INFORMATION TRANSFER IN COMPLEX SYSTEMS, WITH APPLICATIONS TO REGULATION

by
ROGER C. CONANT

AF-33(615)-3890
AF Grant 7-67
NASA Grant
NGR-15-005-411

Technical Report No. 13

January 1968

Jointly Sponsored by
Air Force Office of Scientific Research
Air Force Systems Engineering Group
National Aeronautics and Space Administration

BIOLOGICAL COMPUTER LABORATORY
ELECTRICAL ENGINEERING RESEARCH LABORATORY
ENGINEERING EXPERIMENT STATION
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

INFORMATION TRANSFER IN COMPLEX SYSTEMS, WITH APPLICATIONS TO REGULATION

by

Roger C. Conant

AF 33(615)-3890
AF Grant 7-67
NASA Grant

Technical Report No. 13

January 1968

Jointly Sponsored by

Air Force Office of Scientific Research

Air Force Systems Engineering Group

National Aeronautics and Space Administration

BIOLOGICAL COMPUTER LABORATORY
ELECTRICAL ENGINEERING RESEARCH LABORATORY
ENGINEERING EXPERIMENT STATION
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

ABSTRACT

This study is concerned with information theory and its relevance to the study of complex systems. When information about every detail of their activity is kept, many systems are too complex to be manageable and can only be dealt with by sacrificing detail. It is shown here that multivariable information theory is capable of eliminating much detail while preserving information about the interrelations between parts of a system, even when those interrelations are very complex. A procedure is described and exemplified, for example, which is helpful in the decomposition of hierarchical systems.

It is shown, among other results, that when two variables are related (in the set theoretic sense) the transmission between them is maximized when their behaviors are isomorphic. This observation leads to an algorithm for the computation of channel capacity for arbitrary finite-state systems of a very general type.

The importance of information in regulatory processes is discussed and quantified, and several basic regulatory schemes are discussed in terms of the information involved, showing in an exact way how information transfer and channel capacity limit the ability of any system to act as a successful regulator.

ACKNOWLEDGEMENT

When at the end of this undertaking it comes time to give credit where credit is due, I realize that all the people involved are as links in a chain, each having been essential for the completion of the thesis. But my thanks go first and foremost to Professor W. Ross Ashby, whose insights were responsible for the inception of this project and whose inspiration, suggestions, and encouragement were invaluable during its development.

The work has been carried out in the Biological Computer Laboratory, a most remarkable scientific group assembled by Professor Heinz Von Foerster; to that group and to Professor Von Foerster in particular I owe many thanks for countless stimulating discussions.

To James Gocking, who prepared the illustrations, to my wife Barbara, who helped with many tedious details, and to Charlene Mock, who speedily and cheerfully typed the text, I am most grateful.

This project was supported by the U. S. Air Force Office of Scientific Research AFOSR Grant 7-67, by the U. S. Air Force Systems Engineering Group, Contract AF33(615)-3890, and by the National Aeronautics and Space Administration Grant 14-005-(111).

TABLE OF CONTENTS

Page

I. INTRODUCTION 1

II. NOTATIONS AND CONVENTIONS 4

2.1. Basic notations 4
2.2. Approximate conversions of discrete to continuous distributions and vice versa . 13
2.3. Discrete-time convention 22

III. SOME RESULTS IN INFORMATION THEORY 23

3.1. Operations on the frequency table which leave H, T, and Q unchanged 23
3.2. Collapsing theorems and their consequences. 24
3.3. Can genuinely complex relationships be broken down? 44
3.4. Maximizing transmission between related variables 47
3.5. Information quantities for dynamic variables 57
3.6. Channel capacity, constraint capacity, and the capacity of automata 62

IV. INFORMATION THEORY AND COMPLEX SYSTEMS 90

4.1. Complex systems 90
4.2. The Diagram of Immediate Effects and some information theory analogs 100
4.3. Decomposition of system constraints 124
4.4. Information transfer 141

V. REGULATION 171

5.1. Information requirements for regulation .. 171
5.2. Quantitative analysis of regulation 178
5.3. Important special cases of regulation 192
5.4. Further remarks 203

REFERENCES 205

I. INTRODUCTION

Norbert Wiener defined Cybernetics as the science of control and communication, in the animal and machine[1]. By that definition, this paper could be called a cybernetic study, for it is concerned with communication within and between systems, and also with the role of communication in control.

When science attempts to gain insight into real-world systems, it invariably begins by dismissing, explicitly or implicitly, many of the variables which might be considered but which are thought to be irrelevant or inconsequential. A scientist studying maze-learning in rats might consider the phase of the moon, the length of the rat's tail, the color of the experimenter's tie, and so on as variables, but in fact he would be silly to do so unless he had reason to think them relevant. Science deals not with real-world "systems" but only with models, i.e., abstracted versions, of them.

Until recently, the systems which were studied were sufficiently simple that after all of the irrelevant variables were discarded, the number remaining was small enough to give a manageable model. When genuinely complex systems are tackled, however, the old procedure doesn't work; either one is forced to discard relevant variables to get a model of manageable complexity, which is then of poor quality, or else one ends up with a model which is of good quality but itself unmanageably complex.

The information theory of complex systems, which is the subject of this paper, can in a sense be viewed as a way of dealing with the latter type of model, by discarding details and only keeping information about its functional structure--which variables affect which and to what degree, which variables are statistically "close" to which others, and so on. Chapters II, III, and IV are concerned with this "communication structure" of systems.

The information theory used here is not the highly specialized theory developed for use in sophisticated communications systems, but rather is an outgrowth of the suggestion by McGill[2], Garner[3], and Ashby[4] that the theory formulated by Shannon[5] could be extended to n variables and could be usefully applied to the study of relations in systems of many variables.

Information theory is important for the study of complex systems in another closely related respect. Most complex systems found in nature, and many of man's complex constructs, survive by acting appropriately on the basis of information they receive; they regulate their actions on the basis of information. That virtually all organisms which have survived the process of natural selection have information sensors bears witness to the importance of information to survival. Indeed, the almost incredible sensitivity and delicacy of the sensory apparati developed in the course of evolution lead one to suspect that primacy in the "struggle for survival" goes to those who can best obtain and use information; we humans have at least five distinct systems for taking in information from the environment, and additional systems for sensing our internal conditions.

The channel capacity of a system is a bound on the ability of the system to accept, transform, and act on incoming information, and as such it is a quantity important for the survival of the system. In chapter III is introduced an algorithm for the calculation of channel capacity for a very general type of system; in chapter IV information transfer in systems is discussed in more general terms.

Chapter V, on Regulation, was inspired by but goes considerably beyond Ashby's Law of Requisite Variety[6]. In that chapter we discuss the relationship between regulation and information transfer and show that the two are closely linked.

II. NOTATIONS AND CONVENTIONS

Introduction

Section 2.1 will set the basic notations to be used hereafter. It does not contain any new material. Section 2.2 will provide conversion techniques between discrete-variable and continuous-variable distributions, allowing us to deal thereafter with discrete distributions only. Section 2.3 will justify our exclusive use of the discrete time variable.

2.1 Basic notations

Matrices will be denoted by underlined Latin capitals, e.g., $\underline{A}$, $\underline{M_3}$.

Constants will be denoted by lower case Latin letters, usually early in the alphabet, e.g., a, h, m_{12}.

Sets will be denoted by Latin capitals or by braces enclosing the elements, e.g., $B = \{b_1, b_2, b_3\}$.

Variables will be denoted by upper case Latin capitals usually toward the end of the alphabet, e.g., X, Y. _Compound variables_ whose components are shown explicitly will be denoted with < and > signs, e.g., $\langle X_1, X_2 \rangle$ or even $\langle X, \langle Y_1, Y_2 \rangle, Z \rangle$. If S is an ordered set of variables $\{X_1, X_2, \ldots, X_M\}$, $\langle S \rangle$ is the compound variable $\langle X_1, X_2, \ldots, X_M \rangle$.

Values taken by a variable will be denoted by lower case versions of the letter representing the variable, possibly with subscripts. The set of values a variable can take will be denoted by the Latin capital representing the variable. For example, the set $X = \{x_1, x_2, x_3\}$ is the set of values taken by variable X. Using the same symbol for the variable and its set of values is often convenient, and the context will always make clear in which sense the symbol is being used.

Values of a variable, being merely the elements of a set associated with a variable, need not be numbers, and no metric is implied. If the set is finite, the elements may be ordered and numbered arbitrarily for convenience, and it is frequently useful to deal with such numbers as equivalent to the values, e.g., to equate "X takes its third value" with "X = 3".

Functions will be denoted by lower case Greek letters, or by f or g. The domain and range sets are a fundamental part of a function's definition; they are displayed as, for instance, $f_1 : Y \rightarrow A$, which is read "Function f_1 maps Y into A".

A system S is an ordered set of variables, and the variables are members of S. By system S we will also mean the product set whose components are the value-sets for the variables in S. If there is a relation (in the set theoretic sense) over the members of S, the subset of the product set implied by that relation will be called the system relation; some authors use the term system to refer to what is here called the system relation. If the variables in S are associated with machines, "the system" can also refer to the collection of machines, if no confusion results. The term system may thus be used in three distinct ways; this should cause no confusion in practice.

A system-value is an ordered N-tuple with one component for each variable in S; e.g., $S = \{X_1, X_2, X_3\}$ has the value $<2, 4, 5>$ when $X_1 = 2$, $X_2 = 4$, and $X_3 = 5$.

A Machine-with Input (MWI) is a sequential machine described by a function of the form $f : S^t \times I^t \rightarrow S^{t+1}$, that is, $s^{t+1} = f(s^t, i^t)$, where s^τ is the "state" at time τ and i^τ the "input". This is usually written $f : S \times I \rightarrow S$ with the understanding that f maps the "present" state and input into the "next" state. A MWI is diagrammatically represented as shown in Figure 1. Both I and S may be product sets.

A Mapper is a machine described by a function of the form $g : I^t \rightarrow O^t$, that is, $o^t = g(i^t)$, in which o^τ is the "output" at time τ and i^τ the "input". This is usually written $g : I \rightarrow O$ with the understanding that g maps the "present" input into the "present" output. A mapper is represented as shown in Figure 2.

A Moore automaton is a machine consisting of a MWI $f : S \times I \rightarrow S$ plus a mapper $g : S \rightarrow O$, as shown in Figure 3.

A frequency table associated with a system $S = \{X_1, X_2, \ldots, X_M\}$ is an M-dimensional matrix whose entries are all nonnegative real numbers. It is denoted $\underline{N}(X_1, X_2, \ldots, X_M)$, $\underline{N}(S)$, or just $\underline{N}$ if the argument is understood. The typical element in $\underline{N}$ is $n_{X_1, X_2, \ldots, X_N}$, with particular subscripts indicating particular system-values. Each element gives the real number (ordinarily, an integer) associated with the frequency of the system-value to which it corresponds; e.g., if $S = \{X_1, Y_1, Y_2\}$, the entry $n_{2, 4, 5} = 3$ indicates three occurrences of the triple $<X_1, Y_1, Y_2> = <2, 4, 5>$. The sum of all entries in a table $\underline{N}(S)$ is denoted by $N(S)$ or just N.

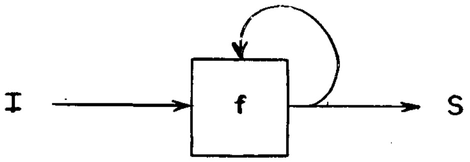


Figure 1.

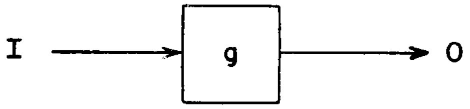


Figure 2.

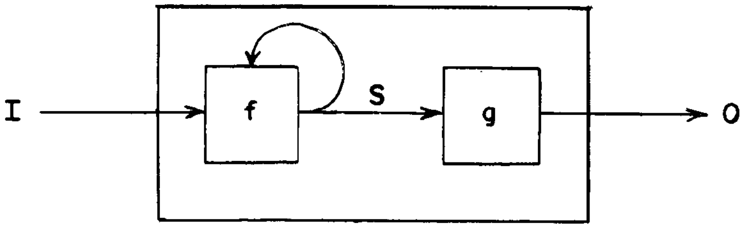


Figure 3.

Thus $\underline{N}$ gives the frequencies of occurrence of all the system-values; the entries of $\underline{N}$ are presumably obtained from some data-gathering process, perhaps by observation of a physical system over a long period. It is not our purpose here to discuss how frequency tables may be obtained, but only to deal with tables already provided.

If a system relation holds over the members of S, some of the entries of $\underline{N}$ will necessarily be zero, and conversely. (If $\underline{N}$ is one-dimensional, the relation becomes a property in set theoretic language.) Somewhat more generally, $\underline{N}$ can be interpreted, after suitable normalization, as the characteristic function, and therefore the descriptor, of an M- ary fuzzy relation[7] on S.

A frequency table associated with $S = \left\{ X_1, X_2, \ldots, X_M \right\}$ can also be associated with other systems, derived from S by grouping the variables in various ways. For example if $S = \left\{ X_1, X_2, X_3 \right\}$ and $Y = < X_2, X_3 >$, the frequency table can be associated with the system $S' = \left\{ X_1, Y \right\}$. This just amounts to noting the obvious fact that an n-tuple of variables can be considered as a single variable with a new name.

An important operation on $\underline{N}(X_1, X_2, \ldots, X_M)$ is that of collapsing the frequency table over one or more of its dimensions (variables). Collapsing over X_i gives a new table $\underline{N}(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_M)$ whose entries are obtained by summing over the X_i dimension:

$$n_{X_1, X_2, \ldots X_{i-1}, X_{i+1}, \ldots, X_M} = \sum_{X_i} n_{X_1, X_2, \ldots, X_M}$$

For example, collapsing $\underline{N}(X, Y)$ over X gives $\underline{N}(Y)$:

	Y		
	0	2	4
X	1	3	1

$\underline{N}(X,Y)$

Y		
1	5	5

$\underline{N}(Y)$

For a one-dimensional frequency table $\underline{N}(X)$, the entropy of X, denoted H(X), is defined to be zero if $N = 0$ and is defined as follows if $N > 0$:

$$H(X) = -\sum_{X} \frac{n_X}{N} \log_2 \frac{n_X}{N}$$

$$= \frac{1}{N} \left[N \log_2 N - \sum n_X \log_2 n_X \right] .$$

The summation runs over all the cells in the frequency table.

Henceforth, in accordance with information theory standards, we will assume logarithms are always to base 2, so that the unit for entropy, etc. is the bit.

With an M-dimensional frequency table $\underline{N}(X_1, X_2, \ldots, X_M)$ for a system $S = \left\{ X_1, X_2, \ldots, X_M \right\}$, the entropy of system S, denoted $H(X_1, X_2, \ldots, X_M)$, H(S), or $H(\underline{N})$, is zero if $N = 0$ and otherwise is defined by

$$H(X_1,X_2,\ldots,X_M) = -\sum_{X_1} \sum_{X_2} \ldots \sum_{X_M} \frac{n_{X_1,X_2,\ldots,X_M}}{N} \log \frac{n_{X_1,X_2,\ldots,X_M}}{N}$$

the summation running over all cells in $\underline{N}(S)$.

The expression n_{X_i}/N may be interpreted as a probability, if this interpretation is useful, but to avoid unnecessary connotations we will generally avoid doing so. The term "probability" carries a connotation of permanence and reference to future events, while the frequency table connotes a reference to events of the past - although the table in the abstract is of course just an array of numbers, with no time reference.

If the assumptions under which a system is being studied allow the probability density function to be meaningfully defined, then the probability density function for a system $S = \left\{ X_1, X_2, \ldots, X_M \right\}$ is denoted $p(X_1, X_2, \ldots, X_M)$ or $p(S)$ and is defined in the ordinary way. In this case, $H(X)$ and $H(S)$ are defined as follows:

$$H(X) = -\int_{-\infty}^{\infty} p(X) \log p(X)\, dX$$

$$H(S) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p(X_1,X_2,\ldots,X_M) \log p(X_1,X_2,\ldots,X_M) \cdot dX_1 dX_2 \ldots dX_M$$

The operation of collapsing a frequency table over a variable X_i corresponds, with probability densities, to integration over X_i:

$$p(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots X_M) = \int_{-\infty}^{\infty} p(X_1, X_2, \ldots, X_M)\, dX_i$$

The relation between discrete and continuous distributions will be considered in more detail in section 2.2.

For $\underline{N}(X,Y)$, the entropy of X conditional on Y is denoted by $H_Y(X)$ and defined by

$$H_Y(X) = H(X, Y) - H(Y)$$

To obtain $H(Y)$ from $\underline{N}(X,Y)$ requires collapsing $\underline{N}$ over the X-dimension, thus obtaining $\underline{N}(Y)$; $H(Y)$ is then obtained from $\underline{N}(Y)$.

The obvious generalization of $H_Y(X)$ is $H_{Y_1,Y_2,\ldots,Y_n}(X_1,X_2,\ldots,X_M)$;

if $S_1 = \left\{X_1, X_2, \ldots, X_m\right\}$ and $S_2 = \left\{Y_1, Y_2, \ldots, Y_n\right\}$, this can be denoted $H_{S_2}(S_1)$, the entropy of S_1 conditional on S_2, and defined by

$$H_{S_2}(S_1) = H(S_1 \cup S_2) - H(S_2).$$

Normally, $H_{S_2}(S_1)$ is of interest only if S_1 and S_2 are disjoint. The set $S_1 \cup S_2$ is an ordered set, just as S_1 and S_2 are ordered sets.

For a two-dimensional table $\underline{N}(X,Y)$, the *transmission between X and Y*, denoted $T(X : Y)$, is defined by

$$T(X : Y) = H(X) + H(Y) - H(X, Y).$$

The expression on the right is equal to $H(X) - H_Y(X)$ and to $H(Y) - H_X(Y)$, but we take the definition above as primary.

$T(X : Y)$ can be generalized in the obvious way to $T(S_1 : S_2)$, but it can be generalized in a more fundamental way by introducing more single variables. The total transmission over the system $S = \left\{X_1, X_2, \ldots, X_M\right\}$, denoted $T(X_1 : X_2 : \ldots : X_M)$, $T(S)$, or $T(\underline{N})$ where $\underline{N}$ is the frequency table for S, is zero if S contains only one variable and otherwise is defined by

$$T(X_1 : X_2 : \ldots : X_M) = H(X_1) + H(X_2) + \ldots + H(X_M)$$
$$- H(X_1, X_2, \ldots, X_M).$$

$T(S)$ is a measure of the total constraint holding between all the variables in S - a measure of the degree to which the variables are statistically interdependent. If $T(S) = 0$, the system relation is of a degenerate type, being merely the conjunction of one-dimensional properties on the several variables. (These statements will be justified later.)

The transmission over a system $S_1 = \left\{ X_1, X_2, \ldots, X_m \right\}$ conditional on $S_2 = \left\{ Y_1, Y_2, \ldots Y_n \right\}$ is denoted by $T_{Y_1, Y_2, \ldots, Y_n}(X_1 : X_2 : \ldots : X_m)$ or $T_{S_2}(S_1)$ and is defined by

$$T_{S_2}(S_1) = H_{S_2}(X_1) + H_{S_2}(X_2) + \ldots + H_{S_2}(X_m) - H_{S_2}(S_1).$$

The transmission between $S_1 = \left\{ X_1, X_2, \ldots, X_m \right\}$ and $S_2 = \left\{ Y_1, Y_2, \ldots, Y_n \right\}$ is denoted by $T(S_1 : S_2)$ and is defined by

$$T(S_1 : S_2) = T(< X_1, X_2, \ldots, X_m > : < Y_1, Y_2, \ldots, Y_n >).$$

All these entropies, conditional entropies, transmissions, and conditional transmissions are non-negative quantities measured in bits, and they all have familiar interpretations discussed in the literature.

A less familiar entity is the interaction. Given a three-dimensional frequency table $\underline{N}(X, Y, Z)$, the interaction between X, Y, and Z is denoted by $Q(X, Y, Z)$ and is defined by

$$Q(X, Y, Z) = T_Z(X : Y) - T(X : Y)$$

It is easy to show, by collecting terms, that

$$Q(X, Y, Z) = T_X(Y : Z) - T(Y : Z)$$

$$= T_Y(X : Z) - T(X : Z)$$

so the definition is actually symmetrical in the variables. $Q(X, Y, Z)$ is a measure of how much the transmission between two of the variables is conditional on the third; Q may be either positive, negative, or zero.

The interaction between X, Y, and Z conditional on W, denoted $Q_W(X, Y, Z)$, is defined like $Q(X, Y, Z)$ but with every H subscripted with a W.

$Q(X, Y, Z)$ may be generalized in an obvious way to $Q(S_1, S_2, S_3)$, or more fundamentally by introducing more variables in the argument. The n-variable interaction over the system $S = \{X_1, X_2, \ldots, X_n\}$, denoted $Q(X_1, X_2, \ldots, X_n)$ or $Q(S)$, is defined iteratively as follows:

$$Q(X_1, X_2, \ldots, X_{n-1}, X_n) = Q_{X_n}(X_1, X_2, \ldots, X_{n-1}) - Q(X_1, X_2, \ldots, X_{n-1})$$

Interactions have been interpreted and discussed in papers by Ashby[4] and McGill[2].

2.2. Approximate conversions of discrete to continuous distributions and vice versa

It is frequently convenient to replace a continuous distribution $p(X)$ on a continuous variable X by a discrete distribution $P(Y)$ $(= \frac{1}{N}\ \underline{N}\ (Y))$ on a discrete variable Y, or to do the reverse. This is because some operations are easier in the discrete domain, some easier in the continuous domain. The problem we attack in this section is, what is the relationship between the entropy of the original distribution and the entropy of the [approximately] transformed distribution? In effect we are looking for a bridge across the gap between continuous - and discrete - variable information theories, a bridge allowing transformations in either direction. We shall show that if the transformation is done with care, the entropies of the original distribution and of its transform differ only by a constant and that transmissions and interactions are unaffected by the transformation.

2.2.1. Transforming a continuous distribution to a discrete distribution

Let $S_c = \left\{ X_1, X_2, \ldots, X_M \right\}$ be a set of continuous variables for which the probability distribution is $p(X_1, X_2, \ldots, X_M) = p(S_c)$, and suppose that for each X_i in S_c, $p(X_i)$ is finite within an interval I_i of finite length L_i and is zero (or may be so approximated) outside I_i. Thus,

$$\int_{I_i} p(X_i) dX_i \cong 1$$

(I_i need not be a connected interval.) Let I_i be divided into N_i subintervals $I_{i1}, I_{i2}, \ldots, I_{iN_i}$, each of length L_i/N_i.

Within the space whose edges are $I_{1j_1}, I_{2j_2}, \ldots, I_{Mj_M}$, a total probability of

$$P(j_1, j_2, \ldots, j_M) = \int_{I_{1j_1}} \int_{I_{2j_2}} \cdots \int_{I_{Mj_M}} p(S_c) dX_1 dX_2 \ldots dX_M$$

is enclosed; the average value of the probability density within that space is

$$\bar{p}(S_c) = \frac{P(j_1, j_2, \ldots, j_M)}{\int_{I_{1j_1}} \int_{I_{2j_2}} \cdots \int_{I_{Mj_M}} dX_1 dX_2 \ldots dX_M} = \frac{P(j_1, j_2, \ldots, j_M)}{V_o}$$

where $V_o = \left(\frac{L_1 \, L_2 \ldots L_M}{N_1 \, N_2 \ldots N_M} \right)$.

If in each such space $p(S_c)$ is replaced by $\bar{p}(S_c)$, the resulting distribution is an approximation to the original, and its quality depends on the numbers N_i, $1 \leq i \leq M$. The entropy of the approximation,

$$H_{appx}(S_c) = -\int_{I_1}\int_{I_2}\cdots\int_{I_M} \bar{\bar{p}}(S_c) \log \bar{\bar{p}}(S_c) dX_1 dX_2 \ldots dX_M$$

will of course equal, in the limit as all N_i go to infinity, the entropy of the original distribution,

$$H(S_c) = -\int_{I_1}\int_{I_2}\cdots\int_{I_M} p(S_c) \log p(S_c) \; dX_1 dX_2 \ldots dX_M.$$

That is,

$$\lim_{\left[\begin{array}{l} N_1 \rightarrow \infty \\ N_2 \rightarrow \infty \\ \ldots \\ N_M \rightarrow \infty \end{array}\right]} H_{appx}(S_c) = H(S_c).$$

Now the numbers $P(j_1, j_2, \ldots, j_M)$ constitute a discrete distribution over a set $S_d = \left\{ Y_1, Y_2, \ldots, Y_M \right\}$ of discrete variables, with Y_i corresponding to X_i:

$$P(Y_1 = j_1, Y_2 = j_2, \ldots, Y_M = j_M) = P(j_1, j_2, \ldots, j_M).$$

The entropy of this discrete distribution is

$$H(S_d) = -\sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2}\cdots\sum_{j_M=1}^{N_M} P(j_1,j_2,\ldots,j_M) \log P(j_1,j_2,\ldots,j_M).$$

Theorem II.1

The relation between $H(S_d)$ and $H_{appx}(S_c)$ is given by

$$H(S_d) = H_{appx}(S_c) + \log \left(\frac{N_1 \; N_2 \; \ldots \; N_M}{L_1 \; L_2 \; \ldots \; L_M} \right).$$

Proof:

Since $\bar{p}(S_c)$ is uniform within each of the volume segments, the integration necessary for finding $H_{appx}(S_c)$ reduces to a summation:

$$H_{appx}(S_c) = -\sum_{j_1=1}^{N_1} \sum_{j_2=1}^{N_2} \cdots \sum_{j_M=1}^{N_M} \left(\frac{P(j_1,j_2,\ldots,j_M)}{V_o} \log \frac{P(j_1,j_2,\ldots,j_M)}{V_o} \right) \cdot V_o$$

$$= -\sum \sum \cdots \sum P(j_1, j_2, \ldots, j_M) \log P(j_1, j_2, \ldots, j_M)$$

$$+ \sum \sum \cdots \sum P(j_1, j_2, \ldots, j_M) \cdot \log V_o$$

$$= H(S_d) + \log V_o.$$

Q. E. D.

Therefore, $H(S_d) \cong H(S_c) + \log \left(\frac{N_1 N_2 \ldots N_M}{L_1 L_2 \ldots L_M} \right)$, with the quality of the approximation depending on the numbers $N_1, N_2, \ldots, N_M$. Clearly this situation holds even when the approximation to $p(S_c)$ varies, within reason, from the rigidly defined $\bar{p}(S_c)$.

As an example, suppose $p(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/8}$ for $X \geqslant 0$; for this distribution $H(X) = 2.04$ bits. See Figure 4. If $p(X)$ is approximated as zero outside the interval $[0, 4) = I$ and the interval is divided into $N = 10$ equal parts, we obtain the following probabilities for the subintervals:

subinterval	probability
[0, 0.4)	.1585
[0.4, 0.8)	.1523
[0.8, 1.2)	.1407
[1.2, 1.6)	.1248
[1.6, 2)	.1064
[2, 2.4)	.0872
[2.4, 2.8)	.0686
[2.8, 3.2)	.0519
[3.2, 3.6)	.0377
[3.6, 4)	.0264

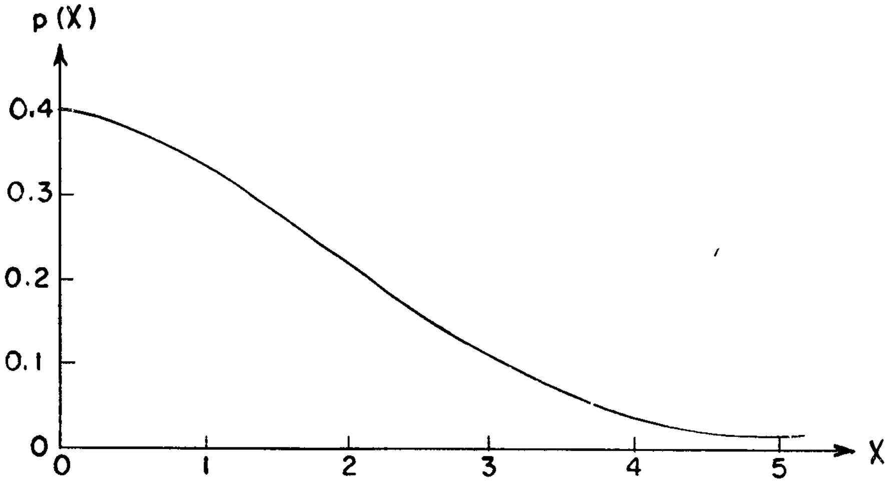


Figure 4.

Calculating $H(S_d)$ with these numbers, and ignoring the fact that they do not total 1.0000, we obtain

$$H(S_d) = 3.14 \text{ bits}$$

and therefore $H(S_c) = H(X) \cong 3.14 - \log\left(\frac{10}{4}\right) = 1.82$ bits.

If the probabilities for the subintervals are not calculated exactly but are only approximated, for instance by multiplying $p(X)$ at one end of the subinterval by $L/N = 0.4$, other estimates for $H(X)$ are obtained.

$$\left\{\text{"Probability" for } [X, X+0.4) = 0.4\ p(X)\right\} \Rightarrow H(X) = 1.85 \text{ bits.}$$

$$\left\{\text{"Probability" for } [X, X+0.4) = 0.4\ p(X+0.4)\right\} \Rightarrow H(X) = 1.78 \text{ bits.}$$

All of these values agree reasonably with the true value of 2.04 bits, considering all the approximations made for the calculation.

2.2.2. Transforming a discrete distribution into a continuous distribution

Given a discrete distribution $P(S_d)$ on set $S_d = \left\{Y_1, Y_2, \ldots Y_M\right\}$, a continuous distribution can be formed by the reverse of the process described above; to do so is of little use, however, unless the continuous distribution thus obtained is subsequently approximated by another continuous distribution which is easier to deal with -- for which integrations are easier, for instance.

2.2.3. The effect of continuous-discrete transformation on transmissions and interactions

The entropy of a continuous distribution and its discrete counterpart differ by a constant (neglecting approximation errors.) Transmissions between continuous variables, and transmissions between their discrete counterparts, are equal; T is unaffected, that is to say, by the transformation. For suppose we have a set of continuous variables, S_c, with a distribution $p(S_c)$, and a corresponding set of discrete variables S_d with the transformed distribution $P(S_d)$:

$$T(S_c) = T(X_1 : X_2 : \ldots : X_M) = H(X_1) + H(X_2) + \ldots + H(X_M) - H(X_1, X_2, \ldots, X_M).$$

$$T(S_d) = T(Y_1 : Y_2 : \ldots : Y_M) = H(Y_1) + H(Y_2) + \ldots + H(Y_M) - H(Y_1, Y_2, \ldots, Y_M).$$

From the theorem,

$$H(Y_1) \cong H(X_1) + \log\left(\frac{N_1}{L_1}\right)$$

$$H(Y_2) \cong H(X_2) + \log\left(\frac{N_2}{L_2}\right)$$

$$\ldots$$

$$H(Y_M) \cong H(X_M) + \log\left(\frac{N_M}{L_M}\right)$$

$$H(Y_1, Y_2, \ldots, Y_M) \cong H(X_1, X_2, \ldots, X_M) + \log\left(\frac{N_1N_2\ldots N_M}{L_1L_2\ldots L_M}\right)$$

Therefore

$$T(S_d) \cong \left[H(X_1) + \log\left(\frac{N_1}{L_1}\right)\right] + \ldots + \left[H(X_M) + \log\left(\frac{N_M}{L_M}\right)\right] - \left[H(X_1, X_2, \ldots X_M) + \log\left(\frac{N_1N_2\ldots N_M}{L_1L_2\ldots L_M}\right)\right]$$

$$\cong T(S_c) + \left[\log\left(\frac{N_1}{L_1}\right) + \ldots + \log\left(\frac{N_M}{L_M}\right) - \log\left(\frac{N_1N_2\ldots N_M}{L_1L_2 \quad L_M}\right)\right]$$

$$\cong T(S_c).$$

Q. E. D.

Interactions, which are defined by differences between transmissions, are therefore also unaffected by the transformation.

2.2.4. General comments on the transformations

Because transformations between discrete and continuous variables and distributions are possible, we do not need to make separate statements for each type but may confine ourselves for the most part to discrete variables, which are generally easier to handle and which fit more readily into the framework of machines-with-input and mappers. When it seems appropriate, we may make explicit statements about the continuous case, but usually that case will be carried along implicitly.

There is usually a certain amount of error involved in approximating a continuous distribution $p(S_c)$ by another, $\overline{p}(S_c)$, which is uniform within each small volume--the more finely the sample space is cut, the smaller will be the error, in general. This error corresponds to "quantization noise," which has been studied elsewhere, and how much error of this type to allow is a pragmatic question which can only be decided from case to case.

Some types of distributions do not allow transformation and in fact are outside the class of distributions information theory can handle, for instance (with μ being the unit step function):

$$p(X) = 0.5\ \delta\ (X-0.5) + 0.5\mu(X)\ -0.5\mu(X-1).$$

See Figure 5.

It is meaningless to talk of H(X) for any distribution which mixes delta "functions" with finite functions.

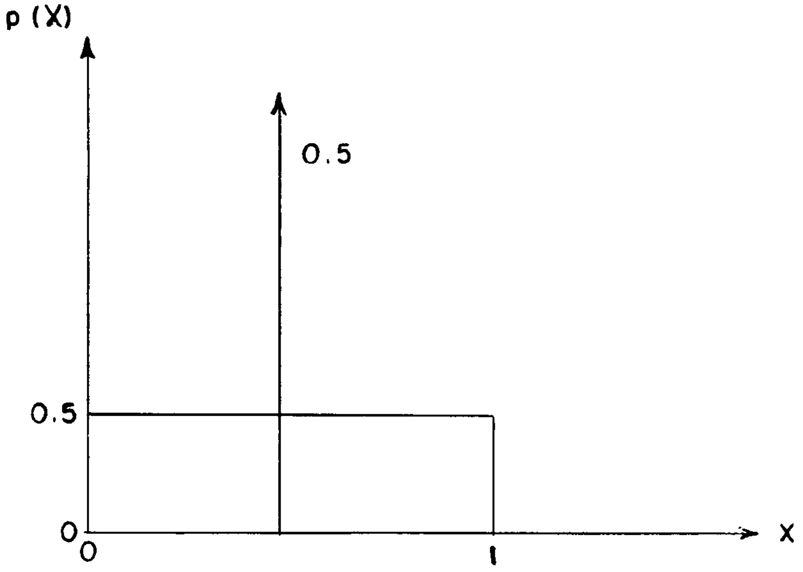


Figure 5.

2.3. Discrete-time convention

Just as it is generally easier to deal with discrete distributions, so is it generally easier to deal with time as a discrete rather than a continuous variable. For one thing, machines-with-input are defined on the basis of discrete time, as are automata, and it is with these that we will deal later. For another, the systems with which one deals in engineering are almost exclusively those for which the approximation of finite bandwidth is appropriate, and to which the Sampling Theorem may therefore be applied to put time on a discrete basis; the errors involved can be made as small as desired by reducing the size of the unit time interval or quantum.

Another reason for treating time as a discrete variable is that we shall frequently be concerned with the values a variable takes over a time span; the value it takes at time τ is in effect a variable; were we to consider all the values over the time span, we should have to deal with an uncountable number of variables and an unmanageable situation. By quantizing the time variable, this problem is avoided.

Finally, much machinery developed for Markov processes is based on the assumption of a discrete time variable, and to take advantage of that machinery we must employ discrete time. So henceforth, unless explicit mention is made to the contrary, we will assume time to be a discrete variable.

III. SOME RESULTS IN INFORMATION THEORY

Introduction

In this chapter we will discuss several results in information theory, whose applications are not limited to the study of complex systems. Since the focus of this paper is on complex systems, the results will be discussed with a bias in that direction, but the results themselves are basically mathematical and applicable to other situations. All of the results, however, are useful in the study of complex systems and find applications, explicitly or implicitly, in the succeeding chapters.

3.1. Operations on the frequency table which leave H, T, and Q unchanged

Given $\underline{N}(S)$, a frequency table for the set of variables S, certain common operations on $\underline{N}$ leave all H's, T's, and Q's unchanged. These are:

1. Permuting the order of the axes (for two variables, transposing $\underline{N}$; for more variables, permuting the order of the variables in $S = \left\{ X_1, X_2, \ldots, X_M \right\}$, which is an ordered set.)
2. Changing the order in which the values for a variable are listed along the axes (for two variables, permuting rows and/or columns.)

3. Multiplication of all the entries in $\underline{N}$ by the same positive constant.

Another operation leaves T's and Q's unchanged but reduces some H's; if there is a variable X_1 in S with two values x_1 and x_1' such that

$$n_{x_1,X_2,\ldots,X_M} = K \cdot n_{x_1',X_2,\ldots,X_M} \quad (K \quad 0)$$

for all values of $X_2, \ldots, X_M$ (for two variables, if two rows or columns are proportional), then $\underline{N}$ may be partially collapsed by summing over those two values, i.e., by setting

$$n'_{x_1,X_2,\ldots,X_M} = n_{x_1,X_2,\ldots,X_M} + n_{x_1',X_2,\ldots,X_M}$$

$$n'_{x_1',X_2,\ldots,X_M} = 0.$$

This last statement is a consequence of the Collapsing Theorem which is proved and discussed in section 3.2 .

We shall use these operations freely in what is to follow, usually without an explicit reminder of their information-preserving property. The fact that variables can be relabeled freely is particularly important in several proofs.

3.2. Collapsing theorems and their consequences

Introduction

The operation of collapsing a frequency table $\underline{N}$ over one of its dimensions, say over the X_M dimension, reduces the H and the T of the table. If $S = \left\{X_1,X_2,\ldots,X_M\right\}$ and $S' = \left\{X_1,X_2,\ldots,X_{M-1}\right\}$ are the original system and the system after collapsing, then

$$H(S') = H(S) - H_{S'}(X_M)$$

$$T(S') = T(S) - T(< X_1, X_2, \ldots, X_{M-1} > : X_M)$$

$$= T(S) - T(S' : X_M)$$

(Ashby[8]), showing that H and T both decline by a nonnegative amount. For interactions,

$$- Q(S') = Q(S) - Q_{X_M}(S)$$

The sign difference between the interaction equation and the others is a consequence of the definition of Q.

The collapse of $\underline{\underline{N}}$ over X_M corresponds to, or implies, complete disregard of the value of X_M; $\underline{\underline{N}}'$, the result, is the table for a system in which X_M is not considered a variable. As such, collapsing is a valuable operation; but what if one wishes to keep X_M as a variable while losing the distinction between *some* of its values? For example, if X_M takes values 1, 2, 3, 4, and 5, one might be interested only in whether the value of X_M is greater than 2, or not. A new variable X'_M with two values could be introduced, related to X by μ,

$$\mu\downarrow \begin{array}{c|ccccc} X_M & 1 & 2 & 3 & 4 & 5 \\ \hline X'_M & 1 & 1 & 2 & 2 & 2 \end{array}$$

and a new system $S' = \{X_1, X_2, \ldots, X_{M-1}, X'_M\}$ defined; this section answers the question of how H(S) and H(S'), T(S) and T(S'), and Q(S) and Q(S') would be related in that case.

From another point of view, this section is important for the situation in which a system (or its frequency table) can be observed only through a mapping which loses information about the variable-values, as would be the case, for example, if an observer were watching the state-changes in a Moore automaton via its many-to-one output function. The Collapsing Theorems give a means of evaluating how much the H's, T's, and Q's would decline (or possibly rise, in the case of interaction) due to the mapping.

