSTR(JJ)
C
        DO 150 I=1,N
        IF(Q(I,J).EQ.0.0) GO TO 150
        KLS=KLASS(I)
        IF(NTYPE(KLS).NE.0) GO TO 150
        NTYPE(IK)=0
        GO TO 200
150     CONTINUE
        NRC=NRC+1
200     CONTINUE
        IF(NRC.LT.IC) GO TO 110
C
999     IS=0
        IK=1
210     DO 220 KLS=1,IC
        IF(NTYPE(KLS).NE.IK) GO TO 220
        I1=KOUNT(KLS)
        I2=KOUNT(KLS+1)-1
        GO TO 230
220     CONTINUE
        TYPE 221,IK
221     FORMAT(1X,I3,'-TH CLASS MISSING')
        GO TO 850
230     KLASS(IK)=IS+1
        DO 240 JJ=I1,I2
        IS=IS+1
240     STATE(IS)=NSTR(JJ)
        IK=IK+1
        IF(IK.LE.IC) GO TO 210
        KLASS(IK)=IS+1
C
        DO 300 I=1,N
        NTYPE(I)=I
        DO 300 J=1,N
300     QQ(I,J)=Q(I,J)
C
        TYPE 800
800     FORMAT(///1X,'ORIGINAL TRANSITION MATRIX')
        CALL REPORT(N,P,NTYPE)
```

```
C
      TYPE 801
801   FORMAT(///1X,'TRANSITION MATRIX IN CANONICAL FORM'//)
      CALL REPORT(N,P,STATE)
C
      TYPE 802
802   FORMAT(1X,'COMMUNICATION CLASSES'//)
      DO 350 IK=1,IC
      I1=KLASS(IK)
      I2=KLASS(IK+1)-1
      TYPE 803,IK,(STATE(I),I=I1,I2)
350   FORMAT(1X,'CLASS ',I3,2X,15I4)
803   RETURN
850   END
```

```
      SUBROUTINE REPORT(N,P,NSTA)
      DIMENSION P(15,15),NSTA(15),TEM(15)
C
      TYPE 100,(NSTA(I),I=1,N)
100   FORMAT(4X,1516)
      DO 20 I=1,N
      II=NSTA(I)
      DO 15 J=1,N
      JJ=NSTA(J)
15    TEM(J)=P(II,JJ)
20    TYPE 104,II,(TEM(J),J=1,N)
104   FORMAT(1X,I2,'*',15F6.3)
      RETURN
      END
```

MATRIX POWER SUM OF ORDER 11

```
1 0 1 0 0 0 0 1 0 0 1
0 1 0 1 1 0 1 1 1 1 0
1 0 1 0 0 0 0 1 0 0 1
0 0 0 1 1 0 0 1 0 1 0
0 0 0 1 1 0 0 1 0 1 0
0 0 0 0 0 1 0 0 0 0 1
0 1 0 1 1 0 1 1 1 1 0
0 0 0 0 0 0 0 1 0 0 0
0 1 0 1 1 0 1 1 1 1 0
0 0 0 0 0 0 0 1 0 1 0
0 0 0 0 0 0 0 0 0 0 0
```

ORIGINAL SQUARE MATRIX

	1	2	3	4	5	6	7	8	9	10	11
1◆	0.500	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500
2◆	0.000	0.330	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000
3◆	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000
4◆	0.000	0.000	0.000	0.000	0.330	0.000	0.000	0.000	0.000	0.330	0.000
5◆	0.000	0.000	0.000	1.000	0.330	0.000	0.000	0.000	0.000	0.000	0.500
6◆	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
7◆	0.000	0.670	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000
8◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000
9◆	0.000	0.000	0.000	0.000	0.340	0.000	0.000	0.250	0.000	0.340	0.000
10◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

SQUARE MATRIX IN CANONICAL FORM

	1	3	2	7	9	6	4	5	10	11	8
1◆	0.500	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.000
3◆	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250
2◆	0.000	0.000	0.330	0.000	1.000	0.000	0.000	0.000	0.330	0.000	0.000
7◆	0.000	0.000	0.670	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9◆	0.000	0.000	0.000	0.750	0.000	0.000	0.000	0.340	0.000	0.000	0.000
6◆	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.500	0.000
4◆	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.330	0.000	0.000	0.250
5◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.340	0.000	0.250
10◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250
8◆	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

COMMUNICATION CLASSES

