STATISTICAL RESEARCH REPORT
Institute of Mathematics
University of Oslo

No 2
January 1976

HOW FAST DOES A MARKOV CHAIN FORGET THE INITIAL STATE? A DECISION THEORETIC APPROACH.

by

Bo Lindqvist

Contents.

ABSTRACT.

Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with finite state space. We study the experiment $\mathcal{E}_n$ of observing X_n in order to obtain information about the initial state X_0. The sequence $\{\mathcal{E}_n\}_{n=1}^{\infty}$ is proved to have a limit $\mathcal{E}_\infty$, which is characterized in terms of properties of the given chain. A measure for the information contained in $\mathcal{E}_n$ is proposed, and an asymptotic expression is derived. If states of the chain are lumped, then some information may get lost. Two methods of measuring the loss are studied, the deficiency and the insufficiency. They are shown to have equal asymptotic properties.

Key words: Markov chain, limit experiment, information, lumping, insufficiency.

1. Introduction.

Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain on the finite state space $\chi = \{1,2,\ldots,r\}$, with transition matrix P. We shall consider a statistical model where X_n is observed in order to obtain information about the initial state X_0. Thus X_0 is the unknown parameter, and parameter set = sample space = the state space of our chain.

Such models are frequently appearing in genetics. One observes some generation of a population and wants information about previous generations. An example is given in section 3.

The main question is, how can we measure the information ? Le Cam has in [7] introduced an information-distance between experiments, the deficiency, which generalizes Blackwell's ([1]) relation "more informative". The deficiency gives rise to a semi-metric in the set of experiments, the Δ-distance.

Let $\mathcal{E}_n$ be the experiment of observing X_n. The information contained in $\mathcal{E}_n$ of course decreases as n grows. It is shown that the sequence $\{\mathcal{E}_n\}$ has a limit $\mathcal{E}_\infty$ (in the Δ-metric). Section 4 is devoted to characterizing the limit experiment in terms of properties of the given Markov chain. It turns out that if the chain is aperiodic, then $\mathcal{E}_\infty$ is simply the experiment of observing the recurrent class into which the chain ends up. We note that the limit $\mathcal{E}_\infty$ exists, even though the sequence $\{P^n\}$ diverges.

In section 5 we show how the Δ-distance provides us with a measure of the information contained in X_n. The quantity studied is $\Delta_n = \Delta(\mathcal{E}_n, \mathcal{E}_\infty)$. It is shown that Δ_n asymptotically behaves like $n^{\tau-1}\rho^n$, where

$$\rho = \max \{ |\lambda| : \lambda \text{ eigenvalue of } P \,,\ |\lambda| < 1 \}$$

and τ is an integer, $\tau \geq 1$.
The computations needed are based on a formula of P^n derived from a result on linear operators in finite dimensional spaces taken from Dunford-Schwarz [2] . The formula involves the spectrum of P and is given in section 2.

Section 8 treats the subject of lumping states of a Markov chain. Kemeny and Snell [6] have defined the concept lumpability of a Markov chain. In this paper we study how information gets lost by lumping. The efficiency of lumping may be measured both by deficiency and insufficiency, a concept introduced by Le Cam in [8].

Some results on insufficiency are given in section 7 for finite experiments. We note that the given results are valid for quite more general experiments. The given proofs, however, turn out to be very "clean" and may illustrate the technique of proving the general results.

It is proved in section 8 that in the present case, the deficiency and the insufficiency behave asymptotically in the same manner. As is pointed out by Le Cam in [8], this may not always be the case.

Sections 1-6 corresponds to Lindqvist [10]. The theory is, however, rewritten and some proofs are simplified.

2. Preliminaries.

If $A = (a_{ij})$ is a $m \times n$-matrix with complex entries, we shall define the norm $\|A\|$ of A by

$$\|A\| = \max_i \sum_j |a_{ij}| .$$

We note that convergence in this norm is equivalent to convergence in each entry.

A $m \times n$-matrix $M = (m_{ij})$ is called a Markov matrix if

$$m_{ij} \geq 0 \text{ for all } i,j \text{ and } \sum_{j=1}^{n} m_{ij} = 1 \text{ for all } i .$$

The set of Markov-matrices with dimension $m \times n$ is denoted $\mathcal{M}_{m,n}$. The set $\mathcal{M}_{m,n}$ is compact in the metric space of $m \times n$-matrices with metric induced from $\|\cdot\|$.

We shall need a suitable representation of powers of matrices. A complex $r \times r$-matrix T may be considered as a linear operator in C^r with scalar field C. Let $\lambda_1,\ldots,\lambda_s$ be the distinct eigenvalues of T. If λ is an eigenvalue, we define the index $\nu(\lambda)$ of λ to be the smallest non-negative integer ν such that $(\lambda I-T)^{\nu}x = 0$ for every vector x such that $(\lambda I-T)^{\nu+1}x = 0$. We shall put $m_i = \nu(\lambda_i)$; $i=1,\ldots,s$. It is seen that $m_i \geq 1$ for all i. The index of an eigenvalue may alternatively be defined as follows: Let $\Psi(t)$ be the minimal polynomial of T, i.e. $\Psi(t)$ is a complex polynomial with leading coefficient 1 and of lowest degree such that $\Psi(T) = 0$. It may be shown (see Gantmacher [4]) that we may write

$$\Psi(\lambda) = (\lambda-\lambda_1)^{m_1}(\lambda-\lambda_2)^{m_2},\ldots,(\lambda-\lambda_s)^{m_s} .$$

Hence the index of λ_i is the multiplicity of λ_i as a root of the minimal polynomial of T.

Theorem 8 of ch. VII in Dunford-Schwarz [2] now states that if f is a polynomial with complex coefficients, then we have the expansion

$$f(T) = \sum_{i=1}^{s} \sum_{j=o}^{m_i-1} \frac{(T-\lambda_i I)^j}{j!} f^{(j)}(\lambda_i) Z_i \qquad (2.1)$$

where $Z_1,\ldots,Z_s$ are complex $r \times r$-matrices which are independent of f and satisfy

(i) $Z_i \neq 0$; $i=1,\ldots,s$

(ii) $Z_i^2 = Z_i$; $i=1,\ldots,s$

(iii) $Z_i Z_k = 0$; for $i \neq k$

(iv) $\sum_{i=1}^{s} Z_i = I$.

If T is given, then the component matrices $Z_1,\ldots,Z_s$ may easily be determined, by inserting suitable $f(t)$ in (2.1). In fact, for any $i=1,\ldots,s$ there exists a polynomial $f_i(t)$ such that $Z_i = f_i(T)$. The idea of the construction is to define $f_i(t)$ so that $f_i^{(j)}(\lambda_k) = 0$ for $k \neq i$, $j=0,\ldots,m_i-1$. Putting $f(t) = t^n$ in (2.1) we obtain an expansion of T^n , which is the fundamental formula used in this paper

$$T^n = \sum_{i=1}^{s} \sum_{j=o}^{m_i-1} \frac{(T-\lambda_i I)^j}{j!} n^{(j)} \lambda_i^{n-j} Z_i \qquad (2.2)$$

where $n^{(r)} = n(n-1)\ldots(n-r+1)$ if $r \geq 1$ and $n^{(o)} = 1$.

A finite experiment will here be defined as an experiment with finite parameter set $\Theta = \{1,2,\ldots,r\}$ and finite sample space $\chi = \{1,2,\ldots,s\}$. Such an experiment is completely given by a matrix $P = (p_{\theta j}) \in \mathcal{M}_{r,s}$; the θ-th row vector defining the probability distribution over χ when θ is the "true" parameter. We shall use the symbol $\mathcal{E}_P$ for the experiment defined by P.

In [7] Le Cam introduced an information distance between experiments: the deficiency. The deficiency is defined in terms of risk-functions. It follows, however, from [7] that for finite experiments an equivalent definition may be given as follows:

Let $\mathcal{E}_P$ and $\mathcal{E}_Q$ be experiments such that $P \in \mathcal{M}_{r,s}$, $Q \in \mathcal{M}_{r,t}$. Then the deficiency of $\mathcal{E}_P$ relative to $\mathcal{E}_Q$ is given by

$$\delta(\mathcal{E}_P, \mathcal{E}_Q) = \inf\{\|PM-Q\| : M \in \mathcal{M}_{s,t}\} .$$

The quantity $\delta(\mathcal{E}_P, \mathcal{E}_Q)$ is interpreted as the maximal loss of information by observing $\mathcal{E}_P$ instead of $\mathcal{E}_Q$.

It follows from the above definition that the deficiency equals 0 if and only if $\mathcal{E}_P$ is sufficient for $\mathcal{E}_Q$ according to Blackwell's definition (see [1]). if $\delta(\mathcal{E}_P, \mathcal{E}_Q) = 0$, then we say that $\mathcal{E}_P$ is more informative than $\mathcal{E}_Q$. We write this $\mathcal{E}_P \geq \mathcal{E}_Q$. Accordingly, $\mathcal{E}_P \geq \mathcal{E}_Q$ if and only if there exists $M \in \mathcal{M}_{s,t}$ such that $PM = Q$. This result will be referred to as the randomization criterion.

Le Cam [7] also introduced a distance in the set of experiments It is defined by

$$\Delta(\mathcal{E}_P, \mathcal{E}_Q) = \delta(\mathcal{E}_P, \mathcal{E}_Q) \vee \delta(\mathcal{E}_Q, \mathcal{E}_P)$$

($\vee$ denotes maximum) and is easily seen to have the properties of a semi-metric. If $\Delta(\mathcal{E}_P, \mathcal{E}_Q) = 0$, then $\mathcal{E}_P$ and $\mathcal{E}_Q$ are said to be equivalent ; in symbols $\mathcal{E}_P \sim \mathcal{E}_Q$.

Let $\{\mathcal{E}^{(n)}\}_{n=1}^{\infty}$ be a sequence of experiments and let $\mathcal{E}$ be an experiment, all with the same parameter set.

Then we say that $\mathcal{E}^{(n)}$ converges to $\mathcal{E}$, in symbols $\mathcal{E}^{(n)} \to \mathcal{E}$ as $n \to \infty$ if $\Delta(\mathcal{E}^{(n)}, \mathcal{E}) \to 0$ as $n \to \infty$. The limit $\mathcal{E}$ is uniquely determined up to Δ-equivalence. In the case of finite parameter set, to each sequence $\{\mathcal{E}^{(n)}\}$ of experiments there exists a sub-sequence $\{\mathcal{E}^{(n')}\}$ and an experiment $\mathcal{E}$ such that $\mathcal{E}^{(n')} \to \mathcal{E}$.

This follows from a remark on p. 228 in Torgersen [11], which states that the set of (standard) experiments is compact in the metric Δ. As a consequence of this we have:

Lemma 2.1.

Let $\{\mathcal{E}^{(n)}\}$ be a sequence of experiments with common finite parameter set Θ. Assume that $\mathcal{E}^{(1)} \geq \mathcal{E}^{(2)} \geq \ldots$. Then there is an experiment $\mathcal{E}$ such that

(i) $\Delta(\mathcal{E}^{(n)}, \mathcal{E}) \downarrow 0$ as $n \to \infty$.

(ii) $\mathcal{E}^{(n)} \geq \mathcal{E}$ for each n.