3.2.1. Collapsing lemmas

We consider a system S = $\{X, Y\}$ and its frequency table $\underline{N}(X, Y)$ or just $\underline{N}$:

$$\begin{array}{cc|ccccc} & & & & Y & & \\ & \underline{N} & y_1 & y_2 & \cdots & y_{m-1} & y_m \\ \hline & x_1 & n_{11} & n_{12} & \cdots & n_{1,m-1} & n_{1,m} \\ & x_2 & n_{21} & n_{22} & \cdots & & \\ X & \vdots & & & & \vdots & \vdots \\ & x_\ell & n_{\ell 1} & & \cdots & n_{\ell,m-1} & n_{\ell,m} \end{array}$$

We will partially collapse $\underline{N}$ over Y by combining the last two columns, representative of combining any two rows or any two columns (see section 3.1). To this end we define a new variable Z, related to Y by the mapping μ: Y → Z:

$$\mu \downarrow \begin{array}{ccccc} y_1 & y_2 & \cdots & y_{m-1} & y_m \\ \hline z_1 & z_2 & \cdots & z_{m-1} & z_{m-1} \end{array}$$

The frequency table for $S' = \left\{ X, Z \right\}$ is $\underline{N}'(X, Z)$ or just $\underline{N}'$:

Z

	$\underline{N}$	z_1	z_2	$\circ\circ\circ$	z_{m-1}
	x_1	$n_{1,1}$	$n_{1,2}$	$\circ\circ\circ$	$n_{1,m-1}$
	x_2	$n_{2,1}$	$n_{2,2}$		
X	$\vdots$				$\vdots$
	x	$n_{\ell,1}$		$\circ\circ\circ$	$n_{\ell,m-1}$

$\underline{N}$ and $\underline{N}'$ are related by

$$n'_{i,j} = \begin{cases} n_{i,j} & \text{if } j < m-1 \\ n_{i,m-1} + n_{i,m} & \text{if } j = m-1. \end{cases}$$

We denote the sum of the entries in the jth column of $\underline{N}$ by N_j, and of course the sum of the N_j's by N. The entropy of the jth column of $\underline{N}$ will be denoted $H_{y_j}(X)$.

The last two columns of $\underline{N}$ constitute a frequency table $\underline{N}* = \underline{N}*(X*,Y*)$:

Y*

	$\underline{N}*$	y_{m-1}	y_m
	x_1	$n_{1,m-1}$	$n_{1,m}$
	x_2		
X*	$\vdots$	$\vdots$	$\vdots$
	x_ℓ	$n_{\ell,m-1}$	$n_{\ell,m}$

with column entropy H(X*) and row entropy H(Y*).

The transmission in $\underline{N^*}$ is $T(\underline{N^*})$, and the sum of its entries is N^*. The Collapsing Lemma for Transmissions in this simplest case is:

Lemma III.1

$$T(S) - T(S') = \frac{N^*}{N} \; T(\underline{N^*})$$

In words, the transmission lost through partial collapsing is the transmission contained in the frequency subtable which is collapsed, times the relative weight of the subtable.

Proof:

$$T(S) = H(X) - H_Y(X)$$

$$= H(X) - \sum_{j=1}^{m} \frac{N_j}{N} H_{y_j}(X)$$

$$= H(X) - \sum_{j=1}^{m-2} \frac{N_j}{N} H_{y_j}(X) - \frac{N_{m-1}}{N} H_{y_{m-1}}(X) - \frac{N_m}{N} H_{y_m}(X)$$

$$T(S') = H(X) - \sum_{j=1}^{m-2} \frac{N_j}{N} H_{z_j}(X) - \frac{N_{m-1}+N_m}{N} H_{z_{m-1}}(X)$$

$$T(S) - T(S') = \frac{N_{m-1}+N_m}{N} H_{z_{m-1}}(X) - \frac{N_{m-1}}{N} H_{y_{m-1}}(X) - \frac{N_m}{N} H_{y_m}(X)$$

$$= \frac{N_{m-1}+N_m}{N} \left[H_{z_{m-1}}(X) - \frac{N_{m-1}}{N_{m-1}+N_m} H_{y_{m-1}}(X) - \frac{N_m}{N_{m-1}+N_m} H_{y_m}(X) \right]$$

$$= \frac{N^*}{N} \left[H(X^*) - \frac{N_{m-1}}{N^*} H_{y_{m-1}}(X) - \frac{N_m}{N^*} H_{y_m}(X) \right]$$

$$= \frac{N^*}{N} \left[H(X^*) - H_{Y^*}(X^*) \right]$$

$$= \frac{N^*}{N} \; T(X^* : Y^*)$$

Q. E. D.

The Collapsing Lemma for Entropy is

Lemma III.2

$$H(S) - H(S') = \frac{N^*}{N} \; H_{X^*}(Y^*)$$

The entropy lost through partial collapsing over Y is the entropy of Y conditional on X in the subtable being collapsed, multiplied by the relative weight of the subtable.

Proof:
$$H(S) = - \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{n_{i,j}}{N} \log \frac{n_{i,j}}{N}$$

$$= - \sum_{i=1}^{\ell} \sum_{j=1}^{m-2} \frac{n_{i,j}}{N} \log \frac{n_{i,j}}{N} - \sum_{i=1}^{\ell} \left[\frac{n_{i,m-1}}{N} \log \frac{n_{i,m-1}}{N} + \frac{n_{i,m}}{N} \log \frac{n_{i,m}}{N} \right]$$

$$H(S') = - \sum_{i=1}^{\ell} \sum_{j=1}^{m-2} \frac{n'_{i,j}}{N} \log \frac{n'_{ij}}{N} - \sum_{i=1}^{\ell} \frac{n'_{i,m-1}}{N} \log \frac{n'_{i,m-1}}{N}$$

$$H(S) - H(S') = \sum_{i=1}^{\ell} \left[\frac{n_{i,m-1} + n_{i,m}}{N} \log \frac{n_{i,m-1} + n_{i,m}}{N} - \frac{n_{i,m-1}}{N} \log \frac{n_{i,m-1}}{N} - \frac{n_{i,m}}{N} \log \frac{n_{i,m}}{N} \right]$$

$$= \frac{N^*}{N} \sum_{i=1}^{\ell} \left[\frac{n_{i,m-1} + n_{i,m}}{N^*} \log \frac{n_{i,m-1} + n_{i,m}}{N^*} - \frac{n_{i,m-1}}{N^*} \log \frac{n_{i,m-1}}{N^*} - \frac{n_{i,m}}{N^*} \log \frac{n_{i,m}}{N^*} \right]$$

$$= \frac{N^*}{N} \left[-H(X^*) + H(X^*, Y^*) \right]$$

$$= \frac{N^*}{N} \; H_{X^*}(Y^*)$$

Q. E. D.

Extending the system to three variables, $S = \{W, X, Y\}$ and partially collapsing over Y to get $S' = \{W, X, Z\}$, we obtain the collapsing Lemma for Interactions. $\underline{N*} = \underline{N*}(W*, X*, Y*)$ is the three-dimensional analog of $\underline{N*}(X*, Y*)$, and the collapsing is understood to be over Y*, i.e., over y_{m-1} and y_m.

Lemma III.3

$$Q(S) - Q(S') = \frac{N*}{N} \quad Q(\underline{N*})$$

The interaction is lowered by the interaction in $\underline{N*}$, suitably weighted.

Proof:

$$Q(S) = Q(W, X, Y) = T_Y(W : X) - T(W : X)$$

$$Q(S') = Q(W, X, Z) = T_Z(W : X) - T(W : X)$$

$$Q(S) - Q(S') = T_Y(W : X) - T_Z(W : X)$$

$$= \frac{N_{m-1}}{N} \; T_{y_{m-1}}(W : X) + \frac{N_m}{N} \; T_{y_m}(W : X)$$

$$- \frac{N_{m-1} + N_m}{N} \; T_{z_{m-1}}(W : X)$$

$$= \frac{N_{m-1} + N_m}{N} \left[\frac{N_{m-1}}{N_{m-1} + N_m} \; T_{y_{m-1}}(W : X) + \frac{N_m}{N_{m-1} + N_m} \; T_{y_m}(W:X) - T_{z_{m-1}}(W : X) \right]$$

$$= \frac{N*}{N} \left[T_{Y*}(W* : X*) - T(W* : X*) \right]$$

$$= \frac{N*}{N} \quad Q(W*, X*, Y*)$$

$$= \frac{N*}{N} \quad Q(N*)$$

Q. E. D.

Since $Q(\underline{N^*})$ may be either positive, negative, or zero, collapsing does not necessarily lower interaction as it does entropy and transmission.

The lemma for entropy can be rewritten, using the identity

$$H(X, Y) - H(X, Z) \equiv H_X(Y) - H_X(Z).$$

in the form

$$H_X(Y) - H_X(Z) = \frac{N^*}{N} H_{X^*}(Y^*)$$

which makes evident the structural similarity between it and the other lemmas; the form of each is

$$f\left(\begin{array}{l}\text{original}\\ \text{table, } N\end{array}\right) - f\left(\begin{array}{l}\text{table after}\\ \text{collapsing, } N'\end{array}\right) = \frac{N^*}{N}\, f\left(\begin{array}{l}\text{collapsed}\\ \text{subtable, } N^*\end{array}\right)$$

with only the operator f differing between the lemmas.

As an example of partially collapsing a two dimensional table, we collapse $\underline{N}(W, X)$ below over its first two rows, which constitute $\underline{N^*}(W^*, X^*)$, and obtain $\underline{N'}(U, X)$.

Example:

Original table: $\underline{N}(W, X)$:

			X	
		1	2	3
	1	1	2	1
W	2	2	1	2
	3	3	0	0

$H(W, X) = 2.689$ bits

$T(W : X) = 0.367$ bits

$N = 12.$

Collapsed subtable: $\underline{N}^*(W^*, X^*)$:

		X*		
		1	2	3
W*	1	1	2	1
	2	2	1	2

$H_{X^*}(W^*) = 0.918$ bits

$T(W^* : X^*) = 0.074$ bits

$N^* = 9.$

Table after collapsing: $N'(U, X)$:

		X		
		1	2	3
U	1	3	3	3
	2	3	0	0

$H(U, X) = 2.000$ bits

$T(U : X) = 0.311$ bits

$N' = 12.$

The entropies for the three tables are related by

$$H(W, X) - H(U, X) = \frac{N^*}{N} \; H_{X^*}(W^*)$$

$$2.689 - 2.000 = \frac{9}{12} \cdot 0.918 = 0.689$$

and the transmissions are related by

$$T(W : X) - T(U : X) = \frac{N^*}{N} \; T(W^* : X^*)$$

$$0.367 - 0.311 = \frac{9}{12} \cdot 0.074 = 0.056.$$

Collapsing $\underline{N}$ over its first two rows lowers the entropy by 0.689 bits and the transmission by 0.056 bits.

The three lemmas hold also when the subtable is collapsed over more than two Y-values. Suppose a table $\underline{N}(X, Y)$ is to be partially collapsed over its last k columns - the columns for $y_{M+1}, y_{M+2}, \ldots, y_{M+k}$ - to get $\underline{N'}(X, Z)$. This could be done by collapsing the last two columns (which we denote submatrix $\underline{M}^{(1)}$ and whose entries sum to $M^{(1)}$), thus obtaining a new matrix $\underline{N}^{(1)}(X_1, Y_1)$; next collapsing the last two columns of $\underline{N}^{(1)}$ (i.e., submatrix $\underline{M}^{(2)}$) to get $\underline{N}^{(2)}(X_2, Y_2)$; and so on, finally getting $\underline{N}^{(k-1)}(X_{k-1}, Y_{k-1})$ or $\underline{N}'(X, Z)$. $T(X : Y)$ and $T(X : Z)$ would be related by

$$\begin{aligned} T(X : Y) - T(X : Z) &= \left[T(X : Y) - T(X_1 : Y_1)\right] + \left[T(X_1 : Y_1) - T(X_2 : Y_2)\right] \\ &\quad + \ldots + \left[T(X_{k-2} : Y_{k-2}) - T(X : Z)\right] \\ &= \left[\frac{M^{(1)}}{N}\, T(\underline{M}^{(1)})\right] + \left[\frac{M^{(2)}}{N}\, T(\underline{M}^{(2)})\right] \\ &\quad + \ldots + \left[\frac{M^{(k-1)}}{N}\, T(\underline{M}^{(k-1)})\right] . \end{aligned}$$

Consider the first two terms in the summation. They can be combined and rewritten as

$$\frac{M^{(2)}}{N}\left[T(\underline{M}^{(2)}) + \frac{M^{(1)}}{M^{(2)}} \quad T(\underline{M}^{(1)})\right]$$

or, since $M^{(1)}$ is the sum of the entries in the last two columns of N, and $M^{(2)}$ is the sum of the entries in the last three columns of N, this quantity may be written as

$$\frac{N_{m+k} + N_{m+k-1} + N_{m+k-2}}{N}\left[T(\underline{M}^{(2)}) + \frac{N_{m+k} + N_{m+k-1}}{N_{m+k} + N_{m+k-1} + N_{m+k-2}}\, T(\underline{M}^{(1)})\right]$$

The Collapsing Lemma for Transmissions states that this quantity is equal to

$$\frac{\left(\begin{array}{l}\text{sum of the entries} \\ \text{in the last three} \\ \text{columns of } \underline{N}\end{array}\right)}{N} \quad \text{x} \quad \left(\begin{array}{l}\text{Transmission in the} \\ \text{submatrix comprising} \\ \text{the last three columns of } \underline{N}\end{array}\right)$$

An argument by induction leads to the conclusion that

$$T(X : Y) - T(X : Z) = \frac{\left(\begin{array}{c}\text{sum of the entries}\\ \text{in the last k columns}\\ \text{of } \underline{N}\end{array}\right)}{N} \left(\begin{array}{c}\text{Transmission in}\\ \text{the submatrix}\\ \text{comprising the}\\ \text{last k columns}\\ \text{of } \underline{N}\end{array}\right)$$

or, more briefly,

$$T(X : Y) - T(X : Z) = \frac{N^*}{N} \; T(\underline{N^*})$$

where $\underline{N^*}$ is the subtable collapsed, with an arbitrary number of columns.

Arguments identical in form to this one easily show that the Collapsing Lemmas for Entropy and Interaction also hold when the subtables collapsed have an arbitrary number of columns.

3.2.2. Collapsing theorems

These lemmas can be further generalized to a system of many variables, $S = \left\{ X_1, X_2, \ldots, X_M, Y \right\}$, for which the frequency table $\underline{N} = \underline{N}(S)$ is to be partially collapsed over the variable Y, with the table $\underline{N}(S')$ representing the resulting system $S' = \left\{ X_1, X_2, \ldots, X_M, Z \right\}$.

We denote by $\underline{N^*}$ the two-dimensional frequency table, with $< X_1, X_2, \ldots, X_M >^*$ the row-variable and Y^* the column-variable, which is to be collapsed by summing over Y^*.

Theorem III.1 (Collapsing Theorem for Transmission, C.T.T.):

$$T(S) - T(S') = \frac{N^*}{N} \; T(\underline{N^*})$$

$$= \frac{N^*}{N} \; T(< X_1, X_2, \ldots, X_M >^* : Y^*)$$

Proof:

$$T(S) = T(X_1:X_2: \ldots :X_M) + T(<X_1,X_2, \ldots ,X_M>: Y)$$

$$T(S') = T(X_1:X_2: \ldots :X_M) + T(<X_1,X_2, \ldots ,X_M>: Z)$$

$$T(S) - T(S') = T(<X_1, \ldots ,X_M>: Y) - T(<X_1, \ldots ,X_M>: Z)$$

$$= \frac{N^*}{N} T(<X_1, \ldots ,X_M>^* : Y^*)$$

Q. E. D.

The last step follows directly from the Lemma for Transmissions.

The C.T.T. says that if a table is partially collapsed over a variable Y, the total transmission is lowered by the transmission between Y and the rest of the variables, in the collapsed portion, weighted appropriately.

From another point of view, the C.T.T. says that viewing a system through a many-to-one mapping can never increase its apparent constraint; if observer A views a system directly and observer B views it via a mapping, the constraint between variables which is apparent to A is always at least as large as the constraint between the variables' images which is apparent to B.

Theorem III.2 (Collapsing Theorem for Entropy, C.T.E.):

$$H(S) - H(S') = \frac{N^*}{N} H_{<X_1, X_2, \ldots, X_M>^*}(Y^*)$$

Proof:

$$H(S) = H(X_1, X_2, \ldots, X_M) + H_{<X_1, \ldots, X_M>}(Y)$$

$$H(S') = H(X_1, \ldots, X_M) + H_{<X_1, \ldots, X_M>}(Z)$$

$$H(S) - H(S') = H_{<X_1, \ldots, X_M>}(Y) - H_{<X_1, \ldots, X_M>}(Z)$$

$$= \frac{N^*}{N} H_{<X_1, \ldots, X_M>^*}(Y^*)$$

The last step follows from the Lemma for Entropy.

The C.T.E. says that collapsing over part of Y lowers the entropy by the entropy of Y conditional on all the other variables, in the collapsed portion, weighted appropriately.

To obtain the Theorem for Interactions, we assume that $\underline{N*} = \underline{N*}(X_1*, X_2*, \ldots, X_M*, Y*)$ is to be collapsed over part of Y*, that is, over the y-values $y_{m+1}, y_{m+2}, \ldots, y_{m+k}$. We denote the (M + 1)-variable interaction in $\underline{N*}$ by $Q(\underline{N*})$.

Theorem III.3 (Collapsing Theorem for Interaction, C.T.I.):

$$Q(S) - Q(S') = \frac{N*}{N} Q(\underline{N*})$$

Proof:

$$Q(S) = Q_Y(X_1, X_2, \ldots, X_M) - Q(X_1, X_2, \ldots, X_M)$$

$$Q(S') = Q_Z(X_1, X_2, \ldots, X_M) - Q(X_1, X_2, \ldots, X_M)$$

$$Q(S) - Q(S') = Q_Y(X_1, \ldots, X_M) - Q_Z(X_1, \ldots, X_M)$$

$$Q(S) - Q(S') = \sum_{j=m+1}^{m+k} \frac{N_j}{N} Q_{y_j}(X_1, X_2, \ldots, X_M) - \frac{N*}{N} Q_{z_{m+1}}(X_1, X_2, \ldots, X_M)$$

$$= \frac{N*}{N} \left[\sum_{j=m+1}^{m+k} \frac{N_j}{N*} Q_{y_j}(X_1, X_2, \ldots, X_M) - Q_{z_{m+1}}(X_1, X_2, \ldots, X_M) \right]$$

$$= \frac{N*}{N} Q(X_1*, X_2*, \ldots, X_M*, Y*)$$

Q. E. D.

Since interactions may be negative, it is possible for Q(S') to be larger than Q(S), in contrast to the situations for H and T. This means that when a system is viewed through a many-to-one mapping, the interaction terms for the image-system may be larger than those for the original system, i.e., the system may appear to be more complex (in some sense) than it really is.

3.2.3. Remarks on the theorems

At this point it should be made clear that although some of the proofs have been stated in terms of "last rows", "last columns", etc. for notational reasons, and have therefore implied that the frequency tables are finite, minor changes in the proofs would remove that implication; the C.T.T., C.T.E., and C.T.I. apply also to nonfinite tables.

Moreover, each of the theorems has a direct analog in terms of continuous variables. For these, collapsing over certain values of a variable Y becomes integration over an interval of Y, and N*/N becomes the probability of the collapsed portion of the distribution. The only place at which care is needed is in the distribution resulting from the collapsing; the probability which becomes concentrated in the collapsing process must be dispersed in a sheet of finite thickness to avoid a distribution which mixes delta "functions" with finite functions, for information theory cannot handle that mixture.

These three theorems - C.T.T., C.T.E., and C.T.I. - have several corollaries, among them the following:

Corollary III.1

$$\text{a)}\quad T(X_1 : X_2 : \ldots : X_M) = T(X_1 : X_2 : \ldots : X_{M-1}) + T(< X_1, X_2, \ldots, X_{M-1} > : X_M)$$

$$\text{b)}\quad H(X_1, X_2, \ldots, X_M) = H(X_1, X_2, \ldots, X_{M-1}) + H_{< X_1, X_2, \ldots, X_{M-1} >}(X_M)$$

These equations, derived elsewhere in the literature, follow from the C.T.T. and C.T.E. by collapsing over all values of X_M.

The following corollary is a very important one for the decomposition of system constraints, to be studied later. It says, for example, that if $X = < X_1, X_2, \ldots, X_m >$ and $Y = < Y_1, Y_2, \ldots, Y_n >$ are independent, then so are any X_i and Y_j.

Corollary III.2

Let $T(X_1 : X_2 : \ldots : X_M) = 0$, where each X_i is a compound variable $< X_{i1}, X_{i2}, \ldots, X_{in_i} >$. If X_i' designates a compound variable whose components are some or all of the X_{ij}'s, then

$$T(X_1' : X_2' : \ldots : X_M') = 0.$$

Proof:

Suppose $T(X_1 : X_2 : \ldots : X_M) = 0$. The previous corollary implies that

$$T(X_1 : X_2 : \ldots : X_{M-1}) = 0$$

and

$$T(< X_1, X_2, \ldots, X_{M-1} > : X_M) = 0.$$

From the identity $T(X : < Y,Z >) \equiv T(X : Y) + T_Y(X : Z)$ it follows that

$$T(< X_1, \ldots, X_{M-1} > : X_M) = T(< X_1, \ldots, X_{M-1} > : X_M') + T_{X_M'}(< X_1, \ldots, X_{M-1} > : < X_M - X_M' >)$$

(where $<X_M - X'_M>$ is the compound variable whose components are the X_{M_j}'s not in X'_M). The left side of the equation is zero, and therefore

$$T(<X_1, \ldots, X_{M-1}> : X'_M) = 0.$$

Consequently $T(X_1 : X_2 : \ldots X_{M-1} : X'_M) = 0$, for

$$T(X_1 : \ldots : X_{M-1} : X'_M) = T(X_1 : \ldots : X_{M-1}) + T(<X_1, \ldots, X_{M-1}> : X'_M)$$

$$= 0 \qquad\qquad + 0.$$

Similar analysis shows that

$$T(X_1 : X_2 : \ldots : X_{M-2} : X'_{M-1} : X'_M) = 0$$

and so on.

Q. E. D.

The next corollary says, to put it picturesquely, that if an observer of a system can sense only some of the values taken by each variable, all other values registering only as "outside the range of the instruments," then he can at least deduce from his observations some minimum values for the entropy and transmission of the whole system.

Corollary III.3

If $\underline{N}(S)$ is a frequency table and $\underline{N^*}$ is any hyperrectangular portion of it, then

a) $T(\underline{N}) \geqslant \frac{N^*}{N} T(\underline{N^*})$

b) $H(\underline{N}) \geqslant \frac{N^*}{N} H(\underline{N^*})$

Proof:

Suppose a two-variable table $\underline{N}(X, Y)$ is collapsed over the submatrix $\underline{M^*}(X^*, Y^*)$ consisting of the last k_1 columns of $\underline{N}$, the result being $\underline{N}'(X, Z)$. Next suppose $\underline{M^*}$ is collapsed over its submatrix $\underline{N^*}(X^{**}, Y^{**})$ consisting of the last k_2 rows of $\underline{M^*}$, the result being $\underline{M}(W, Y^*)$.

(a) The following two equations follow from the C.T.T.:

$$T(\underline{N}) - T(\underline{N'}) = \frac{M^*}{N} \; T(\underline{M^*})$$

$$T(\underline{M^*}) - T(\underline{M}) = \frac{N^*}{M^*} \; T(\underline{N^*})$$

Therefore,

$$\begin{aligned} T(\underline{N}) &= T(\underline{N'}) + \frac{M^*}{N} \left[T(\underline{M}) + \frac{N^*}{M^*} \; T(\underline{N^*}) \right] \\ &= T(\underline{N'}) + \frac{M^*}{N} \; T(\underline{M}) + \frac{N^*}{N} \, T(\underline{N^*}) \end{aligned}$$

$$T(\underline{N}) \geqslant \frac{N^*}{N} \; T(\underline{N^*})$$

where $\underline{N^*}$ is the rectangular portion of $\underline{N}$ in the last k_1 columns and last k_2 rows. The generalization to more than two variables is obvious, proving part (a).

(b) The following two equations follow from the C.T.E.:

$$H(X, Y) - H(X, Z) = \frac{M^*}{N} \left[H(X^*, Y^*) - H(X^*) \right]$$

$$H(X^*, Y^*) - H(W, Y^*) = \frac{N^*}{M^*} \left[H(X^{**}, Y^{**}) - H(Y^{**}) \right]$$

Therefore,

$$\begin{aligned} H(X, Y) &= H(X, Z) + \frac{M^*}{N} \Big[H(W, Y^*) + \frac{N^*}{M^*} \Big(H(X^{**}, Y^{**}) - H(Y^{**}) \Big) \\ &\qquad\qquad - H(X^*) \Big] \\ &= \left[H(X, Z) - \frac{M^*}{N} H(X^*) \right] + \frac{M^*}{N} \left[H(W, Y^*) - \frac{N^*}{M^*} H(Y^{**}) \right] \\ &\qquad + \frac{N^*}{N} \; H(X^{**}, Y^{**}) \\ &= H(Z) + \left[H_Z(X) - \frac{M^*}{N} H(X^*) \right] + \frac{M^*}{N} \; H(W) \\ &\qquad + \frac{M^*}{N} \left[H_W(Y^*) - \frac{N^*}{M^*} \; H(Y^{**}) \right] \\ &\qquad + \frac{N^*}{N} \; H(\underline{N^*}) \end{aligned}$$

The first bracketed quantity is nonnegative, for $H_Z(X)$ is the average entropy is the columns of $\underline{N}'$, obtained by a weighted summation of the individual column entropies; $\frac{M^*}{N} H(X^*)$ is the last term in the summation, and the first quantity in brackets is thus a weighted sum (of nonnegative quantities) over all but the last column. Therefore it is nonnegative. The second bracketed quantity is nonnegative for similar reasons, and thus

$$H(X, Y) = \left(\text{a nonnegative quantity}\right) + \frac{N^*}{N} H(\underline{N^*})$$

proving part b for the two-variable case. The generalization to more than two variables is simple.

Q. E. D.

3.2.4. The equivalence of transmission and statistical dependence

Corollary III.4, which uses the next Lemma, shows that if a two-dimensional table has zero transmission, its columns are proportional, i.e., that zero transmission implies statistical independence.

Lemma III.4

Let $\underline{N}$ be a 2-by-2 frequency table with $T(\underline{N}) = 0$. Then one column of $\underline{N}$ is a non-negative multiple of the other.

Proof:

The distribution $\underline{N}$ may be typified by

1	a
b	abc

$(c \geq 0)$

The second column is a multiple of the first if $c = 1$. $T(\underline{N})$ can be expressed in terms of a, b, and c as follows.

$$T(\underline{N}) = \frac{1}{1 + a + b + abc} \Big\{ (1 + a + b + abc) \log (1 + a + b + abc)$$
$$+ 1 \log 1 + a \log a + b \log b + abc \log abc$$
$$- (1 + a) \log (1 + a) - (1 + b) \log (1 + b)$$
$$- a(1 + bc) \log a(1 + bc) - b(1 + ac) \log b(1 + ac) \Big\} .$$

Assuming $T(\underline{N}) = 0$, expanding, rearranging, and cancelling, we obtain

$$(1 + a + b + abc) \log (1 + a + b + abc) + abc \log c$$
$$= (1 + a) \log (1 + a) + (1 + b) \log (1 + b)$$
$$+ a(1 + bc) \log (1 + bc) + b(1 +ac) \log (1 + ac).$$

Calling the left side $f(c)$ and the right $g(c)$, this equation $f(c) = g(c)$ has a solution at $c = 1$, i.e., when the second column of $\underline{N}$ is a multiple of the first. To show that there are no other finite solutions, we note that

$$\frac{\partial f(c)}{\partial c} = ab \left\{ 2 \log_2 e + \log_2(c + ac + bc + abc^2) \right\}$$
$$\frac{\partial g(c)}{\partial c} = ab \left\{ 2 \log_2 e + \log_2(1 + ac + bc + abc^2) \right\} .$$

$f(c)$ equals $g(c)$ at $c = 1$, and for $c > 1$, $f(c)$ has a steeper slope than $g(c)$; this implies that $f(c) > g(c)$ for $c > 1$. Similarly, $f(c) < g(c)$ for $c < 1$. Therefore, $c = 1$ is the only finite solution to $f(c) = g(c)$, i.e., to $T(\underline{N}) = 0$.

Q. E. D.

Corollary III.4 (to C.T.T.):

Let $\underline{N}(X : Y)$ be a frequency table with m rows (of x_i) and n columns (of y_j). If $T(\underline{N}) = 0$, then the columns of $\underline{N}$ are all nonnegative multiples of $\underline{N}(X)$. Thus zero transmission implies statistical independence.

Proof:

If $\underline{N}$ has zero-rows or zero-columns, they may be permuted to the bottom and the right, and columns may then be permuted to put a positive element in the (1, 1) position; this permuted form of $\underline{N}$ we call $\underline{N}'$. Clearly if the Corollary is true for $\underline{N}'$, it is true for $\underline{N}$. Suppose $T(\underline{N}') = T(\underline{N}) = 0$.

Corollary III.3 says that the upper left 2-by-2 submatrix of $\underline{N}'$ (in fact, *any* rectangular submatrix) has zero transmission. The last Lemma says that the columns of this submatrix are proportional, i.e., that the elements in the second column are k_{12} times their rowmates in the first column, with $k_{12} > 0$. The same argument shows that in the submatrix of rows 2 and 3 and columns 1 and 2, the same proportionality holds, and so on for all elements in columns 1 and 2; all elements in column 2 are k_{12} times their rowmates in column 1. Similarly, the elements in column 3 are k_{23} times their rowmates in column 2, and so on. Finally, each of the columns is proportional to the column-table $\underline{N}'(X)$ formed by collapsing $\underline{N}'$ over its rows.

Q. E. D.

Of course if $\underline{N}(X, Y)$ has proportional columns it also has proportional rows; this condition is equivalent to statistical independence of X and Y.

It is well known that if X and Y are statistically independent variables, $T(X : Y) = 0$. Corollary III.4 shows that the converse also holds; that if $T(X : Y) = 0$, then X and Y are statistically independent. Thus transmission and statistical dependence are *equivalent* concepts couched in different languages.

The argument easily generalizes to many variables; if T(S) = 0, then any subset of variables in S is independent of any other (disjoint) subset.

If the frequency table on hand is the record of an actual experiment, the transmission must of course be interpreted in light of the vagaries of random sampling. To date an adequate test for the significance level of T has not been produced.

3.3. Can genuinely complex relationships be broken down?

If a system contains many variables interacting in a complex way, it is frequently impossible for a human observer to keep track of all of them simultaneously. When this happens, it is common for the human to observe a few variables at a time and then try to piece together the behavior of the whole from those observations. Such an attempt sometimes succeeds and sometimes fails; we want to ask if there is any theoretical limitation on such an attempt, specifically with regard to the information-theoretic quantities involved.

To put the question vividly: suppose an observer capable of observing any N or fewer variables at a time is faced with a system of N + 1 variables. Can he deduce the entropy, total transmission, or highest-order interaction of the system? To approach the problem we define a few terms.

By a simple expression we will mean a single entropy, transmission, or interaction term explicitly involving variables - e.g., H(X),

$T_Z(<W, X> : Y)$, $Q_{X_1}(<X_1, X_2>, X_3, X_4)$. An expression is a sum of simple expressions.

Any simple expression is either identically zero (such as $T_X(X:Y)$) or may be reduced to a proper simple expression, in which no variable appears explicitly in both subscript and argument; for example, the third example above is identically equal to $Q_{X_1}(X_2, X_3, X_4)$, which is proper. The order of a simple expression is zero if the expression is identically zero; otherwise it is equal to the number of distinct variables appearing explicitly in the expression, whether or not they are considered to be components of compound variables. The examples above have orders one, four, and four. The order of an expression is the largest of the orders of its simple expressions.

It would be useful to find order-reducing identities - identities which would express a simple expression as a sum of lower-order expressions, thereby allowing one to view a complex relationship as merely a summation of simpler relations. This is indeed possible through the device of an auxiliary equation; e.g., if $<X,Y> = W$ then $H(X, Y) \equiv H(W)$. However, barring the use of auxiliary equations, no order-reducing identity can exist; relationships which genuinely involve many variables can not be broken down.

Theorem III.4

Let $f \equiv g$ be an identity in which f is a simple expression of finite order M and in which g is an expression of order $K \leq M$ (and involving the same variables). Then $K = M$, i.e., g contains a simple expression of order M.

Proof:

(a) We first prove the theorem when f is an unsubscripted entropy, $f = H(X_1, X_2, \ldots, X_M)$, by supposing $K < M$ and obtaining a contradiction. We define two distributions on $S = \left\{X_1, X_2, \ldots, X_M\right\}$, where each X_i has two values, 1 and 2.

The first, $\underline{N}(S)$, is defined by

$$n_{X_1,X_2,\ldots,X_M} = 2\left[(X_1 + X_2 + \ldots + X_M), \bmod 2\right]$$

and the second, $\underline{N}'(S)$, is defined by

$$n'_{X_1,X_2,\ldots,X_M} = 1.$$

For example, with $M = 3$ they are as follows:

$$\underline{N}\left\{\begin{array}{ccc}
\begin{array}{cc|cc|} & & \multicolumn{2}{c}{X_1} \\ & & 1 & 2 \\ \hline & 1 & 2 & 0 \\ X_2 & & & \\ & 2 & 0 & 2 \\ \hline \end{array} & \quad & \begin{array}{cc|cc|} & & \multicolumn{2}{c}{X_1} \\ & & 1 & 2 \\ \hline & 1 & 0 & 2 \\ X_2 & & & \\ & 2 & 2 & 0 \\ \hline \end{array} \\
X_3 = 1 & & X_3 = 2
\end{array}\right.$$

$$\underline{N'}\left\{\begin{array}{ccc}
\begin{array}{|cc|} \hline 1 & 1 \\ 1 & 1 \\ \hline \end{array} & \quad & \begin{array}{|cc|} \hline 1 & 1 \\ 1 & 1 \\ \hline \end{array}
\end{array}\right.$$

To calculate any simple expression involving fewer than M variables necessitates collapsing $\underline{N}$ and $\underline{N'}$ over the variables omitted; when thus collapsed, $\underline{N}$ and $\underline{N'}$ yield identical distributions and consequently identical values for g. The two distributions yield different values for f, however - an impossible condition if $f \equiv g$ is an identity.

(b) If f is any simple expression of order M, identities of the following form exist[8]:

$$f \equiv h \pm H(X_1, X_2, \ldots, X_M)$$

where h is an expression of order less than M. Thus $f \equiv g$ may be rewritten as

$$\pm H(X_1, X_2, \ldots, X_M) \equiv g - h.$$

Part (a) showed that the expression on the right is of order M; since the order of h is less than M, the order of g must be M.

Q. E. D.

The theorem does not say that both sides of any identity must have equal order, and in fact that is not true; for example,

$$H(X,Y) - H_X(Y) \equiv H(X).$$

It does mean that if a set of variables are actually related in a holistic manner, the relation cannot be broken into a sum of simpler relations without something being lost. While this is perfectly true in general, in many cases of practical interest a high-order relation can be broken down without losing "too much." In section 4.3 we will study systems which lend themselves to such decompositions.

3.4 Maximizing transmission between related variables

Introduction

An important problem is the following. Suppose X and Y are variables taking values from sets $X = \{ x_i \mid 1 \leq i \leq m \}$ and $Y = \{ y_j \mid 1 \leq j \leq n \}$, and suppose $R \subset X \times Y$ is a relation between X and Y. How should the frequencies in $\underline{N}(X, Y)$ be distributed exclusively over the couples in R so that $T(X : Y)$ is maximized? In other words, how can the transmission be maximized with respect to the constraint R?

While this is an interesting problem in its own right, the answer is really crucial for the understanding of channel capacity. For as will be explained in the section on that topic, the description of a channel linking supervariables $\overline{X}$ and $\overline{Y}$ is in fact the description of a relation between $\overline{X}$ and $\overline{Y}$, and the problem of maximizing $T^L(\overline{X} : \overline{Y})$ (i.e., finding the channel capacity) is the same as the problem considered here, only with limits involved.

It will be shown in the chapter on regulation that the transmission between the regulator, R, and the variable it is regulating against, X, is of prime importance in regulation. This section is therefore also of importance to regulation, particularly when there is a relation between R and X.

3.4.1. The theorem

We start by denoting the matrix version of R by

$\underline{R} = \left[r_{ij}\right]_{m,n}$ with

$$r_{ij} = \begin{cases} 1 \text{ if } <x_i, y_j> \text{ is in R,} \\ 0 \text{ otherwise.} \end{cases}$$

We consider here only frequency matrices $\underline{N}(X,Y) = \left[n_{ij}\right]_{m,n}$ compatible with R, i.e., such that couples not in R occur with zero frequency. Nothing is lost by restricting attemtion to cases in which $m \leq n$ and $\underline{R}$ has no zero-rows or zero-columns. Since the argument involves permutations of the rows and columns of $\underline{N}$ and $\underline{R}$, it will be assumed henceforth that when one matrix is permuted, the other is permuted in the same way. We denote a permuted form of a matrix with primes.

For every R, there is at least one "largest one-to-one mapping" μ having the following properties:

i) $\mu \subset R$,

ii) μ has domain $Z \subset X$, where Z contains k elements and $k \leq m$,

iii) μ maps Z one-to-one onto a subset of Y,

iv) no other mapping exists which obeys (i), (ii), and (iii) but on a larger domain than μ.

The number k, giving the number of elements in μ's domain, is dictated by R and may be denoted k(R).

The distribution $\underline{N_o}$, with

$$n_{ij} = \begin{cases} 1 \text{ if } \langle x_i, y_j \rangle \text{ is in } \mu \\ 0 \text{ otherwise,} \end{cases}$$

gives $T(\underline{N_o}) = \log k(R)$. It is always possible to make $T(X : Y) = \log k(R)$, by assigning equal frequencies to the couples in μ; however, by that assignment it is possible that certain values of X and Y, not excluded by R, would be assigned zero frequency. Consequently H(X) and H(Y) would be lower with the assignment $\underline{N_o}$ than with some other distributions, and since

$$T(X : Y) = H(X) + H(Y) - H(X, Y)$$

there is good reason to suspect that some distribution other than $\underline{N_o}$ will maximize the transmission.

The answer to the question posed above is given by:

Theorem III.5

Suppose R is a subset of X x Y. Then for any $\underline{N}(X, Y)$ compatible with R,

$$T(\underline{N}) \leq \log k(R).$$

and thus $\underline{N_o}$ above maximizes $T(\underline{N})$.

To state the theorem somewhat picturesquely, X and Y can communicate best through a one-to-one mapping, even if the price of the biuniqueness is that some of their values never get used. It doesn't pay, as far as transmission is concerned, to introduce more values if their introduction brings in ambiguity.

Proof:

If k = m, the theorem is obviously true since $T(\underline{N}) \leq \log m$ for any distribution $\underline{N}$; the smaller dimension of a matrix limits the transmission. If k < m, we need the following Lemma:

Lemma III.5

If k < m, $\underline{R}$ may be permuted to a form $\underline{R}^*$, which in partitioned form is

$$\underline{R}^* = \left[\begin{array}{c|c|c} \underline{A} & \underline{B} & \underline{C} \\ \hline \underline{D} & \underline{E} & \underline{F} \\ \hline \underline{G} & \underline{H} & \underline{I} \end{array}\right]$$

and in which the square submatrix $(\underline{A}, \underline{B}, \underline{D}, \underline{E})$ has an ascending diagonal of k(R) 1's and the submatrix $(\underline{E}, \underline{F}, \underline{H}, \underline{I})$ is a zero matrix.

Proof of the Lemma:

The mapping μ prescribes in a natural way a permutation of $\underline{R}$ which displays an ascending diagonal of k(R) 1's across the upper left corner of the resulting matrix, these 1's corresponding to couples in the set μ. Pictorially, $\underline{R}'$ is then as shown in Figure 6, with the diagonal line representing a string of 1's.

The submatrix $\underline{L}$ must be a zero matrix, because if there were a 1 in $\underline{L}$, row and column permutations could append it to the existing diagonal. Henceforth we will show zero matrices by shading.

The rows which contain 1's in $\underline{J}$ may be moved to the top of $\underline{R}'$, and appropriate column permutations, always possible, may be performed to preserve the diagonal of 1's intact. This done, $\underline{R}''$ is as shown in Figure 7, where $\underline{J_1}$ has no zero rows. Now $\underline{K_2}$ must be a zero matrix, since otherwise a column permutation could put a 1 in $\underline{L}$ while preserving the diagonal.

Next, the columns which contain 1's in $\underline{K_1}$ may be moved to the left and appropriate row permutations performed to preserve the diagonal. This gives $\underline{R}''$, shown in Figure 8, in which $\underline{K_3}$ has no zero columns.

The process is now repeated with $\underline{M}$, $\underline{N}$, and $\underline{P}$ playing the parts of $\underline{J}$, $\underline{K}$, and $\underline{L}$; $\underline{P}$ must be a zero matrix, for if it were not, a sequence of column permutations could put a 1 in $\underline{L}$ while preserving the diagonal. If there are no rows with 1's in $\underline{M}$, the Lemma is satisfied; if there are such rows, they may be moved to the top of $\underline{M}$ and the diagonal may be preserved through column permutations. Next the columns with 1's in $\underline{N}$, if any, are moved to the left side of $\underline{N}$ while preserving the diagonal, giving $\underline{R^{(4)}}$, shown in Figure 9, where $\underline{J_1}$ and $\underline{M_1}$ have no zero rows, and $\underline{K_3}$ and $\underline{N_3}$ have no zero columns.

