PRINCIPLES OF NEURODYNAMICS

ERRATA

Page 9, line 10: Change Kohler to Köhler

Page 111, footnote (line 4): Change ; to ,

Page 121, line 12: Change "remaining" to "following"

Page 124, line 14: Change $b_i X^{\#}$ to $b_i (X^{\#})$

Page 125, line 18: Change $b_i X^{\#}$ to $b_i (X^{\#})$

Page 126, line 4: Change $b_i X^{\#}$ to $b_i (X^{\#})$

Page 127, line 7: Change $b_i X^{\#}$ to $b_i (X^{\#})$

Page 139, fourth label under "Typical A-units" should read (Probability $Q_j - Q_{ij}$) instead of (Probability $Q_j - Q_i$).

Page 142, Figure 10c: Change R = 1.0 to R = .10

Page 147, Equation 6.10: Include E_{ijk} in the argument of P_x .

Page 149, footnote (line 4); Change "displayed to "displaced"

Page 155, line 8: Change α_i^* to a_i^*

Page 159, first two equations: Change e_{ir}^* to c_{ir}^*

Page 160, line 5: Change first word to "close"

Page 165, heading of first table: Correct spelling of "excitatory"

Page 166, line 14: Change page reference from 75 to 102

Page 326, line 1: continuous

Page 356, line 8: Change first symbol from $M_i^{(t)}$ to $M^{(i)}(t)$

Page 356, line 19: Change (t_0) to (0)

Page 362, line 11: fifth word should be "region"

Page 391, line 19: Insert the sentence: "Signal transmission times are again taken to be instantaneous."

Page 391, line 22: Change t+1 to t+τ

Page 395, label of last matrix, change < to >

Page 399: In label of last matrix, change 302 to 303

Page 443: In subheading of table, change $x = 2$ to $x = 3$

Page 622: Change name in eleventh entry to "Chiu"

PRINCIPLES OF NEURODYNAMICS

PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS

By FRANK ROSENBLATT

SPARTAN BOOKS
6411 CHILLUM PLACE, N.W. • WASHINGTON, D.C

Library of Congress Card Number 62-12882

Printed in the United States of America

"Perception, then, emerges as that relatively primitive, partly autonomous, institutionalized, ratiomorphic subsystem of cognition which achieves prompt and richly detailed orientation habitually concerning the vitally relevant, mostly distal aspects of the environment on the basis of mutually vicarious, relatively restricted and stereotyped, insufficient evidence in uncertainty-geared interaction and compromise, seemingly following the highest probability for smallness of error at the expense of the highest frequency of precision." ----- From "Perception and the Representative Design of Psychological Experiments," by Egon Brunswik.

"That's a simplification. Perception is standing on the sidewalk, watching all the girls go by." ----- From "The New Yorker", December 19, 1959.

PREFACE

It is only after much hesitation that the writer has reconciled himself to the addition of the term "neurodynamics" to the list of such recent linguistic artifacts as "cybernetics", "bionics", "autonomics", "biomimesis", "synnoetics", "intelectronics", and "robotics". It is hoped that by selecting a term which more clearly delimits our realm of interest and indicates its relationship to traditional academic disciplines, the underlying motivation of the perceptron program may be more successfully communicated. The term "perceptron", originally intended as a generic name for a variety of theoretical nerve nets, has an unfortunate tendency to suggest a specific piece of hardware, and it is only with difficulty that its well-meaning popularizers can be persuaded to suppress their natural urge to capitalize the initial "P". On being asked, "How is Perceptron performing today?" I am often tempted to respond, "Very well, thank you, and how are Neutron and Electron behaving?"

That the aims and methods of perceptron research are in need of clarification is apparent from the extent of the controversy within the scientific community since 1957, concerning the value of the perceptron concept. There seem to have been at least three main reasons for negative reactions to the program. First, was the admitted lack of mathematical rigor in preliminary reports. Second, was the handling of the first public announcement of the program in 1958 by the popular press, which fell to the task with all of the exuberance and sense of discretion of a pack of happy bloodhounds. Such headlines as "Frankenstein Monster Designed by Navy Robot That Thinks" (Tulsa, Oklahoma Times) were hardly designed to inspire scientific confidence. Third, and perhaps most significant, there has been a failure to comprehend the difference in motivation between the perceptron program and the various engineering projects concerned with automatic pattern recognition, "artificial intelligence", and advanced computers. For this writer, the perceptron program is not primarily concerned with the inven-

tion of devices for "artificial intelligence", but rather with investigating the physical structures and neurodynamic principles which underlie "natural intelligence". A perceptron is first and foremost a brain model, not an invention for pattern recognition. As a brain model, its utility is in enabling us to determine the physical conditions for the emergence of various psychological properties. It is by no means a "complete" model, and we are fully aware of the simplifications which have been made from biological systems; but it is, at least, an analyzable model. The results of this approach have already been substantial; a number of fundamental principles have been established, which are presented in this report, and these principles may be freely applied, wherever they prove useful, by inventors of pattern recognition machines and artificial intelligence systems.

The purpose of this report is to set forth the principles, motivation, and accomplishments of perceptron theory in their entirety, and to provide a self-sufficient text for those who are interested in a serious study of neurodynamics. The writer is convinced that this is as definitive a treatment as can reasonably be accomplished in a volume of managable size. Since this volume attempts to present a consistent theoretical position, however, the student would be well advised to round out his reading with several of the alternative approaches referenced in Part I. Within the last year, a number of comprehensive reviews of the literature have appeared, which provide convenient jumping-off points for such a study.*

The work reported here has been performed jointly at the Cornell Aeronautical Laboratory in Buffalo and at Cornell University in Ithaca. Both programs have been under the support of the Information Systems Branch of the Office of Naval Research -- the Buffalo program since July, 1957, and the Ithaca

*See, for example, Minsky's article, "Steps Toward Artificial Intelligence", Proc. I.R.E., 49, January, 1961, for an entertaining statement of the views of the loyal opposition, which includes an excellent bibliography.

program since September, 1959. A number of other agencies have contributed to particular aspects of the program. The Rome Air Development Center has assisted in the development of the Mark I perceptron, and we are indebted to the Atomic Energy Commission for making the facilities of the NYU computing center available to us.

A great many individuals have participated in this work. R. D. Joseph and H. D. Block, in particular, have contributed ideas, suggestions, and criticisms to an extent which should entitle them to co-authorship of several chapters of this volume. I am especially indebted to both of them for their heroic performance in proofreading the mathematical exposition presented here, a task which has occupied many weeks of their time, and which has saved me from committing many a mathematical felony. Carl Kesler, Trevor Barker, David Feign, and Louise Hay have rendered invaluable assistance in programming the various digital computers employed on the project, while the engineering work on the Mark I was carried out primarily by Charles Wightman and Francis Martin at C. A. L. The experimental program with the Mark I was carried out by John Hay. In addition to all of those who have contributed directly to the research activities, the writer is indebted to Professors Mark Kac, Barkley Rosser, and other members of the Cornell faculty for their administrative support and encouragement, and to Alexander Stieber, W. S. Holmes, and the administrative staffs of the Cornell Aeronautical Laboratory and the Office of Naval Research whose confidence and support have carried the program successfully through its infancy.