Proof: There is a subsequence $\{\mathcal{E}^{(n')}\}$ and an $\mathcal{E}$ such that

$$\Delta(\mathcal{E}^{(n')}, \mathcal{E}) \to 0 \qquad (2.3)$$

Given n, choose $n' > n$. Then since $\mathcal{E}^{(n)} \geq \mathcal{E}^{(n')}$,

$$\delta(\mathcal{E}^{(n)}, \mathcal{E}) \leq \delta(\mathcal{E}^{(n)}, \mathcal{E}^{(n')}) + \delta(\mathcal{E}^{(n')}, \mathcal{E}) = \delta(\mathcal{E}^{(n')}, \mathcal{E}) .$$

Letting $n' \to \infty$ it is seen that $\delta(\mathcal{E}^{(n)}, \mathcal{E}) = 0$ and hence $\mathcal{E}^{(n)} \geq \mathcal{E}$.

It follows that

$$\begin{aligned} \Delta(\mathcal{E}^{(n+1)}, \mathcal{E}) &= \delta(\mathcal{E}, \mathcal{E}^{(n+1)}) \\ &\leq \delta(\mathcal{E}, \mathcal{E}^{(n)}) + \delta(\mathcal{E}^{(n)}, \mathcal{E}^{(n+1)}) \\ &= \delta(\mathcal{E}, \mathcal{E}^{(n)}) = \Delta(\mathcal{E}^{(n)}, \mathcal{E}) . \end{aligned}$$

Hence the sequence $\{\Delta(\mathcal{E}^{(n)}, \mathcal{E})\}$ is monotonically decreasing and thus convergent.

By (2.3) we must have $\Delta(\mathcal{E}^{(n)}, \mathcal{E}) \downarrow 0$.

If the $\mathcal{E}^{(n)}$ are <u>finite</u> experiments, then under some conditions, the limit experiment will also be a finite experiment.

<u>Lemma 2.2.</u>

Let $\{\mathcal{E}^{(n)}\}$ be a sequence of finite experiments which may be defined by matrices $P^{(n)} \in \mathcal{M}_{r,s_n}$.

Assume that the sequence $\{s_n\}$ is bounded. If $\{\mathcal{E}^{(n)}\}$ converges, then the limit experiment $\mathcal{E}$ is a finite experiment (i.e. may be given by a finite Markov matrix).

<u>Proof.</u> Let $s = \max s_n$. Extend the matrices $P^{(n)}$ to $r \times s$-matrices $Q^{(n)}$ by adding colums of zeros. Of course

$$\mathcal{E}_{Q(n)} \sim \mathcal{E}_{P(n)} .$$

By compactness of $\mathcal{M}_{r,s}$, there is a subsequence $Q^{(n')}$ which converges to a matrix $Q \in \mathcal{M}_{r,s}$. It remains to prove that $\mathcal{E}_{Q(n)} \to \mathcal{E}_Q$. Clearly $\Delta(\mathcal{E}_{Q(n)}, \mathcal{E}_Q) \leq \|Q^{(n)} - Q\|$. Hence $\mathcal{E}_{Q(n')} \to \mathcal{E}_Q$. By uniqueness of limits of experiments, this implies $\mathcal{E} \sim \mathcal{E}_Q$.

Some other notations which will be used are:

If $B_1, \ldots, B_n$ are quadratic matrices, then we denote by $\operatorname{diag}(B_1, \ldots, B_n)$ the matrix

$$\begin{pmatrix} B_1 & 0 & \ldots & 0 \\ 0 & B_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \\ 0 & \ldots & \ldots & B_n \end{pmatrix}$$

$\mathcal{M}_o$ is the set of $r \times r$ Markov matrices $M = (m_{ij})$ with identical row vectors, i.e. there exists $m_1, \ldots, m_r \geq 0$, $\Sigma m_j = 1$ such that $m_{ij} = m_j$ for all i,j.

$\mathcal{M}_{.o}$ is the set of real $r \times r$-matrices $N = (n_{ij})$ with identical rows and with row sums equal to 0, i.e. $\sum_{j=1}^{r} n_{ij} = 0$.

3. The model.

We shall in this paper consider a finite, discrete Markov chain $\{X_n\}_{n=0}^{\infty}$ with state space $\chi = \{1,2,\ldots,r\}$ and transition matrix $P = (p_{ij})$. The elements of P^n will be denoted $p_{ij}^{(n)}$, so that for $i,j = 1,2,\ldots,r$

$$p_{ij}^{(n)} = \Pr(X_n=j \mid X_0=i) \qquad (3.1).$$

We let $\theta = X_0$ be the parameter and let our experiment consist in observing X_n. Thus the parameter set $\Theta = \chi$. Since by (3.1) the θ-th row vector of P^n defines the probability distribution of X_n given that $X_0 = \theta$, it follows that <u>the experiment of observing</u> X_n is given by P^n. We shall denote if by $\mathcal{E}_n$.

It seems reasonable that the information obtained about X_0 by observing X_n decreases as n grows. In fact, if $0 < m \leq n$ then $P^n = P^m P^{n-m}$ and the randomization criterion shows that $\mathcal{E}_m \geq \mathcal{E}_n$. Hence $\mathcal{E}_1 \geq \mathcal{E}_2 \geq \ldots$ and by lemma 2.1 there exists an experiment $\mathcal{E}_\infty$ such that $\Delta(\mathcal{E}_n, \mathcal{E}_\infty) \downarrow 0$ and $\mathcal{E}_n \geq \mathcal{E}_\infty$. We remark that the limit experiment $\mathcal{E}_\infty$ exists for any P. Hence the experiment given by P^n will converge even if P^n does not converge. In section 4 we shall characterize and give an interpretation of the experiment $\mathcal{E}_\infty$ in terms of the properties of the given Markov chain.

At first sight one may think that the information about X_0 may be increased by observing the chain on times $n_1 < n_2 < \ldots < n_k$ instead of merely observing at n_1. However, by the Markov property, the conditional distribution of $X_{n_1},\ldots,X_{n_k}$ given X_{n_1} is independent of θ, so X_{n_1} is sufficient for the vector

$(X_{n_1},\ldots,X_{n_k})$. A question that may arise concerning the given model is : Let P and Q be transition matrices of the same dimension r . Let $\{\mathcal{E}_n\}$ and $\{\mathcal{F}_n\}$ be the corresponding sequences of experiments according to the model given here. Assume that P and Q represent equivalent experiments, i.e. $\mathcal{E}_1 \sim \mathcal{F}_1$. Will this imply that $\mathcal{E}_n \sim \mathcal{F}_n$ for all n ? That this is not true is seen from the following simple example :

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \quad , \qquad Q = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} \tag{3.2}$$

$$P^2 = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad , \qquad Q^2 = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ 0 & 1 \end{pmatrix} .$$

Clearly $\mathcal{E}_1 \sim \mathcal{F}_1$. That $\mathcal{E}_2 \not\sim \mathcal{F}_2$ is seen from the ψ-criterion (see Torgersen [11] , for computation see e.g. Lindqvist and Torgersen [9]).

Defining $\psi(x_1,x_2) = |x_1 - 2x_2|$ yields

$$\psi(\mathcal{E}_2) = 1 \neq \tfrac{3}{2} = \psi(\mathcal{F}_2) .$$

We remark that P and Q , given in (3.2) defining equivalent experiments, have quite different properties as transition matrices for Markov chains. P defines an aperiodic irreducible chain, whence Q defines a chain where state 2 is absorbing.

Example 3.1.

This example is taken from Feller [3] . We consider a particular pair of genes, say a and A . Each individual belongs to one of the three genotypes (a,a) , (a,A) or (A,A) . Our population consists of N individuals. However, we shall introduce the genes themselves as elements of the population, so that we deal with a population of 2N elements which are either a or A . Here N is a fixed number.

Under the assumption of random mating we may assume that the 2N genes of any generation are formed in 2N independent trials : if the parent population consists of j a-genes and 2N-j A-genes, then each trial results in a or A with probabilities

$$p_j = j/2N \quad , \quad q_j = 1-p_j$$

respectively.

Hence we have a Markov chain with state space $\chi = \{0,1,\ldots,2N\}$ and transition probabilites given by the binomial distribution :

$$p_{jk} = \binom{2N}{k} p_j^{\,k} \, q_j^{\,2N-k} \, .$$

The states 0 and 2N are absorbing, i.e. $\{0\}$ and $\{2N\}$ are recurrent classes, whence $1,2,\ldots,2N-1$ are transient states.

In genetics one is often faced with the following problem :

A process of the above type is observed after n generations (we assume that the number of a- and A-genes may be counted by some method), and one wants to draw conclusions about the initial population. This is a special case of the model treated in this paper. By the previous results, the information about the initial generation decreases as n grows.

4. The limit experiment $\mathcal{E}_\infty$:

Let the sequence $\{\mathcal{E}_n\}$ be given as in section 3 . As is noted there, for any transition matrix P , the corresponding sequence $\{\mathcal{E}_n\}$ converges to some experiment $\mathcal{E}_\infty$. By lemma 2.2 , $\mathcal{E}_\infty$ is a finite experiment. It is the purpose of this section to determine the Markov matrix defining $\mathcal{E}_\infty$. The results might have been developed using the expansion (2.2) together with some facts on eigenvalues of transition matrices. However, we shall make a probabilistic approach, using merely elementary properties of Markov chains. The following lemma will be useful:

Lemma 4.1.

Assume that Q is a limit at some subsequence $\{P^{n'}\}$ of $\{P^n\}$. Then $\mathcal{E}_\infty$ may be represented by Q . In particular, if P^n converges, then $\mathcal{E}_\infty$ may be represented by the limiting matrix.

Proof: Analogous to the proof of lemma 2.2

The totality of states of a Markov chain may be partitioned into equivalence classes, where the states in an equivalence class are those which communicate with each other. A class of states is called recurrent if the probability of leaving the class is 0 ; it is otherwise called transient. A Markov chain is called irreducible if there is only one equivalence class (which is then recurrent if the state space is finite). The period of a state is a constant in each equivalence class, so it makes sense to deal with the concept period of a class. A class is called aperiodic

if each state has period 1.

Given a Markov chain with state space $\chi = \{1,2,\ldots,r\}$ we may (eventually by rearranging the states) write P on the "canonical" form

$$P = \left(\begin{array}{cccc} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & P_v & 0 \\ \hline R & & & Q \end{array}\right) \qquad (4.1)$$

where $P_1,\ldots,P_v$ represent irreducible Markov chains, corresponding to the v recurrent classes of the chain.

If the chain has no transient states, then R and Q may be removed from (4.1). Otherwise, $R \neq 0$ and Q is a quadratic non-negative matrix (which is not a Markov matrix).

It is seen that

$$P^n = \left(\begin{array}{cccc} P_1^n & 0 & \ldots & 0 \\ 0 & P_2^n & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & P_v^n & 0 \\ \hline R_n & & & Q^n \end{array}\right) \qquad (4.2)$$

We shall at the outset make the assumption that the v recurrent classes are all aperiodic. Then by the usual limit theorems of irreducible aperiodic chains, for each k, P_k^n converges to a Markov matrix A_k such that all rows of A_k are equal.

Since there is probability 0 of remaining in some transient state, we conclude that $Q^n \to 0$.

It remains to study the properties of R_n as n tends to infinity. For simplicity, let the transient states be labled $1,2,\ldots,\beta$ and let q_{jk} be the probability of being absorbed in the k'th recurrent class, given that one starts in state j .

For each j , $\sum_{k=1}^{v} q_{jk} = 1$.

Let the row vector of A_k be $\bar{a}_k$. By the limit theorem for transition probabilities (see Karlin [5] thm. 3.1) it follows that R_n converges and that the row vector in $\lim R_n$ corresponding to transient state j may be written

$$(\bar{b}_{j1} , \bar{b}_{j2},\ldots,\bar{b}_{jv})$$

where $\bar{b}_{jk} = q_{jk}\bar{a}_k$.