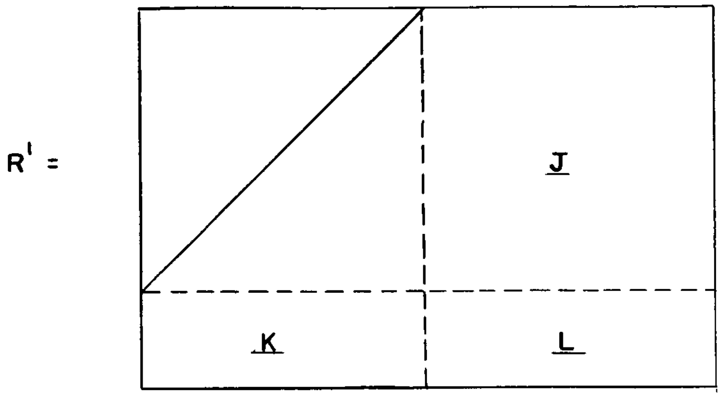


Figure 6.

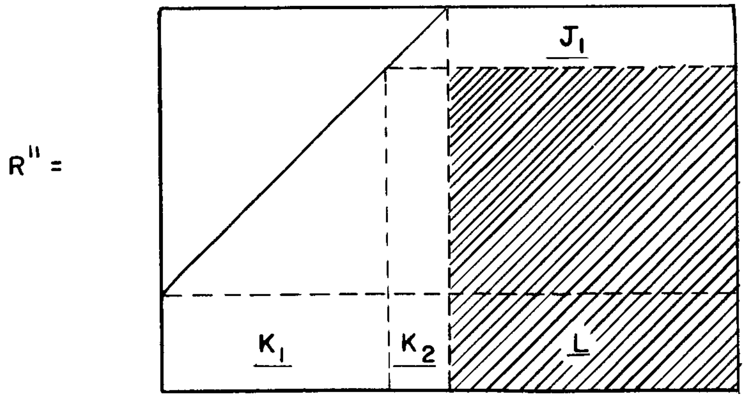


Figure 7.

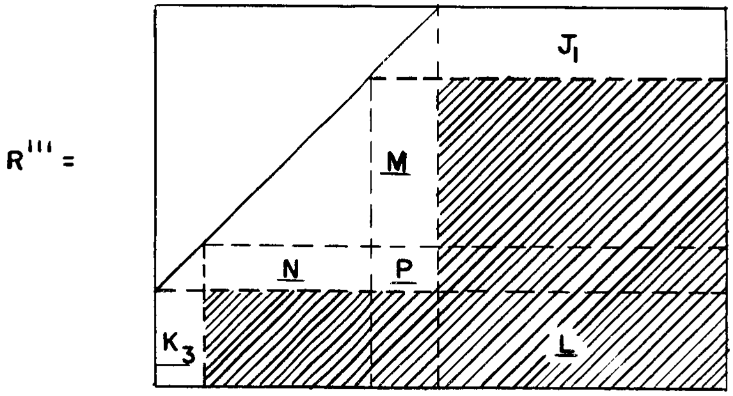


Figure 8.

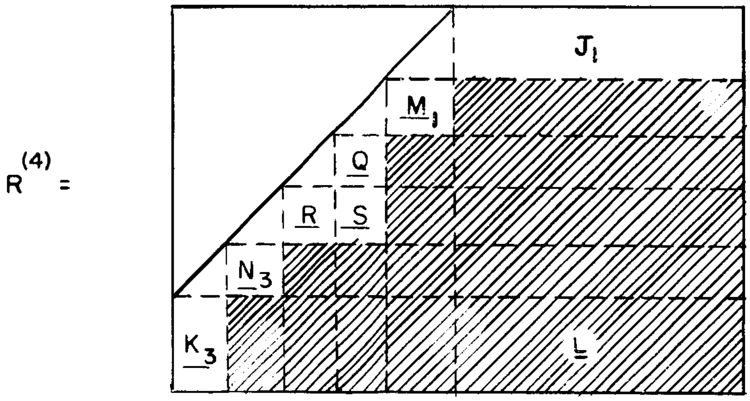


Figure 9.

This process, iteratively applied to any matrix $\underline{R}'$, must end either by disclosing a 1 in a position incompatible with the hypothesis that the diagonal is maximally long, or else by completion of a rectangle of zeroes which "touches" the diagonal. This proves the Lemma.

Note that the process described amounts to an effective procedure for finding the largest one-to-one mapping contained in R (or one of them, if there are more than one).

Returning to the proof of the theorem, we assume $\underline{R}$ has been permuted to the standard form $\underline{R}^*$, and that the distribution $\underline{N}$, unspecified as yet, has been similarly permuted (so that it too has the large rectangle of zeroes).

$$\underline{N}^* = \left[\begin{array}{c|c|c} \underline{N_{11}} & \underline{N_{12}} & \underline{N_{13}} \\ \hline \underline{N_{21}} & \underline{N_{22}} & \underline{N_{23}} \\ \hline \underline{N_{31}} & \underline{N_{32}} & \underline{N_{33}} \end{array}\right] \quad \text{where} \quad \left[\begin{array}{c|c} \underline{N_{22}} & \underline{N_{23}} \\ \hline \underline{N_{32}} & \underline{N_{33}} \end{array}\right] = \underline{0}.$$

Suppose now that $\underline{N}^*$ is partially collapsed by adding together the columns in the right-hand submatrices, obtaining $\underline{N_a}$:

$$\underline{N_a} \quad \left[\begin{array}{c|c} \underline{N_{11}} & \underline{M_1} \\ \hline \underline{N_{21}} & \underline{0} \\ \hline \underline{N_{31}} & \underline{0} \end{array}\right] \quad \text{where} \quad \left[\begin{array}{c} \underline{M_1} \\ \hline \underline{0} \\ \hline \underline{0} \end{array}\right] \quad \text{has one column.}$$

We recall that permutations do not alter transmission; therefore $T(\underline{N^*}) = T(\underline{N})$. The C.T.T. (theorem III.1) consequently states that

$$T(\underline{N}) = T(\underline{N_a}) + \frac{N_{12} + N_{13}}{N} \quad T\left(\left[\begin{array}{c|c} \underline{N_{12}} & \underline{N_{13}} \\ \hline \underline{0} & \underline{0} \\ \hline \underline{0} & \underline{0} \end{array}\right] \right)$$

$$= T(\underline{N_a}) + \frac{N_{12} + N_{13}}{N} \quad T\left(\left[\begin{array}{c|c} \underline{N_{12}} & \underline{N_{13}} \end{array}\right] \right) .$$

Suppose we are given an arbitrary $\underline{N_a}$ and we set about to maximize $T(\underline{N})$ by adjusting the frequencies in $\underline{N_{12}}$ and $\underline{N_{13}}$. The row sums are fixed (by $\underline{M_1}$, which is in $\underline{N_a}$). Recall that $\underline{B}$, the submatrix of $\underline{R}$* corresponding to $\underline{N_{12}}$, has an ascending diagonal of 1's; hence, the row totals for $(\underline{N_{12}}, \underline{N_{13}})$ can be assigned to the diagonal positions in $\underline{N_{12}}$. That assignment maximizes $T(\underline{N})$ without assigning any frequency to $\underline{N_{13}}$. If $\underline{N_{13}}$ is not needed for an arbitrary $\underline{N_a}$, it is not needed for the $\underline{N_a}$ which maximizes $T(\underline{N})$, i.e., there is an $\underline{N}$ which maximizes $T(\underline{N})$ and for which $\underline{N_{13}} = 0$.

The last conclusion is the heart of the proof, for maximizing $T(\underline{N})$ when $\underline{N}$ is

$$\underline{N} = \left[\begin{array}{c|c|c} \underline{N_{11}} & \underline{N_{12}} & \underline{0} \\ \hline \underline{N_{21}} & \underline{0} & \underline{0} \\ \hline \underline{N_{31}} & \underline{0} & \underline{0} \end{array}\right]$$

is easily accomplished by setting

$$n_{ij} = \begin{cases} 1 \text{ if } n_{ij} \text{ is on the ascending diagonal,} \\ \quad \text{i.e., if } i + j = 1 + k(R), \\ 0 \text{ otherwise,} \end{cases}$$

yielding $T(\underline{N}) = \log k(R)$.

Q. E. D.

3.4.2. An attempt to generalize the theorem

Let R be a system relation on $S = \{ X_1, X_2, \ldots, X_i, \ldots, X_M \}$. Then R contains a largest subset μ such that (i) for every X_i in S, the projection mapping pr_i maps μ one-to-one onto a subset of X_i, and (ii) no other mapping satisfies (i) and has more elements than μ. Letting k(R) denote the number of elements in μ, it is tempting to conjecture, as an M-dimensional generalization of the above theorem, that for any $\underline{N}(S)$ compatible with R,

$$T(\underline{N}) \leq (M-1) \log k(R).$$

However, the generalization does not always hold for $M > 2$. For example with $R = S = \{X, Y, Z\}$ the following $\underline{N}(X, Y, Z)$ has $T(\underline{N}) = 3$, but $(M-1) \log k(R) = 2 \log 2 = 2$.

$$\begin{array}{cc|cccc|}
 & & \multicolumn{4}{c}{X} \\
 & & 1 & 2 & 3 & 4 \\ \cline{3-6}
 & 1 & 0 & 0 & 0 & 0 \\
Y & 2 & 0 & 0 & 0 & 0 \\
 & 3 & 0 & 1 & 0 & 0 \\
 & 4 & 1 & 0 & 0 & 0 \\ \cline{3-6}
 & & \multicolumn{4}{c}{Z = 1}
\end{array}
\qquad\qquad
\begin{array}{cc|cccc|}
 & & \multicolumn{4}{c}{X} \\
 & & 1 & 2 & 3 & 4 \\ \cline{3-6}
 & 1 & 0 & 0 & 0 & 1 \\
Y & 2 & 0 & 0 & 1 & 0 \\
 & 3 & 0 & 0 & 0 & 0 \\
 & 4 & 0 & 0 & 0 & 0 \\ \cline{3-6}
 & & \multicolumn{4}{c}{Z = 2}
\end{array}$$

After introducing some new notations to deal with dynamic variables, we will apply the results of this section to the problem of finding the channel capacity for networks of automata.

3.5. Information quantities for dynamic variables

Introduction

We normally think of a dynamic variable, e.g., $X(t)$, as one which changes in time. A semantic and notational confusion results when we wish to consider both (1) the variable $X(t_o)$, i.e., the variable whose values are the possible values of X at the specific time t_o (but with t_o arbitrary), and (2) the variable $X(t)$, i.e., the variable whose values are the possible trajectories X can take over an extended time interval. To distinguish the instantaneous - from the trajectory-variables, we call the first simply a variable and the second a super-variable. The two are of course related, and in this section we will explore that relation as regards the information quantities involved.

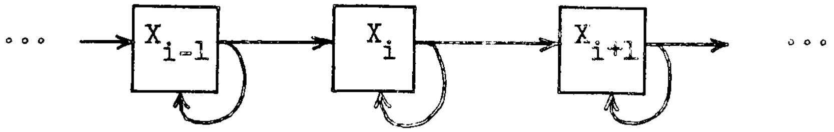
In later sections on channel capacity, information transfer, and regulation we shall rely heavily on the concepts of this chapter.

3.5.1. Definitions for limit-quantities

It frequently happens that a system $S = \left\{ X_1, X_2, \ldots \right\}$ is composed of variables all having the same statistical distribution, or is composed of groups of variables, all within each group having the same distribution. A stationary regular Markov sequence

$$\ldots, X^{i-1}, X^{i}, X^{i+1}, \ldots$$

where the superscripts denote successive instants in time, and the states in a chain of identical MWI'S,



where the subscripts denote successive positions in the chain, are one-dimensional examples. For such a system, certain limit-expressions are meaningful and have profound interpretations in the study of complex systems. We will denote these limit-expressions with a superscript L. At the start, a word about notation is in order. We will use subscripts in this section and elsewhere to distinguish variables or super-variables which are being thought of as different in nature; we will use superscripts, on the other hand, as indices for time. For example, X_1 and X_2 might be a set of temperature-values and a set of humidity-values respectively; the variable "temperature at time τ" would be denoted X_1^{τ} and the variable "humidity at time 7" would be denoted X_2^7.

To simplify notation, we define the super-variable $\overline{X}$ or the s-variable $\overline{X}$ as follows:

$$\overline{X} = < X^1, X^2, \ldots, X^i, \ldots >$$

$\overline{X}$ corresponds to an indefinitely long strip of a protocol,

time:	1	2	3	4	5	...	i	...
$\overline{X}$:	X^1	X^2	X^3	X^4	X^5	...	X^i	

and one value of $\overline{X}$ is one possible way to fill in the protocol. Of course X^{τ} may have components, say if $X^{\tau} = < U^{\tau}, V^{\tau} >$; then $\overline{X} = \overline{<U,V>}$ is a supervariable with components.

We define a super-system $\overline{S}$ as an ordered set of super-variables:

$$\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$$

It is important to distinguish $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2 \right\}$, a supersystem of supervariables, from $\overline{<X_1, X_2>}$, a supervariable with components. The latter corresponds to a protocol of two strips,

$$\begin{array}{|c|c|c|c|c|}\hline X_1^1 & X_1^2 & X_1^3 & X_1^4 & X_1^5 \\ \hline X_2^1 & X_2^2 & X_2^3 & X_2^4 & X_2^5 \\ \hline \end{array}$$

whereas the former corresponds to a set of protocols:

$$\left\{ \begin{array}{|c|c|c|c|}\hline X_1^1 & X_1^2 & X_1^3 & X_1^4 \\ \hline \end{array} \, , \; \begin{array}{|c|c|c|c|}\hline X_2^1 & X_2^2 & X_2^3 & X_2^4 \\ \hline \end{array} \right\}.$$

Thus the prefix super- or s- and the overbar imply variables which are really infinite vector-variables. We will use the term "variable", henceforth, to include both ordinary- and super-variables, using the prefix only when a super-variable is expressly implied; likewise the term "system" will include both types, so that a super-system is also a system.

We denote the limit-entropy of $\bar{X}$ or L-entropy of $\bar{X}$ by $H^L(\bar{X})$ and define it by

$$H^L(\bar{X}) = \lim_{n\to\infty} \frac{1}{n} H(X^1, X^2, \ldots, X^n)$$

if the limit exists. The n-th term in the sequence is what Shannon[5] calls G_N. Similarly, the limit-entropy of a super-system $\bar{S} = \left\{ \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_M \right\}$ is defined as

$$H^L(\bar{S}) = \lim_{n\to\infty} \frac{1}{n} H\Big(<X_1^1, X_2^1, \ldots, X_M^1>, <X_1^2, X_2^2, \ldots, X_M^2>, \ldots, <X_1^n, X_2^n, \ldots, X_M^n> \Big)$$

$$= \lim_{n\to\infty} \frac{1}{n} H(<S^1>, <S^2>, \ldots, <S^n>).$$

The notation is slightly redundant in that L-entropies are defined only for s-variables, but this redundancy will be kept, for emphasis. Continuing the definitions in their general version, we define the L-entropy of $\overline{S}_a$ conditional on $\overline{S}_b$ by

$$H^L_{\overline{S}_b}(\overline{S}_a) = H^L(\overline{S}_a \cup \overline{S}_b) - H^L(\overline{S}_b).$$

And so on. The definitions for all simple limit expressions, except for $H^L(\overline{X})$ and $H^L(\overline{S})$ which are primary expressions, are obtained from the analogous non-limit definitions by superscripting with L, and overlining all variables. For example, the L-transmission over $\overline{S} = \left\{\overline{X}_1, \overline{X}_2, \ldots \overline{X}_M\right\}$ is defined by

$$T^L(\overline{S}) = H^L(\overline{X}_1) + H^L(\overline{X}_2) + \ldots + H^L(\overline{X}_M) - H^L(\overline{X}_1,\overline{X}_2,\ldots,\overline{X}_M).$$

By a simple limit-expression we will mean a single L-entropy, L-transmission, or L-interaction term explicitly involving s-variables, e.g., $H^L(\overline{X})$, $T^L_{\overline{X}}(\overline{W}:\overline{Y})$. A limit-expression is a sum of simple limit-expressions.

3.5.2. The relation between non-limit identities and limit-identities

One of the post powerful theorems in information theory is the one which states that an identity in simple expressions remains an identity if the same subscript is added to each simple expression[9].

The reader might be tempted to suppose that an identity in simple non-limit expressions remains an identity if each term is superscripted with L and all variables are overlined. Since the definitions for all L-transmissions and L-interactions are related to the non-limit definitions by precisely that operation, the supposition is clearly true for identities not involving entropies. If entropies are involved, however, the supposition is by no means obviously true, for a limit-identity has on its two sides the limits of two distinct sequences, and to establish the identity these limits must be shown to be equal.

Theorem III.6

An identity in simple expressions remains an identity if superscript L is added to each simple expression in it and every variable is overlined. That is, every non-limit identity implies a corresponding limit-identity.

Proof:

Let $f \equiv g$ be an identity in non-limit expressions, involving variables $X_1, X_2, \ldots, X_M$:

$$f(X_1, X_2, \ldots, X_M) \equiv g(X_1, X_2, \ldots, X_M).$$

Substituting $<X_1^1, X_1^2, \ldots, X_1^n>$ for X_1, $<X_2^1, \ldots, X_2^n>$ for X_2, etc., and $<X_M^1, X_M^2, \ldots, X_M^n>$ for X_M, another identity is obtained:

$$f\left(<X_1^1,\ldots,X_1^n>,\ldots,<X_M^1,\ldots,X_M^n>\right) \equiv g(<X_1^1,\ldots,X_1^n>,\ldots, <X_M^1,\ldots,X_M^n>).$$

The identity is preserved if both sides are divided by n; therefore, for all $n \geqslant 1$ we have

$$\frac{1}{n}\ f\left(<X_1^1,\ldots,X_1^n>,\ldots,<X_M^1,\ldots,X_M^n>\right) \equiv \frac{1}{n}\ g\left(<X_1^1,\ldots,X_1^n>,\ldots, <X_M^1,\ldots,X_M^n>\right).$$

Our goal is to show that $f^L(\overline{X}_1,\overline{X}_2,\ldots,\overline{X}_M)$ and $g^L(\overline{X}_1,\overline{X}_2,\ldots,\overline{X}_M)$ are identically equal; each of these limit-expressions represents the limit of a sequence, and to show them equal we must show that the two sequences converge to the same limit. That they do follows from the fact that the two sequences are equal in every term, and that is the case since in the last identity above, the expression on the left is just the n-th term in the sequence whose limit is $f^L(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M)$ while the expression on the right is the n-th term in the sequence whose limit is $g^L(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M)$.

Q. E. D.

Deeper exploration of limit-expressions and their profound impprtance for complex systems will be deferred to a later section; here it will suffice to state that $H^L(\overline{X})$ is the information (per step) carried in the sequence $\{X\}$ and $T^L(\overline{X} : \overline{Y})$ is a measure of the linkage between the sequences $\{X\}$ and $\{Y\}$, per step. When $\overline{X}$ is the input and $\overline{Y}$ is the output of an information channel, $T^L(\overline{X} : \overline{Y})$ is the amount of information usually thought of as "transferred through" the channel, and it is bounded by the channel capacity. We will take up the subject of channel capacity in the following section.

3.6. Channel capacity, constraint capacity, and the capacity of automata

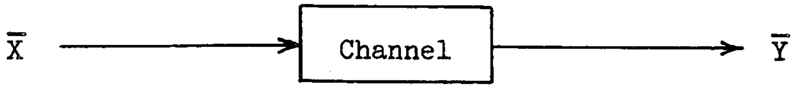
Introduction

The notion of channel capacity is one of the most fundamental in information theory. It applies, classically, to an "input-output" system and is the limit on how much information can be pushed through it per unit time. We will show here that the notion need not be restricted to "input-output" systems nor to systems with only two "terminals;" the generalized notion will be referred to as constraint capacity, to eliminate the connotation of unidirectional flow that the word "channel" carries. Constraint capacity will reappear in a later chapter, when we discuss the decomposition of constraints in a dynamic system, as an upper bound for the linkage between two or several dynamic variables in a dynamic system.

In later chapters on regulation in dynamic systems, it will become apparent that the channel capacity of a regulator is of fundamental importance for its capacity as a regulator. Since a regulator is not always describable as either a machine with input or a mapper alone but can usually be described as an automaton, the calculation of the capacity of automata is of prime interest to this study, and a method is presented in this chapter by which that calculation can be made. The method allows calculation of the capacity for any network of interconnected automata, in fact, and it produces as a by-product the information necessary to construct a source matched to the network so as to realize the maximum information flow.

3.6.1. Channel capacity and constraint capacity

We consider a super-system $\overline{\overline{S}} = \overline{\overline{X}},\overline{\overline{Y}}$ in which $\overline{X}$ is the input s-variable for a channel and $\overline{Y}$ is the output s-variable:



A particular value of $\overline{X} = < X^1, X^2, \ldots >$ is a particular sequence of input symbols to the channel, X^i being the input symbol at time i. The channel specification is in fact specification of a relation R in the product set $\overline{X} \times \overline{Y}$, and the channel capacity is defined by

$$C = \max \left\{ T^L(\overline{X} : \overline{Y}) \right\}$$

where the maximum (or least upper bound, if there is no maximum) is over the various distributions $\underline{N}(\overline{X}, \overline{Y})$ compatible with R.

For many channels of practical interest, the order of maximization and limit-taking may be inverted, giving

$$C = \lim_{n \to \infty} \frac{1}{n} \left[\max T(< X^1, X^2, \ldots, X^n > : < Y^1, Y^2, \ldots Y^n >) \right]$$

The maximization is that considered in section 3.4, namely maximizing transmission under constraint by a relation.

The relation specified by a deterministic input-output channel is normally a mapping from $\overline{X}$ (and perhaps the channel's initial state) into $\overline{Y}$; for such a channel, $H^L_{\overline{X}}(\overline{Y}) = 0$ and therefore

$$C = \max \left\{ H^L(\overline{Y}) \right\} .$$

The characterization of the channel as "input-output" derives from the relation R, not from $\overline{X}$ or $\overline{Y}$. By considering arbitrary relations on arbitrarily many super-variables, we can generalize C to the notion of "constraint capacity" of an object. Supposing there is a super-system $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$, and the object specifies a relation R;

$$R \subset \overline{X}_1 \times \overline{X}_2 \times \ldots \times \overline{X}_M,$$

the constraint capacity of the object is denoted C and defined by

$$C = \max \left\{ T^L(\overline{S}) \right\}$$

with the maximum (or l. u. b.) taken over all the possible distributions $\underline{N}(\overline{S})$ compatible with R.

It may strike the reader as presumptous to speak of a relation in a set of infinite size. In practice, of course, R is usually a highly iterated version of a very simple relation on a finite set. For example, if $\overline{X}$ and $\overline{Y}$ are the input and state supervariables for a MWI with mapping $f : X^i \times Y^i \rightarrow Y^{i+1}$, then

$$\langle \overline{X}, \overline{Y} \rangle \text{ is in } R \Leftrightarrow \text{ for every } i \geqslant 1, \langle X^i, Y^i, Y^{i+1} \rangle \text{ is in } f,$$

where f is viewed as a relation in $(X^i \times Y^i) \times Y^{i+1}$. R is thus shown to be an expanded version of the three-variable relation f.

The treatment thus far has not differentiated between "noisy" and "noiseless" channels. That topic will be taken up in section 3.6.4.

3.6.2. An example of constraint capacity

As an example of constraint capacity in more than two dimensions (variables), we define a relation R on $\overline{S} = \left\{ \overline{X}, \overline{Y}, \overline{Z} \right\}$, where each of the s-variables takes, at each step, one of the values 1, 2, or 3:

$$\langle \overline{X}, \overline{Y}, \overline{Z} \rangle \in R \Leftrightarrow \text{for every } i \geqslant 1, X^i, Y^i, \text{ and } Z^i \text{ all take different values, and } Y^i > Z^i \text{ if } i \text{ is even, } Y^i < Z^i \text{ if } i \text{ is odd.}$$

This is equivalent to

$$\langle \overline{X}, \overline{Y}, \overline{Z} \rangle \in R \Leftrightarrow \text{if } i \text{ is even, } \langle X^i, Y^i, Z^i \rangle \text{ is } \langle 2, 3, 1 \rangle, \langle 3, 2, 1 \rangle, \text{ or } \langle 1, 3, 2 \rangle; \text{ if } i \text{ is odd, } \langle X^i, Y^i, Z^i \rangle \text{ is } \langle 3, 1, 2 \rangle, \langle 1, 2, 3 \rangle, \text{ or } \langle 2, 1, 3 \rangle.$$

The distribution $\underline{N}(X^i, Y^i, Z^i)$, with

$$n_{3,2,1} = n_{1,3,2} = 1; \text{ others zero}$$

when i is even, and

$$n_{3,1,2} = n_{1,2,3} = 1; \text{ others zero}$$

when i is odd, maximizes both $T(X^i : Y^i)$ and $T(<X^i, Y^i> : Z^i)$; therefore it maximizes $T(X^i : Y^i : Z^i)$, at 2 bits. The extension of that distribution maximizes $T^L(\overline{X} : \overline{Y} : \overline{Z})$ at 2 bits/unit time, so the constraint capacity associated with R is 2 bits/unit time. The relation represents a real constraint, since with no constraint ($R = \overline{S}$), the constraint capacity would be log 9 = 3.17 bits/unit time.

3.6.3. Channel capacity of Moore automata

3.6.3.1. The theorem

Viewing the object (the "channel") as a set relation has led to the solution of an outstanding problem - that of finding the channel capacity of an arbitrarily connected network of MWI's, mappers, and Moore automata.

Consider a finite network of arbitrarily interconnected Moore automata, as in Figure 10 where the circles represent automata and an arrow from one circle to another indicates that the output symbols from the first automaton are input symbols to the second. Further suppose that the network acts as a communication channel from a Source to a Receiver, the "input automaton" accepting only Source symbols as input and the Receiver observing the output symbols of the "output automaton" only. This section will provide a procedure for evaluation of the channel capacity of such a network and of its component automata.

There is no loss of generality in assuming that only one automaton accepts inputs from outside of the network, that there is

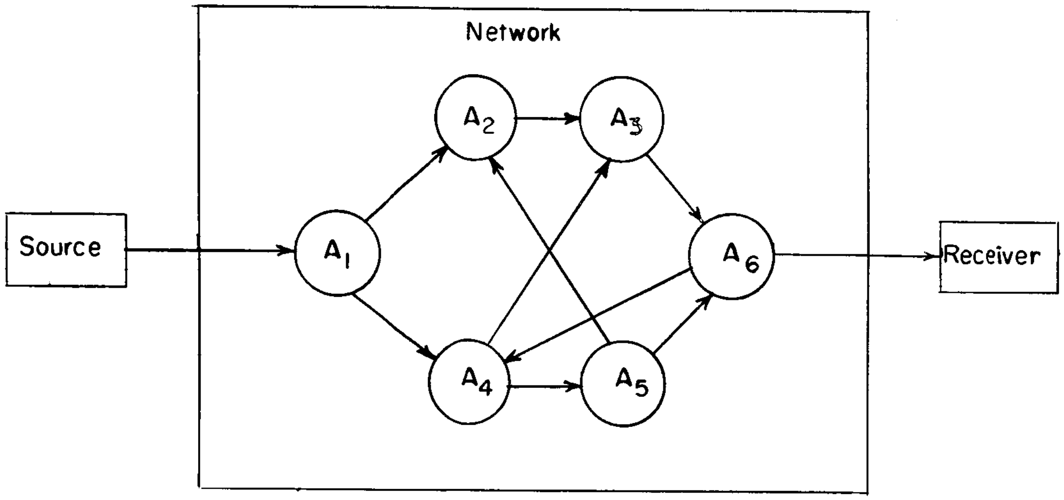


Figure 10.

only one Source, that the input automaton accepts only Source symbols as input, or that the Receiver observes only one automaton; all other cases may be reduced to this one by nominally combining elements, recoding the descriptions of elements, or introducing one "delay automaton." None of these modifications affects the channel capacity of the network.

The network itself may be viewed as a Moore automaton, of course, so that the problem of finding the capacity of a network reduces to that of finding the capacity of a single automaton. On the other hand, each arrow in Figure 10 can be thought of as a unidirectional channel and may be labeled with its channel capacity, which is the capacity of the automaton from which the arrow emanates. One upper bound for the network capacity[10] is the minimum value among all simple cut sets, where the cut sets separate the "input automaton" from the Receiver and where the value of a cut set is the sum of the capacities of branches in the set (but only counting branches directed from the input toward the receiver). Thus the calculation of this upper bound for network capacity also requires the calculation of capacities of single automata, to which we now turn. The method, in essence, is an application of theorem III.5, setting the input and output sequences in biunique correspondence.

We consider a Moore automaton A with a finite input alphabet $\left\{x_1, x_2, \ldots, x_k\right\} = X$, a finite state set $\left\{s_1, s_2, \ldots, s_m\right\} = S$, a finite output set $\left\{y_1, y_2, \ldots, y_n\right\} = Y$, a state function f: $X \times S \rightarrow S$, and an output function g: $S \rightarrow Y$. See Figure 11.

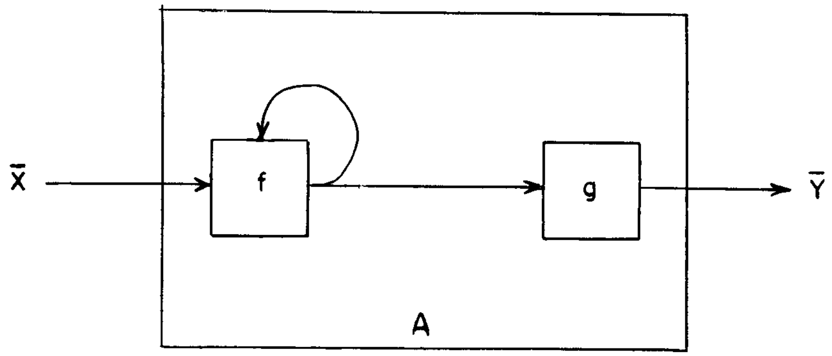


Figure 11.

The state-transition matrix $\underline{\Lambda} = [\lambda_{ij}]_{m,m}$ for A is defined by

$$\lambda_{ij} = \begin{cases} 1 & \text{if } \exists x \in X \text{ s.t. } f(x, s_i) = s_j \\ 0 & \text{otherwise} \end{cases}$$

and the related matrices $\underline{\Lambda_p} = \left[\lambda_{ij_p}\right]_{m,m}$, $1 \leq p \leq n$, by

$$\lambda_{ij_p} = \begin{cases} \lambda_{ij} & \text{if } g(s_i) = y_p \\ 0 & \text{otherwise} \end{cases}$$

Row s_i of $\underline{\Lambda}$ indicates with a 1 every state-transition $s_i \rightarrow s_j$ allowed by f, and $\underline{\Lambda_p}$, $1 \leq p \leq n$, copies those rows of $\underline{\Lambda}$ representing states which g maps to y_p.

For a discrete channel such as A,

$$C = \lim_{T\to\infty} \frac{1}{T} \left[\max T(<X_1, X_2, \ldots, X_T> : <Y_1, Y_2, \ldots, Y_T>) \right].$$

There is at least one sequence $\{X_1, X_2, \ldots, X_T\}$ for each sequence $\{Y_1, Y_2, \ldots, Y_T\}$, and from Theorem III.5 it follows that

$$C_y = \lim_{T\to\infty} \frac{\log N_y(T)}{T}$$

where $N_y(T)$ is the number of output-sequences of length T allowed by the input and the set relation prescribed by A. Shannon gives the expression above as the definition of C for a discrete channel[5].

We denote by $N_s(T)$ the number of state-sequences of length T; $N_s(T)$ and $N_y(T)$ yield capacities C_s and C_y respectively. C_y is the capacity of A.

C_s may be calculated from $\underline{\Lambda}$ by a method due to Shannon; he shows[5] that if $\underline{\Lambda}$ represents the allowed state-transitions, $\underline{I}$ the identity matrix, and W_o the largest real root of the determinantal equation

$$\det \left[\underline{\Lambda} - W\underline{I} \right] = 0$$

then C_s is given by

$$C_s = \log_2 W_o.$$

If g is a one-to-one mapping, each state-sequence yields exactly one output-sequence; in such a case $N_s(T)$ equals $N_y(T)$ for all T, $C_s = C_y$, and the capacity of A may be calculated directly from $\underline{\Lambda}$. If g is not one-to-one the convergence introduced by g will force $N_y(T)$ to be smaller than $N_s(T)$. To find C_y in such a case we systemmatically substitute new automata A', A", etc., with their relations R', R", etc. in $\overline{X} \times \overline{Y}$ being each a proper subset of its predecessor, until an automaton A* is found for which

$$\lim_{T\to\infty} \frac{1}{T} \log N_{s*}(T) = \lim_{T\to\infty} \frac{1}{T} \log N_y(T).$$

That is, $$C_{s*} = C_y.$$

The sequence of automata A, A', A", ..., A* can be formed in such a way that the state-transitions become increasingly constrained while the output-transitions do not, so that C_y may be found from the state-transition matrix for A*, which we will call $\underline{\Lambda}*$.

We define a parallel set P as a set containing two or more state-subsequences of the form

$$\left\{ S_i, S_\alpha, S_\beta, \ldots, S_{i+n} \right\} \quad (n \geq 2)$$

all compatible with $\underline{\Lambda}$, all identical in first and last states, and all of which are mapped by g into the same output-subsequence.

If a parallel set P exists, an observer seeing only the corresponding output-subsequence is unable to determine which state-subsequence in P has caused it, but the observer's uncertainty can be minimized as follows. Given A, one can generate all the state-sequences of length T allowed by $\underline{\Lambda}$. If a parallel set P is found, the constraints on state-transitions can be increased, eliminating members of P until exactly one sequence in P remains allowed; this is always possible, and it amounts to the substitution of a new automaton A' capable of the same number of output sequences as A but a smaller number of state-sequences. One can next generate all the state-sequences of length T allowed for A', and so on. Reiteration of this process will eliminate all parallel sets of length T and will lead to a collection of no more than $m^2N_y(T)$ state-sequences, since for each first-state, last-state pair (of which there are at most m^2) an observer of the output-sequence (of which there are $N_y(T)$) would correctly assign one state-sequence. Moreover, the collection will contain no fewer than $N_y(T)$ sequences, since the elimination process always leaves, for each allowed output-sequence of A, one state-sequence capable of generating it. This process, then provides a sequence of numbers, $N_o(T)$, which give a capacity C_o:

$$C_o = \lim_{T\to\infty} \frac{1}{T} \log N_o(T).$$

From the inequality

$$N_y(T) \leq N_o(T) \leq m^2N_y(T) \quad \text{for all } T \geq 1$$

it follows that $C_y = C_o$. Since C_o may be found from $\underline{\Lambda}*$ by Shannon's method, the foregoing justifies the following theorem:

Theorem III.7

Let W_o be the largest real root of the determinantal equation $\det \left[\underline{\Lambda}* - W\underline{I} \right] = 0$. Then the capacity of A is $\log W_o$.

$\underline{\Lambda}$* embodies the original state-transition constraints and the ones introduced by the elimination procedure, at the point where no further elimination is necessary.

This calls for several comments. First, unless the transition eliminated is a first-order one (e.g., $s_1 \rightarrow s_5$) the states must be recoded and the transition matrix redrawn before the elimination can be made. For example, elimination of a third-order transition (e.g., $< s_2, s_4, s_1 > \rightarrow s_5$) requires that the states be recoded into triples (e.g., $(s_2, s_4, s_1) = s_{241}$) and that the corresponding matrix be constructed before elimination of the transition (e.g., $s_{241} \rightarrow s_{415}$). Corresponding changes in the domain and range of g must be made. The effect of this relabeling is to increase the size of the matrix at each step unless certain simplifications are possible; in the Example, some common simplifications will be illustrated.

Second, if at the Mth iteration of the process the matrix, call it $\underline{\Lambda(M)}$, has become too large to make continuation feasible, an approximation to C_o can be obtained by using $\underline{\Lambda(M)}$ in place of $\underline{\Lambda}$*; such an approximation, C_M, satisfies the inequalities

$$C_y \leq C_M \leq C_{M-1} \leq C_S \qquad (M \geq 1).$$

Finally, there exists a procedure, given below, for deciding whether or not further eliminations are necessary, i.e., whether or not $\underline{\Lambda(M)} = \underline{\Lambda}$*.

We proceed next to outline the process in terms of matrix operations.

3.6.3.2. Calculation of capacity

Sets X, S, and Y and functions f and g are presumed given. As the iterations proceed to substitute new automata for the original, S, Y, f, and g will change accordingly. To simplify the notation we will assume, however, that S has m elements and Y has n ($m > 1$, $n > 1$) at the start of each iteration, signaled by a pass through Step 1, and we will call the transition matrix $\underline{\Lambda}$ throughout.

Preliminary

If S can be partitioned into disjoint subsets such that no state in any subset has any transition to any state in another subset, then A is a merely nominal conjunction of smaller automata, one of which is selected by choice of the initial state. The capacity of A is then the largest of the capacities for the smaller automata.

Transient states, which cannot be reached from any other state, as well as persistent states, which cannot lead to any state other than themselves, may be dropped from S without affecting the capacity. If S is empty after all such states have been dropped, the automaton has a capacity of zero.

Construct $\underline{\Lambda}$ and $\underline{\Lambda_p}$: $1 \leq p \leq n$ as previously defined.

Step 1.

Observe the $\underline{\Lambda_p}$ matrices to see if there exists any column of any $\underline{\Lambda_p}$ containing more than a single 1. If so, proceed to Step 2. If not, no further eliminations are necessary, as the comments for Step 2 will explain; proceed to Step 5.

Comment on Step 2

The successive postmultiplications of a row vector $\underline{E_j}$ (with e_{1j} equal to 1 and the other elements all zero) by $\underline{\Lambda}$, $\underline{\Lambda_{p_2}}$, $\underline{\Lambda_{p_3}}$, ..., $\underline{\Lambda_{p_T}}$ corresponds to the construction of state-sequences starting with s_j and passing through states in the sets $g^{-1}(y_{p_2})$, $g^{-1}(y_{p_3})$, ..., $g^{-1}(y_{p_T})$. For $\underline{E_j \Lambda}$ indicates by its nonzero components the set of states reached in one step from s_j, $\underline{E_j \Lambda \Lambda_{p_2}}$ indicates those states reached in two steps from s_j via some s in $g^{-1}(y_{p_2})$, and so on. If a vector component equal to K $>$ 1 results from the multiplication, there must exist a related parallel set containing K sequences. Conversely, if a parallel set never occurs, it must be the case that no vectors ever arise from the multiplications which, when multiplied by any $\underline{\Lambda_p}$, yield a vector component greater than 1. Clearly, if no column of $\underline{\Lambda_p}$ contains more than a single 1, multiplication of a vector of zeroes and ones by $\underline{\Lambda_p}$ can give rise only to components of zero and one.

Step 2.

Define T_1, a set of row vectors, as follows:

$$T_1 = \left\{ \underline{V_1}, \underline{V_2}, \ldots, \underline{V_m} \right\} \text{ where } \underline{V_i} = \left[\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im} \right].$$

Start the following substeps with N = 1.

Step 2a.

Generate the set of vectors $\underline{Q_N} = \left\{ \underline{V_i \Lambda_p} \mid 1 \leq p \leq n, \underline{V_i} \in T_N \right\}$. For N = 1, these vectors are simply the rows of the matrices $\underline{\Lambda\Lambda_1}$, $\underline{\Lambda\Lambda_2}$, ..., $\underline{\Lambda\Lambda_n}$.

If any vector in Q_N has a component greater than 1, go to Step 3. If none has, go to Step 2b.

Step 2b.

Form the set $T_{N+1} = T_N \cup Q_N$. If $T_{N+1} = T_N$, go to Step 5. If $T_{N+1} \neq T_N$, increase N by 1 and return to Step 2a.

Comment on Step 3

Entry to Step 3 results from the production of at least one vector in Q_N containing, in say its jth column, a number K greater than 1. The vector, produced on the Nth pass through Step 2a, corresponds to the existence of a parallel set P containing K distinct state-sequences, each of length N + 2 and each ending with s_j. All but one of the sequences in P must be eliminated. To every component greater than 1, of every vector in Q_N, there corresponds such a parallel set requiring eliminations.