Frank Rosenblatt
15 March 1961

TABLE OF CONTENTS

Page Number

PART IV - BACK-COUPLED PERCEPTRONS AND PROBLEMS FOR FUTURE STUDY

PART I

DEVELOPMENT OF BASIC CONCEPTS

1. INTRODUCTION

The theory to be presented here is concerned with a class of " brain models" called perceptrons. By "brain model" we shall mean any theoretical system which attempts to explain the psychological functioning of a brain in terms of known laws of physics and mathematics, and known facts of neuroanatomy and physiology. A brain model may actually be constructed, in physical form, as an aid to determining its logical potentialities and performance; this, however, is not an essential feature of the model-approach. The essence of a theoretical model is that it is a system with known properties, readily amenable to analysis, which is hypothesized to embody the essential features of a system with unknown or ambiguous properties -- in the present case, the biological brain. Brain models of different types have been advanced by philosophers, psychologists, biologists, and mathematicians, as well as electrical engineers (c.f., Refs. 17, 31, 33, 54, 59, 61, 74, 91, 105, 109). The perceptron is a relative newcomer to this field,having first been described by this writer in 1957 (Ref. 78). Perceptrons are of interest because their study appears to throw light upon the biophysics of cognitive systems: they illustrate, in rudimentary form, some of the processes by which organisms, or other suitably organized entitites, may come to possess "knowledge" of the physical world in which they exist, and by which the knowledge that they possess can be represented or reported when occasion demands. The theory of the perceptron shows how such knowledge depends upon the organization of the environment, as well as on the perceiving system.

At the time that the first perceptron model was proposed, the writer was primarily concerned with the problem of memory storage in biological systems, and particularly with finding a mechanism which would account for the "distributed memory" and "equipotentiality" phenomena found by Lashley and others (Refs. 48, 49, 95). It soon became clear that the problem of memory mechanisms could not be divorced from a consideration of what it is that is remembered, and as a consequence the perceptron became a model of a more general cognitive system, concerned with both memory and perception..

A perceptron consists of a set of signal generating units (or "neurons") connected together to form a network. Each of these units, upon receiving a suitable input signal (either from other units in the network or from the environment) responds by generating an output signal, which may be transmitted, through connections, to a selected set of receiving units. Each perceptron includes a sensory input (i.e., a set of units capable of responding to signals emanating from the environment) and one or more output units, which generate signals which can be directly observed by an experimenter, or by an automatic control mechanism. The logical properties of a perceptron are defined by:

1. Its topological organization (i.e., the connections among the signal units);

2. A set of signal propagation functions, or rules governing the generation and transmission of signals;

3. A set of memory functions or rules for modification of the network properties as a consequence of activity.

A perceptron is never studied in isolation, but always as part of a closed experimental system, which includes the perceptron itself, a defined environment, and a control mechanism or experimenter capable of applying well-defined rules for the modification, or "reinforcement" of the perceptron's memory state. In most analyses, we are not concerned with a single perceptron, but rather with the properties of a class of perceptrons, whose topological organizations come from some statistical distribution. A perceptron, as distinct from some other types of brain models, or "nerve nets", is usually characterized by the great freedom which is allowed in establishing its connections, and the reliance which is placed upon acquired biases, rather than built-in logical algorithms, as determinants of its behavior.

Because of a common heritage in the philosophy, psychology, physiology, and technology of the last few centuries, there are bound to be similarities between the points of view and the basic assumptions of the theory presented here, and of other theories. The writer makes no claim to uniqueness in this respect. In particular, the neuron model employed is a direct descendant of that originally proposed by McCulloch and Pitts; the basic philosophical approach has been heavily influenced by the theories of Hebb and Hayek and the experimental findings of Lashley; moreover, the writer's predilection for a probabilistic approach is shared with such theorists as Ashby, Uttley, Minsky, MacKay, and von Neumann, among others.

This volume is divided into four main sections. Part I, commencing with this introduction, attempts to review the background, basic sources of data, concepts, and methodology to be employed in the study of perceptrons. In Chapter 2, a brief review of the main alternative

approaches to the development of brain models is presented. Chapter 3 considers the physiological and psychological criteria for a suitable model, and attempts to evalute the empirical evidence which is available on several important issues. Sufficient references to the literature are included throughout these chapters so that the reader who requires additional background in any of the areas discussed can use this as a guide for further reading. Part I concludes with Chapter 4, in which basic definitions and some of the notation to be used in later sections are presented. Parts II and III are devoted to a summary of the established theoretical results obtained to date. In these sections, the strategy will be to present a number of models of increasing complexity and sophistication, with theorems and analytic results on each model to indicate its capabilities and deficiencies. Wherever possible, established mathematical results will be presented first, followed by empirical evidence from simulation and hardware experiments. Part II (Chapters 5 through 14) deals with the theory of three-layer series-coupled perceptrons, on which most work has been done to date. These systems are called "minimal perceptrons". Part III (Chapters 15 through 20) deals with the theory of multi-layer and cross-coupled perceptrons, where a great deal still remains to be done, but where the most provocative results have begun to emerge. Part IV is concerned with more speculative models and problems for future analysis. Of necessity, the final chapters become increasingly heuristic in character, as the theory of perceptrons is not yet complete, and new possibilities are continually coming to light.

Part I (except for the chapter on definitions) is entirely non-mathematical. In Part II, and most of the remainder of the text, familiarity with the elements of modern algebra and probability theory is assumed, and

should be sufficient for most of the material. In several proofs in Part II, and to a greater extent in Part III, analytic methods are employed, assuming knowledge of the calculus and differential equations; an elementary acquaintance with differential geometry would also be useful. Symbolic logic is not required here, but the student will find it necessary for reading much of the ancillary literature in the field.

Several appendices are included which may prove helpful for cross-referencing equations, definitions, and experimental designs which are described in different chapters. Appendix A is a list of all symbols used in a standard manner throughout the volume. Appendix B is a consolidated list of theorems and corollaries. Appendix C lists the principal equations used in the analysis of performance, and basic quantitative functions. Appendix A contains a summary of the experiments used for testing and comparing different perceptrons. These experiments are referred to by number, throughout the text, and are described in detail as they are first introduced.

2. HISTORICAL REVIEW OF ALTERNATIVE APPROACHES

2.1 Approaches to the Brain Model Problem

There are at least two basic points, which are fundamental to a theory of brain functioning, on which most of the present-day theorists seem to be in agreement. First is the assumption that the essential properties of the brain are the topology and the dynamics of impulse-propagation in a network of nerve cells, or neurons. This has been contested by a few theorists who hold that the individual cells and their properties are less important than the bulk properties and electrical currents in the cortical medium as a whole (c.f. Kohler, Ref 45). The "neuron doctrine", however, has now been accepted with sufficient universality that it need not be considered as an issue in this report (Bullock, Ref. 11). It will be assumed that the essential features of the brain can be derived in principle from a knowledge of the connections and states of the neurons which comprise it. Secondly, there is general agreement that the information-handling capabilities of biological networks do not depend upon any specifically vitalistic powers which could not be duplicated by man-made devices. This also has occasionally been questioned, even today, by such neurologists as Eccles (Ref. 18) who advocate a dualistic approach in which the mind interacts with the body. Nonetheless, all currently known properties of a nerve cell can be simulated electronically with readily available devices. It is significant that the individual elements, or cells, of a nerve network have never been demonstrated to possess any specifically psychological functions, such as "memory", "awareness", or "intelligence". Such properties, therefore, presumably reside in the organization and functioning of the network as a whole, rather

than in its elementary parts. In order to understand how the brain works, it thus becomes necessary to investigate the consequences of combining simple neural elements in topological organizations analogous to that of the brain. We are therefore interested in the general class of such networks, which includes the brain as a special case.

While there is substantial agreement up to this point, theorists are divided on the question of how closely the brain's methods of storage, recall, and data processing resemble those practised in engineering today. On the one hand, there is the view that the brain operates by built-in algorithmic methods analogous to those employed in digital computers, while on the other hand, there is the view that the brain operates by non-algorithmic methods, bearing little resemblance to the familiar rules of logic and mathematics which are built into digital devices (c.f. von Neumann, Ref. 105). The advocates of the second position (this writer included) maintain that new fundamental principles must be discovered before it will be possible to formulate an adequate theory of brain mechanisms. It is suggested that probabilistic and adaptive mechanisms are particularly important here. This does not mean that the actual biological nervous system is strictly one type of device or the other; the issue concerns the matter of emphasis, as to whether the brain is primarily a more or less conventional computing mechanism, in which statistical or adaptive processes play an incidental and non-essential role, or whether the brain is so dependent upon such processes that a model which fails to take them into account will find itself unable to account for psychological performance.