We have now shown that under the given conditions of aperiodicity, P^n converges to a matrix B of the form

$$B = \left(\begin{array}{cccc} A_1 & 0 \ldots & 0 & 0 \\ 0 & A_2 \ldots & 0 & 0 \\ \vdots & & & \vdots \\ 0 & & A_v & 0 \\ \hline \bar{b}_{11} & \bar{b}_{12}\ldots & \bar{b}_{1v} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \bar{b}_{\beta 1} & \bar{b}_{\beta 2}\ldots & \bar{b}_{\beta v} & 0 \end{array}\right) \qquad (4.3)$$

By lemma 4.1 $\mathcal{E}_n \to \mathcal{E}_B$. The limit experiment $\mathcal{E}_B$ may be reduced to a minimal sufficient form:

Theorem 4.2.

Let P be given as in (4.1) and assume that $P_1,\dots,P_v$ all represent aperiodic classes of states. Let q_{jk} be given as before. Then the limit experiment $\mathcal{E}_\infty$ is given by the $r\times v$-matrix

$$A = \begin{pmatrix} I_1 & 0 & \dots & 0 \\ 0 & I_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & I_v \\ q_{11} & q_{12} & \dots & q_{1v} \\ \vdots & & & \vdots \\ q_{\beta 1} & q_{\beta 2} & \dots & q_{\beta v} \end{pmatrix}$$

where for $k=1,\dots,v$, I_k is a column vector of the same dimension as P_k, with all entries equal to 1.

Proof: We have to prove that $\mathcal{E}_A \sim \mathcal{E}_B$. The relation $\mathcal{E}_B \geq \mathcal{E}_A$ is established by observing that $BA = A$.
Define now a matrix $M \in \mathcal{M}_{v,r}$ by

$$M = \begin{pmatrix} \bar{a}_1 & 0 & \dots & 0 \\ 0 & \bar{a}_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & \dots & \dots & \bar{a}_v \end{pmatrix}$$

A simple computation shows that $AM = B$, which implies $\mathcal{E}_A \geq \mathcal{E}_B$.

Corollary 4.3.

If the Markov chain is irreducible and aperiodic, then $\{\mathcal{E}_n\}$ converges to the minimal informative experiment.

Remark: The minimal informative experiment is characterized by the fact that the probability distribution does not depend on the parameter θ. Any Markov matrix with equal rows will represent the minimal informative experiment.

The experiment $\mathcal{E}_A$ given in theorem 4.2 may be given a reasonable interpretation as follows : Let P be given as in the theorem. We note that absorbtion into one of the v recurrent classes occurs with probability 1. Thus we may define an experiment consisting in observing (the label of) the absorbing class. It is not difficult to see that the probability distribution of this experiment is exactly the one given by the matrix A.

We shall now discuss periodic chains. At first we assume that P is the transition matrix of an irreducible, aperiodic chain with period $d > 1$. P may then (eventually by permuting the states) be represented on the form

$$P = \begin{pmatrix} 0 & P_1 & 0 \dots & 0 \\ 0 & 0 & P_2 \dots & 0 \\ \vdots & & & \\ 0 & 0 & 0 \dots & P_{d-1} \\ P_d & 0 & 0 \dots & 0 \end{pmatrix}$$

where the matrix in block no. (i,j) has dimension

$\alpha_i \times \alpha_j$; $i,j = 1,\ldots,d$; $\alpha_i > 0$ for all i and $\sum_{i=1}^{d} \alpha_i = r$.

By induction (or by a probabilistic line of argument) we show that

$$P^d = \text{diag}\,(Q_1,\ldots,Q_d) \tag{4.4}$$

where Q_i has dimension $\alpha_i \times \alpha_i$

and

$$Q_1 = P_1 P_2 \ldots P_d$$

$$Q_2 = P_2 P_3 \ldots P_d P_1$$

$$\vdots$$

$$Q_d = P_d P_1 P_2 \ldots P_{d-1}$$

It is easily verified that $Q_1,\ldots,Q_d$ correspond to irreducible, aperiodic chains, so (4.4) shows that P^d is of the form (4.1) with $R=0$, $Q=0$. Hence $\{P^{nd}\}_{n=1}^{\infty}$ converges to a matrix of the form (4.3) with no transient states and $v=d$. Furthermore, by lemma 4.1 and theorem 4.2 $\{\mathcal{E}_n\}$ converges to the experiment given by the matrix

$$\begin{pmatrix} I_1 & 0 & \ldots & 0 \\ 0 & I_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \ldots\ldots & & I_d \end{pmatrix}$$

where I_i is a column vector of all 1's with dimension α_i ; $i=1,\ldots,d$.

We remark at once that the limit experiment in the case of an irreducible chain with period d has the same form as the limit

experiment of a chain with d recurrent classes (the recurrent classes correspond to the cyclic classes in the periodic case). This may be explained as follows:

In the former case we may by observing X_n, since n is known, draw the conclusion as to which cyclic class X_o belongs. This corresponds to, in the latter case, knowing that X_o belongs to the same recurrent class as X_n.

We now return to the general case where P is given as in (4.1) and P_i represents a periodic chain with period $d_i \geq 1$; $i=1,\ldots,v$. By (4.2) and the remarks about irreducible periodic chains we may write (possibly by permuting states in each recurrent class)

$$P^d = \left(\begin{array}{cccc|c} P_1' & 0 & \cdots & 0 & 0 \\ 0 & P_2' & & 0 & \vdots \\ \vdots & & & & \vdots \\ 0 & \cdots\cdots & & P_\gamma' & 0 \\ \hline R_d & & & & Q^d \end{array}\right) \qquad (4.5)$$

with d equal to the least common multiple (l.c.m.) of $d_1,\ldots,d_v$ (i.e. d is the least positive integer so that $d_1,\ldots d_v$ all divide d)

and $\gamma = \sum_{i=1}^{v} d_i$.

Now $P_1',\ldots,P_\gamma'$ all represent irreducible, aperiodic chains, so it is seen that $\{P^{nd}\}_{n=1}^{\infty}$ converges to a matrix of the form B given in (4.3) .

Thus by lemma 4.1 we have (this is the main result of this section):

Theorem 4.4.

Consider a Markov chain with state space $\chi = \{1,2,\ldots,r\}$. We assume that the transition matrix P is given by (4.1), where P_i has period $d_i \geq 1$ and P_i is written on cyclic form

$$P_i = \begin{pmatrix} 0 & P_{i1} & 0 & \ldots & 0 \\ 0 & 0 & P_{i2} & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & P_{i,d_i-1} \\ P_{i,d_i} & 0 & 0 & \ldots & 0 \end{pmatrix}$$

where the matrix in block no. (k,l) has dimension $\alpha_{i,k} \times \alpha_{i,l}$ $(k,l=1,\ldots,d_i)$ $(i=1,\ldots,v)$.

We assume that Q has dimension $\beta\times\beta$, i.e. that the chain has β transient states (β may be 0).

Let d be the l.c.m. of $d_1,\ldots,d_v$. Then P^d (see (4.5)) is the transition matrix of a Markov chain with

$\gamma = \sum_{i=1}^{v} d_i$ recurrent classes and β transient states.

For $j=1,\ldots,\beta$; $k=1,\ldots,\gamma$ let q_{jk} be the probability of being absorbed in the k-th recurrent class of this chain, given that one starts in state j .

The experiment sequence $\{\mathcal{E}_n\}$ defined by P now converges to the experiment $\mathcal{E}_A$ given by the matrix

$$A = \begin{pmatrix} I^{\alpha_{11}} & & & & & & 0 \\ 0 & I^{\alpha_{12}} & & & & & \vdots \\ 0 & & \ddots & & & & \\ \vdots & & & I^{\alpha_{1d_1}} & & & \\ & & & & \ddots & & \\ & & & & & I^{\alpha_{v1}} & \\ & & & & & & I^{\alpha_{v2}} \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & I^{\alpha_{vd_v}} \\ q_{11} & q_{12} \cdots & q_{1d_1} & \cdots & \cdots & \cdots & q_{1\gamma} \\ q_{\beta 1} & q_{\beta 2} \cdots & q_{\beta d_1} & \cdots & \cdots & \cdots & q_{\beta\gamma} \end{pmatrix} \qquad (4.6)$$

where I^h denotes a column vector of dimension h with all entries equal to 1.

Conversely, to each matrix A of the above form (with $q_{jk} \geq 0$, $\sum_{k=1}^{\gamma} q_{jk} = 1$) there is a transtition matrix P such that $\{\mathcal{E}_n\}$ corresponding to this P converges to $\mathcal{E}_A$.

Proof: The last assertion is proved by noting that for any A on the form (4.6) $\mathcal{E}_A$ is equivalent to an experiment $\mathcal{E}_B$ with B on the form given in (4.3).

Now B is a transition matrix and it is seen that $B^n = B$ for $n=1,\ldots,$. Hence $\mathcal{E}_1 \sim \mathcal{E}_2 \sim \ldots \sim \mathcal{E}_B$ so $\mathcal{E}_n \to \mathcal{E}_B$.

Remark: Since the states of a Markov chain may be permuted, any matrix which by a permutation of rows becomes a matrix of the form (4.6) , is a limit experiment of a sequence $\{\mathcal{E}_n\}$. The representation of the limit experiment $\mathcal{E}_\infty$ given by (4.6) is minimal sufficient since an addition of columns in A and a succeeding multiplication with a Markov matrix would not lead us back to A . Hence the given representation is minimal in the sense that any other minimal sufficient representation is given by a permutation of columns in A .

What can be said about the original Markov chain by examining the limit experiment ?

By the discussion in this section it follows that the limit experiment depends on the recurrent classes, the periods and the transient states. Given the limit experiment, however, it turns out that to select say the recurrent classes, we need additional information about the chain. As an example, suppose $r=3$ and $A = (1,1,1)'$.
Then $\{1,2\}$ is recurrent and $\{3\}$ transient, or $\{1\}$ is recurrent and $\{2,3\}$ transient and so on.

If $A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$, then

$\{1,2\}$ and $\{3\}$ may be two recurrent classes, or $\{1,2,3\}$ may be a recurrent class with period 2 etc.

We close this section by taking up the example given in section 3 .

Example 4.5.

By Feller [3] , the probability ϵ_m of being absorbed into state 0 given that the initial population has m a-genes, is given by

$$\epsilon_m = 1 - m/2N \quad ; \quad m=0,1,\ldots,2N \ .$$

Hence the limit experiment is defined by

$$A = \begin{pmatrix} 1 & 0 \\ 1-\frac{1}{2N} & \frac{1}{2N} \\ 1-\frac{2}{2N} & \frac{2}{2N} \\ \vdots & \\ 1-\frac{2N-1}{2N} & \frac{2N-1}{2N} \\ 0 & 1 \end{pmatrix} .$$

5. Rate of convergence.

In this section we set $\Delta_n = \Delta(\mathcal{E}_n, \mathcal{E}_\infty)$.

Assume for a moment that the Markov chain under consideration is irreducible and aperiodic. By corollary 4.3 , the limit experiment $\mathcal{E}_\infty$ is the least informative experiment. Intuitively speaking, a decision based on this experiment may be considered as the guessing of a parameter without taking observations. Hence Δ_n measures the maximum loss when "guessing" X_o instead of basing a decision about X_o on X_n . It may thus seem reasonable to consider Δ_n as a measure of the information contained in X_n .