Step 3.

Find the parallel sets by retracing the steps of multiplication which led to the vectors in question and by consulting the function g. Once the sets are known, all but one member in each set must be declared examples of illegitimate transitions (of order N + 1). Rewrite the transition matrix to show the previously allowed transitions of order N + 1 and modify it (by substituting zeroes for the ones corresponding to the newly illegal transitions) to form the state-transition matrix $\underline{\Lambda}$ for Step 4. S, Y, f, and g must be modified to reflect the relabeling of states described earlier.

Step 4.

Remove transient, persistent, and isolated states from S as follows. If there exists a state s_k in S such that row s_k or column s_k in $\underline{\Lambda}$ contains only zeroes, except perhaps on the main diagonal, remove s_k from S and revise $\underline{\Lambda}$ accordingly. Continue removing states and revising $\underline{\Lambda}$ until every row and column contains at least one off-diagonal 1.

From the resulting $\underline{\Lambda}$ and g, construct the $\underline{\Lambda_p}$ matrices and return to Step 1.

Comment on Step 5

Entry to Step 5 indicates that the state-transitions, as represented by the current $\underline{\Lambda}$, are sufficiently constrained as to guarantee that

$$N_y(T) \leq N_o(T) \leq m^2 N_y(T)$$

for all T. Thus the current $\underline{\Lambda}$ is $\underline{\Lambda}*$.

Step 5.

Solve the equation

$$\det \left[\underline{\Lambda}* - W\underline{I} \right] = 0$$

for its largest real root W_o; calculate $C = \log_2 W_o$ = capacity of A.

The state-transition probabilities which maximize the output entropy at C bits per second are given by

$$\text{Prob } (s(t+1) = s_j \mid s(t) = s_i) = P_{ij} = \frac{B_j}{B_i} \cdot \frac{\lambda_{ij}}{W_o}$$

in which $\underline{B}$ is the eigenvector associated with the eigenvalue W_o in the equation

$$\left[\underline{\Lambda}* - W\underline{I} \right] \underline{B} = \underline{0}.$$

This result is from Shannon[5], and it leads easily to the construction of a source which is optimal for the channel.

3.6.3.3. An example

This example will illustrate how the process typically proceeds and what simplifications are often possible. Let A be an automaton described by sets $X = \{x_1, x_2, x_3\}$, $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$, $Y = \{y_1, y_2, y_3\}$ and functions f and g given in Table I.

Next - state function				Output function	
f	x_1	x_2	x_3	g	
s_1	s_2	s_3	s_5	s_1	y_1
s_2	s_3	s_2	s_3	s_2	y_2
s_3	s_1	s_2	s_3	s_3	y_2
s_4	s_2	s_3	s_1	s_4	y_1
s_5	s_5	s_5	s_5	s_5	y_3
s_6	s_1	s_4	s_4	s_6	y_3

TABLE I. - State and output functions of A.

Preliminary. State s_6 cannot be entered from any $s \in S$, so it can be dropped; with s_6 gone, s_4 cannot be entered, so it can be dropped. State s_5 cannot be abandoned once entered, so it can be dropped; note that this means that the couple (s_1, x_3) must never be allowed to arise. With $S = \{s_1, s_2, s_3\}$ we can proceed.

$$\underline{\Lambda} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad \underline{\Lambda_1} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \underline{\Lambda_2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Step 1. $\underline{\Lambda_2}$ contains columns with more than one 1.

Step 2. $T_1 = \left\{ \underline{V_1}, \underline{V_2}, \underline{V_3} \right\}$ with $\underline{V_1} = \underline{V_2} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$ and $\underline{V_3} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$.

2a. $$\underline{\Lambda\Lambda_1} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\underline{\Lambda\Lambda_2} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

The rows of $\underline{\Lambda\Lambda_1}$ and $\underline{\Lambda\Lambda_2}$ are the vectors in Q_1.

Step 3. To each 2 in the matrix product there corresponds a parallel set containing two sequences, and if the 2 is in the (i, j) position of $\underline{\Lambda\Lambda_p}$, the sequences must start with s_i, pass through an s in $g^{-1}(y_p)$, and end with s_j, since

$$\left[\text{row } i \text{ of } \underline{\Lambda\Lambda_p}\right] = \underline{E_i}\ \underline{\Lambda\Lambda_p}.$$

The parallel sets, subscripted with i and j, are as follows:

$$P_{12} = \left\{ (s_1,s_2,s_2),\ (s_1,s_3,s_2) \right\}$$

$$P_{13} = \left\{ (s_1,s_2,s_3),\ (s_1,s_3,s_3) \right\}$$

$$P_{22} = \left\{ (s_2,s_2,s_2),\ (s_2,s_3,s_2) \right\}$$

$$P_{23} = \left\{ (s_2,s_2,s_3),\ (s_2,s_3,s_3) \right\}$$

$$P_{32} = \left\{ (s_3,s_2,s_2),\ (s_3,s_3,s_2) \right\}$$

$$P_{33} = \left\{ (s_3,s_2,s_3),\ (s_3,s_3,s_3) \right\}.$$

The second order transition matrix, after relabeling states as indicated on page 73 of the text, is given in tabular form below.

s(t + 1)

		s_{31}	s_{12}	s_{22}	s_{32}	s_{13}	s_{23}	s_{33}
	s_{31}	0	1	0	0	1	0	0
	s_{12}	0	0	1	0	0	1	0
	s_{22}	0	0	1	0	0	1	0
s(t)	s_{32}	0	0	1	0	0	1	0
	s_{13}	1	0	0	1	0	0	1
	s_{23}	1	0	0	1	0	0	1
	s_{33}	1	0	0	1	0	0	1

The elimination of a sequence from a parallel set P is accomplished by substituting a zero for the corresponding 1 in this matrix. The sequence in P to be eliminated may be selected arbitrarily, although a good choice will minimize the subsequent computations. We choose in this Example to eliminate the following sequences:

$$s_1, s_3, s_2;\ s_1, s_3, s_3;\ s_2, s_3, s_2$$

$$s_2, s_2, s_3;\ s_3, s_3, s_2;\ s_3, s_2, s_3$$

(this is in fact not the best choice). The result is given below.

s(t + 1)

		s_{31}	s_{12}	s_{22}	s_{32}	s_{13}	s_{23}	s_{33}
	s_{31}	0	1	0	0	1	0	0
	s_{12}	0	0	1	0	0	1	0
	s_{22}	0	0	1	0	0	0	0
s(t)	s_{32}	0	0	1	0	0	0	0
	s_{13}	1	0	0	0	0	0	0
	s_{23}	1	0	0	0	0	0	1
	s_{33}	1	0	0	0	0	0	1

$$S = \left\{ s_{31}, s_{12}, s_{22}, s_{32}, s_{13}, s_{23}, s_{33} \right\}.$$

Step 4. Observation of column s_{32} and row s_{22} indicates that s_{32} and s_{22} can be eliminated from S. Frequently the second-order transition matrix at this point is merely an expanded version of a first-order matrix, allowing a further simplification, but in this Example that is not the case. Table II gives the matrix, in tabular form, resulting from the foregoing eliminations and also redefines the output function g on the relabeled states.

State Transitions		s(t + 1)					Output function g	
		s_{31}	s_{12}	s_{13}	s_{23}	s_{33}		
s(t)	s_{31}	0	1	1	0	0	s_{31}	y_{21}
	s_{12}	0	0	0	1	0	s_{12}	y_{12}
	s_{13}	1	0	0	0	0	s_{13}	y_{12}
	s_{23}	1	0	0	0	1	s_{23}	y_{22}
	s_{33}	1	0	0	0	1	s_{33}	y_{22}

TABLE II. State transitions and output functions after simplification

$$S = \left\{ s_{31}, s_{12}, s_{13}, s_{23}, s_{33} \right\}.$$

$$\underline{\Lambda} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \underline{\Lambda_{12}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\underline{\Lambda_{21}} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{\Lambda_{22}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

With these matrices we return to Step 1.

Step 1. $\underline{\Lambda\Lambda_{22}}$ contains columns with more than one 1.

Step 2. $T_1 = \left\{ \underline{V_1}, \underline{V_2}, \underline{V_3}, \underline{V_4}, \underline{V_5} \right\}$ with $\underline{V_1} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix}$,
$\underline{V_2} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$, $\underline{V_3} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$, and
$\underline{V_4} = \underline{V_5} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \end{bmatrix}$.

2a.

$$\underline{\Lambda\Lambda_{12}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{\Lambda\Lambda_{21}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\underline{\Lambda\Lambda_{22}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The rows of these matrices are the vectors in Q_1.

$Q_1 = \left\{ \underline{V_1}, \underline{V_4}, \underline{V_6}, \underline{V_7} \right\}$ with $\underline{V_1}$ and $\underline{V_4}$ as above and with
$\underline{V_6} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \end{bmatrix}$, $\underline{V_7} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}$.

2b. $T_2 = T_1 \cup Q_1 = \left\{ \underline{V_1}, \underline{V_2}, \underline{V_3}, \underline{V_4}, \underline{V_5}, \underline{V_6}, \underline{V_7} \right\} \neq T_1$

2a. $Q_2 = \left\{ \underline{V_1}, \underline{V_4}, \underline{V_6}, \underline{V_7} \right\}$.

2b. $T_3 = T_2 = T_2 \cup Q_2$

<u>Step 5.</u> The equation $\det \left[\underline{\Lambda} \;\; -W\underline{I} \right] = 0,$

$$\begin{vmatrix} -W & 1 & 1 & 0 & 0 \\ 0 & -W & 0 & 1 & 0 \\ 1 & 0 & -W & 0 & 0 \\ 1 & 0 & 0 & -W & 1 \\ 1 & 0 & 0 & 0 & 1-W \end{vmatrix} = 0$$

has $W_o = 1.618$ as its largest real solution.

$C = \log 1.618 = 0.693$ bits/unit time.

The eigenvector $\underline{B}$ is easily calculated to be

$$\underline{B} = \begin{bmatrix} 0.618 \\ 0.618 \\ 0.382 \\ 1.000 \\ 1.000 \end{bmatrix}$$

The second-order state transition probabilities are given below.

		s(t + 1)				
	P_{ij}	s_{31}	s_{12}	s_{13}	s_{23}	s_{33}
s(t)	s_{31}	0.000	0.618	0.382	0.000	0.000
	s_{12}	0.000	0.000	0.000	1.000	0.000
	s_{13}	1.000	0.000	0.000	0.000	0.000
	s_{23}	0.382	0.000	0.000	0.000	0.618
	s_{33}	0.382	0.000	0.000	0.000	0.618

A source to realize these transition probabilities can be constructed by enabling it to follow the states of A (returning to the original single-subscript notation, in which the set of states is $S = \{s_1, s_2, s_3\}$), and to emit symbols as follows.

If preceeding state and present state of A are		Source emits x_1, x_2, x_3 with these probabilities:		
$s(t - 1)$	$s(t)$	x_1	x_2	x_3
s_3	s_1	0.618	0.382	0.000
s_1	s_2	1.000	0.000	0.000
s_1	s_3	1.000	0.000	0.000
s_2	s_3	0.382	0.000	0.618
s_3	s_3	0.382	0.000	0.618

With this source, the output sequence is a Markov process and the transition probabilities are as follows:

		$y(t + 1)$	
		y_1	y_2
$y(t)$	y_1	0.000	1.000
	y_2	0.382	0.618

The entropy of the sequence is 0.693 bits/unit time.

Before leaving the subject of automata capacity, we will make one final observation which has been deferred to avoid confusing the reader. This is that when $\underline{\Lambda}^*$ has been found, one need not solve the equation

$$\det \left[\underline{\Lambda}^* - W\underline{I} \right] = 0$$

for its largest real root but may solve instead the simpler equation

$$\det \left[g(\underline{\Lambda}^*) - W\underline{I} \right] = 0$$

for its largest real root; the two roots will be the same. In the second equation, $g(\underline{\Lambda}^*)$ is the matrix of allowable output transitions, and it may be deduced directly from $\underline{\Lambda}^*$ and g. For the example, this is illustrated graphically in Figure 12. Arrows indicate allowed transitions in $\underline{\Lambda}^*$, above, and in $g(\underline{\Lambda}^*)$, below. The output transition matrix in tabular form is:

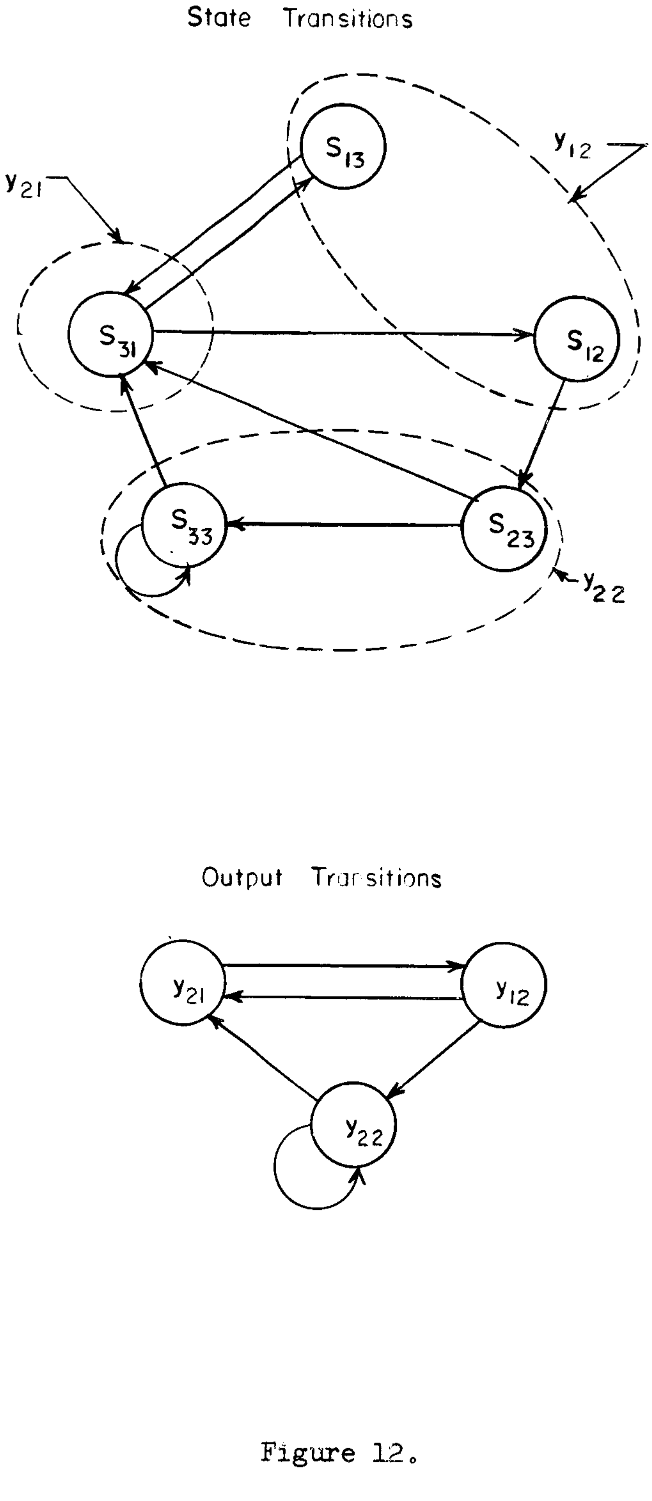
		y(t-1, t)		
		y_{21}	y_{12}	y_{22}
	y_{21}	0	1	0
y(t-2,t-1)	y_{12}	1	0	1
	y_{22}	1	0	1

The determinantal equation $\det \left[g(\underline{\Lambda}^*) - W\underline{I} \right] = 0$,

$$\begin{vmatrix} -W & 1 & 0 \\ 1 & -W & 1 \\ 1 & 0 & 1-W \end{vmatrix} = 0,$$

has $W_o = 1.618$ as its largest real solution, and $\log W_o = 0.693$ as before.

The reason this simplification is possible is that when the output sequence carries just as much information as the state sequence, one gains nothing by maintaining the distinction between states which map to the same output; the exact state sequence could be deduced from the output sequence if needed. Therefore we can deal with a homomorphism of the automaton A*, and using $g(\underline{\Lambda}^*)$ amounts to doing just that.



State Transitions

Output Transitions

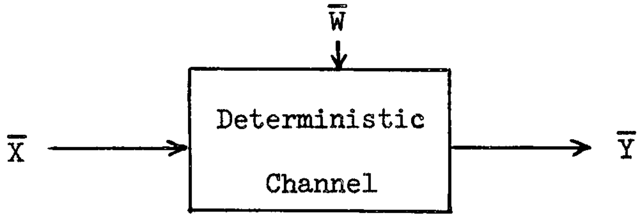
Figure 12.

3.6.3.4. Further remarks

This section has provided a means of calculating, or at least approximating, the capacity of any arbitrarily complex (but finite) network of MWI's, mappers, and Moore automata. Since a great many mechanisms can be approximately modeled by networks of this type, we can now calculate the capacities of many systems. In the chapter on regulation we will show that the power of a regulatory system to regulate is limited by its channel capacity; consequently this section is of substantial importance to the theory of regulation.

3.6.4. Capacity of noisy channels

A Moore automaton is an example of a deterministic channel - a channel for which $H^L_{\bar X}(\bar Y) = 0$. A nondeterministic channel may be viewed as a deterministic channel with an unknown input, $\bar W$, so that $H^L_{<\bar W,\bar X>}(\bar Y) = 0$ although $H^L_{\bar X}(\bar Y) > 0$.



If we think of $\bar X$ as "message input," $\bar Y$ as "output," and $\bar W$ as "noise input," and the channel as a relation R between the three s-variables, this adequately characterizes the situation of the noisy channel.

$H^L(\bar Y)$ is the information rate for the output sequence. The identity

$$H^L(\bar Y) \equiv T^L(\bar X : \bar Y) + H^L_{\bar X}(\bar Y)$$

shows that the information rate at the channel output is the sum of the rate at which information is passed from message input to output and the rate at which the noise contributes to the output, since the last term,

$$H^L_{\overline{X}}(\overline{Y}) = T^L_{\overline{X}}(\overline{W} : \overline{Y})$$

is the rate at which the noise "corrupts" the output in spite of the message.

The last term is zero for noiseless channels. If the contribution of noise is regarded as a nuisance, so that $T^L(\overline{X} : \overline{Y})$ is the rate of "useful" information, then the <u>channel capacity for useful information</u> is

$$C_{\text{useful}} = \max \left\{ T^L(\overline{X} : \overline{Y}) \right\}$$

with the maximum taken over the distributions $\underline{N}(\overline{W}, \overline{X}, \overline{Y})$ compatible with both R and the assumed characteristics of the noise source.

What one regards as message and what as noise is arbitrary; $\overline{W}$ and $\overline{X}$ play symmetric roles, and the equation

$$H^L(\overline{Y}) = T^L(\overline{X} :, \overline{Y}) + T^L(\overline{W} : \overline{Y}) + Q^L(\overline{W}, \overline{X}, \overline{Y})$$

shows this clearly. If $T^L(\overline{W} : \overline{X}) = 0$, i.e., the noise is independent of the message, then

$$H^L(\overline{Y}) = T^L(\overline{X} : \overline{Y}) + T^L(\overline{W} : \overline{Y}) + T^L_{\overline{Y}}(\overline{W} : \overline{X})$$

$$H^L(\overline{Y}) \geqslant T^L(\overline{X} : \overline{Y}) + T^L(\overline{W} : \overline{Y})$$

and the output information rate is at least the sum of the message-to-output rate and the noise-to-output rate.

IV. INFORMATION THEORY AND COMPLEX SYSTEMS

Introduction

In this chapter we will focus attention on information theory as it applies to complex systems. After a brief consideration of what is meant by complexity, we will consider several information theoretic tools for dealing with complexity in systems and will show how these tools can lead to a better understanding of such systems, by discarding excess information. The basic point of the chapter is that to understand a complex system, one must discard much nonessential information, and the methods and measures of information theory throw away a great deal while preserving that related to the structure of the system.

4.1. Complex systems

4.1.1. Measuring complexity

We will deal briefly in this section with some of the difficulties which arise in attempts to measure the complexity of a system, and we will propose two measures which, although not perfect, nevertheless are consistent with many of our intuitions. No attempt will be made to deal with "systems" in the vague, general sense of that word, but rather only with systems as ordered sets of s-variables and as networks of machines, probabilistic or not, embodying those variables. Moreover we will consider only dynamic systems, in which the s-variables represent time sequences, and the focus will be on the complexity of the system's behavior rather than on the complexity of the system per se.

Complexity is a poorly-defined notion in which the subjective component so predominates that it is probably impossible to produce a definition, much less a measure, acceptable to all people in all circumstances. Yet few would disagree that there is a strong link between complexity and information; the more information one has to take in to "understand" the system, (i.e., its behavior), or to describe it, the more complex it seems.

We speak of the complexity of a system as if it were a property of the system, and that semantic usage obscures the fact that complexity is really a relation between the system and its observer, as is apparent from the fact that the same "thing" (say a watch) may appear quite complicated to one observer (a housewife) while not nearly as complicated to another (a jeweler). When a "thing" appears less complex to one observer than to another, the two may actually be considering different systems (i.e., different variables) or, if not, one observer may understand the system better - have a more adequate mental model of it, that is, so that it appears more predictable and less mysterious.

One contention of this section is that it is to the observer's "model" of the system, rather than to the system itself, that any measure of complexity should be applied. By his model we mean the ordered set of variables comprising the system, together with his best current guess as to the internal dynamics of the system - what system-values are most likely, which variables are causally linked to which others, what functional relations obtain, and so on - embodied in his a priori "probability" distribution $\underline{P_i}$, giving for each possible past history of the system, the "probabilities" for the ensuing system-value $S^i = < X_1^i, X_2^i, \ldots, X_M^i >$:

$$\underline{P_i} = \underline{P_i}\ (S^i \mid S^{i-1}, S^{i-2}, \ldots).$$

Dealing with the observer's model rather than with the system itself serves to remove the problem of the observer, to some extent, by making objective his knowledge (or ignorance, or intuition) about the system. Having made clear that we will deal hereafter with models of systems rather than with systems themselves, we can revert to use of the word "system" as a convenient shorthand for "model of a system," bearing the distinction in mind.

An apparently reasonable axiom to adopt with respect to a measure of complexity is the following:

> If one system is a homomorphism of another, then the complexity of the former should be less than the complexity of the latter.

This appears to be well in line with our intuitions, for a homomorphism of a system is usually thought of as a simplified (i.e., less complex) version. If this axiom is accepted, then the following is a direct consequence of it:

> If two systems are isomorphic, their complexities should be equal.

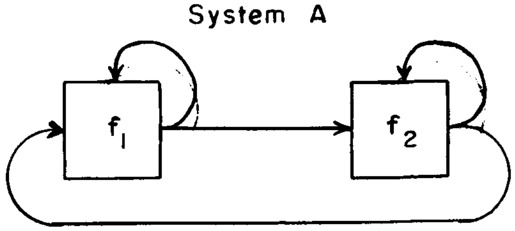
For if a pair of systems are homomorphisms of each other, they are isomorphic, i.e., relabeled versions of each other. We feel that if we understand one system, we understand another isomorphic to it (indeed, this is a common teaching device), and that therefore the two are equally complex. The axiom is quite strong in that it states that the two systems in Figures 13 and 14, which are isomorphic, are equally complex. In some sense, the system of two parts seems more complex than the other; yet our intuitions on this point are contradictory, for it is commonly thought that a system which can be "broken down" into parts is less complex than another, having the same number of states, which cannot - at least that is a common attitude with respect to really large systems.

The axiom rules out reduction of complexity through mere relabeling of states and allows us to view every system as a one-variable system, through relabeling. This may seem to conflict with the observation that relabeling a system sometimes does in fact make it appear less complex, as when one notices that a system which is under study is isomorphic with another system which one "understands." This is not necessarily a weakness of the axiom, but rather further support for our insistence on measuring complexity of one's model of the system; for what apparently happens when the isomorphism is noticed is a revision of the model, making the model for the one system match that of the other.

Another axiom is the following:

> If a system is composed of a number of independent parts, the complexity of the whole system should be the sum of the complexities of its parts.

If one is to be able to relate the complexities of the parts to that of



X

	f_1	0	1
Y	0	0	1
	1	1	0

$f_1: X \times Y \rightarrow X$

Y

	f_2	0	1
X	0	1	0
	1	1	0

$f_2: Y \times X \rightarrow Y$

$X^{i-1} \times Y^{i-1}$

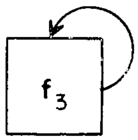
	$\underline{P_i}$	<0,1>	<1,0>	<1,1>	<0,0>
$X^i \times Y^i$	<0,1>	0	0	0	1
	<1,0>	1	0	0	0
	<1,1>	0	1	0	0
	<0,0>	0	0	1	0

probability distribution

$\underline{P}^i\ (S^i, S^{i-1}, S^{i-2}, \ldots)$

Figure 13.

System B



	Z			
f_3	1	2	3	4
	2	3	4	1

$f_3 : Z \to Z$

		Z^{i-1}			
	P_i	1	2	3	4
	1	0	0	0	1
	2	1	0	0	0
Z^i	3	0	1	0	0
	4	0	0	1	0

Figure 14.

the whole, this would seem to be the most natural relation at least when the parts are independent. Yet it is open to the objection that if the parts are "similar" or even isomorphic, even though independent, then the whole is in some sense not much more complex than one of its parts. To counter that objection would require bringing in some notion of similarity or else scrapping the axiom; we will do neither, just regarding the weakness which results as the unfortunate consequence of trying to find a simple, relatively unsophisticated measure of complexity.

The entropy function is consistent with these axioms, and we therefore propose two measures of complexity related to the distribution $\underline{P_i}\ (S^i \mid S^{i-1},S^{i-2},\dots)$. We define *static complexity* C_S as the uncertainty as to which system-value will occur at any instant, if the past history is not known,

$$C_S = H(S^i)$$

and the *dynamic complexity* C_D as the same uncertainty, if the past history *is* known,

$$C_D = H_{\dots,S^{i-2},S^{i-1}}\ (S^i).$$

Both C_S and C_D are obtained from $\underline{P_i}$, the observer's model at time i, and therefore they change, in general, as the observer revises his model. If the observer starts with a model admitting of complete ignorance, then C_S and C_D start at log N, where N is the number of possible system-values S^i, and the complexities decrease thereafter, although not necessarily monotonically, presumably until the model represents the objective system well.

The dynamic complexity C_D is zero if the model is deterministic; this is consistent with the feeling that deterministic systems, although they may be complex (via C_S), are not complex in their style of dynamic progression.

These measures of complexity have much to recommend them, although they have apparent weaknesses; the contention of this section is that the notion of complexity is sufficiently vague that any measures will be found wanting in some respects, but that C_S and C_D are good measures at least for many purposes.

4.1.2. Relevance of information theory to the study of complex systems

We will mention in this section some common attributes of complex systems and the relevance of information theoretic methods to their study.

Perhaps the most obvious feature of really complex systems is that they are large - not physically, but in the number of system-values possible; frequently there are many variables, interdependent in a non-simple way, with each variable taking many values. As larger and larger systems are considered, the point is soon reached beyond which the human, or even the fastest computer, cannot practically cope with the whole system in detail, and the complexity must be "reduced" by substituting a new system, related to the original system but simpler than it. A way of doing this which is frequently possible is to view the original system as composed of parts, each of manageable complexity and all related in a not-too-complex manner. Another is to deal with a homomorphism, or an approximate homomorphism, of the system, thus giving up some detail. To use information theory is yet another way, in which most details of the system are ignored and what remains is essentially a picture of the "activity" of the variables and of the statistical linkages and causal connections between them. These linkages will be explored in later sections of this paper.

Another feature which complex systems often display is a hierarchical structure - a structure in which the whole consists of interrelated subsystems, and in which the subsystems are themselves hierarchical, down to the lowest level of elementary subsystem. By the term hierarchical we mean to include, but not necessarily imply, the case in which each subsystem has a "boss" in the system. The ubiquity of hierarchical structures is discussed by Simon[11]. For the view of a system as composed of parts to be a useful view, the parts must interact with each other in a more or less global way - that is, in a way which is not highly dependent on the internal details of the parts. The interactions in a communications system, in which the parts are represented by blocks and the whole as a "block diagram", is a common example. In section 4.3 we will demonstrate that information theory can be usefully applied to effect a conceptual breakdown of a system into subsystems, and to measure the constraints holding between the subsystems as well as within each subsystem.

Many complex systems can be viewed as goal-seeking; that is, they act in an apparently purposive manner, interacting with their environment so as to "get their way," i.e., so as to maintain certain essential variables within acceptable limits. If the environment represents a real threat, so that the purposive action requires actual action on the part of the system, then information theory is relevant in several ways. First, there are certain quantitative statements which can be made about the coordination required between the environment and the system if the latter is to attain its goal; these will be developed fully later, in the information theoretic analysis of regulation. Second, if internal coordination between parts of the system is necessary to achieve the goal, this coordination is also subject to quantitative constraints, of the same nature. Third, the system must usually take in information about the environment with which it interacts, if it is to achieve the requisite coordination, and the rate at which this information can be taken in is governed by the well-developed laws of information transfer through channels.

Complex systems commonly display another feature; their actions are commonly conditioned by their past history. This feature, which we can refer to loosely as memory, means that the past has a demonstrable effect on the present, and this effect can be studied with the tools of information theory; coordination between variables displaced in time is just as amenable to information theoretic techniques as coordination between simultaneously observed variables. Most complex systems do not have the property of ergodicity, and therefore many specialized theorems of information theory do not apply; nevertheless, much can still be said.

In short, information theory is useful in the study of complex systems when one is willing to sacrifice the minute details involved and to look instead at the variables and their interrelations. The next section will discuss two devices for doing just that. These are the Diagram of Immediate Effects, suggested by Ashby, and an information theory analog to it, the Diagram of Immediate Transmissions.

4.2. The Diagram of Immediate Effects and some information theory analogs

Introduction

The Diagram of Immediate Effects (DIE) described by Ashby in Introduction to Cybernetics is a useful device for displaying the cause-effect relations between parts of a system, and in particular for displaying independence of parts, feedback relations between parts, and so on. The price paid for its extreme simplicity, however, includes the following drawbacks:

(1) The DIE measures the linkage between two parts of a system with only two values - either the two parts are causally linked, or are *not*.

(2) To construct the DIE, one must in general either know the mappings joining them, or else be able to force the system into every conceivable system state.

(3) The DIE is applicable only to state-determined systems.

The coarse-grained character of the DIE means that its quantitative information about relations between parts of a system is insufficient for many purposes, and the requirements listed under (2) and (3) are impossible to meet in many cases of practical interest, e.g. in complex biological systems.

The Diagram of Immediate Transmissions (DIT) described in this section minimizes these problems; it measures the cause-effect linkage between parts to as fine a degree as desired, and it demands for its construction only that the variables of the system be observable as the system follows its natural mode of activity. It is applicable to both deterministic and nondeterministic systems.

One of the chief advantages of the Transmission measures over the Effect measures is that the former are better suited for networks in which there are changing patterns of communication, as in networks displaying "learning", "adaptation", and the like. This is because the transmissions will in general change during the history of the network, whereas the "effect" measures will not, being derivatives of the system's mapping which is assumed fixed. The "effect" measures deal with what communication possibilities are inherent in the network, while the "transmission" measures deal with what actually happens.

We have investigated the DIE, the DIT, and several closely related diagrams in detail and have reported the results elsewhere[12]; here only the major results of that investigation will be given. The next part of this section deals with the DIE, the following deals with the DIT, and the last part offers comments on the usefulness and weaknesses of the diagrams.

4.2.1 The Diagram of Immediate Effects (DIE)

This section defines the DIE and other related diagrams and introduces several theorems about them. Although the DIE is of interest in its own right, it is included here primarily as an introduction to the DIT of the next section.

The DIE is applicable to a state-determined system $\overline{S} = \left\{\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M\right\}$ in which each "part" $\overline{X}_i$ represents a machine with input. We denote by X_i the set of allowed values for the variables X_i^1, X_i^2, ... comprising $\overline{X}_i$, and we let the superscripts indicate time. The mapping f_i maps the state of the system, S, into the next state of part X_i;

$$f_i : X_1 \times X_2 \times \ldots \times X_M \to X_i$$

The mapping for the whole system is f_σ ; f $: S \to S$.

We will find it convenient to use the <u>projection mapping</u> $pr_i : S \to X_i$ which selects the $\overline{X}_i$ component from a vector, or more generally the mapping $pr_{s_a} : S \to S_a$ which selects the ordered n-tuple of components corresponding to variables in $\overline{S}_a$. We will also use $pr_{S-S_a} : S \to S\text{-}S_a$. For example, with $\overline{S} = \left\{\overline{X}_1, \overline{X}_2, \overline{X}_3\right\}$ and $\overline{S}_a = \left\{\overline{X}_1, \overline{X}_3\right\}$ we have $pr_3(<2, 3, 5>) = 5$, $pr_{S_a}(<2, 3, 5>) = <2, 5>$, $pr_{S-S_a}(<2, 3, 5>) = 3$.

We say $\overline{X}_i$ <u>has an immediate effect on</u> $\overline{X}_j$ if there is a pair of system-values s_a and s_b for which $pr_{S-X_i}(s_a) = pr_{S-X_i}(s_b)$ and $pr_i(s_a) \neq pr_i(s_b)$, such that $f_j(s_a) \neq f_j(s_b)$; that is, if there are two system-values different only in their $\overline{X}_i$-components, which lead to different $\overline{X}_j$-values at the next step.

It is convenient to use an arrow, as in $\overline{X}_i \to \overline{X}_j$, as shorthand for the phrase "has an immediate effect upon," and a canceled arrow, as in $\overline{X}_i \not\to \overline{X}_j$, for the contrary.

We define the Matrix of Immediate Effects $\underline{A} = \left[a_{ij}\right]_{M,M}$ as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } \overline{X}_i \rightarrow \overline{X}_j, \\ 0 & \text{otherwise.} \end{cases}$$

The Diagram of Immediate Effects (DIE) is a pictorial representation of $\underline{A}$. It has an open and a closed form. For example, with $\left\{\overline{X}_1, \overline{X}_2, \overline{X}_3\right\} = \overline{S}$ and

$$\underline{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

the closed form is shown in Figure 15, and the open form, with arrowheads assumed but not drawn on the right end of each line, is shown in Figure 16. The DIE is an excellent device for displaying certain cause-effect relations between the variables in a system, giving as it does an easily grasped overview of what parts affect which, what feedback relations may be present, and so on. The open form, while not as simple as the closed form, has certain advantages, notably that it may be iterated to display cause-effect "chains" as illustrated in Figure 17. The DIE displays effects between individual variables in $\overline{S}$. More generally, a subsystem $\overline{S}_a = \left\{ \overline{X}_{a1}, \overline{X}_{a2}, \ldots, \overline{X}_{am} \right\} \subset \overline{S}$ has an immediate effect on another subsystem $\overline{S}_b = \left\{ \overline{X}_{b1}, \overline{X}_{b2}, \ldots, \overline{X}_{bn} \right\} \subset \overline{S}$ if there exists a pair of system-values s_c and s_d for which $pr_{S-S_a}(s_c) = pr_{S-S_a}(s_d)$ and $pr_{S_a}(s_c) \neq pr_{S_a}(s_d)$, such that $pr_{S_b}(f_\sigma(s_c)) = pr_{S_b}(f_\sigma(s_d))$; i.e., if there are two system-values different only in their S_a-components which

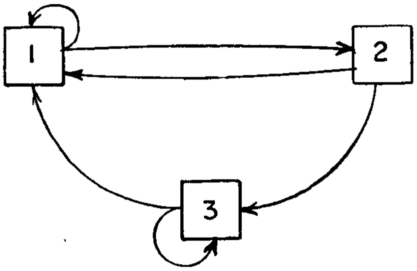


Figure 15.

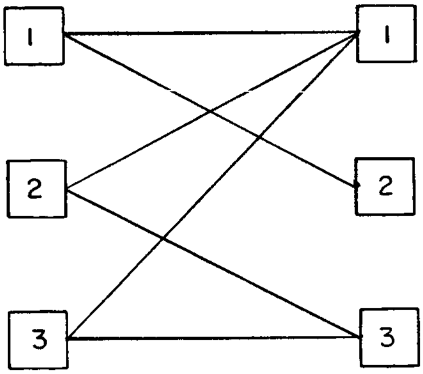


Figure 16.

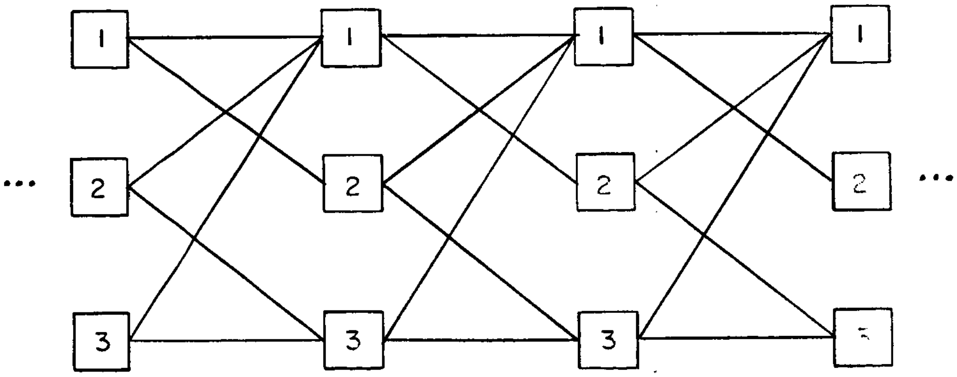


Figure 17.

lead to different S_b-values. If $\bar{\bar{S}}_a$ and $\bar{\bar{S}}_b$ are not disjoint, the closed form DIE is not usable, but the open form is; for example, with $\bar{\bar{S}}_a = \left\{ \bar{\bar{X}}_1, \bar{\bar{X}}_2 \right\}$, $\bar{\bar{S}}_b = \left\{ \bar{\bar{X}}_2, \bar{\bar{X}}_3 \right\}$, and $\underline{A}$ as before, the DIE is as shown in Figure 18.

A convenient feature of the DIE is that when several variables are grouped into subsystems, the DIE for the subsystems can be deduced directly from the DIE for the individual variables.

<u>Theorem IV.1</u>

Let $\bar{\bar{S}}_a$ and $\bar{\bar{S}}_b$ be subsystems of $\bar{\bar{S}}$. Then $\left\{ \bar{\bar{S}}_a \rightarrow \bar{\bar{S}}_b \right\} \Leftrightarrow \left\{ \exists X_i \in S_a, X_j \in S_b \text{ s.t. } \bar{\bar{X}}_i \rightarrow \bar{\bar{X}}_j \right\}$.

<u>Proof:</u>

The direction $\Leftarrow$ is obvious. To show $\Rightarrow$, suppose $\bar{\bar{S}}_a \rightarrow \bar{\bar{S}}_b$ as evidenced by system-values s_c and s_d which are identical except for [some or all of] their $\bar{\bar{S}}_a$-components and which are mapped by f_j into different $\bar{\bar{X}}_j$-values, for some $\bar{\bar{X}}_j$ in $\bar{\bar{S}}_b$. If s_c and s_d differ in only one component, the theorem is automatically satisfied. Suppose s_c and s_d differ in exactly two components, those for $\bar{\bar{X}}_{a1}$ and $\bar{\bar{X}}_{a2}$. Then

$$f_j(s_c) = f_j(x_{a1}, x_{a2}, \ldots) = x_1$$

$$f_j(s_d) = f_j(x'_{a1}, x'_{a2}, \ldots) = x_2 \neq x_1$$

where the dots indicate that the remaining components of s_c and s_d are identical. The theorem states that either $\bar{\bar{X}}_{a1} \rightarrow \bar{\bar{X}}_j$ or $\bar{\bar{X}}_{a2} \rightarrow \bar{\bar{X}}_j$ (or both); we will assume $\bar{\bar{X}}_{a1} \not\rightarrow \bar{\bar{X}}_j$ and $\bar{\bar{X}}_{a2} \not\rightarrow \bar{\bar{X}}_j$ and obtain a contradiction.

Consider s_e:

$$s_e = \langle x'_{a1}, x_{a2}, \ldots \rangle.$$

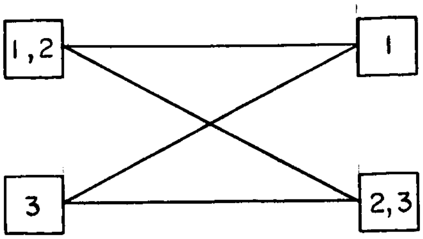


Figure 18.