These two points of view are associated with two basically different procedures for studying the mechanisms of the brain and for the development of brain models. The first procedure will be called the monotypic model approach; it amounts to the detailed logical design of a special-purpose computer to calculate some predetermined "psychological function" such as the result of a recognition algorithm, or a stimulus transformation, which is postulated as a plausible function for a nerve net to calculate. The physical properties of this computer are then compared with those of the brain, in the hopes of finding resemblances. The second procedure will be called the genotypic model approach. Instead of beginning with a detailed description of functional requirements and designing a specific physical system to satisfy them, this approach begins with a set of rules for generating a class of physical systems, and then attempts to analyse their performance under characteristic experimental conditions to determine their common functional properties. The results of such experiments are then compared with similar observations on biological systems, in the hopes of finding a behavioral correspondence. It is the purpose of this chapter to review the historical development and current status of these two alternative "philosophies of approach" to the brain model problem.

2.2 Monotypic Models

In the monotypic model approach, the theorist generally begins by defining as accurately as possible the performance required from his model. For example, he may specify a data processing operation, an input-output or stimulus-response function, or a remembering and

regenerating operation. In one typical model, the system is required to normalize the size and position of a visual image, and to compare functions of this normalized image with certain stored quantities required for identification (Ref. 71). Given a description of the required performance in sufficiently precise terms, the theorist then proceeds to design a computing machine or control system embodying the required function, generally limiting himself to the use of a set of modular switching devices which are analogous to biological neurons in their properties. It is this last constraint which distinguishes the nerve net theorist from any other designer of special purpose computers confronted with the same problem. It is hoped that a network which consists of neuron-like elements, and is capable of computing the required functions, will be found to resemble a biological nerve-net in its organization and the computational principles employed.

While the simulation of animals, saints, and chessplayers by animated machines and clockwork devices goes back many centuries, the idea of constructing such devices out of simple logical elements with neuron-like properties is a relatively recent one, and received its first impetus from two sources: First, Turing's paper "On Computable Numbers", in 1936, and the subsequent development of stored-program digital computers by von Neumann and others during the 1940's (Refs. 12, 100)gave rise to an impressive family of "universal automata", capable of executing programs which would enable them to perform any computation whatsoever with only the simplest of logical devices being employed as "building blocks". Second, the Chicago group of mathematical biophysicists which grew up about Rashevsky after the publication of his "Mathematical Biophysics" in 1938,

(Ref. 73) began to investigate the manner in which "nerve nets" consisting of formalized neurons and connections might be made to perform psychological functions. Householder, Landahl, Pitts, and others made notable contributions to this effort during the late 1930's and early 1940's (Refs. 35, 69, 70).

In 1943, the doctrine and many of the fundamental theorems of this approach to nerve net theory were first stated in explicit form by McCulloch and Pitts, in their well-known paper on "A Logical Calculus of the Ideas Immanent in Nervous Activity". The fundamental thesis of the McCulloch-Pitts theory is that all psychological phenomena can be analyzed and understood in terms of activity in a network of two-state (all-or-nothing) logical devices. The specification of such a network and its propositional logic would, in the words of the writers, "contribute all that could be achieved" in psychology, "even if the analysis were pushed to ultimate psychic units or 'psychons', for a psychon can be no less than the activity of a single neuron... The 'all-or-none' law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions." (Ref. 57). Despite the apparent adherence to an outdated atomistic psychological approach, there is an important contribution in the recognition that the proposed axiomatic representation of neural elements and their properties permits strict logical analysis of arbitrarily complicated networks of such elements, and that such networks are capable of representing any logical proposition whatever. As von Neumann states in a summary of the McCulloch-Pitts model, (Ref. 103) "The 'functioning' of such a network may be defined by singling out some of the inputs of the entire system and some of its outputs, and

then describing what original stimuli on the former are to cause what ultimate stimuli on the latter...McCulloch and Pitts' important result is that any functioning in this sense which can be defined at all logically, strictly, and unambiguously in a finite number of words can also be realized by such a formal neural network."

A great variety of subsequent models have made use of this axiomatic representation, which we now refer to as the "McCulloch-Pitts neuron". As stated in the original paper (Ref. 57), the basic assumptions in this representation are:

" 1. The activity of the neuron is an 'all-or-none' process.

2. A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous acitivy and position on the neuron.

3. The only significant delay within the nervous system is synaptic delay.

4. The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.

5. The structure of the net does not change with time."

These postulates are such as to rule out memory except in the form of modifications of perpetual activity or circulating loops of impulses in the network. Any non-volatile memory, such that the functioning of the network at a given time depends upon previous activity even though a period of total inactivity has intervened, is impossible in a McCulloch-Pitts network. However, a McCulloch Pitts network can always be constructed which will embody whatever input-output relations might be realized by a system with an arbitrary memory mechanism, provided activity is allowed to persist in the network.

Later writers, notably Kleene (Ref. 43) have considered in more detail the kinds of events which can be represented by networks of McCulloch-Pitts neurons. The only important limitation is that events whose definition depends upon the choice of a temporal origin point, or events which extend infinitely into the past, may not be representable by outputs from finite networks. Any event which can be described as one of a definite set of possible input sequences over a finite period of time can be represented. In particular, any events which might conceivably be recognized by a biological system can be represented by outputs of networks of McCulloch-Pitts neurons.

In later papers by Pitts and McCulloch (Ref. 71) and by Culbertson (Refs. 16, 17) specific automata designed to perform actual "psychological" functions such as pattern recognition, have been described. Culbertson, in particular, has carried out such designs in explicit detail for a large number of interesting problems. The approach which he advocates

is expounded in his 1950 work on "Consciousness and Behavior" as follows:

> "Neuroanatomy and neurophysiology have not yet developed far enough to tell us the detailed interconnections holding within human or animal nets...Consequently, ... we cannot start with specified nerve nets and then in a straightforward way determine their properties. Instead, it is the reverse problem which always occurs in dealing with organic behavior. We are given at best the vaguely defined properties of an unknown net and from these must determine what the structure of that net might possibly be. In other words, we know, at least in a rough way, what the net does (as this appears in the behavior of the animal or man) and from this information we have to figure out what structure the net must have...Our investigation passes through two stages. In the first stage--the behavioristic inquiry--we ignore the inner constituents, i.e., the nervous system and its activity, and concentrate our attention instead on the observable relations between the stimuli affecting the organism and the responses to which these stimuli give rise...This makes the second stage--the functional inquiry--possible. Here, as Northrop says, we concentrate our attention on the inner (throughput) constituents of the system and point out the ways in which the receptor cells, central cells, and effector cells could be interconnected so that the input and output relations...would be those discovered in stage 1."

While such a program can hardly be criticized on logical grounds, it appears pragmatically to have fallen short of the proposed goals. Starting rather suddenly, with the development of automata theory in the late 1930's, the ready applicability of symbolic logic brought this approach to early mathematical sophistication. After the first flood of proposed models, further progress has been disappointingly trivial, and returns seem to be diminishing rapidly. The promised biological "explanations" have been particularly lacking. In this writer's opinion, there are at least five main reasons for this:

(1) There is a lack of sufficiently well defined psychological functions as a starting point. The approach requires essentially full knowledge of input-output relations for the behavior of an organism, and such knowledge is not available for any biological species.

(2) Constructed solutions generally show poor correspondence to known conditions of neuroanatomy and neuroeconomy; the numbers of neurons required often exceed those in biological nervous systems, and the logical organization generally requires a precision of connections which appears to be absent in the brain. In some cases, a single misconnection would be sufficient to make the system inoperable.