If the chain is not necessarily irreducible and aperiodic, then Δ_n may be given a similar interpretation. More precisely, Δ_n measures the additional information of observing X_n instead of merely observing the recurrent class into which the chain ends up.

We shall make use of the expansion of P^n given in (2.2) . To begin with, we state some results on eigenvalues of non-negative matrices and in particular of transition matrices. (See e.g. Karlin [5]) . It is tacitly understood that every matrix is of finite dimension.

Theorem 5.1.

(i) Any non-negative matrix $A \neq 0$ has a real eigenvalue $\mu > 0$ such that $|\lambda| \leq \mu$ for any eigenvalue λ of A .

(ii) If P is a transition matrix, then $\mu = 1$ and the multiplicity of the eigenvalue 1 is equal to the number of recurrent classes associated with P .

(iii) If P is the transition matrix of an irreducible periodic Markov chain with period d , then the d'th roots of unity are eigenvalues of P , each with multiplicity one, and there are no

other eigenvalues of modulus 1 .

We remark that multiplicity here means multiplicity as a root of the characteristic polynomial, whence the formula (2.2) involves the multiplicity as a root of the minimal polynomial (which may be smaller) .

Let P be given as in theorem 4.4. The set of eigenvalues of P is equal to the collection of eigenvalues of $P_1,\dots,P_v, Q$. Since the chain associated with P has v recurrent classes, it follows from theorem 5.1 that the maximum eigenvalue of Q is $\mu < 1$ and that any eigenvalue of P with modulus 1 is a root of unity.

We number the distinct eigenvalues of modulus 1 :

$$\varphi_1 = 1 , \varphi_2,\dots,\varphi_u .$$

Let d be given as in theorem 4.4. Then $\varphi_i^d = 1$ for $i=1,2,..,u$. Since P^n is bounded as $n \to \infty$, (2.2) shows that the index of $\varphi_1,\dots,\varphi_u$ are each equal to 1 . (This may also be concluded from theorem 9, ch. VII of [2]). Hence, if $\lambda_1,\dots,\lambda_s$ are the eigenvalues of P with modulus less than 1 , and with indexes $m_1,\dots,m_s$, respectively, then by (2.2)

$$P^n = \sum_{i=1}^{u} \varphi_i^n Z_i + \sum_{i=1}^{s} \sum_{j=0}^{m_i-1} \frac{(P-\lambda_i I)^j}{j!} n^{(j)} \lambda_i^{n-j} C_i \qquad (5.1) .$$

As is proved in section 4 by a probabilistic consideration, (5.1) shows that P^{md} converges to $\sum_{i=1}^{u} Z_i$ as $m \to \infty$.

More general,

$$P^{md+k} \to \sum_{i=1}^{u} \varphi_i^k Z_i \quad \text{as} \quad m \to \infty$$

for $k=0,1,\ldots,d-1$.

We are now in position to derive upper and lower bounds for Δ_n .

Definition 5.2.

If P is a transition matrix, then we define the root of P by

$$\rho = \max \{|\lambda| : \lambda \text{ eigenvalue of } P,\ |\lambda| < 1\} .$$

The maximal index τ of eigenvalues with modulus ρ is called the index of P .

In the following theorems we shall assume that $\rho > 0$. If $\rho = 0$, then by (5.1) from some n on $P^n = \sum_{i=1}^{u} \varphi_i^n Z_i$ so $\Delta_n \equiv 0$, i.e. $\mathcal{E}_n \sim \mathcal{E}_\infty$.

Theorem 5.3.

Let ρ and τ be, respectively, the root and index of P . Then there are constants $0 < k \leq K < \infty$ such that for any $n=1,2,\ldots$

$$k n^{\tau-1} \rho^n \leq \Delta_n \leq K n^{\tau-1} \rho^n .$$

Proof: Let n be given. By the Euclidean algorithm there exist non-negative integers m,k such that $n=md+k$ and $0 \leq k < d$. Since $\mathcal{E}_\infty$ by lemma 4.1 may be represented by

$\sum_{i=1}^{u} \varphi_i^j Z_i$, it follows from (5.1) that

$$\Delta_n = \inf_M \| \sum_{i=1}^{u} \varphi_i^k Z_i M - P^n \|$$

$$\leq \| \sum_{i=1}^{u} \varphi_i^{md+k} Z_i - P^n \|$$

$$\leq \sum_{i=1}^{s} \sum_{j=0}^{m_i-1} n^{(j)} |\lambda_i|^{n-j} \| \frac{(P-\lambda_i)^j}{j!} C_i \|$$

$$\leq K n^{\tau-1} \rho^n .$$

Obviously K may be chosen independent of n.

It remains to prove the left inequality. Let n, m and k be given as before. Without loss of generality we may assume that $|\lambda_1| = \rho$ and that the index of λ_1 is τ. Then, applying to the inequality $\|AB\| \leq \|A\| \, \|B\|$, for any $M \in \mathfrak{M}_{r,r}$:

$$\| \sum_{i=1}^{u} \varphi_i^k Z_i M - P^n \| \geq \frac{1}{\|C_1\|} \| C_1 \sum_{i=1}^{u} \varphi_i^k Z_i M - C_1 P^n \|$$

$$= \frac{1}{\|C_1\|} \| C_1 P^n \| \qquad (5.2)$$

since $C_1 Z_i = 0$ for $i=1,\ldots,u$.

Since $C_1 = f_1(P)$ for some polynomial $f_1(t)$ (see section 2), we have

$$C_1 P^n = P^n C_1 = \sum_{j=0}^{\tau-1} \frac{(P-\lambda_1 I)^j}{j!} n^{(j)} \lambda_1^{n-j} C_1 .$$

The right hand side of (5.2) is independent of M so

$$\left.\begin{aligned} \frac{\Delta n}{n^{\tau-1}\rho^n} &\geq \frac{\|P^n C_1\|}{\|C_1\| n^{\tau-1}\rho^n} \\ &= \frac{1}{\|C_1\|} \Big\| \sum_{j=o}^{\tau-1} \frac{(P-\lambda_1 I)^j}{j!} \frac{n(j)}{n^{\tau-1}} \frac{\lambda_1^{n-j}}{\rho^n} C_1 \Big\| . \end{aligned}\right\} \quad (5.3)$$

The right hand side of (5.3) converges to

$$\frac{1}{\rho^{\tau-1}\|C_1\|} \Big\| \frac{(P-\lambda_1 I)^{\tau-1}}{(\tau-1)!} C_1 \Big\| > 0 .$$

But this assures the existence of $k > 0$ such that

$$\frac{\Delta_n}{n^{\tau-1}\rho^n} \geq k > 0 \quad \text{for all} \quad n .$$

Corollary 5.4.

$$\sqrt[n]{\Delta_n} \to \rho \quad \text{as} \quad n \to \infty .$$

The result of theorem 5.3 may be written

$$\Delta_n = k_n \, n^{\tau-1} \, \rho^n$$

with $k \leq k_n \leq K$ for any n.

What can be said about the value of k_n ? We shall derive an asymptotic expression for Δ_n in the case of an irreducible, aperiodic chain, and show that in some cases the value of k_n is independent of n.

Let $\mathcal{M}_o$ and $\mathcal{N}_o$ be defined as in section 2.

Theorem 5.5.

Let P represent an irreducible, aperiodic chain. Let the eigenvalues of P be $1,\lambda_1,\ldots,\lambda_s$. Assume that

$$1 > |\lambda_1| = |\lambda_2| = \ldots = |\lambda_p| = \rho > |\lambda_{p+1}| \geq \ldots \geq |\lambda_s|$$

and $m_1 = m_2 = \ldots = m_q = \tau > m_{q+1} \geq \ldots \geq m_p$ $(q \leq p)$.

Let $\lambda_j = \rho e^{i\varphi_j}$; $\varphi_j \in [0,2\pi]$; $j=1,\ldots,q$.

Define $T_n = \sum_{k=1}^{q} e^{i(n-\tau+1)\varphi_k} \frac{(P-\lambda_k I)^{\tau-1}}{(\tau-1)!} C_k$.

Then

$$\Delta_n = \inf_{N \in \mathcal{M}_o} \|N - T_n\| + o(n^{\tau-1}\rho^n) .$$

Proof: As is well known from the theory of Markov chains, P^n converges to a matrix $Z_1 \in \mathcal{M}_o$ such that all entries of Z_1 are strictly positive.

Since $Z_1 \in \mathcal{M}_o \Rightarrow Z_1 M \in \mathcal{M}_o$ for any square Markov matrix M, it is seen that

$$\Delta_n = \inf_M \|Z_1 M - P^n\| = \inf_{M \in \mathcal{M}_o} \|M - P^n\| .$$

Put $$R_n = \sum_{k=1}^{q} \frac{(P-\lambda_k I)^{\tau-1}}{(\tau-1)!} n^{(\tau-1)} \lambda_k^{n-\tau+1} C_k \qquad (5.4)$$

$$= n^{(\tau-1)} \rho^{n-\tau+1} T_n$$

and define Q_n so that

$$P^n = Z_1 + R_n + Q_n . \qquad (5.5)$$

Now

$$\Delta n = \inf_{M \in \mathcal{M}_o} \|M - Z_1 - R_n - Q_n\| .$$

Let $\mathcal{M}^*$ be the set of matrices $M - Z_1$ obtained as M varies in $\mathcal{M}_o$. Then

$$\Delta n = \inf_{N \in \mathcal{M}^*} \|N - R_n - Q_n\| .$$

By (5.1), (5.4) and (5.5),

$$\|Q_n\| = o(n^{\tau-1} \rho^n)$$

and it is thus seen that

$$\Delta_n = \inf_{N \in \mathcal{M}^*} \|N - R_n\| + o(n^{\tau-1} \rho^n)$$

$$= n^{(\tau-1)} \rho^{n-\tau+1} \inf_{N \in \mathcal{M}^*} \left\| \frac{N}{n^{(\tau-1)} \rho^{n-\tau+1}} - T_n \right\| \qquad (5.6)$$

Since $\mathcal{N}_o$ is a closed set in the set of r×r-matrices, there exists $N_o^{(n)} \in \mathcal{N}_o$ such that

$$\inf_{N \in \mathcal{N}_o} \|N - T_n\| = \|N_o^{(n)} - T_n\| .$$

We shall show that from a certain n on, the infimum in (5.6) is attained by

$$N^{(n)} = n^{(\tau-1)} \rho^{n-\tau+1} N_o^{(n)} . \qquad (5.7)$$

By definition of $N_o^{(n)}$ it suffices to prove that from some n on, $N^{(n)}$ given by (5.7) is a member of $\mathcal{N}^*$.

The matrices $N_o^{(n)}$ are uniformly bounded in n. In fact for any n

$$\|N_o^{(n)}\| \leq \|N_o^{(n)} - T_n\| + \|T_n\|$$

$$= \inf_{N \in \mathcal{N}_o} \|N - T_n\| + \|T_n\|$$

$$\leq 2\|T_n\| \qquad \text{by putting } N=0$$

$$\leq 2 \sum_{k=1}^{q} \left\| \frac{(P-\lambda_k I)^{\tau-1}}{(\tau-1)!} C_k \right\| .$$

Since all entries of Z_1 are strictly positive, it is seen that there exists $0 < \gamma < 1$ such that $\mathcal{N}^*$ contains any member of $\mathcal{N}_o$ with all entries of absolute value less than γ. Hence the fact that $n^{(\tau-1)} \rho^{n-\tau+1} \to 0$ as $n \to \infty$ implies that

$N^{(n)} \in \mathcal{M}^*$ from a certain n on.