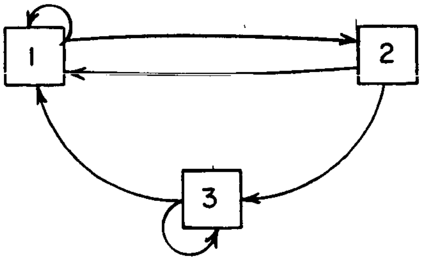


Figure 19.

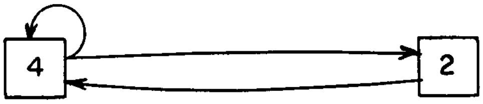


Figure 20.

Since $f_j(s_c) = x_1$ and $\bar{X}_{a1} \not\rightarrow \bar{X}_j$, $f_j(s_e) = x_1$ because s_c and s_e differ only in their $\bar{X}_1$-component. And since $f_j(s_e) = x_1$ and $\bar{X}_{a2} \not\rightarrow \bar{X}_j$, $f_j(s_d) = x_1$ because s_e and s_d differ only in their $\bar{X}_2$-component. But $f_j(s_d) = x_1$. The contradiction implies that either $\bar{X}_{a1} \rightarrow \bar{X}_j$ or $\bar{X}_{a2} \rightarrow \bar{X}_j$ (or both).

The theorem is true, then, if s_c and s_d differ in two components only. The obvious extension of the foregoing, when s_c and s_d differ in arbitrarily many components, shows that at least one variable in $\bar{S}_a$ must have an immediate effect on $\bar{X}_j$.

Q. E. D.

It follows from Theorem IV.1 that if some variables are grouped, i.e., considered as components of a new, compound variable, the DIE for the new system can be deduced directly from the DIE for the old system. For example, if the DIE for $\bar{S}_1 = \left\{ \bar{X}_1, \bar{X}_2, \bar{X}_3 \right\}$ is as shown in Figure 19, and if $\bar{X}_1$ and $\bar{X}_3$ are grouped to form $\bar{X}_4 = \overline{\langle X_1, X_3 \rangle}$, the DIE for $\bar{S}_2 = \left\{ \bar{X}_4, \bar{X}_2 \right\}$ is as shown in Figure 20.

The immediate-effect set of $\bar{X}_i$, denoted $A(\bar{X}_i)$, is defined by

$$A(\bar{X}_i) = \left\{ \bar{X}_j \in \bar{S} \;\middle|\; \bar{X}_i \rightarrow \bar{X}_j \right\}$$

In the DIE, it is the set to which $\bar{X}_i$ sends arrows. The immediate-effect set of $\bar{S}_a \subset \bar{S}$, denoted $A(\bar{S}_a)$, is

$$A(\bar{S}_a) = \left\{ \bar{X}_j \in \bar{S} \;\middle|\; \bar{S}_a \rightarrow \bar{X}_j \right\}.$$

It follows from theorem IV.1 that

$$A(\bar{S}_a) = \bigcup_{\bar{X}_i \in \bar{S}_a} A(\bar{X}_i).$$

If $\bar{S}_a$ and $\bar{S}_b$ are disjoint sets whose union is $\bar{S}$, they are independent if and only if $\bar{S}_a \not\rightarrow \bar{S}_b$ and $\bar{S}_b \not\rightarrow \bar{S}_a$, i.e., if $A(\bar{S}_a) \subset \bar{S}_a$ and $A(\bar{S}_b) \subset \bar{S}_b$.

$\overline{X}_i$ may have a delayed effect on $\overline{X}_j$ even if $\overline{X}_i \not\rightarrow \overline{X}_j$, for the effect may be passed through a third variable or even a whole chain, as if $\overline{X}_i \rightarrow \overline{X}_k$, $\overline{X}_k \rightarrow \overline{X}_l$, $\overline{X}_l \rightarrow \overline{X}_m$, ..., $\overline{X}_n \rightarrow \overline{X}_j$. For this reason it is useful to define the k-effect of $\overline{X}_i$ on $\overline{X}_j$; $\overline{X}_i$ has a k-effect on $\overline{X}_j$, symbolized $\overline{X}_i \xrightarrow{k} \overline{X}_j$, if there is a pair of system-values s_a and s_b for which $pr_{S-X_i}(s_a) = pr_{S-X_i}(s_b)$ and $pr_i(s_a) \neq pr_i(s_b)$, such that $f_j f_\sigma^{k-1}(s_a) \neq f_j f_\sigma^{k-1}(s_b)$, where f_σ^{k-1} stands for k-1 interations of the system's mapping. Thus $\overline{X}_i \xrightarrow{k} \overline{X}_j$ if variations in $\overline{X}_i$ by themselves can sometimes induce variations in $\overline{X}_j$, k steps later. The Matrix of k-effects $\underline{A_k} = \left[a_{ij_k}\right]_{M,M}$ is defined by

$$a_{ij_k} = \begin{cases} 1 & \text{if } \overline{X}_i \xrightarrow{k} \overline{X}_j, \\ 0 & \text{otherwise.} \end{cases}$$

The Diagram of k-effects, DKE, is a pictorial representation of $\underline{A_k}$.

Definitions for the k-effect of $\overline{S}_a$ on $\overline{S}_b$, $\overline{S}_a \xrightarrow{k} \overline{S}_b$, $A_k(\overline{X}_i)$, and $A_k(\overline{S}_a)$ will be omitted since they are strictly analogous to the earlier definitions.

Theorem IV.1 holds if "k-effect" is everywhere substituted for "immediate effect," and as before,

$$A_k(\overline{S}_a) = \bigcup_{\overline{X}_i \in \overline{S}_a} A_k(\overline{X}_i).$$

The operator μ maps all positive real numbers to 1 and all other real numbers to zero. Operating on a matrix, it creates a matrix of zeroes and ones.

The fundamental relation between immediate effects and k-effects follows.

Theorem IV.2

$$\mu(\underline{A_k}) \leq \mu(\underline{A}^k) \text{ for all } k \geqslant 1.$$

That is, if $\bar{\bar{X}}_i \xrightarrow{k} \bar{\bar{X}}_j$ then there must be a chain of exactly k arrows in the DIE leading from $\bar{\bar{X}}_i$ to $\bar{\bar{X}}_j$.

Proof:

For k = 1, the theorem holds. Let k = 2, and suppose $\bar{\bar{X}}_i \xrightarrow{2} \bar{\bar{X}}_j$ as evidenced by a pair of system-values s_a and s_b satisfying the requirements. If $f_\sigma(s_a) = s_1$ and $f_\sigma(s_b) = s_2$ are identical, then $f_j f_\sigma(s_a) = f_j f_\sigma(s_b)$ and $\bar{\bar{X}}_i \not\xrightarrow{2} \bar{\bar{X}}_j$, contrary to our supposition; therefore $s_1 \neq s_2$.

The components of s_1 and s_2 which differ correspond to a set of variables $\bar{\bar{S}}_c \subset A(\bar{\bar{X}}_i)$;

$$\bar{\bar{S}}_c = \left\{ \bar{\bar{X}}_\ell \mid pr_\ell(s_1) \neq pr_\ell(s_2) \right\}.$$

Now s_1 and s_2 differ only in their $\bar{\bar{S}}_c$-components, and $f_j(s_1) \neq f_j(s_2)$; thus $\bar{\bar{S}}_c \rightarrow \bar{\bar{X}}_j$. By theorem IV.1 there is an $\bar{\bar{X}}_\ell$ in $\bar{\bar{S}}_c$ such that $\bar{\bar{X}}_\ell \rightarrow \bar{\bar{X}}_j$, and therefore there is an $\bar{\bar{X}}_\ell$ such that $\bar{\bar{X}}_i \rightarrow \bar{\bar{X}}_\ell$ and $\bar{\bar{X}}_\ell \rightarrow \bar{\bar{X}}_j$. This proves the theorem for k = 2.

Suppose the theorem is true for k = n - 1, so that there is a chain of n - 1 arrows from $\bar{\bar{X}}_i$ to each variable in $A_{n-1}(\bar{\bar{X}}_i)$. If $\bar{\bar{X}}_i \xrightarrow{n} \bar{\bar{X}}_j$, there exist system-values s_a and s_b differeing only in their $\bar{\bar{X}}_i$-components and such that $f_j f_\sigma^{n-1}(s_a) \neq f_j f_\sigma^{n-1}(s_b)$. This can only be the case if $f_\sigma^{n-1}(s_a) \neq f_\sigma^{n-1}(s_b)$; the components which differ define a set $\bar{\bar{S}}_d$ as before:

$$\bar{\bar{S}}_d = \left\{ \bar{\bar{X}}_\ell \mid pr_\ell(f_\sigma^{n-1}(s_a)) \neq pr_\ell(f_\sigma^{n-1}(s_b)) \right\}.$$

As before, $\bar{\bar{S}}_d$ must have an immediate effect on $\bar{\bar{X}}_j$. Therefore, there must be an $\overline{X}_\ell$ in $\overline{S}_d$ such that the DIE has a chain of n - 1 arrows from $\overline{X}_i$ to $\overline{X}_\ell$ and also an arrow from $\overline{X}_\ell$ to $\overline{X}_j$.

By induction, then, the theorem is true for any $k \geqslant 1$.

Q. E. D.

Theorem IV.2 has an obvious corollary.

Corollary IV.1

$$A_k(\overline{X}_i) \subset A^k(\overline{X}_i) = A(\ldots A(A(A(\overline{X}_i))) \ldots)$$

That is, the k-effect set of $\overline{X}_i$ is included in the set of variables reached from $\overline{X}_i$ on the DIE by following all the chains of k arrows.

In fact, if $(n_1, n_2, \ldots, n_m)$ is any partition of k,

$$\mu(\underline{A_k}) \leqslant (\underline{A}^{n_1} \cdot \underline{A}^{n_2} \cdot \ldots \cdot \underline{A}^{n_m})$$

and

$$A_k(\overline{X}_i) \subset A^{n_1}(A^{n_2}(\ldots(A^{n_m}(\overline{X}_i))\ldots)).$$

This fact leads to a simple procedure for estimating high-order $\underline{A_k}$ matrices from lower-order ones. The procedure has been reported elsewhere[12].

The next section will develop the Diagram of Immediate Transmissions, which is strictly analogous to the DIE, and will compare the two Diagrams as the development proceeds.

4.2.2. The Diagram of Immediate Transmissions (DIT)

The DIT is applicable to a system $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots \overline{X}_M \right\}$, deterministic or not, in which each variable $\overline{X}_i$ represents a "part." We will use the same notation in this section as in the preceding, as far as possible.

The DIE contains information about the system's mapping and shows which parts, acting alone, can affect which others. The DIT contains information about the system's behavior, as recorded in a frequency table; since the behavior may depend on changing external factors or, as in the case of a learning or adapting system, on time, the DIT will in general change as the behavior changes, and in this sense it is a more dynamic characterization of the system than the DIE. It shows which parts, acting alone, affect which others, and it shows the magnitude of the effect on a continuous scale, so that one can see which effects are strong and which are weak. These advantages of the DIT over the DIE are obtained, however, at the price of certain complications which do not arise in the DIE. These will be pointed out as they arise in this section.

The immediate effect of $\overline{X}_i$ on $\overline{X}_j$ is naturally associated with what happens to $\overline{X}_j$ when $\overline{X}_i$ varies and all the other parts do not; this is the basis of the DIE and of the DIT as well. But while the DIE gives the answer to the simple query, Does $\overline{X}_j$ ever vary, or not?, the DIT gives the answer to, How much of the variation in $\overline{X}_j$ can be attributed to $\overline{X}_i$? In other words, how much of the variation in $\overline{X}_j$ is due to $\overline{X}_i$ alone, on the average? We denote the measure of this quantity by t_{ij}, call it the <u>immediate transmission</u> from $\overline{X}_i$ to $\overline{X}_j$, and define it by

$$t_{ij} = T_{S-X_i}(X_i : X_j').$$

The prime is used to indicate that we are interested in the transmission between X_i at one moment and X_j at the following moment, i.e., as shorthand for

$$t_{ij} = \sum_{\tau} \text{Prob}(\tau) \cdot T_{S^{\tau}-X_i^{\tau}}(X_i^{\tau} : X_j^{\tau+1}).$$

Put operationally, t_{ij} is the result of the following observations and calculations on $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$. By observation one obtains one or more protocols which list the successive system-values taken by $\overline{S}$ during a finite time span. Some particular set of values for all variables except $\overline{X}_i$ is chosen; that is, an element in the set $pr_{S-X_i}(S)$ is selected, and the protocol is scanned for system-values matching that element (in all but X_i, of course). Whenever one is found, the value of X_i and the subsequent value of X_j are recorded, and eventually a frequency table for (X_i, X_j') is thus constructed. The transmission in that table is a measure of the effect of $\overline{X}_i$ on $\overline{X}_j$ when the other variables are constant at the selected value. The process is repeated for all the other elements in $pr_{S-X_i}(S)$, and a weighted average of all the resulting transmissions gives t_{ij}. Thus t_{ij} is a measure of the effect $\overline{X}_i$ has on $\overline{X}_j$ when the effect of all other parts on $\overline{X}_j$ is blocked.

As an example, we will calculate t_{13} from this short protocol of $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \overline{X}_3 \right\}$.

time	1	2	3	4	5	6	7	8	9
X_1	1	3	2	1	3	2	2	2	1
X_2	3	1	3	1	1	3	3	1	1
X_3	2	1	2	1	1	2	2	2	1

Below are the frequency tables for the $< X_2, X_3 >$ couples which occur once or more.

$< X_2, X_3 > = < 1, 1 >$

		X_j'	
		1	2
X_i	1	1	0
	2	0	0
	3	0	2

$< 3, 2 >$

		X_j'	
		1	2
X_i	1	1	0
	2	1	2
	3	0	0

$< 1, 2 >$

		X_j'	
		1	2
X_i	1	0	0
	2	1	0
	3	0	0

Frequency tables for other $< X_2, X_3 >$ combinations contain only zeroes and hence have zero transmission. The tables shown have transmissions of 0.918, 0.311, and 0.000 bits, and thus

$$t_{13} = \frac{3}{8}\ (0.918) + \frac{4}{8}\ (0.311) + \frac{1}{8}\ (0.000)$$

$$= 0.500.$$

When $t_{ij} > 0$ we say that $\bar{X}_i$ has an immediate transmission to $\bar{X}_j$; this will be symbolized $\bar{X}_i \longrightarrow t \longrightarrow \bar{X}_j$ in general or by substituting the numerical value for t, as by $\bar{X}_1 \longrightarrow 0.500 \longrightarrow \bar{X}_3$ for the example.

The matrix $\underline{T} = [t_{ij}]_{M,M}$ is the Matrix of Immediate Transmissions, and its pictorial representation is the Diagram of Immediate Transmissions (DIT). The DIT is just like the DIE except that with each arrow or line is associated the numerical value of the transmission. The matrix $\underline{T}$ and the DIT in both forms are given below, for the example.

$$\underline{T} = \begin{bmatrix} 0.75 & 0.41 & 0.50 \\ 0.16 & 0.06 & 0.16 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$$

The closed form is shown in Figure 21 and the open form in Figure 22.

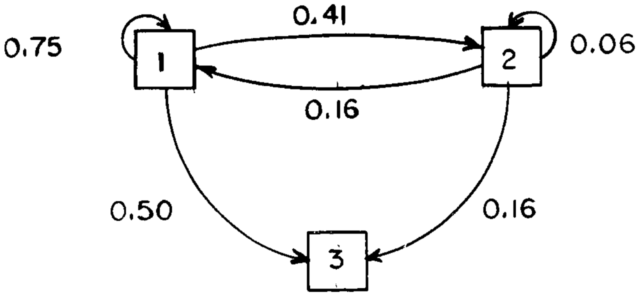


Figure 21.

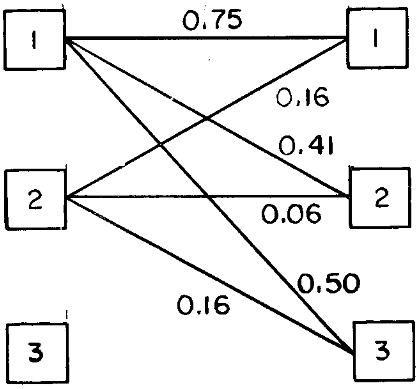


Figure 22.

The following theorem gives the most fundamental relation between the DIE and the DIT.

<u>Theorem IV.3</u>

If $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$ is state-determined, $\mu(\underline{I}) \leq \underline{A}$, or alternately, $\left\{ \overline{X}_i \rightarrow t \rightarrow \overline{X}_j \right\} \Rightarrow \left\{ \overline{X}_i \rightarrow \overline{X}_j \right\}$.

<u>Proof:</u>

Suppose $\overline{X}_i \not\rightarrow \overline{X}_j$; then for every pair of system-values s_a and s_b differing only in their $\overline{X}_i$-components, $f_j(s_a) = f_j(s_b)$. This implies that $H_{S-X_i}(X_j') = 0$ and thus that $t_{ij} = 0$.

Q. E. D.

For a state-determined system, then, the absense of an immediate effect of $\overline{X}_i$ on $\overline{X}_j$ forces the corresponding immediate transmission to be zero. The presence of an immediate effect does not, of course, imply that the immediate transmission must be positive.

Just as in the previous section, we can generalize the definition by allowing it to include transmissions of subsystems on other subsystems, and also transmissions across more than one time interval. We define <u>the k-transmission from $\overline{S}_a$ to $\overline{S}_b$</u> as $t_{S_a S_b, k}$:

$$t_{S_a, S_b, k} = T_{S-S_a}(S_a : S_b^k).$$

The k, like the prime used earlier, indicates a time gap of k time units or steps. We say that $\overline{S}_a$ <u>has a k-transmission to</u> $\overline{S}_b$ if $t_{S_a S_b, k} > 0$; this is symbolized as $\overline{S}_a \longrightarrow t \xrightarrow{k} \overline{S}_b$, or with the numerical value in place of the t.

Theorem IV.3 can be strengthened considerably as follows.

Theorem IV.4

Let $\overline{S}_a$ and $\overline{S}_b$ be subsystems of a state-determined system $\overline{S}$.

$$\text{Then } \left\{ \overline{S}_a \rightarrow t \xrightarrow{k} \overline{S}_b \right\} \Rightarrow \left\{ \overline{S}_a \xrightarrow{k} \overline{S}_b \right\}.$$

The proof is identical in form to that for theorem IV.3 and will not be given here.

Recall from the last section :

$$\left\{ \exists\, \overline{X}_i \in \overline{S}_a,\ \overline{X}_j \in \overline{S}_b \text{ s.t. } \overline{X}_i \xrightarrow{k} \overline{X}_j \right\} \Leftrightarrow \left\{ \overline{S}_a \xrightarrow{k} \overline{S}_b \right\}.$$

The corresponding statement for k-transmissions is only half true, that is:

Theorem IV.5

Let $\overline{S}$ and $\overline{S}$ be subsystems of $\overline{S}$. Then

$$\left\{ \exists\, \overline{X}_i \in \overline{S}_a,\ \overline{X}_j \in \overline{S}_b \text{ s.t. } \overline{X}_i \rightarrow t_1 \xrightarrow{k} \overline{X}_j \right\} \Rightarrow \left\{ \overline{S}_a \rightarrow t_2 \xrightarrow{k} \overline{S}_b \right\}$$

and $t_2 \geqslant t_1$.

Proof:

$$t_2 = t_{S_a, S_b, k} = H_{S-S_a}(S_b^k) - H_S(S_b^k).$$

By using the identity $H(X, Y) = H(X) + H_X(Y)$ and by adding and subtracting $H_{S-X_i}(X_j^k)$, we obtain

$$t_2 = H_{S-S_a}(X_j^k) + H_{X_j^k, S-S_a}(S_b^k - X_j^k) - H_S(X_j^k)$$

$$+ H_{X_j^k, S}(S_b^k - X_j^k) + H_{S-X_i}(X_j^k) - H_{S-X_i}(X_j^k).$$

Grouping the fifth and third, the first and last, and the second and fourth terms,

$$t_2 = t_{X_i, X_j, k} + T_{S-S_a}(S_a - X_i : X_j^k) + T_{X_j^k, S-S_a}(S_a : S_b^k - X_j^k)$$

$$\geqslant t_{X_i, X_j, k} = t_1.$$

Q. E. D.

In fact the theorem holds if subsystems $\bar{S}_i$ and $\bar{S}_j$ are substituted throughout for $\bar{X}_i$ and $\bar{X}_j$; this is also the case in the statement for k-effects.

That the converse of theorem IV.5 fails can be shown by an example. The frequency table below gives the frequencies $\underset{\sim}{N}(X^\tau, Y^\tau, X^{\tau+1}, Y^{\tau+1})$ for a system $\bar{S} = \{\bar{X}, \bar{Y}\}$.

$<X^\tau, Y^\tau> = S^\tau$

		<1,1>	<1,2>	<2,1>	<2,2>
$<X^{\tau+1}, Y^{\tau+1}> = S^{\tau+1}$	<1,1>	1	0	0	1
	<1,2>	0	1	1	0
	<2,1>	0	1	1	0
	<2,2>	1	0	0	1

Calculations based on these frequencies give the following values:

$t_{S,S} = 1$; $t_{S,X} = t_{S,Y} = t_{X,S} = t_{Y,S} = t_{X,X} = t_{X,Y} = t_{Y,X} = t_{Y,Y} = 0$.

From this example we see that one subsystem may have an immediate transmission to another subsystem without there being any lower-order transmissions at all.

The strongest statement it is possible to make regarding the converse of theorem IV.5 is given by the following theorem.

<u>Theorem IV.6</u>

Let $\bar{S}_a$ and $\bar{S}_b$ be subsystems of a state-determined system $\bar{S}$. Then

$$\left\{\bar{S}_a \rightarrow t \xrightarrow{k} \bar{S}_b\right\} \Rightarrow \left\{\exists\, \bar{X}_j \in \bar{S}_b \text{ s.t. } \bar{S}_a \rightarrow t \xrightarrow{k} \bar{X}_j\right\}.$$

Proof:

Suppose that for every $\overline{X}_j$ in $\overline{S}_b$, $t_{S_a,X_j,k} = 0$. From the definition of k-transmission, this implies that for every $\overline{X}_j$ in $\overline{S}_b$,

$$H_{S-S_a}(X_j^k) = H_{S_a,S-S_a}(X_j^k)$$

The term on the right is $H_S(X_j^k)$, and it is zero since $\overline{S}$ is state-determined; therefore, $H_{S-S_a}(X_j^k) = 0$ for every $\overline{X}_j$ in $\overline{S}_b$. The following is an identity.

$$H_{S-S_a}(S_b^k) \equiv \left[\sum_{X_j \text{ in } S_b}\left(H_{S-S_a}\ (X_j^k)\right)\right] - T_{S-S_a}(S_b^k)$$

Thus,

$$H_{S-S_a}(S_b^k) = -\ T_{S-S_a}(S_b^k)$$

and since entropies and transmissions are always nonnegative, $H_{S-S_a}(S_b^k) = 0$.

Consequently,

$$\begin{aligned} t_{S_a,S_b,k} &= H_{S-S_a}(S_b^k) - H_{S_a,S-S_a}(S_b^k) \\ &= H_{S-S_a}(S_b^k) - H_S(S_b^k) \\ &= 0 - 0 \\ &= 0. \end{aligned}$$

When $t_{S_a,S_b,k} > 0$, therefore, the supposition that $t_{S_a,X_j,k} = 0$ for all $\overline{X}_j$ in $\overline{S}_b$ must be false.

Q. E. D.

Even in a state-determined system, one cannot in general infer from $\overline{S}_a \rightarrow t \rightarrow \overline{X}_j$ that there is some $\overline{X}_i$ in $\overline{S}_a$ such that $\overline{X}_i \rightarrow t \rightarrow \overline{X}_j$. The situation is somewhat different, then, for the DIE and the DIT. There is a simple relation between the DIE of a system and the DIE of a related system formed by grouping variables into subsystems; the relation is more complex for the DIT.

Next, recall theorem IV.2, which said that if $\overline{X}_i \xrightarrow{k} \overline{X}_j$, there must be a chain of k arrows in the DIE linking $\overline{X}_i$ to $\overline{X}_j$. The corresponding statement for transmissions is not true; $\overline{X}_i \rightarrow t \xrightarrow{k} \overline{X}_j$ is possible when there is no chain of arrows in the DIT from $\overline{X}_i$ to $\overline{X}_j$. One would expect that if $\overline{X}_i$ were to have a k-transmission to $\overline{X}_j$, this would have to come about by $\overline{X}_i$ having an immediate transmission to the whole system $\overline{S}$, $\overline{S}$ having an immediate transmission to itself, and $\overline{S}$ having an immediate transmission to $\overline{X}_j$, so that $\overline{S}$ would be a "channel" for the k-transmission. Surprisingly, this is not necessary; below is a frequency table $\underline{N}(X^{\tau}, X^{\tau+1}, X^{\tau+2})$ for a system $\overline{S} = \left\{ \overline{X} \right\}$, and

		$\langle X^{\tau}, X^{\tau+1} \rangle$			
		<1,1>	<1,2>	<2,1>	<2,2>
$X^{\tau+2}$	1	1	1	0	0
	2	0	0	1	1

from this table one calculates that $\overline{X}$ has no immediate transmission to itself, but that it does have a k-transmission to itself (for k = 2) of 1 bit.

(The table could represent the transition frequencies for a Markov chain, if zero in the table were replaced by ϵ and 1 were replaced by 1-ϵ ; $t_{X,X}$ could then be made arbitrarily small, and zero in the limit.)

For state-determined systems, however, $\overline{S}$ may be viewed as a "channel" for the k-transmission.

Theorem IV.7

Let $\overline{S}_a$ and $\overline{S}_b$ be subsystems of a state-determined system $\overline{S}$. If $\overline{S}_a \rightarrow t_0 \xrightarrow{k} \overline{S}_b$, then $\overline{S}_a \rightarrow t_1 \rightarrow \overline{S}$, $t_1 \geqslant t_0$, and $\overline{S} \rightarrow t_2 \rightarrow \overline{S}$, $t_2 \geqslant t_0$, and $\overline{S} \rightarrow t_3 \rightarrow \overline{S}_b$, $t_3 \geqslant t_0$.

Proof:

$$t_0 = t_{S_a,S_b,k} = H_{S-S_a}(S_b^k).$$

Now $$t_1 = H_{S-S_a}(S')$$

$$= H_{S-S_a}(S') + H_{S',S-S_a}(S_b^k)$$

The last term is zero, since $\overline{S}$ is state-determined.

$$t_1 = H_{S-S_a}(S', S_b^k)$$

$$= H_{S-S_a}(S_b^k) + H_{S_b^k,S-S_a}(S')$$

$$t_1 \geqslant H_{S-S_a}(S_b^k) = t_0 \ .$$

Next, $$t_2 = H_{S-S}(S') = H(S')$$

$$= H(S') + H_{S'}(S_b^k)$$

$$= H(S', S_b^k)$$

$$= H(S_b^k) + H_{S_b^k}(S')$$

$$t_2 = H_{S-S_a}(S_b^k) + T(S-S_a : S_b) + H_{S_b^k}(S')$$

$$\geqslant H_{S-S_a}(S_b^k) = t_0 \ .$$

Last, $t_3 = H_{S-S}(S_b^k) = H(S_b^k)$

$$= H(S_b^k) + H_{S'}(S_b^k)$$

$$= H(S' , S_b^k)$$

$$= t_2 \geqslant t_0 \text{ .}$$

Thus t_1, t_2, and t_3 are all at least as large as t_0.

Q. E. D.

In summary, the DIT is similar to the DIE in many ways when the system diagrammed is state-determined, but otherwise its properties are quite different and only weak generalizations may be made about it. Even so, it is a useful device for displaying cause-effect relations in a system of parts. The next section will discuss the strengths and weaknesses of the DIE and DIT.

4.2.3. Comments on the DIE and DIT

In the same way that a hammer is well suited to driving nails while useless for tightening nuts, the DIE and DIT are tools which are well suited to a particular class of problems and naturally poorly suited to others. Both diagrams have arisen from the question, which parts of this system affect which others? But the emphases in the two cases are slightly different, for the DIE deals with which parts *might* affect which others (within the constraints imposed by the system's mapping), whereas the DIT deals with which variables actually *do* affect which others, and how much. Both display the answer in a pictorial way which allows one to get a grasp of the system-as-a-whole; the DIT can be drawn with the thickness of the arrows proportional to the corresponding transmissions, making the representation even more vivid. When this is done with the example on page 114 the result is as shown in Figure 23.

Moreover for a system whose behavior slowly changes, a movie-style sequence of DIT's (one for each epoch in the system's history) could represent gross features of the changes in a similarly vivid way.

The major drawback of the DIT is its inability to adequately represent cause-effect relations in which the "effect" is caused by several variables acting in concert, unless these variables are explicitly grouped as components of a compound variable represented in the diagram. For the variables may only have an effect via their participation in the group (as in the example on page 117), and equally, variables which individually have effects may have none as a group, if some cancel the effects of others. Indeed the latter phenomenon is the essense of regulation, and it will be discussed more fully later.

There is another disadvantage of the DIT which is important if the diagram is based on observation of a real system; the length of the protocol required to minimize the effects of random sampling grows [roughly] exponentially with the number of variables. For this reason and others, $T(X_i : X_j')$ is in some ways a more practical measure of the effect $\overline{X}_i$ has on $\overline{X}_j$ than is $T_{S-X_i}(X_i : X_j')$; in the next section we will explore that transmission and its uses.

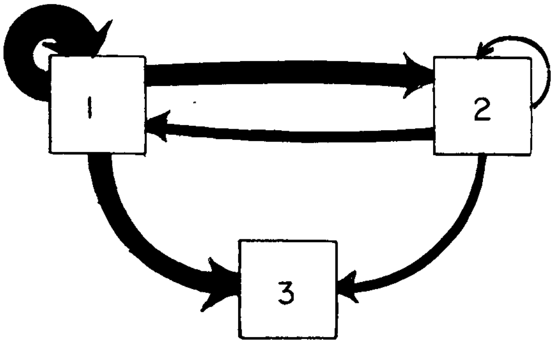


Figure 23.

4.3. Decomposition of system constraints

Introduction

A dynamic system of M supervariables, observed over n time intervals, provides values for Mn variables. The total constraint over this set of variables cannot, in general, be decomposed into a sum of constraints over proper subsets; this was shown in section 3.3. The total constraint can, however, be decomposed into constraints holding within subsets and between these subsets, and various decompositions of this type will be discussed in this section.

After a general consideration of such decompositions, a method of decomposing hierarchical systems will be proposed and illustrated.

4.3.1. Total constraint

In this section we will be considering the constraint over the set of variables $\left\{ X_j^i \mid 1 \leq i \leq n, \overline{X}_j \in \overline{S} \right\}$ representing a dynamic system of M super-variables over a duration of n consecutive time intervals. These variables correspond to the values which might appear in a protocol of length n:

$$\begin{array}{cr|c|c|c|c|c|}
 & \text{time:} & 1 & 2 & 3 & \cdots & n \\
\hline
 & \overline{X}_1: & X_1^1 & X_1^2 & & & X_1^n \\
\hline
 & \overline{X}_2: & X_2^1 & & & & \\
\hline
\overline{S} & \vdots & & & & & \\
\hline
 & \overline{X}_M: & X_M^1 & & & & X_M^n \\
\hline
\end{array}$$

We will denote the above set by Σ_n, with additional identifying subscripts, when necessary, preceeding the n:

$$\Sigma_{a,n} = \left\{ X_j^i \,\middle|\, 1 \leq i \leq n, \bar{X}_j \in \bar{S}_a \right\}.$$

The quantity of primary importance for a dynamic system $\bar{S} = \left\{ \bar{X}_1, \bar{X}_2, \ldots \bar{X}_M \right\}$ is the <u>total transmission in $\bar{S}$ over n time intervals</u>, $T(\Sigma_n)$. It is the grand transmission measuring the constraint over all nM variables - M variables for each of the n time intervals. $T(\Sigma_n)$ is an upper bound for the magnitudes of all transmissions and interactions involving any or all of the variables. The following sections in this chapter are concerned primarily with different ways of decomposing this grand transmission into additive components, by viewing the super-system first as composed of interacting super-variables, next as a system with memory, and last as a group of interacting subsystems.

Normally, $T(\Sigma_n)$ increases without bound as $n \to \infty$, so we will use the superscript L as before to denote the normalizing-and-limiting operation:

$$T^L(\Sigma) = \lim_{n\to\infty} \frac{1}{n} T(\Sigma_n)$$

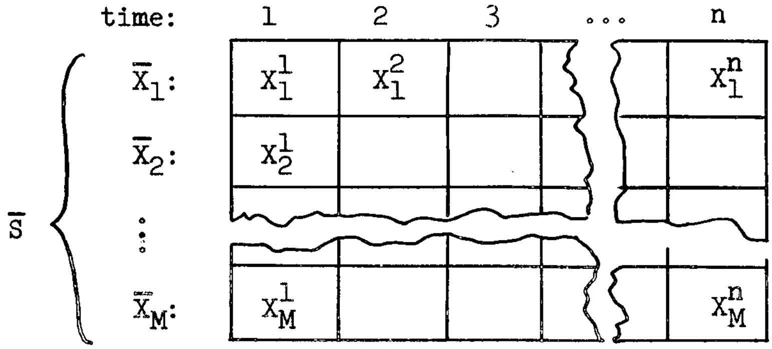
when the limit exists. $T^L(\Sigma)$ is the total transmission in the system per unit time interval.

<u>4.3.2. Two primary decompositions</u>

By decomposition of $T(\Sigma_n)$ we will mean expressing it as a sum of other transmissions. The primary <u>Decomposition Identity</u> is as follows:

$$T(S) \equiv \sum_{k=1}^{N} T(S_k) + T(S_1 : S_2 : \ldots : S_N)$$

where S is any set of variables and is the union of the disjoint subsets S_k, $1 \leq k \leq N$. The set Σ_n for a super-system $\bar{S} = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_M\}$ can be displayed in the manner shown below, which is meant to suggest a sample protocol of $\bar{S}$.



time:	1	2	3	∘∘∘	n
$\bar{X}_1$:	X_1^1	X_1^2			X_1^n
$\bar{X}_2$:	X_2^1				
⋮					
$\bar{X}_M$:	X_M^1				X_M^n

There are two primary ways to "slice" this display: into M horizontal strips representing the super-variables, and into n vertical strips representing the system at the different times.

We denote the set representing a horizontal "slice", $\{X_j^1, X_j^2, \ldots, X_j^n\}$, by $\chi_{j,n}$. The horizontal partitioning suggests the following version of the Decomposition Identity:

$$T(\Sigma_n) \equiv \sum_{j=1}^{N} T(\chi_{j,n}) + T(\chi_{1,n} : \chi_{2,n} : \ \cdots \ : \chi_{M,n}).$$

Consider first the terms $T(\chi_{j,n})$, representing constraints internal to the several super-variables. When we say a dynamic system exhibits memory, we mean that there is a constraint holding over the variables displaced in time. For memory implies a constraint, an effect of past system-values on the present value; a system without memory is one for which knowledge of the past and present is of no use in predicting the future. The constraint representing memory (over a finite time span) in the super-variable $\bar{X}_j$ is, in this view, just $T(X_j^1 : X_j^2 : \ldots : X_j^n)$

or $T(\mathcal{X}_{j,n})$. The summation in the identity therefore represents the memory-constraints in the M super-variables (over n time intervals).

The last term represents the constraint over the set $\left\{ < X_1^1, X_1^2, \ldots, X_1^n >, \ldots, < X_M^1, X_M^2, \ldots, X_M^n > \right\}$. It is the constraint, that is, binding the super-variables together (but over only a finite time span).

This decomposition would be appropriate, for instance, in studying the behavior of a married couple, with the "family" constraint decomposed into one memory constraint for the husband, another for the wife, and a term representing the bond between them.

Denoting, as before, the normalizing-and-limiting operation with a superscript L, we have

$$T^L(\mathcal{X}_j) = \lim_{n \to \infty} \frac{1}{n} T(\mathcal{X}_{j,n})$$

and

$$T^L(\Sigma) \equiv \sum_{j=1}^{N} T^L(\mathcal{X}_j) + T^L(\overline{X}_1 : \overline{X}_2 : \ldots : \overline{X}_M).$$

The last term is bounded by the constraint capacity of the super-system $\overline{S}$.

The previous decomposition was appropriate to the view of a system as a collection of interacting parts, each with memory. The next decomposition fits the view of a system as a number of parts mutually constrained at each instant, with memory being attributed to the system as a whole. Denoting, as before, the set $\left\{ X_1^i, X_2^i, \ldots, X_M^i \right\}$ by S^i and the set $\left\{ <S^1>, <S^2>, \ldots, <S^n> \right\}$ by $<\mathcal{S}_n>$, it is

$$T(\Sigma_n) \equiv \sum_{i=1}^{n} T(S^i) + T(<\mathcal{S}_n>).$$

The terms in the summation are the instantaneous constraints holding in each of the n time intervals, and the last term is the memory constraint for the compound super-variable $\overline{<S>}$ (note the difference between $\bar{S}$, a set of super-variables, and $\overline{<S>}$, a super-variable with components.) The term $T(<\mathcal{S}_n>)$ might be called the system memory constraint.

This decomposition is appropriate for structures of the form shown, for instance, in piano music, where the restriction to "harmonious chords" implies an instantaneous constraint while the restriction to "melodious chord sequences" implies a system memory constraint.

Application of the normalizing-and-limiting operation gives

$$T^L(<\mathcal{S}>) = \lim_{n\to\infty} \frac{1}{n} T(<\mathcal{S}_n>)$$

and

$$T^L(\Sigma) = \lim_{n\to\infty} \left(\frac{\sum_{i=1}^{n} T(S^i)}{n} \right) + T^L(<\mathcal{S}>).$$

The total constraint, per step, is the sum of the average instantaneous constraint and the system memory constraint (per step).

The two primary decompositions of $T(\Sigma_n)$ are by no means the only ones possible, and in the next section we turn to a hybrid type, decomposition of a system into subsystems with memory. First, however, it should be emphasized that the memory constraint for a compound variable may be less than, equal to, or greater than the sum of the memory constraints of the components. For example, if $\bar{S} = \left\{ \bar{X}, \bar{Y} \right\}$ and $y^\tau = x^{\tau-1}$ we can have $T(<\mathcal{S}_n>)$ less than $\left[T(\mathcal{X}_n) + T(\mathcal{Y}_n) \right]$ by having $\bar{X}$ be cyclic:

$$\begin{array}{ll} & \text{time:} \quad \ldots, \tau, \tau+1, \tau+2, \tau+3, \tau+4, \tau+5, \ldots \\ \bar{S} & \left\{ \begin{array}{l} \bar{X}: \quad \ldots, 1, 2, 1, 2, 1, 2, \ldots \\ \bar{Y}: \quad \ldots, 2, 1, 2, 1, 2, 1, \ldots \end{array} \right. \end{array}$$

$$T(<\mathcal{S}_n>) = n\text{-}1 \text{ bits}$$

$$T(\mathcal{X}_n) = T(\mathcal{Y}_n) = n\text{-}1 \text{ bits}$$

Or we can have $T(<\mathcal{S}_n>)$ greater than $\left[T(\mathcal{X}_n) + T(\mathcal{Y}_n)\right]$ by having $\bar{X}$ take values 1 and 2 equiprobably and independently:

$$\begin{array}{llllllllll} & \text{time:} & \ldots, & \tau, & \tau+1, & \tau+2, & \tau+3, & \tau+4, & \tau+5, & \ldots \\ \bar{S}\left\{\begin{array}{l} \\ \\ \end{array}\right. & \begin{array}{l} \bar{X}: \\ \bar{Y}: \end{array} & \begin{array}{l} \ldots, \\ \ldots, \end{array} & \begin{array}{l} 1, \\ , \end{array} & \begin{array}{l} 1, \\ 1, \end{array} & \begin{array}{l} 2, \\ 1, \end{array} & \begin{array}{l} 1, \\ 2, \end{array} & \begin{array}{l} 2, \\ 1, \end{array} & \begin{array}{l} , \\ 2, \end{array} & \begin{array}{l} \ldots \\ \ldots \end{array} \end{array}$$

$$T(<\mathcal{S}_n>) = n \text{ bits}$$

$$T(\mathcal{X}_n) = T(\mathcal{Y}_n) = 0 \text{ bits.}$$

If the supervariables are independent over the n time intervals, i.e., if $T(\mathcal{X}_{1,n} : \mathcal{X}_{2,n} : \ldots : \mathcal{X}_{M,n}) = 0$, then the system memory constraint exactly equals the sum of the individual memory constraints:

$$T(<\mathcal{S}_n>) = \sum_{j=1}^{M} T(\mathcal{X}_{j,n}) = T(\Sigma_n).$$

This follows immediately from corollary III.2, which gives

$$T(\mathcal{X}_{1,n}:\ldots:\mathcal{X}_{M,n}) = 0 \Rightarrow \sum_{i=1}^{n} T(S^i) = 0,$$

and from the decomposition identities for $T(\Sigma_n)$.