(3) The models fail to yield general laws of organization. A monotypic model is in general overdetermined, corresponding at best to a biological phenotype, rather than a species as a whole; its specification in the form of a detailed "wiring diagram" frequently misses essentials in a plethora of detail. Unique solutions for the proposed functions are generally lacking and an enormous variety of models can be generated which appear to solve the same problem equally well. Therefore, unless the system is actually tested against its biological counterpart, nothing is gained by a detailed construction of the model except a further confirmation of an existence theorem which is already well established.

(4) The models lack predictive value. Once a particular model has been proposed, further analysis can reveal little that is not included in the functional description with which we began.

(5) The models are not biologically testable in detail. Specific connections cannot be traced with sufficient precision in nervous tissue to say whether or not a particular wiring diagram is exactly realized. Consequently, the models are fated to remain purely speculative unless histological techniques are improved to a highly improbable degree.

In the foregoing, we have concentrated on the line of models which have attempted to represent the brain as a symbolic logic calculator, in which events of the outside world are represented by the firing or non-firing of particular neurons. It is in these models that rigorous mathematical treatment has been most successfully achieved. Not all monotypic models are of this variety, however. Field theorists such as Köhler have taken exception to the idea that psychological phenomena can be represented in this fashion. Köhler, arguing for an isomorphic representation of perceptual phenomena, asks (Ref. 46): "How can a cortical process such as that of a square give rise to an apparition with certain structural characteristics, if these characteristics are not present in the process itself? According to Dr. McCulloch, this is actually the case. But if we follow the example of physics, we shall hesitate to accept his view. In physics, the structural

characteristics of a state of affairs are given by the structural properties of the factors which determine that state of affairs... Situations in physics which depend upon the spatial distribution of given conditions never have more, and more specific, structural characteristics than are contained in the conditions". While Köhler's own model is not generaly considered plausible today, his criticism is a significant one, and a number of theorists, such as Lashley (Ref. 50) MacKay (Refs. 55, 56) and Green (Ref. 28) have been concerned with possible forms of representation of perceptual information which would preserve the intrinsic structural features of the perceived event rather than merely assigning an arbitrary symbol to it.

The main line of monotypic models, although failing to provide a satisfactory brain model, has left us a number of important analytic tools and concepts, including the McCulloch-Pitts neuron, and the theorems concerning the existence of networks representing arbitrary functions. For the actual design of plausible organizations, however, the genotypic approach appears to hold more promise.

2.3 Genotypic Models

In the monotypic approach, the properties of the components, or neurons, which comprise the networks are fully specified axiomatically, and the topology of the network is fully specified as well. In the genotypic approach, the properties of the components may be fully specified, but the organization of the network is specified only in part, by constraints and probability distributions which generate a class of systems rather than a

specific design. The genotypic approach, then, is concerned with the properties of systems which conform to designated laws of organization, rather than with the logical function realized by a particular system.

This difference in approach leads to important differences in the types of models which are generated, and the kinds of things which can be done with them. In the case of monotypic models, for example, the propositional calculus is applicable and probability theory is poorly suited to the analysis of performance, since a single fully deterministic system is under consideration which either does or does not satisfy the required functional equations. In dealing with genotypic models, on the other hand, sumbolic logic is apt to prove cumbersome or totally inapplicable (even though, in principle, any particular system which is generated might be expressed by a set of logical propositions). In the analysis of such models, the chief interest is in the properties of the class of systems which is generated by particular rules of organization, and these properties are best described statistically. Probability theory therefore plays a prominent part in this approach. A second major difference is in the method of determining functional characteristics of the models. In the monotypic approach, the functional properties are generally postulated as a starting point. In the genotypic approach, they are the end-objective of analysis, and the physical system itself (or the statistical properties of the class of systems) constitutes the starting point. This means that psychological functions need not be determined in full detail before setting out to construct a model, and, indeed, it is hoped that such models may help in answering open psychological questions.

While the monotypic approach arose rather suddenly with the advent of modern computers and control system theory, and rapidly advanced to a high level of mathematical sophistication, the genotypic approach has been much more gradual in its development, and has not yet developed all of the mathematical tools required to deal adequately with its problems. The genotypic models have been influenced less by the engineering sciences, and more by physiology and neuroanatomy. The descriptive anatomy of the nineteenth century laid the groundwork for modern studies of localization of function in the brain, and neurologists such as John Hughlings Jackson noted the apparent plasticity of the system -- the ability of neighboring regions to take over the function of damaged areas. Pavlov and others speculated about possible mechanisms for adaptive modification of the central nervous system in the early part of this century, and various hypotheses for the deposition of "memory traces" were of interest to psychologists and physiologists alike. The doctrine of equipotentiality, propounded by Lashley (Ref. 49), went even further in claiming complete interchangeability of most parts of the cerebral cortex, and evidence for "distributed memory" which suggested that "traces" must be more or less uniformly dispersed throughout the cortical tissue began to accumulate. All of this neurological evidence engendered a picture of the brain as a relatively undifferentiated structure, capable of undergoing radical reorganization by means of unspecified adaptive mechanisms, and showing only gross anatomical equivalence from one individual to another. While recent work on localization (Refs. 51, 65, 66, 94, 108) has shown some surprisingly precise mapping of functions, modern morphological investigations (Refs. 8, 52, 93) have borne out the apparently statistical organization of the "fine structure" of neurons and their interconnections. It now seems reasonable to suppose that while there are many constraints

on the organization of neurons in the brain, which are undoubtedly essential to the system's functioning, these constraints take the form of prohibitions, biases, and directional preferences, rather than a specific blueprint which must be followed to the last detail. In order words, there are enormous numbers of functionally equivalent systems, all obeying the same rules of organization, and all equally likely to be generated by the genetic mechanisms of a particular species.

While the neurologists mentioned above had a great deal to say about the observed and hypothetical organization of the brain, they were not concerned with the construction of models in the sense of detailed theoretical systems from which precise deductions could be made. Psychologists and philosophers, more willing to indulge in speculation, were the first to attempt detailed conjectures on the maturation of psychological functions in systems which might justifiably be called "brain models". Hebb (Ref. 33) and Hayek (Ref. 32), following the tradition of James Stuart Mill and Helmholtz, have attempted to show how an organism can acquire perceptual capabilities through a maturational process. For Hayek, the recognition of the attributes of a stimulus is essentialy a problem in classification, and his point of view has inspired Uttley (Refs. 101, 102) to design a type of classifying-automaton which attempts to translate the approach into more rigorous mathematical form. Hebb's model is more detailed in its biological description, and suggests a process by which neurons which are frequently activated together become linked into functional organizations called "cell assemblies" and "phase sequences" which, when stimulated, correspond to the evocation of an elementary idea or percept. While Hebb's

work is far more complete in its specification of a "model" than most preceding suggestions along this line, it is still too programmatic and too loose in its definitions to permit a rigorous testing of hypotheses. It should be considered more as a description of what a satisfactory model might ultimately look like than as a fully formulated model in its own right. Nonetheless, it comes sufficiently close to a detailed specification so that Rochester and associates, using an IBM computer, were able to propose enough of the missing detail to put the cell assembly hypothesis to an empirical test (Ref. 77). Unfortunately, with a theory so loosely specified, the inconclusive results of the IBM experiments carry little weight in evaluating Hebb's original system. Milner, in a recent paper (Ref. 58) has attempted to update the Hebb theory, and it may be that his model can be more readily translated into analyzable form, although this has not yet been done.

It is interesting that one of the first applications of probability theory to brain models is due to Landahl, McCulloch, and Pitts, appearing in 1943 along with the McCulloch-Pitts symbolic logic model (Ref. 47). In this paper, the topology of the network is still assumed to be a strictly deterministic, fully known organization, but impulses are assumed to be propagated with known frequencies but with uncertainties in their precise timing. A theorem is stated which permits the substitution of frequencies for symbols in the logical equations of the network, in order to obtain the expected frequency with which different cells will respond. This statistical treatment is related to the work of von Neumann (Ref. 104) on the probability of error in networks with fallible components.