The proof is now complete.

Corollary 5.6.

Assume that λ_1 is real (i.e. $\lambda_1 = \pm \rho$) and that $q=1$. Then there is a constant $c > 0$ such that

$$\frac{\Delta_n}{n^{\tau-1}\rho^n} \to c \quad \text{as} \quad n \to \infty .$$

Proof: By theorem 5.5

$$\Delta_n = k\, n^{(\tau-1)}\, \rho^{m-\tau+1} + o(n^{\tau-1}\rho^n)$$

with $$k = \inf_{N \in \mathcal{M}_0} \| N - \frac{(P-\lambda_1 I)^{\tau-1}}{(\tau-1)!}\, C_1 \| .$$

That $k > 0$ is a consequence of theorem 5.3.

The corollary follows.

If the Q_n defined in the proof of theorem 5.5 equals exactly 0 from a certain n on (this happens if $\tau=1$ and P has no non-zero eigenvalues with modulus less than ρ), then theorem 5.5 gives rise to exact expressions for Δ_n from a certain n on.

The proof of theorem 5.5 also shows that if $Q_n = 0$ and $N_0^{(n)} \in \mathcal{M}^*$ for $n=1,2,\ldots$ then we obtain exact expressions for Δ_n for $n=1,2,\ldots$. To give an example, we shall derive Δ_n for a general P of dimension 2×2.

Example 5.7.

Let $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$

where $0 \leq \alpha , \beta \leq 1$.

To avoid trivial chains, we shall assume that $\alpha+\beta \neq 0,1,2$.
The eigenvalues of P are 1 and $1-\alpha-\beta$ and

$$P^n = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} + \frac{(1-\alpha-\beta)^n}{\alpha+\beta}\begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}$$

i.e. by the terminology of theorem 5.5, $\rho = |1-\alpha-\beta|$, $\tau=1$,

$$C_1 = \frac{1}{\alpha+\beta}\begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}$$

and $\varphi_1 = 0$ or π according to whether $1-\alpha-\beta$ is > 0 or < 0 .
We shall let N_o be the $N \in \mathcal{M}_o$ minimizing $\|N - C_1\|$.

Putting $N = \begin{pmatrix} c & -c \\ c & -c \end{pmatrix}$ yields

$$\|N - C_1\| = 2\left|c - \frac{\alpha}{\alpha+\beta}\right| \vee \left|c + \frac{\beta}{\alpha+\beta}\right| .$$

A geometrical consideration shows that infimum taken over c of this expression is obtained with

$$c = \frac{\alpha-\beta}{2(\alpha+\beta)}$$

which yields

$$\|N_o = Z_1\| = 1 .$$

Furthermore $N_o \in \mathcal{N}^*$ so by theorem 5.5 and the remarks preceding this example we have

$$\Delta_n = \rho^n \quad \text{for} \quad n=1,2,\ldots .$$

Example 5.8. (Example 3.1 continued).
Feller has computed the eigenvalues of the transition matrix of the chain. In our notations we have

$$\rho = 1 - \frac{1}{2N} \quad \text{and} \quad \tau = 1 .$$

Furthermore, the requirements of corollary 5.6 are satisfied, so that

$$\Delta_n \sim c(1 - \frac{1}{2N})^n$$

for some constant $c > 0$.

To end this section, let $\{\mathcal{E}_n\}$ be a sequence of experiments determined by the matrix P , and let $\{\mathcal{F}_n\}$ be determined by Q . Assume further that $\{\mathcal{E}_n\}$ and $\{\mathcal{F}_n\}$ have a common limit experiment $\mathcal{G}$.

It is obvious that $\Delta(\mathcal{E}_n, \mathcal{F}_n) \to 0$, since $\Delta(\mathcal{E}_n, \mathcal{F}_n) \leq \Delta(\mathcal{E}_n, \mathcal{G}) + \Delta(\mathcal{F}_n, \mathcal{G})$.
We shall prove a result on the rate of convergence of $\Delta(\mathcal{E}_n, \mathcal{F}_n)$.

Let ρ , τ be the root and index of P ; σ , ν be the root and index of Q .

Then we have

Theorem 5.9.

If (i) $\rho > \sigma$

or (ii) $\rho = \sigma$ and $\tau > \nu$

then

$$\frac{\Delta(\mathcal{E}_n, \mathcal{F}_n)}{n^{\tau-1}\rho^n}$$ is bounded away from 0 and ∞ as $n \to \infty$.

In particular, $\sqrt[n]{\Delta(\mathcal{E}_n, \mathcal{F}_n)} \to \rho$.

If $\rho = \sigma$ and $\tau = \nu$, then the above results may not hold. In fact we may have

$$\sqrt[n]{\Delta(\mathcal{E}_n, \mathcal{F}_n)} \to \delta \quad \text{with} \quad 0 < \delta < \rho .$$

Remark: If $P = Q$, then of course $\Delta(\mathcal{E}_n, \mathcal{F}_n) = 0$ for all n.

Proof: Assume that (i) or (ii) holds.

Then

$$\frac{\Delta(\mathcal{E}_n, \mathcal{G}) - \Delta(\mathcal{F}_n, \mathcal{G})}{n^{\tau-1}\rho^n} \leq \frac{\Delta(\mathcal{E}_n, \mathcal{F}_n)}{n^{\tau-1}\rho^n} \leq \frac{\Delta(\mathcal{E}_n, \mathcal{G}) + \Delta(\mathcal{F}_n, \mathcal{G})}{n^{\tau-1}\rho^n}$$

The first part of the theorem now follows from theorem 5.3, since

$$\frac{\Delta(\mathcal{F}_n, \mathcal{G})}{n^{\tau-1}\rho^n} \to 0 \quad \text{as} \quad n \to \infty .$$

To prove the last part, we shall apply to an example.

Let $P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$

P represents an irreducible, aperiodic chain and $\{\mathcal{E}_n\}$ converges to the minimal informative experiment. The eigenvalues of P are 1, $-\frac{3}{8}$ $-\frac{1}{4}$ and

$$P^n = Z_o + (-\tfrac{3}{8})^n Z_1 + (-\tfrac{1}{4})^n Z_2$$

with

$$Z_o = \frac{1}{11}\begin{pmatrix} 3 & 4 & 4 \\ 3 & 4 & 4 \\ 3 & 4 & 4 \end{pmatrix}$$

$$Z_1 = \frac{1}{11}\begin{pmatrix} 8 & -4 & -4 \\ 8 & -4 & -4 \\ -14 & 7 & 7 \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}$$

Define $Q = Z_o - \frac{3}{8} Z_1$

i.e. $Q = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$

Then $Q^n = Z_o + (-\frac{3}{8})^n Z_1 \qquad (n=1,2,\ldots)$

and $\qquad \rho = \sigma = \frac{3}{8}, \quad \tau = v = 1 .$

Now

$$\Delta(\mathcal{E}_n, \mathcal{F}_n) \leq \|P^n - Q^n\| = \| (-\tfrac{1}{4})^n Z_2\| = 2(\tfrac{1}{4})^n .$$

Since $Z_2 Z_o = Z_2 Z_1 = 0$, $Z_2^2 = Z_2$ we get for any 3×3 Markov matrix M

$$\|Q^n M - P^n\| \, \|Z_2\| \geq \|Z_2 Q^n M - Z_2 P^n\| = \|(-\tfrac{1}{4})^n Z_2\| = (\tfrac{1}{4})^n \|Z_2\|$$

so that

$$\Delta(\mathcal{E}_n, \mathcal{F}_n) \geq \delta(\mathcal{F}_n, \mathcal{E}_n) = \inf_M \|Q^n M - P^n\| \geq (\tfrac{1}{4})^n$$

Hence

$$(\tfrac{1}{4})^n \leq \Delta(\mathcal{E}_n, \mathcal{F}_n) \leq 2(\tfrac{1}{4})^n$$

so $\qquad \sqrt[n]{\Delta(\mathcal{E}_n, \mathcal{F}_n)} \to \frac{1}{4} < \frac{3}{8} = \rho .$

In [12], Torgersen investigates the situation where $\mathcal{E}$ is a dichotomy (i.e. the parameter set contains <u>two</u> points) and where $\mathcal{E}^n$ is the experiment of taking n independent observations of $\mathcal{E}$. It is shown that if $\mathcal{M}$ is the maximal informative experiment, then there exists a number $c(\mathcal{E})$ such that

$$\sqrt[n]{\Delta(\mathcal{M}, \mathcal{E}^n)} \to c(\mathcal{E}) \quad \text{as} \quad n \to \infty.$$

Further it is shown that if $\mathcal{E}$, $\mathcal{F}$ is a pair of non-equivalent dichotomies, then $\sqrt[n]{\Delta(\mathcal{E}^n, \mathcal{F}^n)} \to c(\mathcal{E}) \vee c(\mathcal{F})$ as $n \to \infty$. The result given in theorem 5.9 is quite analogous, apart from the fact that we may have

$$\sqrt[n]{\Delta(\mathcal{E}_n, \mathcal{F}_n)} \to \delta < \rho \vee \sigma \quad \text{if} \quad \rho = \sigma .$$

6. The minimizing Markov matrix.

We shall in this section assume that P^n converges to a matrix Z_o. Then Z_o is the form (4.3).

Since the set of Markov-matrices is closed in the topology induced by the matrix norm, we may for each n pick out a Markov-matrix M_n such that

$$\Delta_n = \|Z_o M_n - P^n\| .$$

Since P^n converges to Z_o, it may seem reasonable that $M_n = I$ and thus $\Delta_n = \|Z_o - P^n\|$.

That this is not true in general, is easily seen in the case $r = 2$. Consider example 5.7. For each n

$$P^n = Z_o + \frac{(1-\alpha-\beta)^n}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}$$

so that

$$\|Z_o - P^n\| = \frac{2(\alpha \vee \beta)}{\alpha+\beta} \rho^n > \rho^n = \Delta_n .$$

From the expression $\Delta_n = \inf_M \|Z_o M - P^n\|$ we observe that M is involved only through $Z_o M$. Hence M_n may be replaced by any matrix M_n' such that $Z_o M_n = Z_o M_n'$. Such M_n' may be defined by $M_n' = A + M_n$ if $Z_o A = 0$. Since $Z_o^2 = Z_o P = Z_o I = Z_o$, A may be chosen as the difference between two of the matrices Z_o, P or I. (In each case one has to check that M_n' is non-negative).

In example 7.11 of [9] it is shown that with

$$M_n = \begin{pmatrix} 1-a_n & a_n \\ b_n & 1-b_n \end{pmatrix}$$

(in the case $r = 2$ from example 5.7) we may choose a_n, b_n as arbitrary numbers in $[0,1]$ satisfying

$$a_n\beta - b_n\alpha = \frac{\beta-\alpha}{2}\rho^n$$

Taking $a_n = b_n = \frac{1}{2}\rho^n$ yields

$$M_n = I + \tfrac{1}{2}\rho^n \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

so that $M_n \to I$.

If $\alpha < \beta$ we may put

$b_n = 1$, $a_n = \frac{\beta-\alpha}{2\beta}\rho^n + \frac{\alpha}{\beta}$ to get

$$M_n = \begin{pmatrix} 1 - \frac{\alpha}{\beta} & \\ 1 & 0 \end{pmatrix} + \frac{\beta-\alpha}{2\beta}\rho^n \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}$$

so that $M_n \not\to I$.

However, M_n converges to a matrix A satisfying $Z_oA = Z_o$.

We have the following general results:

Theorem 6.1.