4.3.3. Hierarchical structures

One of the most time-honored and successful approaches to the study of complex systems has been to view them as composed of interrelated subsystems, to study each subsystem individually, and then to study the interrelation between them. The fact that this approach has been so successful for so long attests to the ubiquity of systems having structures amenable to the approach - structures in which the subsystems can be understood more or less adequately in isolation and in which the subsystems interact on a more or less global basis. Simon, in his delightful paper[11], deals at length with such systems and with a reason for their prevalence; he uses the word "hierarchical," as do we, to mean not only the type of structure in which each subsystem has a "boss," as in the organization of a business firm, but to include any type of structure in which the system is decomposable into interrelated subsystems (and perhaps the subsystems into sub-subsystems, and so on), as exemplified by a book which is composed of chapters, which are in turn composed of sections, which are divided into paragraphs, and so on.

Simon points out that the subsystems of most physical and biological hierarchies can be differentiated spatially, whereas social hierarchies are most easily decomposed by noting "who interacts with whom." This difference is largely irrelevant, however, for we note that in both cases, what allows a collection of parts to be reasonably called a subsystem is that those parts exercise a stronger effect on one another than on outsiders; that is, the cause-effect links or communication ties are disproportionately strong <u>within</u> the subsystem.

The Decomposition Identity is admirably suited to the decomposition of $\bar{\bar{S}}$ into N subsystems $\bar{\bar{S}}_k$, $1 \leq k \leq N$:

$$T(\Sigma_n) \equiv \sum_{k=1}^{N} T(\Sigma_{k,n}) + T(\Sigma_{1,n}:\Sigma_{2,n}: \dots :\Sigma_{k,n}).$$

The identity expresses the total constraint over $\bar{\bar{S}}$ as the sum of the individual constraints within the N subsystems, plus the constraint holding between the subsystems (considered as basic units); it thus corresponds precisely to viewing $\bar{\bar{S}}$ as a whole, on the left, and as a collection of N interacting subsystems, on the right. Furthermore, each term $T(\Sigma_{k,n})$ on the right may be decomposed by the same identity (or the earlier ones) into terms which correspond to viewing subsystem $\bar{S}_k$ as composed of interacting parts (or variables, etc.). And so on.

When n = 1, the identity is not well suited to decomposition of dynamic systems, for if one variable in a system has a direct effect on another that effect will usually show up most strongly one time interval later. On the other hand, the limiting form of the above identity,

$$T^L(\Sigma) \equiv \sum_{k=1}^{N} T^L(\Sigma_k) + T^L(\bar{S}_1 : \bar{S}_2 : \dots : \bar{S}_N),$$

while it represents the decomposition well, contains quantities difficult to estimate on the basis of experimental protocols unless those protocols are very long. For these reasons the identity for n = 2,

$$T(\Sigma_2) \equiv \sum_{k=1}^{N} T(\Sigma_{k,2}) + T(\Sigma_{1,2} : \Sigma_{2,2} : \dots : \Sigma_{N,2}),$$

is often the most useful.

We will next suggest a practical method for decomposing systems assumed to be hierarchical and then illustrate it with an example.

When one is confronted with a mass of data in the form of a protocol for a system $\bar{S}$, decomposing $\bar{S}$ into parts $\bar{S}_1, \bar{S}_2, \dots, \bar{S}_N$ in a "reasonable" way is a formidable undertaking, especially if little is known about the variables. The DIT is sometimes useful for detecting which variables strongly affect which others, i.e., for detecting a natural decomposition of $\bar{S}$, but a more generally useful measure is $T(X_i : X_j')$, the transmission between variable $\bar{X}_i$ at one moment and some other variable at the next. Of course the best measure of the interdependence of two super-variables is $T^L(\overline{X}_i : \overline{X}_j)$, but estimation of that number from a protocol leads to sampling problems unless the protocol is very long; $T(X_i : X'_j)$ is more convenient statistically and and also implies a direction - the effect of $\overline{X}_i$ *on* $\overline{X}_j$.

To illustrate how $T(X_i : X'_j)$ can be used to suggest a decomposition of $\overline{S}$ into parts, we simulated on a computer a simple network of one Markov source, one mapper, and three MWI's. We then obtained a 1000-step protocol of the system. The first fourteen steps of the protocol, a not atypical segment, are shown below.

time:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
S	1	1	1	1	2	2	1	1	2	2	1	1	2	2	1
X_1	1	1	3	3	3	2	2	1	1	2	2	1	1	2	2
X_2	1	1	2	2	1	1	1	2	2	2	1	1	2	2	2
X_3	1	3	3	3	3	3	2	2	1	3	2	2	1	3	2
X_4	1	2	2	1	1	1	2	2	2	1	2	2	2	2	1
X_5	1	1	2	1	1	1	1	2	1	2	2	1	2	1	2

Next, frequency tables were compiled and the transmissions $T(X_i : X'_j)$ were calculated. These were as follows:

$T(X_i : X'_j)$	X'_1	X'_2	X'_3	X'_4	X'_5
X_1	.124	.013	1.057	.131	.073
X_2	.002	.023	.002	.118	.012
X_3	.138	.012	.541	.036	.017
X_4	.002	.405	.002	.007	.017
X_5	.000	.182	.002	.210	.194

If the parts $\bar{X}_i$ are represented by ⓘ and arrows representing transmissions are drawn in one at a time, starting with the largest transmission $T(X_1 : X_3')$, the sequence shown in Figure 24 is obtained. The sequence suggests that $\bar{S}$ can be naturally decomposed into $\bar{S}_a = \left\{\bar{X}_1, \bar{X}_3\right\}$ and $\bar{S}_b = \left\{\bar{X}_2, \bar{X}_4, \bar{X}_5\right\}$.

In fact this suggestion is well in line with the facts. The DIE for the network is shown in Figure 25. Note the similarity between the DIE and the ninth diagram of the sequence.

The mappings for the mapper and MWI's are as follows:

MWI, #1:

			X_1		
	μ_1	1	2	3	
$\langle I, X_3\rangle$	1,1	1	1	1	
	1,2	1	1	1	
	1,3	3	1	3	
	2,1	2	2	2	
	2,2	2	2	2	
	2,3	2	2	2	(X_1')

MWI, #2:

		X_2		
	μ_2	1	2	
$\langle X_4, X_5\rangle$	1,1	1	1	
	1,2	1	1	
	2,1	2	2	
	2,2	1	2	(X_2')

MWI, #3:

		X_3			
	μ_3	1	2	3	
	1	3	1	3	
X_1	2	2	2	2	
	3	2	2	3	(X_3')

Mapper (with delay) #4:

		X_2		
	μ_4	1	2	
	1,1	2	1	
	1,2	2	2	
	2,1	2	1	
	2,2	2	2	
	3,1	1	1	
$\langle X_1,X_5\rangle$	3,2	1	1	(X_4')

MWI, #5:

		X_5		
	μ_5	1	2	
	1,1	1	2	
	1,2	2	1	
	2,1	1	2	
$\langle X_1,X_4\rangle$	2,2	2	1	
	3,1	1	1	
	3,2	1	1	(X_5')

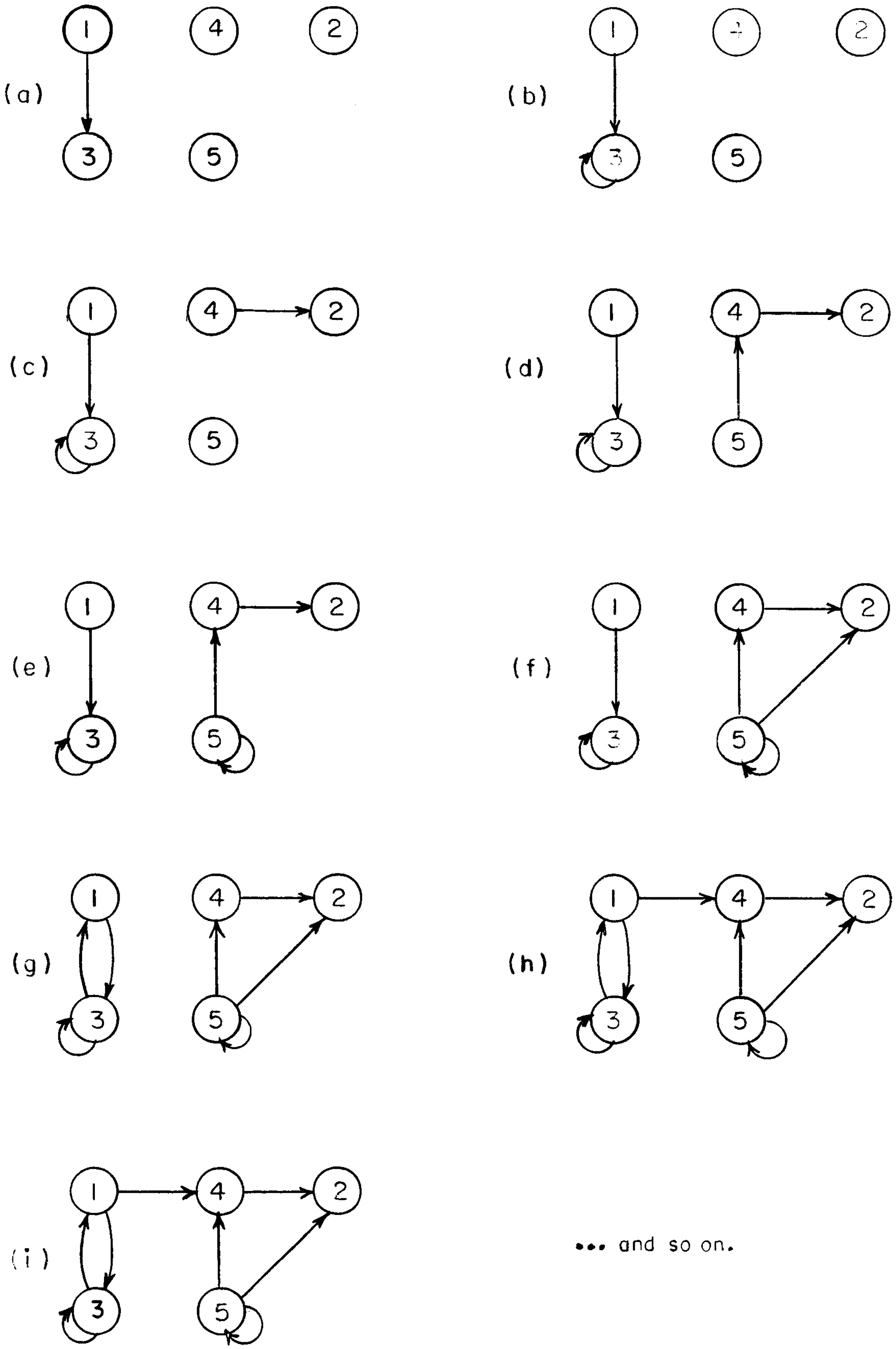

(a)
1
4
2
3
5
(b)
1
4
2
3
5
(c)
1
4
2
3
5
(d)
1
4
2
3
5
(e)
1
4
2
3
5
(f)
1
4
2
3
5
(g)
1
4
2
3
5
(h)
1
4
2
3
5
(i)
1
4
2
3
5
••• and so on.

Figure 24.

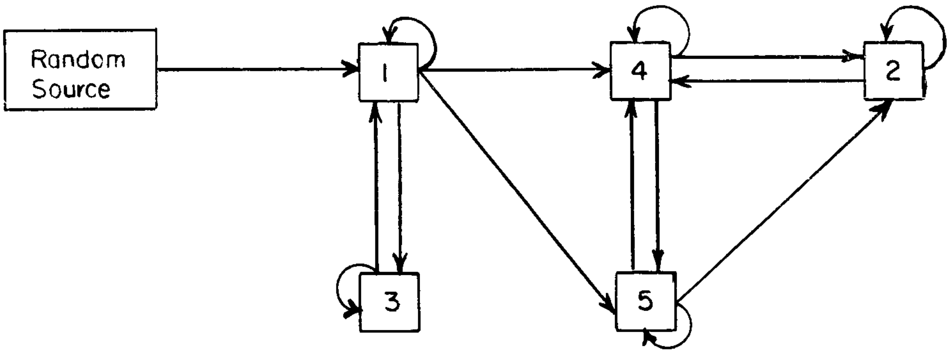


Figure 25.

Evidently a "good" decomposition of $\overline{\overline{S}}$, other things being equal, is one for which the number of parts is "reasonable" (perhaps approximately the square root of the number of variables) and the interpart constraint is small compared to the total - as small as possible, in fact. The identity and the associated experimental values for the decomposition $\overline{\overline{S}} = \overline{\overline{S}}_a \cup \overline{\overline{S}}_b$ are given below.

$$T(\Sigma_2) \equiv [T(\Sigma_{a,2}) + T(\Sigma_{b,2})] + T(\Sigma_{a,2} : \Sigma_{b,2})$$

$$5.101 = [1.957 + 2.722] + 0.422.$$

The transmission between the subsystems is only about 8% of the total, indicating that the choice of $\overline{\overline{S}}_a$ and $\overline{\overline{S}}_b$ is a reasonable one. By way of contrast, if $\overline{\overline{S}}$ is decomposed into $\overline{\overline{S}}_c = \left\{ \overline{\overline{X}}_1, \overline{\overline{X}}_2 \right\}$ and $\overline{\overline{S}}_d = \left\{ \overline{\overline{X}}_3, \overline{\overline{X}}_4, \overline{\overline{X}}_5 \right\}$, a decomposition which the $T(X_i : X_j')$ values imply is inappropriate, the following values result:

$$T(\Sigma_2) \equiv [T(\Sigma_{c,2}) + T(\Sigma_{d,2})] + T(\Sigma_{c,2} : \Sigma_{d,2})$$

$$5.101 = [0.168 + 1.966] + 2.967.$$

Here the transmission between subsystems accounts for 58% of the total, evidence that $\overline{\overline{S}}_c$ and $\overline{\overline{S}}_d$ do not constitute good choices for subsystems.

To continue the analysis, $\overline{\overline{S}}_a$ can be decomposed two ways - into individual memories plus intervariable constraint,

$$T(\Sigma_{a,2}) \equiv [T(\chi_{1,2}) + T(\chi_{3,2})] + T(\chi_{1,2} : \chi_{3,2})$$

$$1.957 = [0.124 + 0.541] + 1.292$$

or into instantaneous constraints plus system memory,

$$T(\Sigma_{a,2}) \equiv [T(S_a^1) + T(S_a^2)] + T(< \delta_{a,2} >)$$

$$1.957 = [0.144 + 0.144] + 1.669.$$

Neither decomposition is very successful; the numbers indicate a strong intervariable constraint and a strong system memory. The same is true of $\bar{S}_b$:

$$T(\Sigma_{b,2}) \equiv \left[T(\chi_{2,2}) + T(\chi_{4,2}) + T(\chi_{5,2})\right] + T(\chi_{2,2}:\chi_{4,2}:\chi_{5,2})$$

$$2.722 = \left[0.023 + 0.007 + 0.194\right] + 2.498$$

$$T(\Sigma_{b,2}) \equiv \left[T(S_b^1) + T(S_b^2)\right] + T(<\mathcal{S}_{b,2}>)$$

$$2.722 = \left[0.201 + 0.201\right] + 2.320$$

The indications are that $\bar{S}_a$ and $\bar{S}_b$ are not readily decomposable by these identities.

Analysis of $\bar{S}_a$ and $\bar{S}_b$ in terms of their kinematic graphs[6] bears out this conclusion. The kinematic graphs of $\bar{S}_a$, with $\boxed{i,j}$ representing the state $< X_1 = i, X_3 = j >$, are given in Figure 26. The arrows from transient states are shown dotted. S_a enters state $< 3,3 >$ only when the input contains a sequence of four or more consecutive 1's, and it leaves $< 3,3 >$ whenever the string of 1's ends.

The kinematic graph of $\bar{S}_b$, with $\boxed{i,j,k}$ representing state $< i, j, k >$, is shown in Figure 27. $\bar{S}_b$ tends to follow the cycle

$$\begin{tikzcd}
<1,2,1> \arrow[r] & <2,2,2> \arrow[r] & <2,2,1> \arrow[r] & <2,1,2> \arrow[r] & <1,2,2> \arrow[llll, bend left]
\end{tikzcd}$$

until $\bar{S}_a$ enters the rare state $< 3,3 >$, at which time $\bar{S}_b$ soon "resets" to $<1,1,1>$ and waits for $\bar{S}_a$ to change state; then $\bar{S}_b$ starts up again.

The decomposition identity suggested that $\bar{S}_a$ and $\bar{S}_b$ were only weakly interconnected, as is the case; $\bar{S}_a$ influences $\bar{S}_b$ only through the rare state $< 3,3 >$, and $\bar{S}_b$ does not affect $\bar{S}_a$ at all. Other identities suggested that the subsystems would be hard to break up. If the

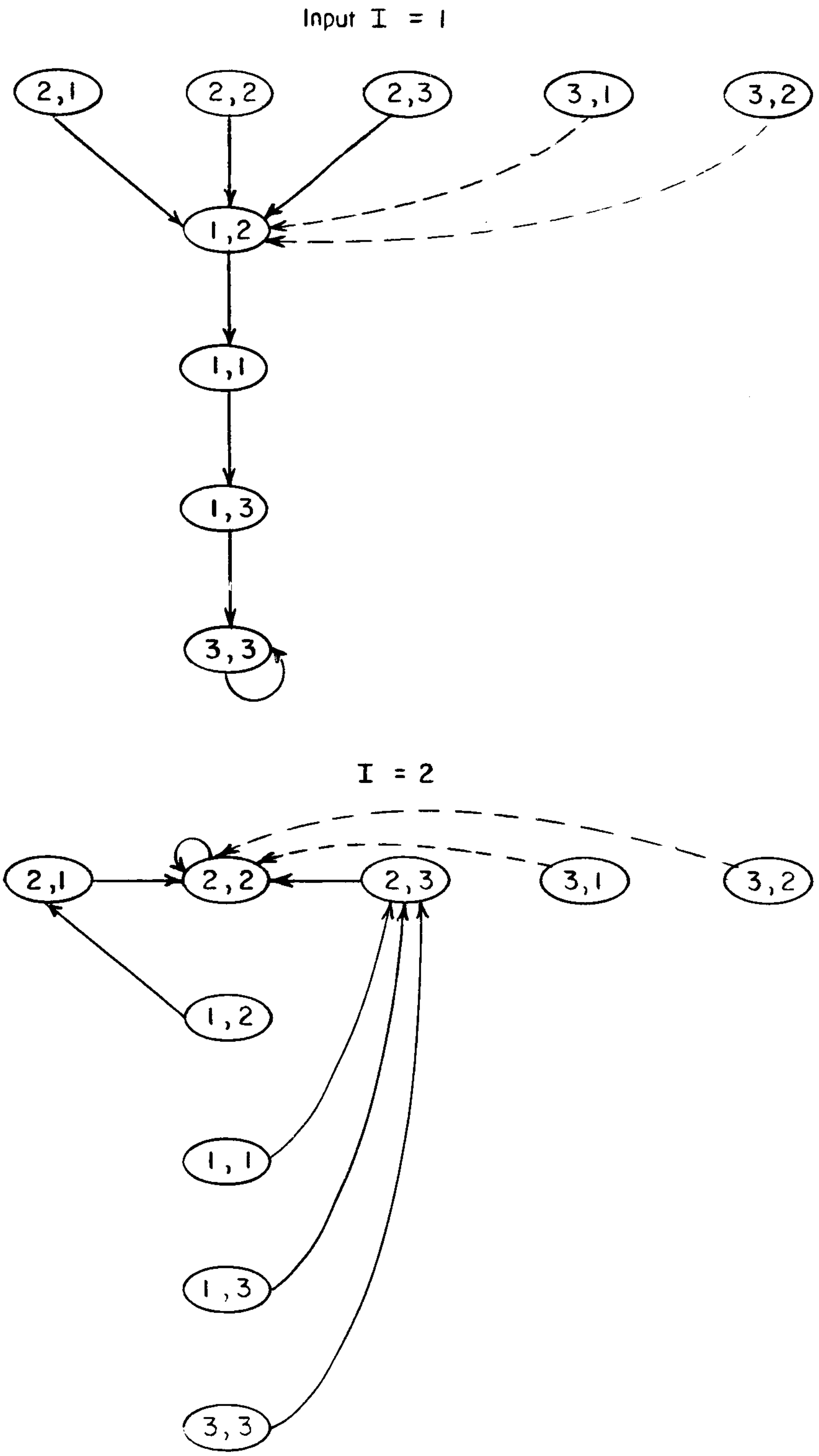
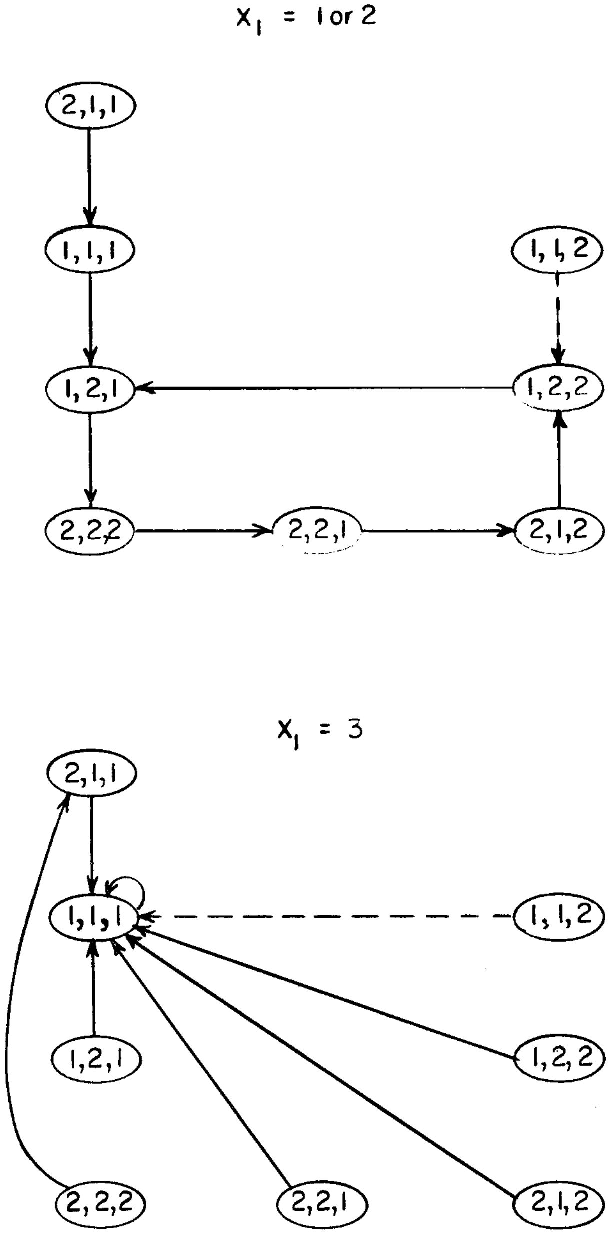


Figure 26.


X_1 = 1 or 2

X_1 = 3

Figure 27.

reader doubts it, let him try!

This example illustrates a method which could well be very useful in the decomposition of complex systems, particularly in situations where the experimenter has very little idea as to which variables can be naturally grouped. It is an all-too-common occurrence in science for an experimenter to be faced with a highly complex system in which data is easy to obtain but hard to "make sense of" because the experimenter does not know which variables are functionally "close" to which others. Faced with the overwhelming complexity of a large system such as a brain or an industrial society, the scientist may easily be defeated by the data unless some sort of simplification is possible. In such a case, the method outlined here may be a useful simplification since it suggests a natural decomposition of hierarchical systems.

The transmission $T(X_i : X_j')$ used in the method is a simple form of what we might call information transfer. The next section will take up in more detail the topic of information transfer.

4.4. Information transfer

Introduction

Frequently complex systems contain both sources of information and passive components which merely react to information. In this section we will comment on the information transfer in such systems, after first exploring information processes in purely stochastic and then in purely deterministic systems. The topic is important to the understanding of regulation in complex systems, since as we shall indicate in the chapter on regulation, regulators often take the form of deterministic subsystems accepting information and transforming it into appropriate regulatory action.

4.4.1. Information in Markov processes

If the process $\left\{X^1, X^2, X^3, \ldots\right\}$ is a Markov chain in which each variable X^k takes values from the finite set $X = \left\{x_1, x_2, \ldots, x_m\right\}$, it is natural to define the Markov super-variable $\overline{X} = < X^1, X^2, X^3, \ldots>$ corresponding to the process. The transition probability matrix for $\overline{X}$ is $\underline{P} = \left[p_{ij}\right]_{n,n}$:

$$p_{ij} = \text{Prob}\left\{ X^k = x_i \mid X^{k-1} = x_j \right\}$$

We deal here with discrete, ergodic Markov processes only.

From the definition of the Markov property,

$$p(X^{n+1} \mid X^1, X^2, \ldots, X^n) = p(X^{n+1} \mid X^n) \qquad \forall\, n \geqslant 1,$$

it follows that

$$H_{X^1, X^2, \ldots, X^n}(X^{n+1}) = H_{X^n}(X^{n+1}) \qquad \forall\, n \geqslant 1.$$

This is well known, but to the author's knowledge it is not well known that for an ergodic process the two statements are actually equivalent.

Theorem IV.8

If the process $\left\{X^1, X^2, \ldots, X^n, \ldots\right\}$ is ergodic, then the two statements below are equivalent.

(1) $\forall\, n \geqslant 1,\ p(X^{n+1} \mid X^1, X^2, \ldots, X^n) = p(X^{n+1} \mid X^n)$,

i.e., the process is Markovian.

(2) $\forall\, n \geqslant 1,\ H_{X^1, X^2, \ldots, X^n}(X^{n+1}) = H_{X^n}(X^{n+1})$.

Proof:

That (1) implies (2) is well established elsewhere[13]; we will show that (2) implies (1). The entropy equation implies the probability equation when n = 1 for any ergodic process, Markovian or not. For any $n \geqslant 2$, suppose that

$$H_{X^1, X^2, \ldots, X^n}(X^{n+1}) = H_{X^n}(X^{n+1}).$$

By definition of conditional transmission, then,

$$T_{X^n}(< X^1, X^2, \ldots, X^{n-1} > : X^{n+1}) = 0$$

which by corollary III.4 implies that

$$p(X^1, \ldots, X^{n-1}, X^{n+1} \mid X^n) = p(X^1 \ldots, X^{n-1} \mid X^n)\, p(X^{n+1} \mid X^n).$$

Multiplying by $p(X^n)$ gives

$$p(X^1, \ldots, X^{n+1}) = p(X^1, \ldots, X^n)\, p(X^{n+1} \mid X^n)$$

$$p(X^1, \ldots, X^n)\, p(X^{n+1} \mid X^1, \ldots, X^n) = p(X^1, \ldots, X^n)\, p(X^{n+1} \mid X^n)$$

Thus

$$p(X^{n+1} \mid X^1, \ldots, X^n) = p(X^{n+1} \mid X^n).$$

Q. E. D.

For ergodic processes, then statement (2) can be used as the definition of the Markov property.

The entropy of and constraint within a finite segment of an ergodic Markov chain are proportional to its length, and they obey the following equations:

$$H(X^1, X^2, \ldots, X^n) = nH_{X^1}(X^2) + T(X^1 : X^2)$$

$$T(X^1 : X^2 : \ldots : X^n) = nT(X^1 : X^2)$$

Moreover any ergodic process satisfying either of the above for all $n \geqslant 1$ is necessarily Markovian. These assertions are proved in the following:

Theorem IV.9

If $\{X^1, X^2, \ldots, X^i, \ldots\}$ is an ergodic process, then the three statements below are all equivalent.

(1) The process is Markovian.

(2) $\forall\, n > 1,\ H(X^1, X^2, \ldots, X^n) = nH_{X^1}(X^2) + T(X^1 : X^2)$

(3) $\forall\, n > 1,\ T(X^1 : X^2 : \ldots : X^n) = (n-1)T(X^1 : X^2)$.

Proof:

To show (1) $\Rightarrow$ (2): The identity

$$H(X^1, X^2, \ldots, X^n) \equiv H(X^1) + H_{X^1}(X^2) + H_{X^1,X^2}(X^3) + \ldots + H_{X^1, X^2, \ldots, X^{n-1}}(X^n)$$

together with the Markov property imply that

$$\begin{aligned} H(X^1, \ldots, X^n) &= H(X^1) + H_{X^1}(X^2) + \ldots + H_{X^{n-1}}(X^n) \\ &= H(X^1) + (n-1)\, H_{X^1}(X^2) \\ &= nH_{X^1}(X^2) + T(X^1 : X^2) \end{aligned}$$

for all n, so (1) $\Rightarrow$ (2). To show (2) $\Rightarrow$ (1), we assume (2) true and show by induction on m that for all $m \geqslant 1$, the following assertion follows:

$$\left\{ H_{X^1, \ldots X^k}(X^{k+1}) = H_{X^k}(X^{k+1}) \text{ for all } k,\ 1 \leqslant k \leqslant m \right\}.$$

The assertion is automatic for m = 1. For m = 2, we actually have only to show that the assertion holds for k = 2. The statement (2) above, with n = 3 and with liberal use of the property of stationarity, yields

$$H(X^1, X^2, X^3) = H(X^1) + H_{X^1}(X^2) + H_{X^2}(X^3).$$

This, with the identity

$$H(X^1, X^2, X^3) \equiv H(X^1) + H_{X^1}(X^2) + H_{X^1,X^2}(X^3)$$

establishes that

$$H_{X^1,X^2}(X^3) = H_{X^2}(X^3).$$

Thus the assertion is true for $m = 2$.

Next, suppose it true for $m - 1$. To show it also true for m requires only to prove it for $k = m$. Statement (2) and the property of stationarity yield

$$H(X^1, \ldots, X^{m+1}) = H(X^1) + H_{X^1}(X^2) + \ldots + H_{X^{m-1}}(X^m) + H_{X^m}(X^{m+1}).$$

The following is an identity:

$$H(X^1, \ldots, X^{m+1}) \equiv H(X^1) + H_{X^1}(X^2) + \ldots + H_{X^1,\ldots,X^{m-1}}(X^m) + H_{X^1,\ldots,X^m}(X^{m+1}).$$

The first m terms on the right of both equations are equal, term by term (since the assertion is true for $m - 1$). Consequently,

$$H_{X^1,\ldots,X^m}(X^{m+1}) = H_{X^m}(X^{m+1}).$$

We have just shown that if the assertion is true for $m - 1$, it is also true for m. Consequently, by induction it is true for all $m \geqslant 1$. Therefore, statement (2) implies that for all $m \geqslant 1$,

$$H_{X^1,\ldots,X^m}(X^{m+1}) = H_{X^m}(X^{m+1})$$

which by theorem IV.8 establishes that the process is Markovian. Thus $(2) \Rightarrow (1)$.

To show $(2) \Rightarrow (3)$ is simple. We assume, for any $n > 1$, that (2) is true.

$$- H(X^1, X^2, \ldots, X^n) = - nH_{X^1}(X^2) - T(X^1 : X^2)$$

Adding $nH(X^i)$ to both sides, we get

$$nH(X^1) - H(X^1, \ldots, X^n) = n\left[H(X^2) - H_{X^1}(X^2)\right] - T(X^1 : X^2)$$

$$T(X^1 : X^2 : \ldots : X^n) = (n - 1)\,T(X^1 : X^2)$$

showing that (2) $\Rightarrow$ (3). Reversing the process shows (3) $\Rightarrow$ (2).

Consequently, (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3).

Q. E. D.

This theorem and the one before provide four equivalent definitions for ergodic Markov processes. The quantifier "for all n $>$ 1" is essential, since non-Markovian processes can satisfy the criteria for all n up to a finite N_o. For example, if one writes down in order the binary equivalents of the series $\{0,1,2,\ldots,15,0,1,\ldots\}$,

$$\{\ 0000 \quad 0001 \quad 0010 \quad \ldots. \quad 1111 \quad 0000 \quad \ldots.\ \},$$

the resulting chain of 0's and 1's, which is certainly not Markovian, satisfies all the criteria for n $\leq$ 4. In fact one cannot conclude from any test based on observations of finite length that a process is Markovian, for one could never eliminate the possibility that the process was cyclic and only part of a cycle had been observed.

From the preceeding theorem it follows immediately that if $\bar{X}$ is a Markov supervariable (and ergodic, the only case we have considered), then

$$H^L(\bar{X}) = H_{X^1}(X^2)$$

and the per-step memory constraint is

$$T^L(\chi) = T(X^1 : X^2).$$

If $\overline{<S>} = < \bar{X}_1, \ldots, \bar{X}_M >$ is a Markov super-variable with components, the components need not themselves be Markov super-variables.

Obviously they may be, for instance if the components are independent, but the following transition matrix shows that they need not be.

$$\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{\langle X_1^k, X_2^k \rangle} \\
 & & 1,1 & 1,2 & 2,1 & 2,2 \\
\hline
 & 1,1 & 0 & 0 & 0 & 1 \\
 & 1,2 & 1 & 0 & 0 & 0 \\
\langle X_1^{k+1}, X_2^{k+1} \rangle & 2,1 & 0 & 1 & 0 & 0 \\
 & 2,2 & 0 & 0 & 1 & 0
\end{array}$$

Sample protocol:

	time:	1	2	3	4	5	6	7	8
$\overline{S}$	$\overline{X}_1$:	1	1	2	2	1	1	2	2
	$\overline{X}_2$:	1	2	1	2	1	2	1	2

Here $\overline{\langle S \rangle}$ and $\overline{X_2}$ are Markovian but $\overline{X_1}$ is not.

Whenever one or more components are not Markovian, however, there must be a constraint between the components if the whole is to be Markovian.

Theorem IV.10

Let $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$ and let $\overline{\langle S \rangle}$ be Markovian.

Then $\left\{ T(\overline{S}) = 0 \right\} \Rightarrow \left\{ \forall\, \overline{X}_j \in \overline{S},\ \overline{X}_j \text{ is Markovian.} \right\}$

Proof:

Suppose $T(\overline{S}) = 0$. By corollary III.2, $T(S^i) = 0$ for all i and consequently the system memory constraint is the sum of the individual memory constraints:

$$T(\langle S^1 \rangle : \langle S^2 \rangle : \ldots : \langle S^n \rangle) = \sum_{j=1}^{M} T(X_j^1 : X_j^2 : \ldots : X_j^n).$$

If $\overline{<S>}$ is Markov,

$$(n - 1)\ T(<S^1> : <S^2>) = \sum_{j=1}^{M} T(X_j^1 : X_j^2 : \ldots : X_j^n)$$

$$(n - 1)\left[\sum_{j=1}^{M} T(X_j^1 : X_j^2)\right] = \sum_{j=1}^{M} T(X_j^1 : X_j^2 : \ldots : X_j^n)$$

Therefore

$$\sum_{j=1}^{n} \left[T(X_j^1 : X_j^2 : \ldots : X_j^n) - (n - 1)\ T(X_j^1 : X_j^2)\right] = 0.$$

For every j, the quantity in brackets is nonegative. To see this, we expand both parts by identities, and use the stationary property freely:

$$T(X_j^1 : \ldots : X_j^n) \equiv (n - 1)H(X^1) - \left[H_{X^1}(X^2) + H_{X^1,X^2}(X^3) + \ldots + H_{X^1,X^2,\ldots,X^{n-1}}(X^n)\right]$$

$$(n - 1)T(X^1 : X^2) \equiv (n - 1)H(X^1) - \left[H_{X^1}(X^2) + H_{X^2}(X^3) + \ldots + H_{X^{n-1}}(X^n)\right]$$

By subtracting the second identity from the first, we obtain on the right a sum of transmissions, for

$$-H_{X^1,X^2,\ldots,X^k}(X^{k+1}) + H_{X^k}(X^{k+1}) = T_{X^k}(<X^1, \ldots, X^{k-1}> : X^{k+1}).$$

A sum of nonnegative quantities is zero only if each term is zero; consequently for every $j \leq M$,

$$T(X_j^1 : X_j^2 : \ldots : X_j^n) = (n - 1)T(X_j^1 : X_j^2)$$

and each $\overline{X}_j$ is Markovian, by theorem IV.9.

Q. E. D.

4.4.2. Information in state-determined systems

The sequence of states in a state-determined system[6] with $< S^i > = < X_1^i, X_2^i, \ldots, X_M^i >$,

$$\left\{ < S^1 >, < S^2 >, \ldots, < S^i >, \ldots \right\},$$

represents a special case of a Markov process, in which all the conditional probabilities are either 0 or 1. The system's mapping, f_σ, maps the set of states into itself; given the present state s^τ in S, the probability that the next state will be $f_\sigma(s^\tau)$ is 1. This of course means that $H_{<S^i>} (< S^{i+1} >) = 0$ for all i and consequently that

$$H(< S^1 >, < S^2 >, \ldots, < S^n >) = H(< S^1 >).$$

We assume the system to have a finite number of states, so that $H(< S^1 >)$ is finite. The < and > marks are actually redundant in H and T expressions and will be omitted henceforth. In the notation of section 4.3,

$$H(\Sigma_n) = H(S^1)$$

The uncertainty in a sequence of length n is precisely the uncertainty as to the initial state of the sequence. The per-step entropy of the sequence (in the limit) is consequently zero, which is to say that the sequence carries no information (except information about the initial state):

$$H^L(\Sigma) = \lim_{n \to \infty} \frac{H(S^1)}{n} = 0.$$

The components $\overline{X}_j$ carry no information either, in the limit.

In fact any deterministic sequence has a per-step entropy of zero.

Any state-determined system (with a finite number of states) will eventually fall into a cycle of behavior[6], and the components, if the state is compound, must then fall into cycles also. The behavior of each component is then deterministic and predictable without reference to any other component, so that when $\overline{<S>}$ is state-determined and finite,

$$H^L(\Sigma) = 0,$$
$$H^L(\chi_j) = 0 \qquad \text{for all } j \leq M,$$
$$T^L(\chi_1 : \chi_2 : \ldots : \chi_M) = 0.$$

Although the observation is somewhat frivolous and not very meaningful, it could be pointed out that since $T(\chi_1 : \chi_2 : \ldots : \chi_M) = 0$ always, any part of a state-determined system, when viewed as a channel between two other parts, has a channel capacity of zero. The Markov super-variable $\overline{<S>}$ suggested by the state-sequence is not necessarily ergodic nor even stationary; in fact the sequence of entropies $H(S^1)$, $H(S^2)$, ..., $H(S^i)$, ... is monotonically decreasing, since

$$H(S^i, S^{i+1}) \equiv H(S^i) + H_{S^i}(S^{i+1})$$
$$H(S^i, S^{i+1}) \equiv H(S^{i+1}) + H_{S^{i+1}}(S^i)$$

and consequently

$$H(S^i) - H(S^{i+1}) \equiv H_{S^{i+1}}(S^i) - H_{S^i}(S^{i+1})$$
$$= H_{S^{i+1}}(S^i) \geq 0.$$

Since the $H(S^i)$ are monotonically decreasing, so are the $T(S^i : S^{i+1})$, for

$$T(S^i : S^{i+1}) \equiv H(S^{i+1}) - H_{S^i}(S^{i+1})$$
$$= H(S^{i+1}).$$

The constraint in the sequence $< S^i >$ is the strongest mathematically possible:

$$T(S^1 : S^2 : \dots : S^n) = \sum_{i=1}^{n} H(S^i) - \left[H(S^1) + H_{S^1}(S^2, S^3, \dots, S^n)\right]$$

$$= \sum_{i=2}^{n} H(S^i).$$

The interactions $Q(S^{i+1}, S^{i+2}, \dots, S^{i+n})$ take a particularly simple form and are also monotonically decreasing in magnitude:

$$Q(S^{i+1}, S^{i+2}, \dots, S^{i+n}) = (-1)^n H(S^{i+n}).$$

To establish this, we let i = 0 for convenience and use induction on n. For n = 3,

$$Q(S^1, S^2, S^3) \equiv T_{S^1}(S^2 : S^3) - T(S^2 : S^3)$$

$$= H_{S^1}(S^3) - H_{S^1S^2}(S^3) - H(S^3) + H_{S^2}(S^3)$$

In a state-determined system, the entropy of any S^n conditional on S^k, with $k < n$, is zero; given the state at any time k, one can calculate with no uncertainty what the state will be at any later time. Therefore all the subscripted terms above are zero and

$$Q(S^1, S^2, S^3) =-H(S^3).$$

Now we suppose that $Q(S^1, S^2, \dots, S^n) = (-1)^n H(S^n)$ or, more conveniently for our purposes, that $Q(S^2, S^3, \dots, S^{n+1}) = (-1)^n H(S^{n+1})$, a mere relabeling. From the iterative definition of Q,

$$Q(S^1, S^2, \dots, S^{n+1}) = Q_{S^1}(S^2, \dots, S^{n+1}) - Q(S^2, \dots, S^{n+1}).$$

The subscripted term could be expanded into a sum of entropy terms, but the subscript of each would contain S^1 and consequently all would be zero. By inspection, then $Q_{S^1}(S^2, \dots, S^{n+1}) = 0$ and

$$\begin{aligned} Q(S^1, S^2, \ldots, S^{n+1}) &= -Q(S^2, \ldots, S^{n+1}) \\ &= -\left[(-1)^n H(S^{n+1})\right] \\ &= (-1)^{n+1} H(S^{n+1}). \end{aligned}$$

Therefore, by induction we conclude that for all $n \geqslant 3$,

$$Q(S^1, \ldots, S^n) = (-1)^n H(S^n).$$

Q. E. D.

If $\overline{\langle S \rangle} = \langle \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \rangle$ is a state-determined Markov super-variable with components, the components need not themselves be state-determined. The example in the last section, with protocol as follows,

	time:	1	2	3	4	5	6	7	8
$\overline{S}$	$\overline{X}_1$:	1	1	2	2	1	1	2	2
	$\overline{X}_2$:	1	2	1	2	1	2	1	2

illustrates this; $\overline{\langle S \rangle}$ and $\overline{X}_2$ are state-determined but $\overline{X}_1$ is not.