The first systematic attempt to develop a family of statistically organized networks, and to analyze these in a rigorous fashion by means of a genotypic approach seems to have been due to Shimbel and Rapoport, in 1948 (Ref. 92). Starting with an axiomatic representation of neurons and connections, similar to that of McCulloch and Pitts, a network is characterized by probability distributions for thresholds, synaptic types, and origins of connections. A general equation is then developed for the probability that a neuron at a specified location will fire at a specified time, as a function of preceding activity and parameters of the net. This is applied to a number of specific classes of networks to determine the possibility of steady-state activity, and changes in the firing distribution with time. This work is a forerunner of a number of stability studies (e.g., Allanson, Ref. 2) which are still of interest.

The use of a digital computer by Rochester and associates was mentioned above in connection with Hebb's model. Simulation of a statistically connected network to investigate possible learning capabilities was first carried out successfully by Farley and Clark in 1954 (Ref. 10). Although mathematical analysis was not attempted in either the Farley-Clark or the Rochester models, they illustrate a convenient method of axiomatizing a network (by means of a computer program) to a degree which makes the investigation of hypotheses possible. While none of these experiments led to very sophisticated systems, they are of considerable historical interest, and the mechanism for pattern generalization proposed by Clark and Farley (Ref. 15) is essentially identical to that found in simple perceptrons.

Statistical models of various types have been proposed during the last decade. In particular, the models of Beurle, Taylor, and Uttley (Refs. 6, 99, 101, 102) are of interest as attempts to analyze models with a clear resemblance to the organization of a primitive nervous system, with receptors, associative elements, and output or motor neurons. Moreover, in some of these models, environments of sufficient complexity to permit the representation of visual and temporal patterns (albeit of a very primitive type) are included in the analysis. Minsky (Ref. 59) has also devised and analyzed several models capable of learning responses to simple stimuli.

A contribution of considerable methodological significance was Ashby's "Design for a Brain", in 1952 (Ref. 3). While Ashby's work (despite its title) does not specify an actual brain model in our present sense, it develops the rationale for an analysis of closed systems which must include the environment as well as the responding organism and rules of interaction as the object of study. Ashby's fields of variables correspond closely to our concept of "experimental systems" which will be defined in Chapter 4. In addition to his conceptual contribution, which is concerned with the general approach to be used rather than with a specific model, Ashby has demonstrated in a number of experiments how statistical mechanisms can yield adaptive behavior in an organism.

While the genotypic approach has found favor among many biologists, it is by no means universally accepted. A typical criticism is

voiced by Sutherland (Ref. 97) in connection with Hebb's system:

> "When Hebb's theory was first put forward, it was hailed as showing how it might be possible to account for behavior in terms of plausible neurophysiological mechanisms... However, a moment's reflection shows that, if he is right, what he has really succeded in doing is to demonstrate the utter impossibility of giving detailed neurophysiological mechanisms for explaining psychological or behavioral findings. According to Hebb the precise circuits used in the brain for the classification of a particular shape will vary from individual to individual with chance variation in nerve connectivity determined by genetic and maturational factors... Different individuals will achieve the same end result in behavior by very different neurological circuits... If Hebb's general system is right, it precludes the possibility of every making detailed predictions about behavior from a detailed model of the system underlying behavior."

While objections such as this seem to stem from a misunderstanding of the possibility of obtaining seemingly deterministic phenomena from a statistical substrate (as in statistical mechanics) the above argument is bolstered by many findings which suggest complicated hereditary mechanisms for the analysis of stimuli in "instinctive" behavior. The work of Sperry and Lettvin has already been cited in connection with the mechanisms for precise localization of connections which seem to exist in the brain. Our conclusion is that the biological system must employ some mixture of specific connection mechanisms and statistically determined structures; just how much constraint is present in the genetic constitution of the brain is an open question.

On most of the specific points of criticism raised in connection with monotypic models, the genotypic approach seems to fare much better. Detailed psychological functions are not required as a starting point. Detailed physiological knowledge of the brain would be helpful, but even a rough parametric description enables us to start off in the right direction, and present models have a considerable way to go before they have assimilated all of the physiological data which are available.

Since this approach begins with the physical model rather than the functions which must be performed, it is easy to guarantee its conformity in size and organization to the general characteristics of a biological system. Most important is the fact that this approach appears to be yielding results of increasing significance and interest, and the models frequently suggest progressive lines of development from simple first approximations to more sophisticated systems. In the application of the genotypic approach to perceptrons, a number of laws of considerable generality have been discovered, as will be seen in subsequent chapters.

2.4 Position of the Present Theory

The groundwork of perceptron theory was laid in 1957, and subsequent studies by Rosenblatt, Joseph, and others have considered a large number of models with different properties (Refs. 7, 30, 31, 40, 41, 76, 79, 80, 81, 82, 84, 85, 86). Perceptrons are genotypic models, with a memory mechanism which permits them to learn responses to stimuli in various types of experiments. In each case, the object of

analysis is an experimental system which includes the perceptron, a defined environment, and a training procedure or agency. Results of such analyses can then be compared with results of comparable experiments on human or animal subjects to determine the functional correspondence and weaknesses of the model. A number of specific psychological tasks and criteria, which will be discussed in the following chapter, are used for the comparison of different systems.

Perceptrons are not intended to serve as detailed copies of any actual nervous system. They are simplified networks, designed to permit the study of lawful relationships between the organization of a nerve net, the organization of its environment, and the "psychological" performances of which the network is capable. Perceptrons might actually correspond to parts of more extended networks in biological systems; in this case, the results obtained will be directly applicable. More likely, they represent extreme simplifications of the central nervous system, in which some properties are exaggerated, others suppressed. In this case, successive perturbations and refinements of the system may yield a closer approximation.

The main strength of this approach is that it permits meaningful questions to be asked and answered about particular types of organization, hypothetical memory mechanisms, and neuron models. When exact analytic answers are unobtainable, experimental methods, either with digital simulation or hardware models, are employed. The model is not a terminal result, but a starting point for exploratory analysis of its behavior.

3. PHYSIOLOGICAL AND PSYCHOLOGICAL CONSIDERATIONS

In the last chapter, a methodological doctrine was proposed, which undertakes to evaluate classes of brainlike systems by comparing their performance with that of biological subjects in behavioral experiments; by gradually increasing the sophistication and varying the axiomatic constraints which define the experimental systems, it is hoped that models which closely resemble the biological prototype can ultimately be achieved. In this chapter, the desiderata for a satisfactory brain model are considered in more detail, from the standpoint of physiology and psychology. What are the parametric constraints, functional properties, and performance criteria which must be met, in order to achieve a model which is a plausible representation of the brain?

The following discussion comes under three main headings: (1) established fundamentals; (2) current issues; and (3) the design of experimental tests of performance. It is not our purpose to review all of the relevant background in biology and psychology, but rather to highlight those points which bear most directly upon the present undertaking, and to suggest certain areas in which investigations might provide decisive evidence for or against some of the models which we shall propose. It will be noted that no attempt has been made to distinguish specifically "psychological" or specifically "physiological" problems in the following sections. Such distinctions are not only arbitrary in a number of the cases considered, but also tend to obscure the fact that we are interested in all of these problems because of their relevance to brain models, rather

than to psychology or physiology per se. In this discussion, attention will be concentrated on the level of complexity which seems most commensurate with that of the proposed models. Psychological material on psychoneuroses, or on attitude formation, for example, while it might be brought to bear on the evaluation of some future models, is hardly likely to be relevant at this time. On the physiological side, we are chiefly concerned with the overall organization of the nervous system, its microstructure, and conditions for impulse transmissions; we are less concerned with details of neuroanatomy and neurochemistry, although such data may become important in more sophisticated models, where a closer correlation with the biological system is sought.