Let $\{M_n\}$ be a sequence of minimizing Markov matrices. Then

(i) $Z_oM_n \to Z_o$ as $n \to \infty$
with exponential speed.

(ii) There is a subsequence $\{M_n'\}$ which converges to a matrix A satisfying $Z_oA = Z_o$.

Proof : Define R_n so that $P^n = Z_o + R_n$.

Then $\|Z_oM_n - Z_o\| \leq \|Z_oM_n - Z_o - R_n\| + \|R_n\| = \Delta_n + \|R_n\|$. (i) follows.

To prove (ii), note that by compactness there is a subsequence $\{M_n'\}$ which converges to a Markov matrix A .
Now $Z_oA = Z_o$, since

$$\|Z_oA - Z_o\| \leq \|Z_o\| \, \|A - M_n'\| + \|Z_oM_n' - Z_o\|$$

and the right hand side tends to 0 .

The result (ii) may easily be improved in a particular case :

Theorem 6.2.

If the Markov chain has no transient states, then we may pick out a minimizing sequence $\{M_n\}$ so that $M_n \to Z_o$.

Proof: In this case,

$$Z_o = \text{diag } (A_1, \ldots, A_v)$$

where $A_1, \ldots, A_v$ are quadratic Markov matrices each with identical row vectors. Let $\mathcal{C}$ be the set of all such matrices.

Furthermore, $P^n = \text{diag } (P_1^n, \ldots, P_v^n)$.
For any $M \in \mathcal{M}_{r,r}$, Z_oM is a matrix of the form

$$Z_oM = \begin{pmatrix} B_1 \\ \vdots \\ B_v \end{pmatrix}$$

where $B_1, \ldots, B_v$ each has identical rows. Since M_n minimizes

$\|Z_o M_n - P^n\|$, a glance at P^n will make it clear that M_n ought to be chosen so that $Z_o M_n \in \mathcal{C}$. In other words, we have

$$\Delta_n = \inf_{M \in \mathcal{C}} \|M - P^n\| \quad .$$

Let M_n' be so that $\Delta_n = \|M_n' - P^n\|$.

Since $Z_o M = M$ whenever $M \in \mathcal{C}$, theorem 6.1 (i) implies that $M_n' \to Z_o$.

That $\{M_n'\}$ is a sequence of minimizing matrices, follows from the fact that $\|Z_o M_n' - P^n\| = \|M_n' - P^n\|$.

7. Sufficiency and insufficiency in finite experiments.

Notations:

Let $A = \{A_1,\ldots,A_k\}$ be a partition of the space $\chi = \{1,2,\ldots,s\}$ $(k \leq s)$.

By an A-Markov matrix $W = (w_{ij})$ we shall mean a matrix $W \in \mathcal{M}_{k,s}$ such that for $j=1,\ldots,s$; $i=1,\ldots,k$

$$w_{ij} = 0 \quad \text{if} \quad j \notin A_i \ .$$

As an example, let $s=3$ and $A = \{\{1,2\},\{3\}\}$.

Then $W = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 \end{pmatrix}$ is an A-Markov matrix.

By V_A we shall denote the element (v_{ji}) of $\mathcal{M}_{s,k}$ defined by

$$v_{ji} = \begin{cases} 1 & \text{if} \quad j \in A_i \\ 0 & \text{if} \quad j \notin A_i \end{cases}$$

In the above example, $V_A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$.

By U_A we shall denote the A-Markov matrix where each non-zero entry of the i'th row equals $(\#A_i)^{-1}$.

In our example $U_A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

When no confusion can arise, we may write V and U instead of V_A and U_A .

Let $\mathcal{E}$ be a finite experiment as defined in section 2, and let $P \in \mathcal{M}_{r,s}$ be the matrix defining $\mathcal{E}$.

We start this section by giving a criterion of sufficiency in terms of matrices.

A statistic defined on the sample space $\chi = \{1,2,\ldots,s\}$ of $\mathcal{E}$ may be identified with a partition $A = \{A_1,\ldots,A_k\}$ of χ. (A consists of the subsets of χ to which the statistic assigns the same value). Hence we shall deal with the concept <u>sufficient partition</u> instead of <u>sufficient statistic.</u>

The experiment of observing A_i is a reduction of the experiment $\mathcal{E}$. It will be denoted by $\mathcal{E}^A$ and is defined by PV_A.

<u>Theorem 7.1.</u>

Let $\mathcal{E}$ be given as above. A partition A is sufficient for P if and only if there is a A-Markov matrix W such that

$$P = PV_AW \tag{7.1}$$

<u>Remark:</u>

The criterion (7.1) may be reformulated as follows : Take out the columns corresponding to a certain partition. The resulting matrix has the property that the elements of each row are equally proportioned. As an example, let

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{2}{3} & \frac{1}{3} & 0 \end{pmatrix}. \quad \text{Then} \quad A = \{\{1,2\},\{3\}\} \quad \text{is sufficient,}$$

with $W = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Proof: PV_A is an element of $\mathcal{M}_{r,k}$, the θ-th row vector defining the probability distribution over $A_1,\ldots,A_k$ when θ is the "true" parameter. By the definition of Δ-sufficiency (see e.g. Lindqvist and Torgersen [9] ch. 8), A is sufficient if and only if $\mathcal{E}^A \sim \mathcal{E}$. If the condition (7.1) holds, then this is obviously true.

Conversely, assume that A is sufficient. By definition, the conditional distribution over χ given A may be specified independent of θ. Let W be the matrix defining this conditional distribution. Then W is an A-Markov matrix and we have $P = PV_AW$.

Le Cam has in [8] introduced a concept called <u>insufficiency,</u> in order to measure the loss of information incurred by restricting ourselves to a function of the observations (statistic). This measure is so defined that a statistic is sufficient if and only if the insufficiency equals 0. As will be seen, the insufficiency happens to be generally larger than the deficiency. We state and prove some results on insufficiency for <u>finite</u> experiments.

<u>Definition 7.2.</u>

Let $\mathcal{E}$ be given by the matrix P. Let A be a partition of χ. By the <u>insufficiency</u> of A, denoted $\eta(A, \mathcal{E})$ we shall mean the quantity

$$\eta(A, \mathcal{E}) = 2 \inf_{Q} \|Q-P\|$$

where infimum is taken over all matrices $Q \in \mathcal{M}_{s,r}$ such that A is sufficient for Q. The set of such matrices is non-void, since it contains any Q with identical rows.

Theorem 7.3.

$$\Delta(\mathcal{E}^A, \mathcal{E}) \leq \eta(A, \mathcal{E}) .$$

Proof: Assume that A is sufficient for Q. By theorem 7.1 there exists an A-Markov matrix W such that

$Q = QVW$. Furthermore,

$$\begin{aligned}
\Delta(\mathcal{E}^A, \mathcal{E}) &= \inf_{M} \|PVM - P\| \\
&\leq \inf_{M} \|PVM - Q\| + \|Q - P\| \\
&\leq \|PVW - QVW\| + \|Q - P\| \\
&\leq \|P - Q\| \cdot \|V\| \cdot \|W\| + \|Q - P\| \\
&= 2\|Q - P\| .
\end{aligned}$$

Hence $\Delta(\mathcal{E}^A, \mathcal{E}) \leq \eta(A, \mathcal{E})$.

Theorem 7.3 gives a lower bound for $\eta(A, \mathcal{E})$. We shall, however, also need an upper bound.

Define $$\gamma(A, \mathcal{E}) = \inf_{W} \|PV_AW - P\|$$

where infimum is taken over all A-Markov matrices W. Then we have:

Theorem 7.4.

$$\gamma(A, \mathcal{E}) \leq \eta(A, \mathcal{E}) \leq 2\gamma(A, \mathcal{E}) .$$

Proof: Let W be an A-Markov matrix and define $Q = PVW$. Then A is sufficient for Q by theorem 7.1, since

$$QVW = PVWVW = PVW = Q \qquad \text{(we have } WV = I) .$$

Hence
$$\eta(A, \mathcal{E}) \leq 2\|PVW-P\| \quad \text{so} \quad \eta(A, \mathcal{E}) \leq \gamma(A, \mathcal{E}) .$$

Conversely, let Q be such that A is sufficient for Q. Let W be given according to theorem 7.1, so that $Q = QVW$.
Now $PVW = Q + (PV-QV)W$. Subtracting P on either side yields

$$\gamma(A, \mathcal{E}) \leq \|PVW-P\| \leq 2\|Q-P\| .$$

Hence $\gamma(A, \mathcal{E}) \leq \eta(A, \mathcal{E})$ and the proof is complete.

Example 7.5.

Assume that $A = \{\chi\}$. This is an extreme case. The experiment $\mathcal{E}^A$ is now the minimal informative experiment. Thus we have

$$\Delta(\mathcal{E}^A, \mathcal{E}) = \inf_{M \in \mathcal{M}_0} \|M-P\| .$$

On the other hand, $V_A = (1,1,\ldots,1)'$, so that any Q such that A is sufficient, has identical rows.

Hence
$$\eta(A, \mathcal{E}) = 2\inf_{M \in \mathcal{M}_0} \|M_0-P\| = 2\Delta(\mathcal{E}^A, \mathcal{E}) .$$

Similarly it is seen that $\gamma(A, \mathcal{E}) \cdot \Delta(\mathcal{E}^A, \mathcal{E})$.

The above ideas may be applied to our Markov chain experiments. We have:

Theorem 7.6.

Consider a Markov chain with v recurrent, aperiodic classes. In other words, we may assume that P is given by

$$P = \text{diag}\ (P_1, \ldots, P_v) .$$

Let the partition A of χ be the collection of recurrent classes. Then

$$\eta(A, \mathcal{E}_n) = 2\Delta_n$$

(where $\Delta_n = \Delta(\mathcal{E}_n, \mathcal{E}_\infty)$ as defined in section 5).

Proof: By the proof of theorem 6.2,

$$\Delta_n = \inf_{M \in \mathcal{C}} \|M - P^n\|$$

($\mathcal{C}$ is defined in the proof).

Now

$$\eta(A, \mathcal{E}_n) = 2 \inf_Q \|Q - P^n\| \qquad (7.2)$$

Obviously (see remark of theorem 7.1), A is sufficient for any $Q \in \mathcal{C}$. On the other hand, let Q be an arbitrary matrix so that A is sufficient for Q. We contend that there is a $Q' \in \mathcal{C}$ such that $\|Q' - P^n\| \leq \|Q - P^n\|$. The method of constructing such a Q' is illustrated by defining the first row of Q'.

Assume that P_1 has dimension $\alpha \times \alpha$ and let $(q_1, \ldots, q_r)$ be the first row of Q. We define the first row

$(q_1', q_2', \ldots, q_\alpha', 0, \ldots, 0)$ of Q' by sharing the amount $q_{\alpha+1}+\ldots+q_r$ between $q_1, \ldots, q_\alpha$ and adding in such a way that the proportions are not altered, i.e.

$$q_1 : q_2 :\ldots: q_\alpha = q_1' : q'_2 :\ldots: q_\alpha' .$$

This is obtained by defining

$$q_j' = \frac{q_j}{q_1+\ldots+q_\alpha} \quad ; \quad j=1,\ldots,\alpha .$$

We remark that $q_1+\ldots+q_\alpha$ may always be assumed to be > 0, since $q_1+\ldots+q_\alpha = 0$ would imply $\|Q-P^n\| = 2$ and we have (see Torgersen [10])

$$\inf_{Q\in\mathcal{C}} \|Q-P^n\| \leq 2 - \frac{1}{r} .$$

The remaining rows of Q' are defined in a similar way, so that we end up with

$$Q' = \operatorname{diag}(Q_1', \ldots, Q_v') .$$

Since A is sufficient for Q, it follows from the remark of theorem 7.1 that A is sufficient for Q'. Furthermore, since for any j the elements of the rows of Q_j are equally proportioned and have sum 1, Q_j' has identical rows. Hence $Q' \in \mathcal{C}$.