In an analogy to theorem IV.10, however, we can prove that if $\overline{\langle S \rangle}$ is state-determined while some component $\overline{X}$ is not, then there must be a constraint between the components which "accounts" for the fact.

<u>Theorem IV.11</u>

Let $\overline{S} = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_M \right\}$ and let $\overline{\langle S \rangle}$ be state-determined. Then

$$\left\{ T(\overline{S}) = 0 \right\} \Rightarrow \left\{ \forall\, \overline{X}_j \in \overline{S},\ \overline{X}_j \text{ is state-determined.} \right\}$$

<u>Proof:</u>

If $\overline{\langle S \rangle}$ is state determined, then $H_{S^i}(S^{i+1}) = 0$ for all $i \geqslant 1$.

Consequently for all $i \geqslant 1$,

$$H(S^{i+1}) - T(S^i : S^{i+1}) = 0.$$

It was shown earlier that when $T(\overline{S}) = 0$, the system memory constraint equals the sum of the memory constraints for the individual variables. Thus for all $i \geqslant 1$,

$$H(S^{i+1}) - \sum_{j=1}^{M} T(X_j^i : X_j^{i+1}) = 0.$$

It was also shown, in corollary III.2, that $T(\overline{S}) = 0$ implies $T(S^i) = 0$ for all $i \geqslant 1$. This in turn implies that

$$H(S^{i+1}) = \sum_{j=1}^{M} H(X_j^{i+1}).$$

Therefore we conclude that for all $i \geqslant 1$,

$$\sum_{j=1}^{M} \left[H(X_j^{i+1}) - T(X_j^i : X_j^{i+1}) \right] = 0$$

$$\sum_{j=1}^{M} H_{X_j^i} (X_j^{i+1}) \qquad = 0.$$

This sum of non-negative quantities is zero if and only if for every $j \leqslant M$, and for all $i \geqslant 1$,

$$H_{X_j^i}(X_j^{i+1}) = 0,$$

that is, if and only if each $\overline{X}_j$ is state-determined.

Q. E. D.

Having considered Markov processes and state-determined systems, we turn in the following section to systems which are part random and part deterministic: systems involving both Markov sources and finite state machines.

4.4.3. Information transfer through finite-state machines

Any arbitrarily complex network involving finite-state machines (machines-with-input, mappers, and automata) and Markov sources may be viewed as a single Moore automaton driven by a single Markov source, both the state of the source and the state of the automaton having, in general, several components (see Figure 28). Although it is not always advantageous to view a network this way, the fact that it is possible makes it evident that we should understand the information transfer in this paradigm case before attempting more complex cases. The understanding of this simple case is also essential to the understanding of later sections on regulation.

The fundamental information quantity associated with any finite-state machine is its channel capacity. The capacity of a mapper is log M, where M is the number of distinct values in the range of the mapping. The channel capacity of a MWI is log W_o, with W_o as defined by Shannon[5]. And section 3.6 of this report has provided a way to calculate the channel capacity of an automaton. That section also provided a procedure for constructing a source which maximizes $T^L(\overline{X} : \overline{Y})$, and therefore also $H^L(\overline{Y})$, at the capacity.

It is interesting and useful to note that if the output (i.e., state) sequence of a machine-with-input has the highest possible limit-entropy (or just "entropy", for this discussion), then the sequence is a Markov chain. Thus if the output is not Markov, one may be sure that the MWI is not operating at capacity. In the case of an information-preserving MWI (a MWI for which one can deduce, by observing any allowable output sequence, exactly which input sequence caused it) this is almost obvious, since the input must be zero-order Markov to realize capacity in that case. That the output must be Markov in the more general case follows from the fact that if a

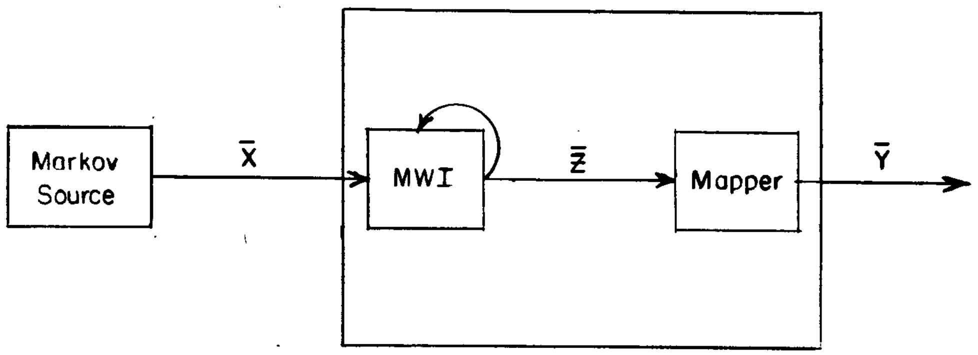


Figure 28.

distribution $P(Z^N \mid Z^1, Z^2, \ldots, Z^{N-1})$ is to maximize the output entropy, it must make all allowable state sequences of length N (as N goes to infinity) equally likely; that fact actually specifies the distribution, which Shannon has shown is Markov[5]. From this point of view, a MWI operating at capacity is a device for transforming an input which is not Markovian (in general) into an output which is.

We will consider now the problem of finding how much information the output sequence carries when driven by a Markov source of known characteristics. We assume that we are given a state-transition matrix $\underline{P} = [p_{ij}]$ for a Markov source, and mappings f_o and g_o for the MWI and mapper;

$$f_o\text{:} \quad X \times W_o \rightarrow W_o$$

$$g_o\text{:} \quad W_o \rightarrow Y.$$

The situation is represented in Figure 29.

If the input to a MWI is Markov, the state-transition sequence is only Markov under exceptional conditions, and information is usually lost in the MWI (that is, one cannot usually deduce what the input sequence was from the state sequence alone). Our job of finding the output entropy is considerably simplified if we break the MWI into two parts - a new MWI which does not lose information, and a mapper which does, as suggested in Figure 30. The new MWI is constructed so that for every z_j in Z, f maps X x z_j one-to-one onto Z, and g_1 is constructed so that the sequence W_o is the same as with the original MWI. This amounts to the introduction of extra states in the MWI, so as to make $\bar{\bar{Z}}$ a noiseless coding of $\bar{\bar{X}}$, and the subsequent elimination of the extra states by an information-losing mapping. For example, if

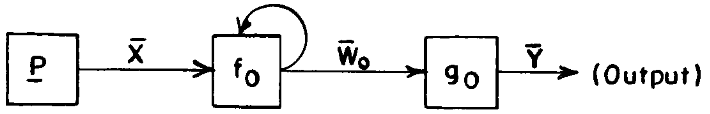


Figure 29.

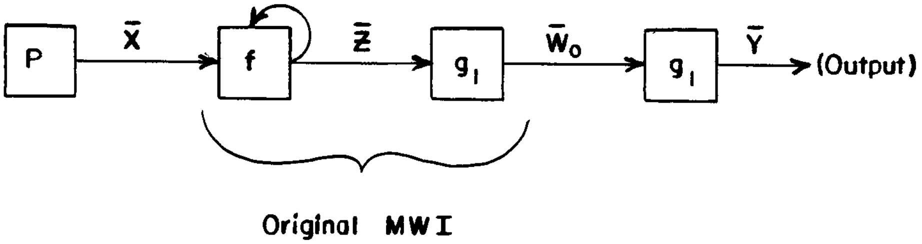


Figure 30.

f_o is given as

				W_o		
	f_o	1	2	3	4	
	1	1	2	4	3	
X	2	3	$\underline{2}$	$\underline{4}$	2	
	3	2	1	$\underline{4}$	1	W_o'

with the multiple entries (which make f_o information-losing) underlined, we could construct f and g_1 as follows:

					W_o				
	f	1	2	3	4	5	6	7	
	1	1	2	4	3	2	3	3	
X	2	3	5	6	2	5	2	2	
	3	2	1	7	1	1	1	1	W_o'

				W_o			
g_1	1	2	3	4	5	6	7
	1	2	3	4	2	4	4
				Z			

When this done, $\overline{Z}$ is a second-order Markov process,

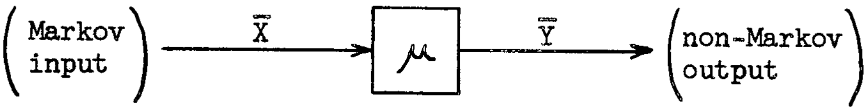
$$p(Z^{i+1} \mid Z^1, \ldots, Z^{i-1}, Z^i) = p(Z^{i+1} \mid Z^{i-1} Z^i) \qquad \forall\, i \geqslant 2,$$

since given Z^{i-1} and Z^i, one can deduce X^{i-1}, and the further uncertainty about Z^{i+1} is exactly the further uncertainty about X^i. To find the output entropy, then, we need only to consider how mapping a second-order Markov process by an information-losing mapping, $(g_0 g_1)$, changes the entropy. By a change of variables,

$$U^i = < Z^{i-1}, Z^i >$$
$$V^i = < Y^{i-1}, Y^i >$$

the problem is simplified still further, since if $\bar{Z}$ is second-order Markov, $\bar{U}$ is [first-order] Markov.

Thus by successive steps we can reduce the original problem to the problem of finding the output entropy which results when a Markov input sequence $\bar{X}$ is mapped by a convergent mapping μ into a non-Markov output sequence $\bar{Y}$.



The exact solution to this problem is not known, but for ergodic chains an approximate answer can be obtained from the inequalities

$$H_{X^1, Y^2, \dots Y^n}(Y^{n+1}) \leq H^L(\bar{Y}) \leq H_{Y^1, Y^2, \dots, Y^n}(Y^{n+1})$$

in which the outside quantities converge monotonically to $H^L(\bar{Y})$ as n goes to infinity[14].

The fact that a finite-state machine with Markov input usually has a non-Markov output does not in any way imply that information is necessarily lost. Indeed, it is possible to have an arbitrarily long chain of finite-state machines, for example MWI's (see Figure 31), and as long as all of them are information-preserving, $H^L(\bar{Y}_n)$ will equal $H^L(\bar{X})$ even though $\bar{Y}_n$ will be (n + 1)-order Markov in general. An information-preserving MWI can be viewed as a coding device which

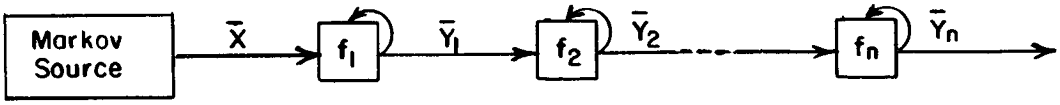


Figure 31.

encodes the input sequence into an output sequence in such a way that the span of intersymbol constraints is lengthened.

In fact, most finite-state machines have a tendency to increase the span of intersymbol constraints as they "transform" a sequence from input to output. By this is meant that if one must take n sequential symbols into account to get a reasonably good approximation for the input entropy,

$$\left| \frac{H_{X^{i+1},\ldots,X^{i+n-1}}(X^{i+n})}{H^{L}(\bar{X})} - 1 \right| < \epsilon \ll 1$$

then one must usually take more than n symbols into account to get an equally good approximation for output entropy. Finite-state machines tend to "spread out" the information, to put it loosely but picturesquely. This is, of course, only a tendency and not a law, the notable exception being when the input is matched to a MWI so as to realize the channel capacity; in that case quite the opposite takes place, for the output ends up Markov although the input seldom is.

In the light of Birch's results[14], and in view of the fact that when a Markov sequence is mapped by a convergent mapping the result is almost never a Markov sequence, it is rather surprising that a mapper may sometimes reduce the span of intersymbol constraints just as a MWI can. The example of section 3.6 shows this clearly; there the MWI part of the automaton transformed a non-Markov input sequence into a second-order Markov state sequence, and the mapper transformed that further into a [first-order] Markov output sequence.

4.4.4. Information transfer in networks of finite-state machines

The fact that the span of intersymbol constraints tends to increase as a message is passed through one or more finite-state machines greatly complicates the analysis of information transfer in complex networks of such machines, unless the network is viewed as a single automaton. One might think that the situation would become completely unmanageable in networks with feedback, for example the classic configuration shown in Figure 32. In this network, the input sequence is combined, by way of the mappings, with various vestiges of its own past; one would expect that the span of intersymbol constraints in the output sequence would be immense. In fact, however, if the MWI denoted by f_2 is operating at its own capacity (or close to it), the output sequence is Markov (or nearly so). We shall have more to say on this topic in later sections on regulation, and here it will suffice to point out that when an input sequence is "processed" by a network of finite state machines, what results need not necessarily have a larger span of constraints than the input.

We can deduce several inequalities relating the input, state, and output entropies for an automaton (see Figure 33).

The inequalities all derive from various decompositions of $H(\Sigma_n)$; for one,

$$\begin{aligned} H(\Sigma_n) \equiv {} & H(X^1, X^2, \ldots, X^n, Z^1, Z^2, \ldots, Z^n, Y^1, Y^2, \ldots, Y^n) \\ & H(X^1, \ldots, X^n) + H_{X^1, \ldots, X^n}(Z^1, \ldots, Z^n) \\ & + H_{X^1, \ldots, X^n, Z^1, \ldots, Z^n}(Y^1, \ldots, Y^n). \end{aligned}$$

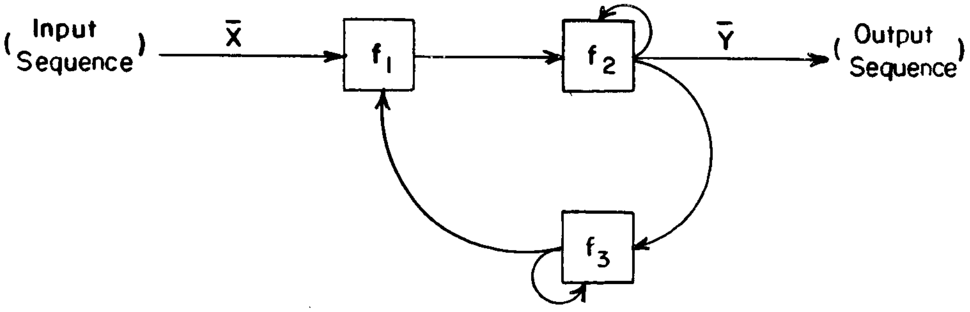


Figure 32.

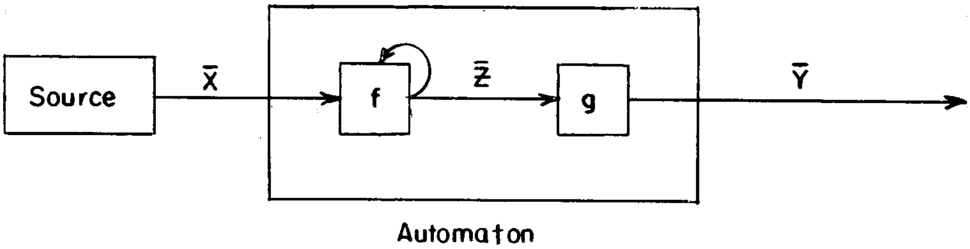


Figure 33.

If $X^1, \ldots, X^n$ and Z^1 are known, there is no uncertainty about $Z^1, \ldots, Z^n$. And if $Z^1, \ldots, Z^n$ are known there is no uncertainty about $Y^1, \ldots, Y^n$. Consequently

$$H(\Sigma_n) = H(X^1, \ldots, X^n) + H_{X^1, \ldots, X^n}(Z^1).$$

Another expansion of $H(\Sigma_n)$ is

$$H(\Sigma_n) = H(Z^1, \ldots, Z^n) + H_{Z^1, \ldots, Z^n}(X^1, \ldots, X^n) + H_{Z^1, \ldots, Z^n, X^1, \ldots, X^n}(Y^1, \ldots, Y^n).$$

The last term is zero as stated before. Putting the two expansions for $H(\Sigma_n)$ together, we obtain

$$H(Z^1, \ldots, Z^n) = H(X^1, \ldots, X^n) + H_{X^1,\ldots,X^n}(Z^1) - H_{Z^1,\ldots,Z^n}(X^1, \ldots X^n).$$

The negative term is the uncertainty about the input sequence which remains after one observes the state sequence. Dropping it gives

$$H(Z^1, \ldots, Z^n) \leq H(X^1, \ldots, X^n) + H_{X^1, \ldots, X^n}(Z^1)$$

or, a less strict inequality,

$$H(Z^1, \ldots, Z^n) \leq H(X^1, \ldots, X^n) + H(Z^1).$$

Of course since $H(Z^1, \ldots, Z^n, Y^1, \ldots, Y^n) = H(Z^1, \ldots, Z^n)$,

$$H(Z^1, \ldots, Z^n, Y^1, \ldots, Y^n) \leq H(X^1, \ldots, X^n) + H(Z^1).$$

In the limit, as $n \longrightarrow \infty$,

$$H^L(\overline{X}) \geqslant H^L(\overline{<Z,Y>}) = H^L(\overline{Z}) \geqslant H^L(\overline{Y}).$$

The entropy of a sequence, as it is transformed to state-sequence and output-sequence, can only fall; if one is more uncertain as to the output than the input (for a finite sequence) this surplus uncertainty is only due to uncertainty about the initial state of the network, and this finite uncertainty is relatively unimportant in the limit. In other words, finite state machines cannot generate information; they can only transform it or lose it.

Generalizing from the automaton to a network of interconnected finite state machines, this has the following consequences:

Theorem IV.12

Let $\overline{S}_X = \left\{ \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m \right\}$ be a set of supervariables which are inputs to a network of finite state machines, and let the state and output supervariables for the machines in that network constitute the set $\overline{S}_V = \left\{ \overline{V}_1, \overline{V}_2, \ldots, \overline{V}_n \right\}$. Then for any $n \geqslant 1$,

(a) $$H(S_V^1, \ldots, S_V^n) = H(S_X^1, \ldots, S_X^n) + H_{S_X^1, \ldots, S_X^n}(S_V^1) - H_{S_V^1, \ldots, S_V^n}(S_X^1, \ldots, S_X^n).$$

(b) $$H(S_V^1, \ldots, S_V^n) \leqslant H(S_X^1, \ldots, S_X^n) + H(S_V^1).$$

(c) $$H^L(\overline{S}_V) \leqslant H^L(\overline{S}_X).$$

The proof is a trivial extension of the foregoing argument. The theorem has some immediate consequences. For one, if $\overline{S}_j$ is any subset of $\overline{S}_V$, then $H(S_j^1, \ldots, S_j^n) \leqslant H(S_V^1, \ldots, S_V^n)$ and $H^L(\overline{S}_j) \leqslant H^L(\overline{S}_V)$, so the entropy of any subset of the machine's supervariables is bounded by the same quantities as the whole. This in turn implies that the limit-transmission between any k disjoint subsets of $\overline{S}_V$ satisfies the inequality

$$T^L(\overline{S}_{V1} : \overline{S}_{V2} : \ldots : \overline{S}_{Vk}) \leqslant (k-1)\, H^L(\overline{S}_X).$$

Limit-interactions are also bounded; all n-th order interactions (those with N variables in the argument) are bounded by $\pm\ 2^{N-2}H^{L}(\overline{S}_{x})$. There are, of course, analogous limits for the non-limit quantities.

Through the preceeding inequalities, the incoming entropy limits all information quantities relevant to the study of the network. We have in the theorem another verification that in a network of state-determined machines, with no information sources pumping in entropy, all limit-entropies, limit-transmissions, and limit-interactions are zero.

Notice that the theorem covers the nonergodic case (in statements (a) and (b)) as well as the ergodic.

The transmission between two complementary parts (whose union is $\overline{S}_{V}$) is bounded by $H^{L}(\overline{S}_{in})$. This fact will be important later when we consider networks decomposable into a regulator and a regulated part; the transmission between these parts is a crucial quantity.

An application of the cut set theorem of Elias et al[10] leads to a possibly smaller upper bound for the entropy of any subset $\overline{S}_{k}$ of $\overline{S}_{V}$. Suppose that a network of finite state machines and information sources (not necessarily Markov) is specified by giving all the mappings, all the interconnections between the parts, and the entropies of all sources. The channel capacities of all the finite state machines can be found, and a graph of the type shown in Figure 34 can be drawn. The graph is essentially a diagram of immediate effects, with the addition of the sources $\overline{X}_{i}$ and arrows showing which of the $\overline{V}_{j}$ in $\overline{S}_{V}$ they affect. Each line leaving a $\overline{V}_{j}$ is labeled with the channel capacity of the associated machine, and each line leaving an $\overline{X}_{i}$ is labeled with the source entropy. We assume for the time being that the graph is connected.

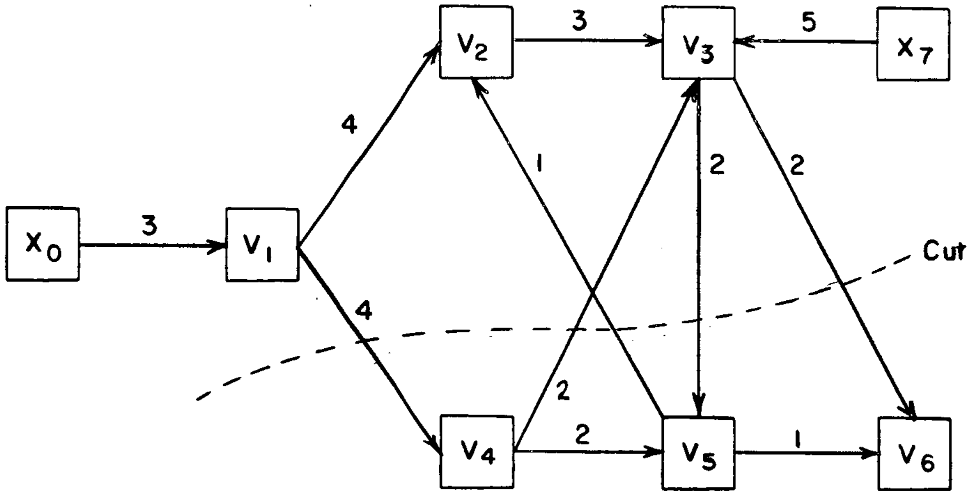


Figure 34.

A cut set on this graph is a set of arrows such that if all arrows were deleted, the graph would fall into two or more unconnected parts. A simple cut set is a cut set such that if any arrow is removed, what remains is not a cut set. For example, the cut shown by a dotted line on Figure 34 prescribes the simple cut set A:

$$A = \{1 \rightarrow 4,\ 4 \rightarrow 3,\ 5 \rightarrow 2,\ 3 \rightarrow 5,\ 3 \rightarrow 6\}.$$

With each set $\overline{S}_k \subset \overline{S}_Y$ there is associated a family of simple cut sets separating $\overline{S}_x$ from $\overline{S}_k$; the value of each simple cut set in the $\overline{S}_k$ family is the sum of the numbers on arrows crossing the cut in the direction of $\overline{S}_k$. If $\overline{S}_k = \{\overline{V}_4, \overline{V}_5\}$ the set A above is in the $\overline{S}_k$ family, and its value is $4 + 2 + 2 = 8$.

By slightly reinterpreting the cut set theorem, we conclude that the channel capacity from $\overline{S}_x$ to $\overline{S}_k$ cannot exceed the minimum value among all simple cut sets in the $\overline{S}_k$ family. With $\overline{S}_k = \{\overline{V}_4, \overline{V}_5\}$ the minimum value is 5, from the cut set B:

$$B = \{0 \rightarrow 1,\ 2 \rightarrow 3,\ 4 \rightarrow 3,\ 3 \rightarrow 5,\ 5 \rightarrow 6\}.$$

It follows that the limit-entropy of any variable or set of variables cannot exceed the minimum value among all simple cut sets separating it from $\overline{S}_x$; for the example $H^L(\overline{S}_k) \leq 5$.

We assumed above that the graph was connected. If it is not connected the same results hold; we need only redefine a cut set as a set of arrows such that their deletion separates the graph into more disconnected subgraphs than originally existed, and so on. If the original graph is not connected, and if we choose two variables in separate parts, it is plausible to conjecture that the transmission between them must be zero. This is indeed the case if the source driving the one part is independent of the source driving the other. For with the prototype graph of Figure 35, we have $H^L(\overline{X}_1, \overline{V}_1) = H^L(\overline{X}_1)$, $H^L(\overline{X}_2, \overline{V}_2) = H^L(\overline{X}_2)$, and $H^L(\overline{X}_1, \overline{V}_1, \overline{X}_2, \overline{V}_2) = H^L(\overline{X}_1, \overline{X}_2)$. Consequently

$$\left\{T^L(\overline{X}_1 : \overline{X}_2) = 0\right\} \Rightarrow \left\{T^L(\overline{<X_1, V_1>} : \overline{<X_2, V_2>}) = 0\right\}$$

and by corollary III.2, this implies $T^L(\overline{V}_1 : \overline{V}_2) = 0$.

Moreover it is reasonable to expect that if there is no chain of arrows leading either from $\overline{V}_i$ to $\overline{V}_j$ or from $\overline{V}_j$ to $\overline{V}_i$ in a connected graph, then $T^L(\overline{V}_i : \overline{V}_j) = 0$. But plausible or not, this conjecture is false, and to see that one need only consider the graph of Figure 36, in which $\overline{V}_1$ and $\overline{V}_2$ are identical machines subject to the same input: $\overline{V}_1$ and $\overline{V}_2$, being identical, behave identically, and $T^L(\overline{V}_1 : \overline{V}_2) = H(\overline{V}_1)$. We shall have more to say later about this important situation, with regard to regulation; for the moment it serves to illustrate the fact that there may be high transmission between two parts which have no direct effect on another via mappings or even via mediating variables.

With this background on information transfer in networks, we turn now to the subjects of regulation and of information transfer in regulatory networks.

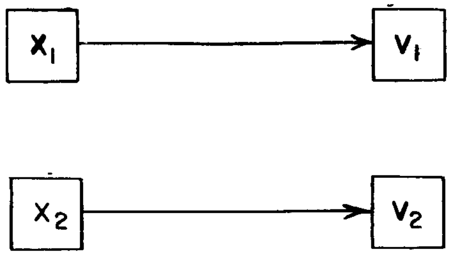


Figure 35.

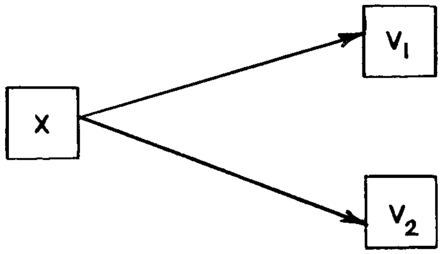


Figure 36.

V. REGULATION

Introduction

The preceeding chapters have been concerned with the relevance of information theory to complex systems in general; in this chapter we specialize to those systems in which one part is trying to regulate another part. Section 5.1 contains general remarks on regulation, and shows in a qualitative way the importance of information to successful regulation. Section 5.2 quantifies and proves more rigorously the results of the preceeding section, and section 5.3 provides an information analysis of three basic regulatory schemes. The paper is concluded with some brief, general remarks on regulation in section 5.4.

5.1. Information requirements for regulation

Up to this point, we have mentioned the topic of regulation only in passing; we have given several results showing how the methods of information theory are useful for the understanding of complex systems, without specifying any particular type of system. We will now turn attention specifically to complex systems in which regulation is involved - where one part of the system can be thought of as attempting to regulate some other part. By this we will mean that the regulator, which we will denote for brevity by R, and the part of the system being regulated against, X, jointly determine an outcome, Z, and that the goal of the regulator is to force the outcome (or outcomes, if the process is an ongoing one) to be favorable to R, by some [pre-established] criterion. The regulator tries to get its own way, in other words, in an outcome in which it is only one of the determining factors. The situation is represented in Figure 37.

We will impose few constraints on this very general formulation, leaving specialization for later. In particular we will leave open the questions of what sort of machinery is in the boxes marked X and R in the diagram above, and of what factors affect X and R, as indicated by the entrant arrows. We will also leave open the question of whether X is passive (as in the case of an automobile being regulated by a human pilot) or antagonistic to R (as in a game-playing situation in which X is trying to regulate R, just as R is trying to regulate X). The only constraints we will impose are as follows:

1. R, X, and Z are variables taking values from the sets $R = \{r_1, r_2, \ldots, r_m\}$, $X = \{x_1, x_2, \ldots, x_n\}$, and $Z = \{z_1, z_2, \ldots, z_p\}$ respectively.
2. The system operates on a discrete time basis.
3. The outcome is determined by R and X through a mapping f_z. That is,

$$f_z : X \times R \rightarrow Z.$$

Seen in this general formulation, regulation is a pervasive feature of everyday life, ranging from simple acts such as taking an aspirin to ward off a cold to highly complex phenomena such as government regulation of interstate commerce. With several examples we will

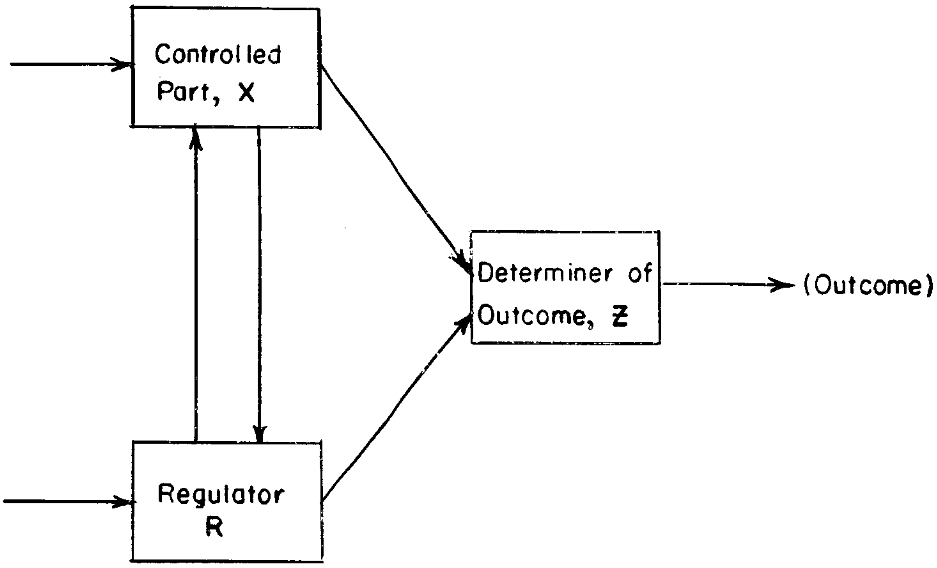

Controlled Part, X
Determiner of Outcome, Z
(Outcome)
Regulator R

Figure 37.

next illustrate different forms regulation can take.

One basic type of regulation is essentially an attempt on R's part to destroy X's ability to affect Z, by cutting off the effect-path from X to Z - to destroy the channel from X to Z, as we might put it. This type of regulation is usually a single-occurrence phenomenon, in which R takes one action to destroy the channel and thereafter need take no further action. The installation of stop signs at a busy intersection to minimize the probability of accidents there, and the deposit of a dime in a parking meter to regulate against ticket-issuing policemen, are examples. Examples of single-occurrence regulation in which the goal is preservation of constancy are: (1) assuring temperature constancy of an object by dropping it in the bottom of the ocean, (2) assuring constancy of room temperature by installing an automatic air conditioner, and (3) stabilizing the political climate in a totalitarian regime by imposing a news blackout on the press and radio. All of these examples illustrate how R can regulate against unwanted disturbances by incapacitating the mechanisms by which they would otherwise affect the outcome.

Regulation of quite a different type, and a type more interesting for this study, takes place when R cannot block the channel from X to Z but can only attempt to counteract the effect of X by appropriate counteraction of its own. This type of regulation is usually more dynamic than the type just mentioned. The goal of R can take the form of maximizing a probability, as when a doctor attempts to maximize the probability of "Patient Lives" when regulating against diseases, or when a fencer tries to maximize the probability of

"Avoids Being Hit" when regulating against his opponent. The goal can also take the form of preservation of constancy, as in (1) a thermostat maintaining constant room temperature despite changing weather, etc., (2) the driver of an automobile maintaining a constant speed despite hills, winds, and the like, and (3) in an open society, a government countering hostile propaganda with propaganda of its own, to preserve domestic tranquility.

The distinction drawn here between single-occurrence regulation and dynamic regulation, while useful, is somewhat artificial and arbitrary. For if a regulator takes a sequence of actions, the sequence may be viewed as many actions in an ongoing, dynamic process, or on the other hand as one choice of strategy or one trajectory. The distinction between the goals of maximizing a probability or preserving constancy is also arbitrary; nevertheless it is useful.

About the case of single-occurrence regulation there is not much to be said other than that if R selects one action out of a set of possible actions, and if that action is appropriate (i.e., is successful) while the others are not, then R needs information to make the selection. If a regulator selects appropriately to a degree better than chance, it must do so on the basis of information about which choice is appropriate. To select one action from a set of N possible actions, when all are equally attractive, requires log N bits of information.

If the selection is recurrent, so that the concepts of information theory become meaningful, much more can be said. We will deal henceforth with this class of "dynamic" regulators, which take on values as steps in a continuing process. Some regulators of this type deserve only brief mention; these are the regulators which take several actions (or values) but do so in an autonomous, deterministic way, such as the traffic lights which regulate traffic flow by their repeated cycles of red and green. We will be concerned, on the other hand, with regulators which must take in information and act appropriately on it in order to satisfy their goal criteria. Among situations which we normally regard as involving regulation, this situation is by far the predominant one.

We characterize the regulatory situation, then, as one in which to achieve its goal the regulator must (1) take in information by sensing some variables outside itself, (2) select from its repertoire of possible actions the one which is appropriate for attaining the goal, and (3) take that action. The process of regulation breaks up naturally into these three components, and the quality of regulation is governed by all three (of which we shall have more to say quantitatively later).

Information plays an important role in all of these steps; this is clear in the example of the fencer. To protect himself from his opponent, he must (1) take in visual information about his opponent's actions, (2) call on his knowledge and past training to select appropriate countermoves, and (3) perform the necessary maneuvers, which serve as input information for the opponent. Clearly the fencer's regulatory ability is dependent on all three ; if his input channel capacity is impaired (by dim lighting, poor eyesight, etc.), or if his selection is impaired (by lack of training, or drug-induced befuddlement), or if his performance of the selected maneuvers is impaired (by fatigue or physical weakness), he will be no match for an opponent not so disabled.

Similarly in the example of an automobile driver; when rain or fog cuts down the necessary input information, or when selection is impaired by fatigue, or when the capability for maneuvers is reduced by ice on the highways, the instinctive reaction is to slow down the vehicle in recognition of the fact that one's ability to regulate effectively is reduced.

The main factors opposing successful regulation, then, can be characterized as

(1) ignorance, or lack of input channel capacity,

(2) lack of insight, or lack of "computational" channel capacity transforming input information into appropriate outputs,

(3) impotence, or inability to influence the outcome successfully due to a lack of options, i.e. lack of output channel capacity.

In the next section we will investigate regulation in greater depth and attempt to quantify the qualitative assertion that information is of primary importance in any analysis of regulation.

5.2. Quantitative analysis of regulation

5.2.1. Regulation when the goal is to maximize a probability

We consider in this section and the next a mapping $f_{z_1} : X \times R \rightarrow Z_1$ and a continuing process (either finite or infinite in length) in which X and R take values at time τ and f_{z_1} determines the outcome at time τ. For example, f_{z_1} might be as follows:

		X			
	f_{z_1}	1	2	3	4
	1	1	2	3	1
R	2	5	1	2	4
	3	3	5	1	3

(Z_1)

Suppose that R's goal is to force the outcome to be "1". We can simplify the problem facing R by mapping Z_1 into Z by the rule: $Z = 1$ if Z_1 is an outcome acceptable to R, $Z = 0$ otherwise. This gives the following mapping $f_z : X \times R \rightarrow Z$.

		X			
	f_z	1	2	3	4
	1	1	0	0	1
R	2	0	1	0	0
	3	0	0	1	0

(Z)

We will assume in this section that the distribution of X's choices is fixed and independent of R; that is, we assume that $\underline{N}(X)$ or $P(X)$ is given. Under this assumption, what can be said about R's ability to force a desirable outcome? For concreteness, suppose X takes its four values equiprobably; then R can force a "1" half the time by perpetually taking the value R = 1. In fact if R chooses values independently of X, so that T(X : R) = 0, it is easy to show that this is the best R can do. To show this, we define the following:

$$P = \text{Prob}\left\{ Z = 1 \right\}$$

$$P_j = \text{Prob}\left\{ Z = 1 \right\} \text{ under the condition } \left\{ R^{\tau} = r_i \text{ for all } \tau \right\}$$

$$P^* = \max_{i=1}^{m} \left\{ P_i \right\}$$

r^* = the numerically lowest value in the set $\left\{ r_k \mid P_k = P^* \right\}$.

The definition of r* is a bit peculiar in order to single out only one of the set of "best" values.

Theorem V. 1

If $f_Z : X \times R \rightarrow Z$ where Z = 1 implies an outcome favorable to R and Z = 0 implies an outcome not favorable, and if P(X) is fixed and T(X : R) = 0, then the expectation of a favorable outcome cannot exceed P*.

Proof:

$$P = \sum_{< x_j, r_i > \in f_z^{-1}(1)} P(x_j, r_i)$$

$$= \sum_{< x_j, r_i > \in f_z^{-1}(1)} P(r_i) \cdot P(x_j)$$

$$= \sum_{i=1}^{m} P(r_i) \cdot P_i$$

$$\leq \sum_{i=1}^{m} P(r_i) \cdot P^* \leq P^*$$

Equalities are established if $P(r^*) = 1$.

Q. E. D.

The theorem says that if R is to choose values independent of X's values, it can do no better than to perpetually choose the value r^*. Thus if P is to exceed P^*, R must take values which are correlated with those of X; i.e., there must be transmission between X and R. Single-occurrence regulation corresponds to the choice of $R^{\tau} = r^*$ for all τ, and if dynamic regulation is to improve on that, there must be a channel linking X and R.

We must next construct a measure for the regulation imposed by R. We denote the measure by ρ_1. The simplest measure would be $\rho_1 = (P - P^*)$; however, this measure would not differentiate between one regulator raising the probability of a favorable outcome from 0.8 to 1.0, and another raising it from 0.05 to 0.25. Intuitively we feel that the latter has attained a more spectacular success, and that ρ_1 should be proportional to $\log\frac{P}{P^*}$. As a compromise between these contradictory demands, we define ρ_1 as follows:

$$\rho_1 = \left| P - P^* \right| \log \frac{P}{P^*}.$$

When $P^* = 0$, that is when no values of R can lead to a favorable outcome, the whole notion of regulation becomes absurd and ρ_1 is undefined.