3.1 Established Fundamentals

3.1.1 Neuron Doctrine and Nerve Impulses

It was only during the first decade of this century that a strong case was developed for regarding the neuron as the basic anatomical unit of the nervous system. The demonstration that this is the case rests largely upon the work of Ramon y Cajal (Ref. 14). Since Cajal's time, a great variety of neurons, differing in size, numbers of dendritic and axonal processes, and the distribution of these, have been described by neuroanatomists (Refs. 8, 52, 93). Today it is generally accepted that in virtually all biological species, the nervous system consists of a network of neurons, each consisting of a cell body with one or more afferent (incoming) processes, or dendrites, and one or more efferent (outgoing) processes, or axons. The axons branch into

small fibers which may make contact with, but remain separate from the surface membrane of cells or dendrites upon which they terminate. Neurons are generally divided into three classes: (1) sensory neurons, which generate signals in response to energy applied to sensory transducers, such as photo-receptors or pressure sensitive corpuscles; (2) motor neurons, (or effector neurons) which transmit signals to muscles or glands and directly control their activity; (3) internuncial neurons, (or associative neurons) which form a network connecting sensory and motor neurons to one another. The brain, or central nervous system, is made up almost entirely of neurons of this last type.

The actual signals carried by these neurons may take one of several forms. Until recently, it was supposed that all information in the nervous system was represented by a code of all-or-nothing impulses, corresponding to on-off states of the neurons. A sufficient input signal was supposed to trigger the receiving cell directly into emitting a spike potential, which was transmitted without decrement from the receiving region of the dendrites to the cell body, and out along the axon to the terminal endbulbs, where it might or might not succeed in triggering later cells in the network. In a recent review (Ref. 11) Bullock has pointed out that this view has been largely supplanted by a far more complicated picture. While it is true that the transmission of signals over long distances is generally accomplished by means of all-or-nothing spike propagation along the axons of nerve cells, the spike impulse is not a direct response to impulses which arrive at the dendrites, and may originate at a point which is separated by a considerable

distance from the site at which incoming impulses are received. Essentially, the currently accepted concept is that the dendritic structure and cell body jointly act as an integrating system, in which a series of incoming signals interact to establish a pre-firing state in a region at the base of the axon, from which impulses originate. If this pre-firing state reaches a threshold level (presumably measured by membrane depolarization) at a point within the critical region, a spike potential is initiated, and spreads without decrement along the axon. The interactions which may occur in the cell body and dendrites, however, involve potential fields in which the effects of impulses received at a given point spread over the surrounding membrane surface in a decrementing fashion. These effects may be graded in intensity, depending on frequency of impulses received, and the state of the receiving membrane at the time. Successions of impulses arriving at the same synapse can sometimes cause an increase in the sensitivity of the receiving membrane (facilitation) and can sometimes cause a progressive diminution in sensitivity (Ref. 11). There is evidence to suggest that different local patches of surface membrane are differently specialized, and respond in different ways to impulses received, even within the same neuron. Some of these regions appear to act as sources of internally generated signals, which may lead to spontaneous activity of the neuron, and the emission of spike impulses without any input signals from outside the cell.

Two main types of synapses are recognized: excitatory and inhibitory. It is generally assumed, although it has not been proven, that a single neuron is either all excitatory or all inhibitory, in its effect upon post-synaptic cells. It remains possible, however, that the individual

synaptic endings are specialized, some of them releasing a depolarizing transmitter substance (excitatory endings) while others release a hyperpolarizing substance (inhibitory endings). A single synapse, so far as is known, remains either excitatory or inhibitory, and is incapable of changing from one to the other.

The nerve impulse itself is a basically non-linear response to stimulation. It is supported by energy-reserves of the axon by which it is transmitted, rather than by a propagation of energy from the sources of excitation. The nerve impulse is manifested by a moving zone of electrical depolarization of the surface membrane of the neuron, the exterior of which is normally 70 to 100 millivolts positive relative to the interior. This zone tends to spread along the axon due to ionic currents which tend to break down the potential difference between the interior and exterior of the neuron, until the membrane is repolarized by metabolic processes (see Eccles, Refs. 18, 19). The resulting "spike potential" takes the form of an electrically negative impulse (measured relative to the normal surface potential of the membrane) which propagates down the fiber with an average velocity of about 10 to 100 meters per second, depending on the diameter of the fibers (c. f., Brink, Ref. 9).

The arrival of a single (excitatory) impulse gives rise to a partial depolarization of the post-synaptic membrane surface, which spreads over an appreciable area, and decays exponentially with time. This is called a local excitatory state (l.e.s.). The l.e.s. due to successive impulses is (approximately) additive. Several impulses arriving in sufficiently close succession may thus combine to touch off

an impulse in the receiving neuron if the local excitatory state at the base of the axon achieves the threshold level. This phenomenon is called temporal summation. Similarly, impulses which arrive at different points on the cell body or on the dedrites may combine by spatial summation to trigger an impulse if the l.e.s. induced at the base of the axon is strong enough.

The passage of an impulse in a given cell is followed by an absolute refractory period during which the cell cannot be fired again, regardless of the level of input activity. This is equivalent to an infinite threshold during this period. The spike potential and absolute refractory period last about 1 millisecond. Finally, there is a relative refractory period which may last for many milliseconds after the initial impulse. During this time, the threshold gradually returns to normal, and may even fall to somewhat below its normal level for a time. While the response of a cell to a single momentary stimulus, such as an electrical pulse, is markedly non-linear (the amplitude of the generated impulse being quite independent of the amplitude of the triggering signal) the effect of a sustained excitatory signal, in many cases, is to evoke a volley of output spikes, the frequency of which may be roughly proportional to the intensity of the stimulus over a wide range. This is particularly true of sensory neurons, where the frequency of firing may be used to determine the intensity of the stimulus energy with considerable accuracy.

The general picture of the nervous system, then, is one of a large set of signal generators, each having one or more outputs, on which nerve impulses may appear. These impulses may vary in frequency, and to some extent in amplitude, but seem to carry information mainly in a pulse-coded form. The signal generators themselves are decision elements of a most intricate type; each one makes its decision to initiate an output impulse according to a complicated function of the series of signals received at each of its synapses or receptor areas, as well as its own internal state. In a brain model, a neuron of this complexity would tend to make the system unintelligible and unmanageable with the analytic and mathematical tools at our disposal. Simplifications will therefore be introduced, as in the manner of the McCulloch-Pitts neuron; but it should be remembered that the biological neuron is considerably more complicated, and may incorporate within itself functions which we require whole networks of simplified neurons to realize.

3.1.2 Topological Organization of the Network

The human brain consists of some 10^{10} neurons of all types. These are arranged in a network which receives inputs from receptor neurons at one end, and conveys signals to the effector neurons at the output end. Different sensory modalities -- vision, hearing, touch, etc. -- communicate with the central nervous system by way of distinct nerve bundles, which enter it at different points. Each of these modalities, after passing its information through a network of cells which respond more or less exclusively to stimuli from that modality, eventually contri-

butes to a common pool of activity in the "association areas" of the central nervous system (CNS). Output signals originate either from the parts of the CNS which are specific to a particular modality (for example, the pupillary reflex mechanism) or from the common activity areas (as in speech). Final outputs may go through a series of stages in which motor patterns or sequences are selected, and detailed coordination is regulated. From these motor control regions, feedback paths re-enter the association areas and sensory integration areas, so that the possibility of an elaborate servo-mechanism for the control of motor activity exists.

While this general picture holds true for most biological organisms, there is considerable variation both in gross and detailed anatomy, from species to species and individual to individual. In undertaking to design a first order approximation to this structure for use in a brain model, we will begin with a network consisting of a single array of sensory units, a layer of association units, and a single effector, or response unit. In later models, more complicated structures will be considered. Even the simplest models, however, are capable of showing a surprising similitude to the functional properties of the brain. It seems reasonable, therefore, to regard the complications of neuroanatomy in the various species as elaborations of a basically simple schema, which is to be found throughout. This basic plan of organization is illustrated in Figure 1.