That $\|Q'-P^n\| \leq \|Q-P^n\|$ follows from the fact that P^n is of the same diagonal form as Q' and has zero's outside the diagonal. The theorem follows from (7.2).

8. Lumping of states.

Kemeny and Snell have in § 6.3 of [6] introduced the concept of lumping states of a Markov chain.

There is given a Markov chain $\{X_n\}$ on the state space $\chi = \{1,2,\ldots,r\}$ with transition matrix P. Let $A = \{A_1,\ldots,A_k\}$ be a partition of the state space.
Define a stochastic process $\{Y_n\}$ on A by taking Y_n as the set A_j containing X_n. In general, $\{Y_n\}$ is no Markov chain ; in a number of situations, however, it will in fact be. For examples, we refer to [6] .

According to Kemeny and Snell, a Markov chain is said to be lumpable w.r.t. a partition A (in short: A-lumpable) if for each initial probability vector π , $\{Y_n\}$ constitutes a Markov chain with transition probabilities independent of π .

If the chain given by P is A-lumpable, then the transition matrix of the lumped chain is given by $\overset{\Delta}{P} = U_A P V_A$ (for notations, see section 6) .

Theorem 6.3.5 of [6] states that a chain is A-lumpable if and only if

$$V_A U_A P V_A = P V_A \quad . \tag{8.1}$$

In this case we have

$$\overset{\Delta}{P}^n = U P^n V \quad .$$

As is seen from section 4, the information contained in a Markov chain depends on the eigenvalues of the transition matrix.
We shall state and prove a result on the eigenvalues of lumped chains:

Theorem 8.1.

Let P represent a chain which is lumpable w.r.t. a partition A. Then any eigenvalue of $\hat{P} = U_A P V_A$ is an eigenvalue of P. In other words, $\hat{P}$ "inherits" eigenvalues from P.

Proof : If λ is an eigenvalue of $\hat{P}$, then there exists a vector x such that $\hat{P}x = \lambda x$. This is equivalent to $UPVx = \lambda x$, which implies $VUPVx = \lambda Vx$. By (8.1) this is equivalent to $PVx = \lambda Vx$ so that λ is an eigenvalue of P with corresponding eigenvector Vx ■

Lumping of the states leads us to a reduction of the experiment $\mathcal{E}_n$. We shall now consider the experiment $\mathcal{F}_n$ with sample space $A = \{A_1, \ldots, A_k\}$ and consisting in observing Y_n. $\mathcal{F}_n$ is defined by the $r \times k$ Markov matrix. $V\hat{P}^n$, the θ-th row vector of which is the row vector of $\hat{P}^n$ corresponding to the A_j to which θ belongs. An easy verification shows that $V\hat{P}^n = P^n V$. The experiment $\mathcal{F}_n$ has the same parameter set as $\mathcal{E}_n$, so that $\mathcal{E}_n$ and $\mathcal{F}_n$ are comparable as regards information.

In some practical situations one may find it convenient observing Y_n instead of X_n. This may happen if r is large and it is difficult (or expensive) to get an accurate value for X_n. As seems reasonable, $\mathcal{E}_n \geq \mathcal{F}_n$. This is seen from the randomization criterion , since $V\hat{P}^n = P^n V$.

Next, it will be of interest to estimate <u>how</u> much information will get lost by observing Y_n instead of X_n. This may be measured by $\Delta(\mathcal{E}_n, \mathcal{F}_n)$, which will be discussed later. We describe now an alternative approach :

For simplicity, assume that $\{\mathcal{E}_n\}$ converges to the least informative experiment $\mathcal{O}$, and let $\{\mathcal{F}_n\}$ be defined as before. Then $\{\mathcal{F}_n\}$ converges to the least informative experiment, since $\mathcal{E}_n \geq \mathcal{F}_n \geq \mathcal{O}$.

As is noted in section 5, $\Delta_n = \Delta(\mathcal{E}_n, \mathcal{O})$ measures the information contained in X_n.

Define $\gamma_n = \Delta(\mathcal{F}_n, \mathcal{O})$. Examining $\hat{V}P^n$ we will see that $\gamma_n = \Delta(\tilde{\mathcal{F}}_n, \tilde{\mathcal{O}})$, where $\tilde{\mathcal{F}}_n$ is the experiment with parameter set $\{1,\ldots,k\}$ defined by $\hat{P}^n$ and $\tilde{\mathcal{O}}$ is the least informative experiment over the same parameter set.

Hence, ρ being the root of P and μ being the root of $\hat{P}$,

$$\sqrt[n]{\frac{\Delta(\mathcal{F}_n, \mathcal{O})}{\Delta(\mathcal{E}_n, \mathcal{O})}} \to \mu/\rho \quad \text{as} \quad n \to \infty .$$

The quantity μ/ρ measures the "goodness" of our lumping as regards the information we obtain about θ.

By theorem 8.1, $\mu/\rho \leq 1$ and $\mu/\rho = 1$ if and only if $\hat{P}$ inherits an eigenvalue with modulus ρ. Hence, we should always lump in such a way that $\hat{P}$ inherits the root of P.

Example 8.2.

We shall consider the chain studied in [6] p.29, ex. 8.
X_n represents the weather in the Land of Oz on the n'th day; the possible values being

1 = rain, 2 = sunshine, 3 = snow.

The transition matrix is assumed to be

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

The eigenvalues of P are $1, \frac{1}{4}, -\frac{1}{4}$. A computation (see Lindqvist [10], p.59) shows that $\Delta_n = \frac{4}{3}(\frac{1}{4})^n$. This chain is lumpable w.r.t. $A = \{\{1,3\}, \{2\}\}$, the lumped chain being given by the transition matrix

$$\hat{P} = UPV = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ 1 & 0 \end{pmatrix}.$$

The root of $\hat{P}$ equals $\frac{1}{4}$, so by example 5.7, $\gamma_n = (\frac{1}{4})^n$. Considering Δ_n and γ_n as measures of information, this shows that very little information is lost by lumping the states.

Of course states may be lumped regardless of the resulting chain being a Markov chain or not.

Let $\{X_n\}$, A and $\{Y_n\}$ have the same meaning as in the beginning of this section.

The Markov-matrix defining the probability distribution of Y_n is obtained from the one defining X_n by adding the columns corresponding to each element of our partition. Hence $\tilde{\mathcal{F}}_n$ (the experiment of observing Y_n) is given by $P^n V_A$. As is seen in the case of lumpability, $\mathcal{E}_n \geq \tilde{\mathcal{F}}_n$.

To ask if $\mathcal{E}_n \sim \mathcal{F}_n$ is the same as asking if Y_n is sufficient for X_n. Applying to theorem 7.1 we may prove by induction that if Y_1 is sufficient for X_1, then Y_n is sufficient for X_n for each n. Hence, if Y_1 is sufficient for X_1, then it makes sense to say that A is sufficient for $\{X_n\}$.

It is noted that sufficiency of a partition A and lumpability of A are quite different concepts. Lumpability will in general not imply sufficiency and conversely.

We recall that in the case of lumpability we had $\mathcal{F}_1 \geq \mathcal{F}_2 \geq \cdots$ (this follows from the randomization criterion since $_n$ is given by $\hat{VP^n}$).

If A is sufficient, then $\mathcal{E}_1 \sim \mathcal{F}_1$, $\mathcal{E}_2 \sim \mathcal{F}_2, \ldots$ so $\mathcal{F}_1 \geq \mathcal{F}_2 \geq \cdots$ since $\mathcal{E}_1 \geq \mathcal{E}_2 \geq \cdots$.

If A is neither sufficient nor the chain is lumpable, then, however, this may not be true. For example, define

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \quad \text{and let} \quad A = \{\{1,2\}, \{3\}\}.$$

$$\text{Then} \quad PV = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad P^2V = \begin{pmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{5}{8} & \frac{3}{8} \end{pmatrix}$$

By the Ψ-criterion (see Torgersen [11]) $\mathcal{F}_1 \geq \mathcal{F}_2$ if and only if $\Psi(\mathcal{F}_1) \geq \Psi(\mathcal{F}_2)$ for all $\Psi \in \psi$. Defining $\Psi(x_1,x_2,x_3) = |x_2-x_3|$, however, we get $\Psi(\mathcal{F}_1) = 0$, $\Psi(\mathcal{F}_2) = \frac{1}{4}$ (for computation of $\Psi(\ _i)$, see [9]).

If the relation $\mathcal{F}_1 \geq \mathcal{F}_2 \geq \ldots$ does not hold, then we shall see that a limit experiment $\mathcal{F}_\infty$ such that $\mathcal{F}_n \to \mathcal{F}_\infty$ may not exist. If P^n converges, then of course $\mathcal{F}_\infty$ exists and is given by $\lim P_n V_A$. Thus a sequence $\{\mathcal{F}_n\}$ having no limit necessarily springs from a periodic chain. A very simple example is given on $\chi = \{1,2,3\}$ by

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Here $\{1\}$ and $\{2,3\}$ are cyclic classes.

We have $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ and it is

seen that

$$P = P^3 = P^5 = \ldots, \qquad P^2 = P^4 = P^6 = \ldots .$$

Hence the limit experiment $\mathcal{F}_\infty$, if it exists, may be given by either $P\,V$ or $P^2\,V$. These experiments are, however, not equivalent. Consequently, the sequence $\{\mathcal{F}_n\}$ has <u>no</u> limit.

As is noted earlier, the effectiveness of a restriction to an experiment $\mathcal{F}_n$ ought to be measured by $\Delta(\mathcal{E}_n, \mathcal{F}_n)$. However, since Y_n is a function of X_n, the <u>insufficiency</u> of Y_n may

be considered (more precisely the insufficiency of the partition A defined by Y_n), $\eta(A, \mathcal{E}_n)$. For simplicity, put $\varepsilon_n = \Delta(\mathcal{E}_n, \mathcal{F}_n)$, $\eta_n = \eta(A, \mathcal{E}_n)$.

To avoid the difficulties occurring if $\{\mathcal{F}_n\}$ does not converge, we will in the following assume that P^n converges. Further we shall assume that $\{\mathcal{E}_n\}$ and $\{\mathcal{F}_n\}$ have a common limit experiment $\mathcal{G}$. If this assumption is not satisfied, then lumping seems to have little practical interest.

Of course $\varepsilon_n \to 0$ as $n \to \infty$, since

$$\varepsilon_n \leq \Delta(\mathcal{E}_n, \mathcal{G}) + \Delta(\mathcal{F}_n, \mathcal{G}) .$$

Theorem 7.3 implies that $\varepsilon_n \leq \eta_n$. The rest of this section will be devoted to showing that $\sqrt[n]{\varepsilon_n}$ and $\sqrt[n]{\eta_n}$ have a common limit as $n \to \infty$, i.e. that ε_n and η_n behave asymptotically in the same manner.

In general, however, it may happen that the Δ-distance is small and the η-insufficiency large. An illustrative example is given in Le Cam [8] p. 46, (where the deficiency behaves like $\frac{C_1}{n}$ and the insufficiency behaves like $\frac{C_2}{\sqrt{n}}$). We shall need the succeeding lemma :

Lemma 8.3.