```
CLASS # 1◆    1  3
CLASS # 2◆    2  7  9
CLASS # 3◆    6
CLASS # 4◆    4  5
CLASS # 5◆   10
CLASS # 6◆   11
CLASS # 7◆    8
EXIT
```

REFERENCE

1. Blanc-Lapierre, A. - Fortet, R.: "Theory of Random Functions", Gordon and Breach Science Publishers, N.Y., 1968.

2. Burmeister, E. - Dobell, A.R.: "Mathematical Theories of Economic Growth", McMillan Co., 1970.

3. Cinlar, E.: "Stochastic Processes", Northwestern University, 1970.

4. Chung, K.L.: "A Course in Probability Theory", Harcourt Brace J. Inc., N.Y., 1968.

5. Chung, K.L.: "Markov Chains with Stationary Transition Probabilities", Springer-Keylag, Berlin, 1967.

6. Chung, K.L.: "Lectures on Boundary Theory for Markov Chains", Annals of Mathematical Studies, Princeston University, 1970.

7. Cox, D.R. - Miller, H.D.: "Theory of Stochastic Processes", John Wiley, N.Y., 1965.

8. Dynkin, E.B.: "Markov Processes", Springer-Keylag, 1965.

9. Faddeev, D.K. - Fadeeva, V.N.: "Computational Methods of Linear Algebra", W.H. Freeman and Co., San Francisco, 1963.

10. Gantmacher, F.R.: "The Theory of Matrices", Chelsea Publishing Co., N.Y., 1959.

11. Kemeny, J.G. - Snell, J.L.: "Mathematical Models in the Social Sciences", Blaisdell Publicshing Co., 1962.

12. Kemeny, J.G. - Snell, J.L. - Knapp, A.W.: "Denumerable Markov Chains", D. Van Nostrand Co., 1966.

13. Kolmogorov, A.N.: "Foundations of the Theory of Probability", Chelsea Publishing Co., 1950.

14. Kolmogorov, A.N. - Fomin, S.V.: "Introductory Real Analysis", Prentice Hall Inc., 1970.

15. Lipschutz, S.: "Linear Algebra", Schaum's Outline Series, McGraw-Hill Book Co., 1968.

16. Von Mises, R.: "Probability, Statistics and Truth", McMillan Co., N.Y., 1957.

17. Papoulis, A.: "Probability, Random Variables and Stochastic Processes", McGraw-Hill Inc., 1965.

18. Parthasarathy, K.R.: "Probability Measures on Metric Spaces", Academic Press, 1967.

19. Ross, S.M.: "Applied Probability Models with Optimization Applications", Holden-Day, 1970.

20. Spiegel, M.R.: "Advanced Calculus", Schaum Publishing Co., N.Y., 1963.

21. Sverdrup, E.: "Laws and Chance Variations", (2 volumes), North-Holland Publishing Co., Amsterdam, 1967.

22. Thrall, R.M. - Tornheim, L.: "Vector Spaces and Matrices", J. Wiley and Sons, London, 1957.

23. Zadeh, L.A. - Polak, E.: "System Theory", McGraw-Hill, 1969.

VITA

Name:	Georgios Dimitrios Mistriotis
Place of Birth:	Halkis, Province of Evia, Greece
Date of Birth:	May 20, 1940
Education:	
High-School	2nd Gymnasium of Athens, Greece
University	National Technical University of Athens, B.A.Sc. (El.Eng.), 1963
Graduate Work	University of Edinburgh, Scotland, Diploma in Electronics and Control, 1967