In the example above, R can guarantee the desired outcome, that is, can make P = 1, by selecting its values according to the following mapping:

$$\downarrow \quad \begin{array}{c|cccc} X^{\tau} & 1 & 2 & 3 & 4 \\ \hline R^{\tau} & 1 & 2 & 3 & 1 \end{array}$$

In this case, $P = 1$, $P^* = 0.5$, $\rho_1 = 0.5$, and $T(X : R) = 1.5$ bits.

With the above definition of ρ_1, it follows immediately from theorem V.1 that $T(X : R) = 0$ implies $\rho_1 \leq 0$. Can ρ_1 and $T(X : R)$ be put in any other quantitative relation? We propose the following:

<u>Conjecture</u>:

$$\rho_1 \leq 2 \cdot T(X : R).$$

The conjecture can be supported as follows. When one tries to construct an f_z and a distribution $\underline{N}(X,R)$ for which the ratio $\rho_1/T(X :R)$ (or ρ_1/T) is as large as possible, it soon appears, through trial and error, that the ratio is largest when both ρ_1 and T are very small. T is made small by making the columns of $\underline{N}(X,R)$ nearly proportional. The mapping most favorable to regulation under these conditions is apparently an f_z of the following form,

$$\begin{array}{cc|cccccc} & & & & X & & & \\ f_z & & 1 & 2 & 3 & \cdots & m-1 & m \\ \hline & 1 & 1 & 0 & 0 & & 0 & 0 \\ & 2 & 0 & 1 & 0 & & 0 & 0 \\ & 3 & 0 & 0 & 1 & & 0 & 0 \\ R & \vdots & & & & & & \\ & m-1 & 0 & 0 & 0 & & 1 & 0 \\ & m & 0 & 0 & 0 & & 0 & 1 \end{array} \qquad (2)$$

since in this case P* can be made small while P can be made considerably larger with only a small T(X : R). The assignment N(X,R), with

$$n_{X,R} = \begin{cases} \frac{1}{m^2} + \epsilon(m-1) \text{ if } X = R; \ \epsilon \ll \frac{1}{m^2} \\ \\ \frac{1}{m^2} - \epsilon \qquad \text{if } X \neq R \end{cases}$$

has the following characteristics:

(1) The columns are nearly proportional, suggesting a minimal T(X : R).

(2) P* is as small as it can be with one 1 in each column of f_z, and (P - P*) is proportional to ϵ.

With this f_z and this N(X,R), ρ_1 is computed as follows:

$$P = m \left[\frac{1}{m^2} + (m-1)\epsilon\right] = \frac{1}{m} + m(m-1)\epsilon$$

$$P^* = \frac{1}{m}$$

$$\rho_1 = m(m-1)\epsilon \log \frac{\frac{1}{m} + m(m-1)\epsilon}{\frac{1}{m}}$$

$$= m(m-1)\epsilon \log \left[1 + m^2(m-1)\epsilon\right]$$

For very small ϵ,

$$\rho_1 \approx m(m-1)\epsilon\left[m^2(m-1)\epsilon\right] \log e$$

$$= m^3(m-1)^2\epsilon^2 \log e.$$

The transmission is computed as follows.

$$T(X : R) = H(R) - H_X(R)$$

$$= \log m + \left\{ \left(\frac{1}{m} + (m-1)m\epsilon\right) \log \left(\frac{1}{m} + (m-1)m\epsilon\right) + (m-1)\left(\frac{1}{m} - m\epsilon\right) \log \left(\frac{1}{m} - m\epsilon\right)\right\}$$

$$
\begin{aligned}
&= \log m - \log m \left[\frac{1}{m} - (m-1)m\epsilon + (m-1)\left(\frac{1}{m} - m\epsilon\right)\right] \\
&+ \frac{1}{m}\left\{\left[1 + (m-1)m^2\epsilon\right]\left[(m-1)m^2\epsilon - \tfrac{1}{2}(m-1)^2m^4\epsilon^2 + \ldots\right]\right. \\
&+ \left.\left[(m-1)\,(1-m^2\epsilon\,)\right]\left[-m^2\epsilon - \tfrac{1}{2}\,m^4\epsilon^2 - \ldots\right]\right\} \log e \\
&= \frac{1}{m}\left\{\tfrac{1}{2}(m-1)m^5\epsilon^2\right\} \log e + \ldots \\
&\approx \tfrac{1}{2}(m-1)m^4\epsilon^2 \log e.
\end{aligned}
$$

The ratio ρ_1/T is

$$
\begin{aligned}
\frac{\rho_1}{T} &\approx \frac{m^3(m-1)^2\epsilon^2 \log e}{\frac{1}{2}\,(m-1)m^4\epsilon^2 \log e} \\
&= 2\left(\frac{m-1}{m}\right).
\end{aligned}
$$

Consequently, ρ_1/T is less than 2 for any m. If this distribution is indeed the type that maximizes ρ_1/T, as there is good reason to believe, then $\rho_1 \leq 2\,T(X : R)$ always.

The transmission between X and R is thus seen to be an upper bound for regulation when the goal is maximizing the probability of a particular outcome or set of outcomes; if the goal is minimization of a probability, the same sort of analysis holds, for to minimize the probability that an event will occur is of course the same as to maximize the probability that it will not.

We will next consider regulation when the goal of R is to preserve constancy.

5.2.2. Regulation when the goal is to maintain constancy

In many situations involving regulation, the goal of the regulator is to preserve a variable or variables at as nearly a constant value as possible. The vast majority of the homeostatic mechanisms occurring in plants and animals are of this type, of course; for example, the mechanisms maintaining temperature and blood sugar levels in humans, or of moisture content in plants. Many mechanical regulators, such as thermostats, automatic volume controls, and automatic airplane pilots, are also of this type.

As has been pointed out by Ashby[6], regulation in such cases can frequently be viewed as blocking the transfer of information from X to Z. X takes various actions which would show up as variations in Z, were it not for appropriate counter-actions taken by R. If R is completely successful, variations in Z are completely eliminated, with the result that an observer of Z would obtain no information at all about the values taken by X or R. The goal of R, maintainence of constancy in Z, can thus also be seen as the suppression of entropy at the output.

We can consequently define a new measure for regulation, ρ, based on how much output entropy is eliminated by R's actions. To meaningfully compare the output entropy with R acting and R not acting (R fixed at some value, in other words) it is necessary to assume, for this section and most of the next, that X is passive and does not change its actions according to how R behaves. We will consider, then, situations in which the distributions for X are fixed, the process is a continuing one (finite or infinite), and the outcome at time τ is determined by X and R at time τ ; $f_z : X \times R \rightarrow Z$. For example, f_z might be as follows:

		X			
	f_z	1	2	3	4
R	1	1	2	3	1
	2	5	1	2	4
	3	3	5	1	3

(Z)

Suppose X takes its values independently and equiprobably, so that $P(x_i) = 1/4$, $1 \leq i \leq 4$. What will be the output entropy if R is fixed at some particular value? If $R^\tau = 1$ for all τ, the outcomes 1,2, and 3 will occur in the frequency ratios 2 : 1 : 1, and the output entropy will be

$$H^1(Z) = - \left[2/4 \log 2/4 + 1/4 \log 1/4 + 1/4 \log 1/4 \right] = 1.5 \text{ bits.}$$

Similarly with $R^\tau = 2$ for all τ we obtain $H^2(Z) = 2.0$ bits, and with $R^\tau = 3$ for all τ, we obtain $H^3(Z) = 1.5$ bits. The regulator can hold the output entropy to 1.5 bits by persistently taking values 1 or 3.

Now we ask, by how much further can R decrease the entropy through appropriate actions? Clearly the output entropy, H(Z), can be dropped to zero if R takes its values in accordance with this mapping:

↓

X^τ	1	2	3	4
R^τ	1	2	3	1

If regulation is measured by this further decrease in entropy, it comes to 1.5 bits. The regulator, by selecting values which are appropriately matched with those of X, can succeed in maintaining the output constant.

Let us define the following:

$$H = H(Z)$$

$$H^i = H(Z) \text{ under the condition } \left\{ R^\tau = r_i \text{ for all } \tau \right\} .$$

$$H^* = \min_{i=1}^{m} \left\{ H^i \right\}$$

$$r^* = \text{the numerically lowest } r_i \text{ in the set } \left\{ r_i \mid H^i = H^* \right\} .$$

$$\rho = H^* - H.$$

ρ, then, is a measure of the amount of output entropy which R suppresses by acting, beyond the amount which it could suppress by perpetually taking the value r*. We will proceed next to expand the expression H* - H, to show the relation of ρ to T(R : X).

We will denote with a superscript * those quantities which obtain when R is fixed permanently at r*. To get another expression equivalent to H*, we proceed as follows.

$$\begin{aligned} H^* &\equiv H^*(Z) \\ &\equiv H^*(X,Z) - H^*_Z(X) \\ &\equiv H^*(X) + H^*_X(Z) - H^*_Z(X) \end{aligned}$$

Now H*(X) = H(X), since we have assumed that the distribution for X is not dependent upon R's values. Also, $H^*_X(Z) = 0$ since Z is a determinate function of R and X. Consequently

$$H^* = H(X) - H^*_Z(X).$$

To get an expression equivalent to H,

$$\begin{aligned} H &\equiv H(R,Z) - H_Z(R) \\ &\equiv T(R : Z) + H_R(Z) \\ &\equiv T(R : Z) + H_R(X) + H_{R,X}(Z) - H_{R,Z}(X). \end{aligned}$$

Since $H_{R,X}(Z) = 0$, this simplifies to

$$H = T(R : Z) + H_R(X) - H_{R,Z}(X).$$

The difference between H* and H is ρ:

$$\rho = \left[H(X) - H^*_Z(X)\right] - \left[T(R : Z) + H_R(X) - H_{R,Z}(X)\right]$$

$$\rho = T(R : X) - T(R : Z) + \left[H_{R,Z}(X) - H^*_Z(X)\right].$$

Let us examine these terms in turn. T(R : X) is of course a measure of the coordination between R and X. It is bounded by H(R) and by H(X), which are indicators of the "activity" of R and X. In fact if

R takes values according to a mapping $\mu : X \longrightarrow R$, then $T(R : X) = H(R)$. The first term in the expression for ρ, therefore, indicates the statistical dependence of R on X.

The next term, $T(R : Z)$, can be interpreted as the amount of information one obtains about R by observing Z. Earlier it was remarked that this quantity is small to the degree that R regulates successfully; $T(R : Z)$ is bounded by $H(Z)$, the output entropy which R tries to minimize.

The last two terms, $H_{R,Z}(X)$ and $H^*_Z(X)$, can best be interpreted in terms of f_z. If f_z has the property that for any r_i, f_z maps $X \times r_i$ one-to-one into Z (that is, no r_i-row of f_z has any repeated entries), then $H_{R,Z}(X) = H^*_Z(X) = 0$, since given R and Z there is no uncertainty about X. In this case,

$$\rho = T(R : X) - T(R : Z)$$

and clearly $\rho \leq T(R : X)$ always. This inequality is closely related to, but not identical with, Ashby's "Law of Requisite Variety".

Back to interpreting the last two terms, it should be clear that $H_{R,Z}(X)$ and $H^*_Z(X)$ are nonzero only when there are rows of f_z (where rows correspond to values of R) with repeated entries, as in the example on page 184. Formally, let

$$k_{ip} = \text{number of X-values in the set } \left\{ x_j \mid f_z(x_j, r_i) = z_p \right\}$$

$$k = \max_{i,p} \left\{ k_{ip} \right\}$$

$$K = \log k.$$

Then no row of f_z has any z repeated more than k times, and consequently $H_{R,Z}(X)$ and $H*_Z(X)$ are both bounded by K. We will occasionally refer to the number k as the multiplicity of the mapping f_Z.

The contribution to ρ is the difference between $H_{R,Z}(X)$ and $H*_Z(X)$; the difference is of course bounded by K, and it can be positive or negative. Whenever $H_{R,Z}(X)$ is positive, $H*_Z(X)$ is necessarily positive, so the difference is in fact always less than K, if $K > 0$. We collect these relationships in the following theorem:

Theorem V.2

$$\rho = T(R : X) - T(R : Z) + \left[H_{R,Z}(X) - H*_Z(X) \right]$$

$$\rho \leq T(R : X) + K$$

The amount of regulation which R can impose is limited by the transmission between R and X, plus a quantity $H_{R,Z}(X) \leq K$.

Theorem V.3

$$T(R : X) = 0 \quad \Rightarrow \quad \rho \leq 0, \text{ regardless of } K.$$

Proof:

We need only to show that $T(R : X) = 0$ implies $H_{R,Z}(X) = H*_Z(X)$. Suppose $T(R : X) = 0$.

$$H_{R,Z}(X) = \sum_{i=1}^{m} P(r_i)\, H_Z^i\,(X)$$

where superscript i is used to indicate quantities which are defined under the condition $\left\{ R^\tau = r_i \text{ for all } \tau \right\}$. The identity

$$H^i(X) + H_X^i\,(Z) \equiv H^i(Z) + H_Z^i(X)$$

together with the fact that $H_X^i(Z) = 0$ gives

$$H_Z^i(X) = H^i(X) - H^i(Z).$$

Since the distribution of X does not depend on R, $H^i(X) = H*(X)$. Substituting in the first equation, we obtain

$$H_{R,Z}(X) = \sum_{i=1}^{m} P(r_i) \left[H^*(X) - H^i(Z) \right]$$

$$= H^*(X) - \sum_{i=1}^{m} P(r_i)\, H^i(Z)$$

On the right is a weighted sum of terms each at least as large as $H^*(Z)$. Thus

$$H_{R,Z}(X) \leq H^*(X) - H^*(Z).$$

The right side of this inequality is $H^*{}_Z(X)$, for

$$H^*_Z(X) \equiv H^*(X) + H^*_X(Z) - H^*(Z)$$

and $H^*_X(Z) = 0$. Q. E. D.

These last two theorems are cental to the understanding of regulation. The first shows that there is a very definite bound on regulation, this bound being the transmission between the regulator and the regulated variable, plus an additional term which can be thought of as indicating the congeniality of f_Z to regulation. The second theorem says that regardless of the mapping, unless the regulator is coordinated with the part it is trying to regulate it can do no better than to perpetually take the value r*; taking any other values can only degrade the regulation when $T(R : X) = 0$.

The situation is similar to that discussed earlier, where the goal of R was to maximize a probability. In both cases the goal can be partly attained by permanently taking a "best" value r*, and any improvement over that can only take place if the regulator is coordinated with the variable it hopes to regulate. Moreover the improvement is limited by the amount of that coordination.

These results can be generalized to include situations in which the goal of the regulator is to cause, at the output, a deterministic cycle of events, and to guard that cycle against disturbances from X. The goal is to preserve constancy of a repetitive output, in other words - a heartbeat cycle, say, or the wing-flapping cycle of a bird. Such situations may be encoded into a form in which the goal is constancy, as before, but it is more convenient to deal with them directly through a generalization of our previous results.

We will consider, therefore, supervariables $\bar{X}$, $\bar{R}$, and $\bar{Z}$ and the mapping $f_Z : X^\tau \times R^\tau \rightarrow Z^\tau$, and we will define quantities analogous to those used earlier in this section. Whereas before we used a superscript i to indicate quantities defined under the condition $\left\{ R^\tau = r_i \text{ for all } \tau \right\}$, here we use superscript j to indicate the condition $\left\{ \bar{R} = (\bar{r})_j \right\}$, i.e., the value $\bar{R}$ takes is the jth member of the set of all possible values for $\bar{R}$. (The members can be numbered, because the set of values is countably infinite as shown by the numbering scheme suggested below, when R^τ takes one of the values 1, 2, or 3:

j	$(\bar{r})_j$
0	1, 1, 1, 1, ...
1	2, 1, 1, 1, ...
2	3, 1, 1, 1, ...
3	1, 2, 1, 1, ...

and so on. In general,

$$j = \sum_{k=1}^{\infty} (r^k - 1)\,(3^{k-1}) \text{ where } r^k = pr_k(\bar{r})_j.)$$

Now, in a manner strictly analogous to the development before, we define

$$H^{L} = H^{L}(\bar{Z})$$

$$H^{Lj} = H^{L}(\bar{Z}) \text{ under the condition } \left\{ \bar{R} = (\bar{r})_j \right\} .$$

$$H^{L*} = \min_{j \geq 1} \left\{ H^{Lj} \right\} \text{ , or g.l.b. }_{j \geq 1} \left\{ H^{Lj} \right\} \text{ if there is no minimum.}$$

$$(\bar{r})^* = \text{the } (\bar{r})_j \text{ with smallest } j, \text{ in the set } \left\{ (\bar{r})_j \mid H^{Lj} = H^{L*} \right\} .$$

$$\rho^{L} = H^{L*} - H^{L}.$$

Some clarification may be helpful here. When we indicate that the output information $H^{L}(\bar{Z})$ is positive, this is subject to two interpretations. One is that even if we are given all preceeding values of Z in the sequence $\left\{ Z^1, Z^2, \ldots, Z^{\tau}, \ldots, Z^n \right\}$ we are nevertheless not certain what will come next, even in the limit as n ⟶ ∞ . Another interpretation is that in a number of "experiments" each yielding an infinite sequence $\left\{ Z^1, Z^2, \ldots. \right\}$, our uncertainty as to which sequence will occur in any particular experiment is infinite; that is, we cannot even designate beforehand a finite set of such sequences into which the new sequence must fall. This second interpretation should make it clear that the condition $\left\{ \bar{R} = (\bar{r})_j \right\}$ implies $H^{L}(\bar{R}) = 0$; that is, the regulator is deterministic. A deterministic regulator, undergoing deterministic behavior, can minimize the information in the output sequence by an auspicious choice of $(\bar{r})_j$. The degree to which the information is further reduced by non-deterministic behavior of the regulator is measured by ρ^{L}.

The reader should have little difficulty in seeing that our development of the expression for ρ serves also to yield an expression for ρ^L; one has only to superscript all the expressions with L throughout. The result is given in the following theorem:

Theorem V.4

$$\rho^L = T^L(\bar{R} : \bar{X}) - T^L(\bar{R} : \bar{Z}) + \left[H^L_{\bar{R},\bar{Z}}(\bar{X}) - H^{L*}_{\bar{Z}}(\bar{X}) \right]$$

$$\rho^L \leq T^L(\bar{R} : \bar{X}) + K$$

The amount of regulation which R can impose is limited by $T^L(\bar{R} : \bar{X})$, plus the quantity $H^L_{\bar{R},\bar{Z}}(\bar{X}) \leq K$. The situation is exactly analogous to that of theorem V.2.

Similarly the proof of theorem V.3, with only minor changes such as the substitution of $P\left[(\bar{r})_j\right]$ for $P(r_i)$, etc., serves as proof for the following:

Theorem V.5

$$T^L(\bar{R} : \bar{X}) = 0 \quad \Rightarrow \quad \rho^L \leq 0, \text{ regardless of K.}$$

This completes our generalization. The point of this chapter is just this: regulation, whether the goal is maximizing or minimizing the expectation of a particular set of outcomes, or is the suppression of entropy, H(Z), or information, $H^L(\bar{Z})$, can be partly attained by the choice of auspicious permanent values or deterministic sequences - by single-occurrence regulation, in other words. But to effect any improvement over that, the regulator must coordinate his actions with the system being regulated against, and the degree of that coordination sets a bound on the regulation which can be achieved.

5.3. Important special cases of regulation

The last section indicated the importance of the quantities

$T(R : X)$ and $T^L(\bar{R} : \bar{X})$ to regulation. Few constraints were placed on the general formulation, and in particular nothing was mentioned about which variables acted as input to the regulator R. In this section we will briefly examine some common regulatory situations in the light of the previous results.

5.3.1. Error-controlled feedback regulation

It is very common in texts on servomechanisms to see a diagram of the sort shown in Figure 38; X(s) is the "command" or reference input, E(s) is the "error" signal, and Y(s) is the "controlled output" signal. The servomechanism is generally considered successful if the error signal is kept within prescribed limits, or its root-mean-square value is lower than a given number, or some other criterion is satisfied.

From our point of view, the goal of the regulatory mechanism is to keep the error signal as nearly constant as possible. Preserving the topology but changing the names of the variables, we can redraw the diagram in our terms as shown in Figure 39. The mapping f_z corresponding to the subtraction device in the servomechanism has multiplicity one, i.e., $H_{R,Z}(X) = 0$. Consequently, from theorems V.2 and V.4,

$$\rho \leq T(R : X)$$

$$\rho^L \leq T^L(\bar{R} : \bar{X}).$$

This configuration has the interesting property that R receives information about X only through Z, and at the same time R is trying to suppress entropy at Z. The regulator thus appears to be cutting off its own source of information and lowering its own efficiency. Clearly it cannot be fully successful at eliminating H(Z), for if H(Z) were zero,

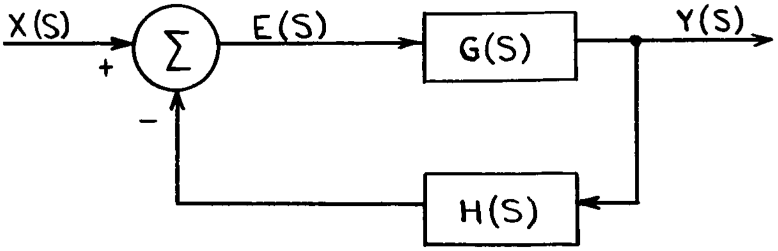


Figure 38.

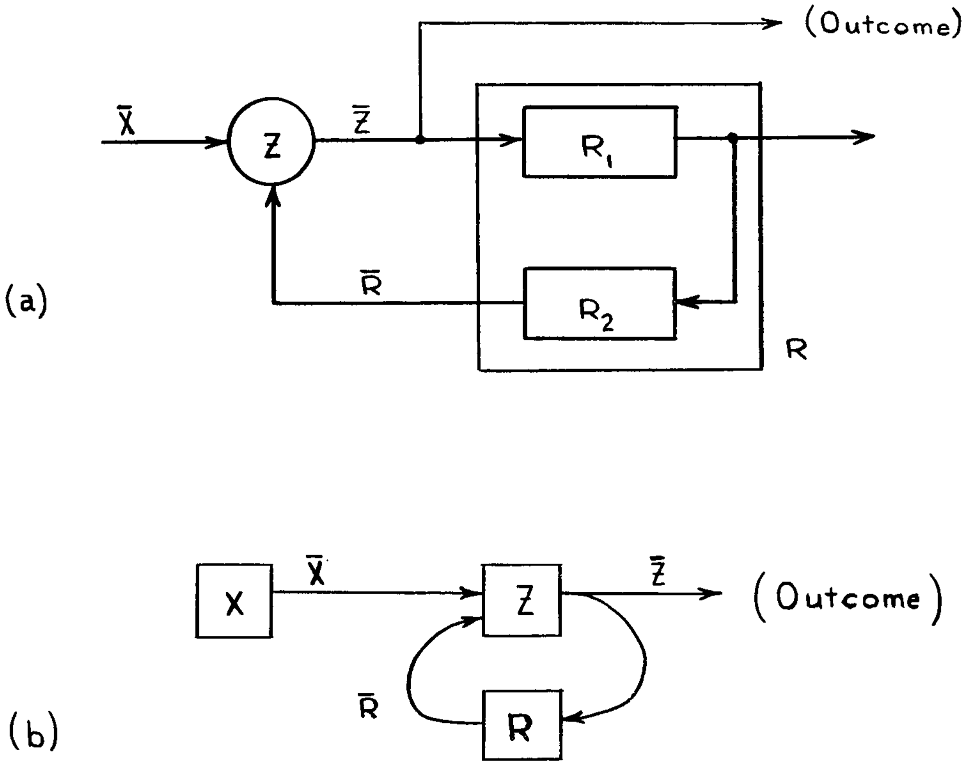


Figure 39.

$H^L(\overline{Z})$ would be zero and so also would $H^L(\overline{R})$, by theorem IV.12. If $H^L(\overline{X})$ were positive we would have a contradiction, because the subtraction device, if one of the two inputs is known, is not an information-losing mechanism. From this we conclude that regulation can never be fully successful in an "error-controlled" regulator, except in the degenerate case of a deterministic input.

What is perhaps more surprising is that ρ^L is necessarily zero! The "error" sequence must contain exactly as much information as the input sequence, regardless of the activity of R. To see this, we note that given a long sequence of Z, one can deduce the corresponding sequence of R (R, being passive, cannot generate information). And since f_z has multiplicity one, knowing $\overline{R}$ and $\overline{Z}$ is sufficient to deduce $\overline{X}$. Consequently from $\overline{Z}$ one can reconstruct $\overline{X}$; the reverse is also true, so $H^L(\overline{X}) = H^L(\overline{Z})$. It is for this reason that we hedged above in saying that R <u>appears</u> to be cutting off its own source of information; in fact, it doesn't. The regulator is a mere recoder, preserving the information but transforming it to a form with possibly lower entropy. The regulation ρ is the difference between the input entropy and the error entropy,

$$\begin{aligned}\rho &= H^* (Z) - H(Z)\\ &= H(X) - H(Z)\end{aligned}$$

since $H^*(Z) = H(X)$ whenever the multiplicity of f_z is one.

If there are no memory-constraints in the input sequence, i.e., if $H^L(\overline{X}) = H(X)$, then the regulator's task is completely hopeless, since such a sequence cannot be converted to a form with lower entropy without losing information. Consequently $\rho = 0$.

This observation can be generalized further: if $H^L(\bar{X}) = H(X) - M$, so that the input sequence has a memory-type constraint of M bits per step, then ρ cannot exceed M, and consequently

$$H(X) - M \leq H(Z) \leq H(X).$$

To show this we need only note that $H^L(\bar{Z}) = H^L(\bar{X}) = H(X) - M$ bits per step; the entropy H(Z) is minimized by encoding the information into a form with no memory constraints, i.e., a form with $H(Z) = H^L(\bar{Z})$, since $H(Z) < H^L(\bar{Z})$ is impossible. Therefore

$$H(Z) \geq H(X) - M$$

and $$\rho \leq H(X) - \left[H(X) - M\right] = M.$$

The regulation is limited by the amount of [per-step] sequential constraint in the input sequence.

It might appear that ρ is limited by the channel capacity of R, and that if the regulator is to achieve the maximum regulation of M bits per step, it must have a channel capacity of M bits per step, or more. This is not necessarily so. If the input is deterministic, for example, then $M = \left[H(X) - H^L(\bar{X})\right] = H(X)$, and R can achieve regulation $\rho = M$ by following a deterministic sequence absolutely identical to that of X. R can be a perfect regulator, that is, and can keep the error sequence absolutely constant, even with a channel capacity of zero.

However it is true that ρ is limited by the entropy of R, since $\rho \leq T(R : X) \leq H(R)$, and therefore if R is to regulate it must take more than one value. We might say that regulation is limited by the "variety" capacity of R.

To summarize: from the point of view of information theory, an error-controlled feedback regulator cannot reduce the information in the error sequence; it can only take advantage of sequential constraints in the input to reduce the entropy of the error sequence. If there are no such constraints, regulation is impossible.

We are led to suppose, therefore, that the great variety of applications in which error-controlled feedback regulators prove useful all have one thing in common: the input sequences have sequential constraints, and probably very strong constraints.

5.3.2. Feed-forward regulation

In the error-controlled regulator, R got its information about X by way of Z. In the configuration we will discuss next, R gets this information directly from X. This configuration, which we will call feed-forward regulation, is represented in Figure 40. This is the type of regulation which occurs when one starts to fall but catches himself, or when an army which has obtained access to the enemy's battle plan takes appropriate countermoves, or when an automobile driver activates his own brakes whenever he notices the car ahead braking.

In most practical applications, there is a delay between the time the regulator obtains information about X and the time it acts on that information. We will take this into account by assuming that X does not have an immediate effect on R but does have an effect on R one time unit later, i.e., that R^{τ} depends on $X^1, X^2, \ldots, X^{\tau-1}$ but not on X^{τ}. We will assume that R^{τ} is in fact determined by $X^1, X^2, \ldots, X^{\tau-1}$.

The constraint between X^{τ} and its predecessors in the X-sequence is $T(<X^1, X^2, \ldots, X^{\tau-1}> : X^{\tau})$; in the limit it is $[H(X) - H^L(\bar{X})] = M$. By the Collapsing Theorem for Transmission,

$$T(R^{\tau} : X^{\tau}) \leq T(<X^1, X^2, \ldots, X^{\tau-1}> : X^{\tau})$$

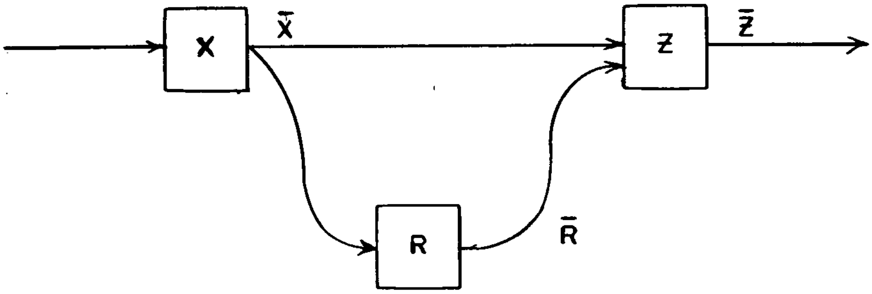


Figure 40.

since R^{τ} is a function of the earlier X's. Thus we have

$$T(R : X) \leq M$$

and consequently $\rho \leq M + K$, where 2^{K} is the multiplicity of the mapping f_{Z}.

The assumed time delay thus leads to the conclusion that ρ can only be positive when there is memory constraint in the input sequence, and ρ is limited by that constraint in the same way it was limited in the error-controlled feedback regulator (except for the additive term K, which in the feedback case we assumed was zero). This is only common sense, of course; if R is to regulate on the basis of the past history of X, there must be some correlation between that past and the present value which R is trying to counteract.

If f_{Z} has multiplicity one, then just as in the case of the feedback regulator R cannot reduce H(Z) to zero except in the degenerate case of a deterministic X. And just as in that case, and for the same reasons, the channel capacity of R is not necessarily a bound for ρ.

If f_{Z} has multiplicity one, then surprisingly enough ρ^{L} is necessarily zero, just as for the feedback regulator. That is,

$$H^{L}(\bar{X}) = H^{L}(\bar{Z})$$

and no action on R's part can reduce the information at Z. To see this, suppose that one has been given the values for X^{1}, X^{2}, ..., $X^{\tau-1}$, and by observing Z^{τ} he wants to deduce X^{τ}. This is always possible, since if X^{1}, ..., $X^{\tau-1}$ are given, R^{τ} can be calculated, and when Z^{τ} and R^{τ} are known, there is no uncertainty about X^{τ} (when f_{Z} has multiplicity one). Consequently if one is given some early values of X and then an indefinitely long sequence of Z-values, one can deduce all the corresponding

X-values. The same is true if the roles of X and Z are interchanged, so the X-sequence and Z-sequence must carry the same amount of information, regardless of R.

The similarities between regulation in the feedback and feedforward cases are striking; in fact there is no substantial point on which they differ. Neither is able to block information, $H^L(Z)$, at all when f_z is of multiplicity one. ρ in each case is limited by sequential constraints in $\bar{X}$, and the regulators in both cases succeed, if they succeed at all, only by making use of those constraints. Neither type is capable of "perfect" regulation, that is, maintainence of absolute constancy at Z, except in degenerate cases.

The close relationship between the two is apparent also in the difficulty of deciding whether to classify a given example of regulation as feed-back or feed-forward. When one is following the motions of a tennis ball with his eyes, for example, are eye-movements guided by information about the position of the ball, or by information about the angular error? It would be difficult to say.

When the quality of regulation achievable by feedback or feedforward regulation is not sufficient, another type which we shall call "parallel" regulation is often used.

5.3.3. Parallel regulation

In parallel regulation the regulator does not wait for X to affect Z before starting to operate; it makes use of information from the same source that affects X, as represented in Figure 41. The box D represents a primary source of disturbances which affect X and R.

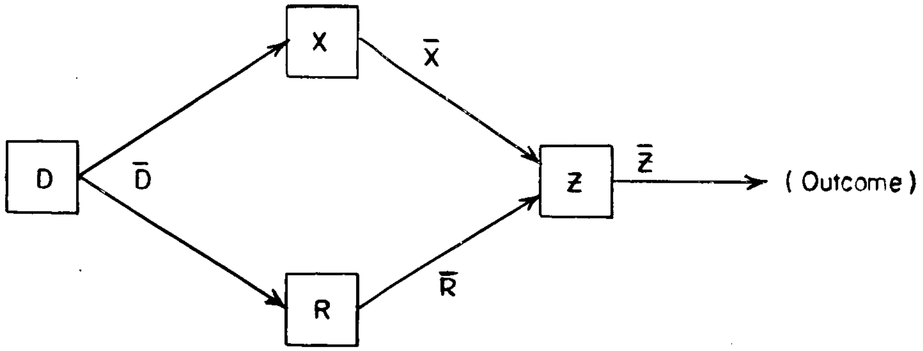


Figure 41.

This is the type of regulation in which R is frequently thought of as "anticipating" X, so that the regulatory action is simultaneous with the action of X. A driver sees a child run into the street and applies his brakes at the same time as the car ahead; a homeowner hears of an imminent cold wave and starts up his furnace; a schoolteacher smells fire and leads her students out of the building. As Ashby has pointed out, many of our senses have been developed precisely to get advance warning of disturbances, so that regulatory steps can be taken _before_ the outcome can be affected.

The job of the regulator, in fact, is to coordinate his actions with those of X in such a way that the outcome is not affected, no matter what disturbances arise, or in other words to match X in such a way that the channel capacity from D to Z is zero. In contrast to the other situations we have studied, this _is_ possible with parallel regulation; H(Z) can sometimes be made equal to zero.

Much depends on f_z, of course. In the worst possible case, f_z maps X x R one-to-one into Z and all regulation is clearly impossible; R can do no better than to pick some value r_i and keep that value always. If on the other hand there is a value z_k and a mapping $\mu : X \rightarrow R$ such that $f_z(x_i, \mu(x_i)) = z_k$ for all $x_i \in X$, then perfect regulation is possible, for whatever value X takes, R need only take the value $\mu(X)$ to keep the output fixed at z_k. In this case R can attain perfect regulation by acting in a manner isomorphic with X, for as was pointed out earlier, if X and R are isomorphic machines subject to the same input, they behave isomorphically and $T(X : R) = H(X) = H(R)$.

To summarize: if for every value x_i there is a corresponding value $r_i = \mu(x_i)$ such that $f_Z(x_i, r_i)$ is the same for all i, then R can attain perfect regulation ($H(Z) = 0$) by being isomorphic with X and subject to the same input.

If f_Z is of multiplicity one, then ρ^L is limited by the channel capacity of R, and in any case, since $T^L(\bar{R} : \bar{X}) \leq H^L(\bar{R})$,

$$\rho^L \leq \text{(channel capacity of R)} + K.$$

Thus in parallel regulation, the channel capacity of the regulator is a fundamental limit on its ability to reduce the output information rate, a fact which is a pleasant complement to the fact that the capacity also limits its ability to increase that rate.

This fact, that parallel regulation ρ^L is limited by the channel capacity of the regulator, is a fundamental link between information and control; it means that unless the situation is especially fortuitous (i.e., f_Z is especially favorable to regulation so that $[H_{R,Z}(X) - H^*_Z(X)]$ is positive), any attempt at regulation can only succeed to the degree that the regulator has access to sufficient information, "knows how" to transform it into appropriate action, and is able to carry out that action. The channel capacity, and thus the regulation, is limited by the weakest link in that chain.

5.4. Further remarks

The major restriction on the quantitative results in this chapter is that they were derived under the assumption that X was not affected by R; yet much of real-world regulation fits that assumption. Regulation in complex systems is frequently in one of the three forms we have discussed, often with X and R being complex systems and Z being a vector with components; the theorems developed above hold just as well in that case as when X, R, and Z are all very simple. Of course it requires little imagination to concoct regulatory schemes which appear to be more complex than any of the three basic forms, but further inspection often shows that a scheme apparently more complex may be recoded into one of the basic three or a simple combination of them.

Our purpose in this chapter, however, has been not to analyze all common schemes but rather to indicate some of the primary relations between information and regulation, to quantify these relations as much as is feasible in a general discussion, and to illustrate these relations by the three important examples. This, we hope, is a good start toward a better understanding of regulation.

REFERENCES

1. Wiener, N. Cybernetics. New York: Wiley, 1948.

2. McGill, W.J. "Multivariate Information Transmission." Psychometrika, June 1954, 19: 97-116.

3. Garner, W.R. Uncertainty and Structure as Psychological Concepts. New York: Wiley, 1962.

4. Ashby, W. Ross. "Measuring the Internal Informational Exchange in a System." Cybernetica, 1965, 1: 5-22.

5. Shannon, C.E. Mathematical Theory of Communication. Urbana: University of Illinois Press, 1964.

6. Ashby, W. Ross. Introduction to Cybernetics. London: Chapman & Hall, 1956.

7. Zadeh, L.A. "Fuzzy Sets." Information and Control, June 1965, 8: 338-53.

8. Ashby, W. Ross. Notes for course "Introduction to Cybernetics." Urbana: University of Illinois, 1963.

9. Powers, S.G. Uncertainty Analysis in Dynamic Systems. BCL Report No. 8.0. Urbana: University of Illinois, 1967.

10. Elias, P., Feinstein, A., and Shannon, C.E. "A Note on the Maximum Flow Through a Network." IRE Transactions on Information Theory, December 1956, 2: 117-19.

11. Simon, H.A. "Architecture of Complexity." Proceedings of the American Philosophical Society, December 1962, 106: 467-82.

12. Conant, R.C. Cause and Effect Relations Within a Network. BCL Report No. 8.1. Urbana: University of Illinois, 1967.

13. Reza, F.M. Introduction to Information Theory. New York: McGraw-Hill, 1961.

14. Birch, J.J. "Approximations for the Entropy for Functions of Markov Chains." Annals of Mathematical Statistics, 1962, 33: 930-38.

15. Garner, W.R. and McGill, W.J. "Relation Between Information and Variance Analyses." Psychometrika, September 1956, 21: 219-28.

REFERENCES (Cont.)

16. Grodins, F.S. Control Theory and Biological Systems. New York: Columbia University Press, 1963.

17. Fano, R.M. Transmission of Information, a Statistical Theory of Communications. Cambridge: M.I.T. Press, 1961.

18. Watanabe, S. and Abraham, C.T. "Loss and Recovery of Information by Coarse Observation of Stochastic Chain." Information and Control, September 1960, 3: 248-78.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* University of Illinois Biological Computer Laboratory Urbana, Illinois 61801	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

INFORMATION TRANSFER ON COMPLEX SYSTEMS, WITH APPLICATIONS TO REGULATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
scientific; technical; interim

5. AUTHOR(S) *(Last name, first name, initial)*

Conant, Roger C.

6. REPORT DATE January 1968	7a. TOTAL NO. OF PAGES 206	7b. NO. OF REFS 18

8a. CONTRACT OR GRANT NO. AF-AFOSR 7-67	9a. ORIGINATOR'S REPORT NUMBER(S) Tech. Report No. 13
b. PROJECT AND TASK NO. c. d.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES Partial sponsorship also under AF 33(615)-3890, NASA	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research Directorate of Information Sciences Arlington, Virginia 22209

13. ABSTRACT

This study is concerned with information theory and its relevance to the study of complex systems. When information about every detail of their activity is kept, many systems are too complex to be manageable and can only be dealt with by sacrificing detail. It is shown here that multivariable information theory is capable of eliminating much detail while preserving information about the interrelations between parts of a system, even when those interrelations are very complex. A procedure is described and exemplified, for example, which is helpful in the decomposition of hierarchical systems.

It is shown, among other results, that when two variables are related (in the set theoretic sense) the transmission between them is maximized when their behaviors are isomorphic. This observation leads to an algorithm for the computation of channel capacity for arbitrary finite-state systems of a very general type.

The importance of information in regulatory processes is discussed and quantified, and several basic regulatory schemes are discussed in terms of the information involved, showing in an exact way how information transfer and channel capacity limit the ability of any system to act as a successful regulator.

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Complex Systems						
Multivariable Information Theory						
Isomorphic Behavior						
Channel Capacity						
System Constraints						
Feedback						
Automata						
Control						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
_______________________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
_______________________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
_______________________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, *or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified
Security Classification