Let $A_1, A_2, \ldots, A_z$, $B_1, B_2, \ldots, B_z \in \mathcal{M}_{r,r}$ and assume that there exists no $M \in \mathcal{M}_{r,r}$ such that

$$A_i M = B_i \quad \text{for} \quad i=1,\ldots,z .$$

Then there is a number $\varphi > 0$ such that for any $M \in \mathcal{M}_{r,r}$

$$\|A_i M = B_i\| \geq \varphi \quad \text{for at least } \underline{\text{one}} \quad i=1,\ldots,z \ .$$

Proof : The real function f on $\mathcal{M}_{r,r}$ defined by $f(M) = \sum_{i=1}^{z} \|A_i M - B_i\|$ is a continuous function on a compact space. Since $f(M) > 0$ on $\mathcal{M}_{r,r}$, there exists $a > 0$ so that $f(M) \geq a$ for all M. Put $\varphi = a/z$.

Theorem 8.4.

There exists $\nu \geq 1$, $0 \leq \delta \leq 1$, $0 < k \leq K < \infty$ so that

$$cn^{\nu-1}\delta^n \leq \epsilon_n \leq Cn^{\nu-1}\delta^n \quad \text{for } n=1,2,\ldots .$$

Proof : We order the eigenvalues of P according to decreasing modulus. Note that since P^n is required to converge, 1 is the only eigenvalues of P with modulus 1. Furthermore it has index 1.

Let $1 < \rho_1 < \rho_2 < \ldots < \rho_t < 0$ be the different moduli of eigenvalues of P.

We have $\epsilon_n = \inf_M \|P^n VM - P^n\|$, the infimum taken over $\mathcal{M}_{k,r}$.

Let $Z_o = \lim P^n$. Then

$$P^n VM - P^n = Z_o(VM-I) + \sum_{i=1}^{s} \sum_{j=0}^{m_i-1} \frac{(P-\lambda_i I)^j}{j!} n^{(j)} \lambda_i^{n-j} Z_i (VM-I) \quad (8.2)$$

The value of δ is found by the following procedure :

Since $\{\mathcal{E}_n\}$ and $\{\tilde{\mathcal{F}}_n\}$ are assumed to have the same limit experiment, and since these limit experiments must be given by Z_o and $Z_o V$, respectively, it follows that there exists at least

one M so that $Z_o VM = Z_o$. Let $\mathcal{M}_o$ be the set of such M's ; $\mathcal{M}_o \subseteq \mathcal{M}_{r,r}$.

Define $\mathcal{M}_1 \subseteq \mathcal{M}_o$ as the set of $M \in \mathcal{M}_o$ so that $Z_i VM = Z_i$ for all Z_i corresponding to eigenvalues with modulus ρ_1 . In the same manner define $\mathcal{M}_2 \subseteq \mathcal{M}_1$ so that $Z_i VM = Z_i$ for all Z_i corresponding to eigenvalues with modulus ρ_2 .

Continuing, one gets a chain

$$\mathcal{M}_{r,r} \supseteq \mathcal{M}_o \supseteq \mathcal{M}_1 \supseteq \cdots .$$

If $\mathcal{M}_t \neq \emptyset$, then it is clear that ϵ_n <u>equals</u> 0 from some n on . This is the case if our lumping is sufficient.

If $\mathcal{M}_t = \emptyset$, let j be the smallest number $(1 \leq j \leq t)$ so that $\mathcal{M}_j = \emptyset$.

We claim that $\delta = \rho_j$.

For each $M \in \mathcal{M}_{j-1}$, let ν_M be the maximum index of eigenvalues λ_i with modulus ρ_j corresponding to Z_i's with $Z_i VM \neq Z_i$. We claim that

$$\nu = \min \{\nu_M : M \in \mathcal{M}_{j-1}\} .$$

By inserting an $M' \in \mathcal{M}_{j-1}$ with $\nu_M = \nu$, we get

$$\epsilon_n = \inf_M \|P^n VM - P^n\| \leq \|P^n VM' - P^n\|$$

which by (8.2) is seen to be $\leq C_n^{\nu-1} \delta^n$ for some number $C > 0$, independent of n .

To prove the left inequality of the theorem, assume that a is the largest label for eigenvalues with modulus $\delta = \rho_j$. We multiply each side of (8.2) by $\sum_{i=1}^{a} Z_i$ to get (for any M)

$$\|\sum_{i=1}^{a} Z_i\| \, \|P^n VM - P^n\|$$

$$\geq \|Z_o(VM-I) + \sum_{i=1}^{a} \sum_{j=o}^{m_i-1} \frac{(P-\lambda_i I)^j}{j!} n^{(j)} \lambda_i^{n-j} Z_i (VM-I)\| \qquad (8.3)$$

Applying to lemma 8.3, there exists a number $\varphi > 0$, independent of M , so that

$$\|Z_i(VM-I)\| \geq \varphi \quad \text{for at least one} \quad i=0,1,\ldots,k \, .$$

Considering the terms occurring on the right hand side of (8.3) we thus conclude that

$$\|P^n VM - P^n\| \geq c n^{\nu-1} \delta^n$$

where $c > 0$ is independent of M .
This completes the proof.

We are now in position to prove our main result.

Theorem 8.5.

Let δ be as in theorem 8.4.
Then $\sqrt[n]{\eta_n} \to \delta$ as $n \to \infty$.

Proof : Let $\gamma_n = \gamma(A, \mathcal{E}_n)$, where γ is defined in section 7 . It suffices to prove that there exists $K > 0$ and $w \geq 0$ such that $\gamma_n \leq K n^w \delta^n$,k since by theorem 7.3 and 7.4 we have $\epsilon_n \leq \eta_n \leq 2\gamma_n$.

Now $$\gamma(A, \mathcal{E}_n) = \inf_W \|P^n VW - P^n\| \qquad (8.4)$$

where infimum is taken over all A-Markov matrices W .

We return to the proof of theorem 8.4 . Let b be the largest label of eigenvalues with modules ρ_{j-1} . Then there exists $M \in \mathcal{M}_{r,r}$ such that

$$Z_o VM = Z_o, \ldots, Z_b VM = Z_b \tag{8.5}$$

Z_o is of the form B given in (4.3), where the matrices $A_1, \ldots, A_v$ have strictly positive entries.

Hence from some n on (which of course may be $n=1$)

$$Q_n = Z_o + \sum_{i=1}^{b} \sum_{j=o}^{m_i-1} \frac{(P-\lambda_i I)^j}{j!} n^{(j)}\lambda_i^{n-j} Z_i \tag{8.6}$$

will have no negative entries and hence define an experiment.

By (8.5) there exists M so that $Q_n VM = Q_n$ for $n=1,2,\ldots$. Hence A is sufficient for Q_n , so by theorem 7.1 M may for each n be replaced by an A-Markov matrix W_n such that $Q_n VW_n = Q_n$. But then, by (8.2), (8.4) and (8.6), $\gamma_n \leq \|P^n VW_n - P^n\| \leq C_n w \delta^n$ and the proof is complete.

We end up this section by a couple of examples.

Example 8.6.

Let $$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \end{pmatrix}$$

The eigenvalues of P are 1 , $\frac{3}{4}$, $-\frac{1}{2}$.

Let $A = \{\{1,2\}, \{3\}\}$. The chain is A-lumpable.

We have $P^n = Z_o + (\frac{3}{4})^n Z_1 + (-\frac{1}{2})^n Z_2$

where

$$Z_o = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

$$Z_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\frac{3}{5} & -\frac{2}{5} & 1 \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{2}{3} & \frac{2}{3} & 0 \\ -\frac{1}{15} & \frac{1}{15} & 0 \end{pmatrix}$$

A simple computation shows that $\mathcal{M}_o$ consists of all matrices of the form

$$\begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ m_1 & m_2 & m_3 \end{pmatrix}$$

$m_1, m_2, m_3 \geq 0$; $m_1 + m_2 + m_3 = 1$.

With $M \in \mathcal{M}_o$ we get

$$Z_1(VM-I) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ m_1 - \frac{1}{15} & m_2 + \frac{1}{15} & m_3 - 1 \end{pmatrix}$$

and no allowable choice of m_1, m_2, m_3 will give $Z_1(VM-I) = 0$.

Hence, $\mathcal{M}_1 = \emptyset$, and by the preceding theorems,

$$\sqrt[n]{\epsilon_n} \quad \text{and} \quad \sqrt[n]{\eta_n} \to \frac{3}{4}$$

i.e. we have convergence to the root of P . That this will not happen in general, is seen from the next example.

Example 8.7.

Let $P = \begin{pmatrix} \frac{7}{16} & \frac{1}{4} & \frac{5}{16} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{5}{16} & \frac{1}{4} & \frac{7}{16} \end{pmatrix}$

The eigenvalues of P are $1, -\frac{1}{4}, \frac{1}{8}$.

Let $A = \{\{1,3\}, \{2\}\}$.

Now $P^n = Z_o + (-\frac{1}{4})^n Z_1 + (\frac{1}{8})^n Z_2$

with

$$Z_o = \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix}, \qquad Z_1 = \begin{pmatrix} \frac{1}{10} & -\frac{2}{10} & \frac{1}{10} \\ -\frac{2}{5} & \frac{4}{5} & -\frac{2}{5} \\ \frac{1}{10} & -\frac{2}{10} & \frac{1}{10} \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

A computation shows that $\mathcal{M}_1 \neq \emptyset$, more precisely, $\mathcal{M}_1$ has one element

$$M = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$$

However, $\mathcal{M}_2 = \emptyset$ since $Z_2 V = 0$.

Hence $\sqrt[n]{\varepsilon_n}$ and $\sqrt[n]{\eta_n} \to \frac{1}{8}$, whence the root of P equals $\frac{1}{4}$.

Acknowledgement.

The author is grateful to Erik N. Torgersen for proposing the subject and for many helpful discussions. Many thanks also to Gerd Salter who typed these notes.

References.

[1] Blackwell, D. (1951). Comparison of experiments. Proc.Sec. Berkeley Symp. Math.statist.Probab. 93-102.

[2] Dunford and Schwarz. (1958). Linear operators. Part I. Interscience Publishers, New York.

[3] Feller, W. (1951). Diffusion Processes in Genetics. Proc. Second Berkeley Symp. Math.statist. Probab. 227-246.

[4] Gantmacher, F.R. (1959). Theory of matrices. Vol. 1-2. Chelsea, New York.

[5] Karlin, S. (1966). A first course in stochastic processes. Academic Press.

[6] Kemeny and Snell. (1960). Finite Markov chains. Van Nostrand.

[7] Le Cam, L. (1964). Sufficiency and approximate sufficiency. Ann.Math.Statist. 35, 1419-1455.

[8] Le Cam, L. (1974). Notes on asymptotic methods in statistical decision theory. Centre de Recherches Mathématiques, Université de Montréal.

[9] Lindqvist and Torgersen (1975). Notes on comparison of statistical experiments. Statist.Memoirs no.1. 1975, Inst. of Math., University of Oslo.

[10] Lindqvist, B. (1975) Anvendelse av teorien for sammenlikning av eksperimenter på Markov-kjeder med ukjent initialtilstand. Graduate thesis. Inst. of Math., University of Oslo.

[11] Torgersen, E.N.(1970). Comparison of experiments when the parameter space is finite Z. Wahrsch.theorie. Verw. Geb. 16, 219-249.

[12] Torgersen, E.N. (1974). Asymptotic behaviour of powers of dichotomies. Statist. Research Rep. no. 6, 1974, Inst.of Math., University of Oslo.