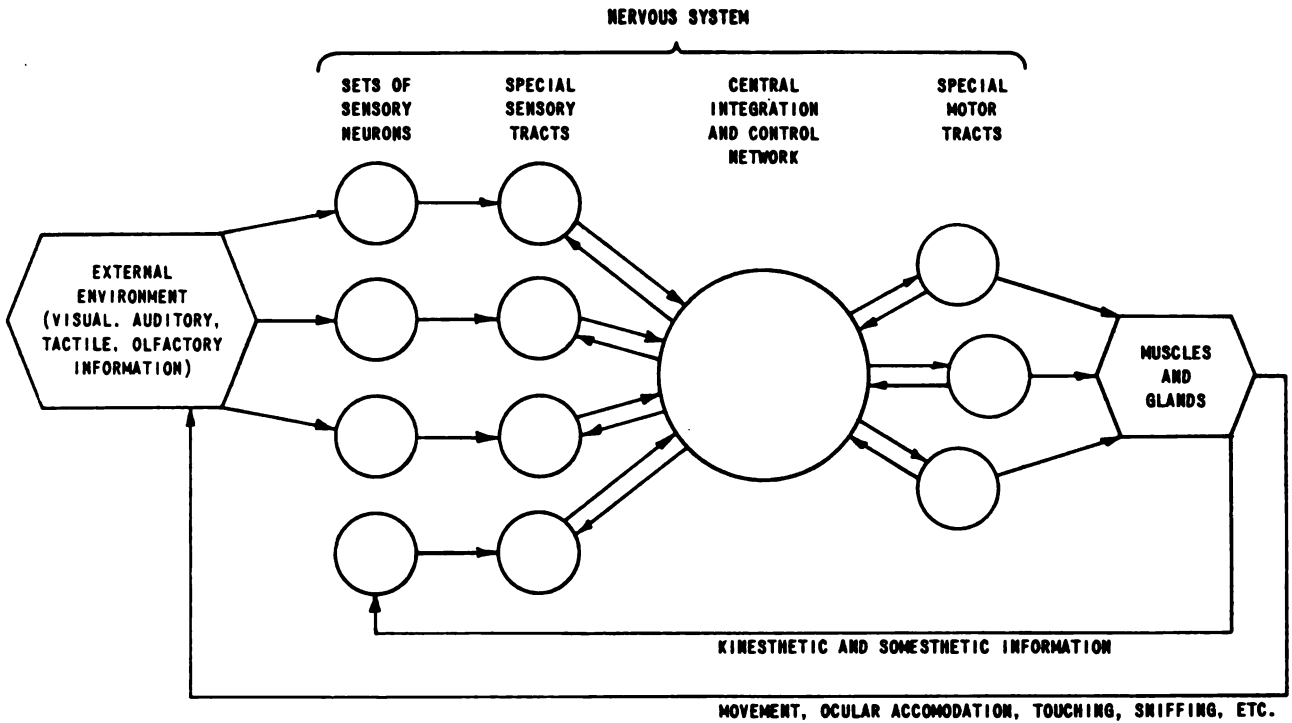


Figure I BASIC TOPOLOGICAL STRUCTURE OF THE NERVOUS SYSTEM AND ITS SOURCES OF INFORMATION

The distribution of cell types and connection patterns has been studied by Lorente de Nô, Sholl, Bok, and others (Refs. 8, 52, 93). A typical cell in the cerebral cortex receives input connections from some hundreds of other cells, which may be located in widely scattered regions, but its output is more likely to be transmitted to a relatively localized region. Cells which receive sensory input signals are likely to have a restricted field of origins in a sensory surface, such as the retina or the skin.

The mapping of the frog retina into the brain has been studied by Lettvin (Ref. 51) who finds a rather precise topographic mapping, in which several different types of information are represented in different layers.* This topographic mapping is established genetically despite the fact that the fibers which transmit the information from the retina are apparently completely "scrambled" in the optic nerve. Moreover, experiments by Sperry (Ref. 94) and more recently by Lettvin (Ref. 51) show that if the optic nerve is severed and allowed to grow together again, the fibers which originally transmitted to a particular terminal location will tend to reconnect to that same terminal location, with surprisingly little loss of precision. This points to a highly specific neural organizing capability, which must be taken into account in considering admissible types of constraints for a brain model. In the mammalian brain, each sensory modality appears to be represented by an orderly topographic mapping analogous to that just described. Auditory stimuli, for example, are mapped into a region which is organized according to pitch; tactile stimuli are mapped according to body location, and so forth. Similarly,

---

* See also Section 3.1.4.

the motor neurons are organized, in the cerebral cortex, in an ordered arrangement which is topologically similar to the organization of the muscles which are controlled.

In contrast to the highly specific regional organization in the gross anatomy of the sensory projection areas of the cortex, the detailed microstructure of the network appears to be essentially random, governed only by directional gradients and preferences, and statistical distributions of fiber lengths for various types of cells (see Sholl, Ref. 93). In the human nervous system, it appears that the most specific and constrained topological organizations are to be found in the sensory and motor systems, while the intervening association network of the CNS is less tightly controlled in its organization, presumably depending more on learning and adaptive modification to establish the required pathways and linkages. The degree of precision in establishing the topological organization of neurons in even the most highly constrained reflex mechanisms is probably far less than that in most artificial data processing devices, and must retain a certain degree of randomness wherever the number and density of connections is appreciable. Unfortunately, no data are available which would indicate the complexity of topological constraints which correspond to the highly complex inherited behavior patterns which are known to exist in many species. Since the nature of such constraints is unknown, we shall avoid gratuitous assumptions about them, as far as possible. In the development of brain models, it will be our general strategy to start out with minimally constrained networks, and examine the consequences of introducing particular types of constraints, one at a time.

3.1.3 Localization of Function

Ever since the brain was first credited with the control of psychological activity, attempts have been made to delineate separate functions for its different parts. In the last century (largely under the influence of Gall) this took the form of an assignment of "mental faculties" such as intelligence, combativeness, amativeness, and religiosity, to special regions of the brain. As techniques for the study of functional anatomy improved, this gave way to a concept of organization into sensory tracts, motor tracts, and association tracts. The functional organization which was revealed has been most firmly established in the case of sensory and motor tracts, where a particular position in the brain is correlated with a particular sensory locus, or a particular set of muscles whose activity it controls. An excellent review of sensory and motor mapping can be found in Ruch (Refs. 88, 89). More recently, a finer breakdown in the localization of sensory functions has been demonstrated by Lettvin and associates (Ref. 51). Four distinct types of information, involving distinct aspects of the visual stimulus (contrast, curvature, movement, and dimming of illumination) have been shown to be mapped into four distinct layers of the tectum of the frog. This suggests localization of analytic functions, of a sort which has been suspected but not previously demonstrated.

In dealing with the so-called "association areas" of the cerebral cortex, and with other parts of the brain which are not clearly related to sensory data processing or motor coordination, something of the old treatment in terms of "mental faculties" still remains; specifically, centers have been found which are commonly attributed with primary

responsibility for temporary and permanent memory, for emotional behavior, for speech recognition and speech production, and (in the frontal lobes) for the integration of complex goal-directed activities. The lack of clear operational tests for such capabilities has been a hindrance to progress in such functional mapping, and the results are considerably more ambiguous than is the case with sensory and motor functions. A discussion of current evidence on brain localization with respect to these "higher faculties" is found in Pribram (Ref. 72). Much of the recent work is concerned with the localization of tracts which influence motivation, alertness, and consciousness in the organism (Refs. 1, 22, 38, 64, 65).

One feature which is of particular importance for brain models is the apparent plasticity of localization in the "association areas" (or "intrinsic systems", to use the terminology advocated by Primbram) in contrast to the relatively fixed and irreplaceable character of the sensory and motor tracts. Loss of function, due to destruction of association cortex, is apt to be transient, with adjacent areas taking over the function after a period of readaptation. Jackson, in his classic studies of the motor cortex, (Ref. 36) observed that even here localization is not rigid and absolute, and that a certain amount of flexibility exists, permitting the functions of damaged tissue to be taken over by neighboring areas. The sensory projection areas, on the other hand, appear to be indispensible to perception; destruction of the optical cortex leads to permanent blindness in an area corresponding to the location of the lesion, and similar phenomena are to be found in other sensory modalities. Thus, the extreme hypothesis of equipotentiality advocated originally by Lashley (Ref. 49), (who observed that cortical

ablation appeared to produce a general deficit in performance proportional to the amount of cortex extirpated, rather than eliminating specific memories and abilities) has been modified in the direction of relative localization, which is quite strict for certain sensory functions, and comparatively weak and readily modified for more complicated control functions, thinking, and memory.

A rather different approach to localization is suggested by the histological studies of cortical tissue, initiated originally by Brodmann, and pursued more recently by Lorente de Nô and Sholl (Refs. 52, 93). The "cytoarchitectonic areas" which have been described in these studies differ in their microstructure and detailed organization, and attempts have been made to relate such differences to the function of the cortex in which they occur. To date, this approach has not led to particularly significant results, although in principle it may ultimately suggest the essential organizational properties which must be incorporated into a brain model.

At the primitive level of organization to which our models will aspire at this time, current data on brain localization are of only secondary interest. The main features of the brain still seem to be adequately described by the general topological structure shown in Fig. 1. The "central integration and control network" indicated in the diagram is known to possess some important internal demarcations in higher organisms, but the precise functions of these parts and their interrelations is still largely speculative. In simpler brains (crustacea, for example) the gross organization is probably no more complex than indicated by the diagram; and it seems likely that in general it is the fine structure, rather than the gross anatomy, which determines the functional properties of the network.

3.1.4 Innate Computational Functions

There is no doubt that mechanisms of considerable complexity, sufficient for perceptual tasks and the control of organized behavior, can be created by genetic control of growth and maturation. This is most dramatically evident in the instinctual patterns of insects (for example, the well known communication system of bees, and the frequently cited behavior patterns of carpenter wasps), but is also clearly present in vertebrates (e.g., the spawning behavior of salmon, and the migratory behavior of birds, as described in Ref. 90). Recently, Gibson and Walk have furnished clear experimental evidence for the innate perception of depth in mammals (Ref. 24). All of these phenomena require "built-in" control mechanisms, of a rather intricate sort. In the cases just cited, these built-in mechanisms are not known in any detail. A number of more elementary functions have been discovered, however, which provide some picture of the types of "computational mechanisms" which are likely to exist throughout the central nervous sytem.

The stimulus analyzing mechanisms discovered by Lettvin and associates for frog vision have already been mentioned.* In these studies, it is found that certain ganglion cells in the frog retina respond only to contours or strong contrast gradients within their sensory field; others respond only to convex images; others to moving boundaries; and still others to a general dimming of illumination over their entire field. Each of these four cell types transmits its information to a distinct layer of the frog's tectum, where its position is mapped topographically. Thus, one layer represents a contour

* Other visual analyzing mechanisms have recently been demonstrated by Hubel and Wiesel (Ref. 113) in the cat's cortex (see Chapter 23).

map, or outline drawing of the stimulus field, another represents a location map for small convex objects or corners, a third represents movement vectors, and a fourth indicates regions of dimming illumination.

At the motor-control end of the nervous system, a number of reflex arcs and servo-control systems have been analyzed. The pupillary reflex, for example, has been analyzed as a typical servomechanism by Stark and Baker (Ref. 96). A considerable amount of work has also been done on the cerebellar servomechanisms which regulate muscular action under the control of cortical decisions and kinesthetic feedback information (c.f. Ruch, Ref. 89). It is probably safe to assume that similar closed-loop control systems, employing familiar servomechanism principles, are employed throughout the central nervous system for such purposes as controlling level of activity, preventing runaway excitation phenomena (such as occur in epileptic seizures), and regulating sensitivity to selected aspects of the sensory input data.

It is worth noting that most of the specific computing mechanisms used in muscular control appear to be of an analog variety, rather than digital; they make use of intensities and frequencies of activity for the direct control of servo-systems, rather than computing a control formula from encoded data and then generating the control signal required. The stimulus analyzing mechanisms found by Lettvin, however, constitute a sort of digital code, in which stimulus properties are represented by presence or absence of signals from particular neurons. It seems likely, as von Neumann has observed (Ref. 105) that the brain makes extensive use of both digital and analog principles in its operation, and it appears that both types of devices may be genetically determined.

An interesting example of theoretical speculations on possible computational functions employed in shape discrimination in the octopus can be found in Sutherland (Ref. 98). Sutherland reviews several alternative theories, and presents evidence in support of his own conjecture that the octopus responds to an analysis of the horizontal and vertical dimensions of the stimulus measured along all possible cross-sections. No attempt is made, however, to tie the computational process to a particular neurological structure, or to indicate a mechanism which might carry out the indicated operations.

3.1.5. Phenomena of Learning and Forgetting

Thus far, we have concentrated on the anatomical and physiological features of the nervous system which appear to be basic for the design of a brain model. We now turn to some of the behavioristic and psychological functions which a brain model should be able to demonstrate.

Phenomena of retention and adaptation in organisms have been studied in a variety of experiments, varying greatly in their design. In traditional usage, "memory" experiments have been concerned more with the retention and recall of experience, while "learning" experiments are concerned with the acquisition and modification of behavior. Both types of investigation, however, are concerned with lasting modifications in the state of the organism, and in complicated problems (e.g., those involving "insight") one tends to merge into the other; accordingly, all of these experiments will be considered together in this discussion.

Quantitative studies of learning and memory in psychology stem from the classical experiments of Ebbinghaus, in 1885, on the learning and retention of nonsense syllables. Using himself as a subject, he obtained learning and forgetting curves, and demonstrated many of the phenomena of recognition and retention which have interested psychologists ever since. Related phenomena have been studied by Bartlett (Ref. 5) using more highly organized material. A second type of experiment, the conditioned reflex experiment, first employed by Pavlov, is characterized by the association of an existing response to a new stimulus, which did not evoke the response prior to the conditioning procedure. A third type of experiment, employed originally by Thorndike and recently studied extensively by Skinner and others, is concerned with the learning of a pattern of behavior which is instrumental to the solution of a problem, or which satisfies a drive. Where such problem-solving behavior appears to depend in a crucial way upon a "cognitive restructuring" of the situation, or the formation of a new "concept", we have an experiment in "insight" or "concept formation", as in the studies of the Gestalt psychologists.

It is possible that these three types of experiments are actually demonstrating fundamentally different mechanisms of learning. The first deals with recognition and recall of previous perceptual experience; the second is concerned with the generalization of responses from initial stimuli to new stimuli by virtue of temporal association; the third is concerned with the discovery and establishment of problem-solving behavior. Still other experiments deal with such phenomena as short-term memory span, acquisition of needs and motives, attitude formation, perfection of a

motor skill, or learning to make fine perceptual judgements. Undoubtedly, the same physiological processes are tapped in many of these tasks; on the other hand, attempts at subsuming all of them under a set of general "laws of learning" does not seem to be particularly helpful for our present purpose. From the standpoint of brain model construction, it seems safest to regard each type of learning experiment as a distinct problem, with its own variables and rules of behavior which we hope that our model will duplicate under equivalent experimental conditions. The main value of such psychological experimentation, then, is to provide us with a set of "calibration experiments", by means of which a model can be compared with known organisms under well defined conditions. The reader who is unfamiliar with the literature of learning experimentation will find the reviews by Hilgard, Brogden, and Hovland (in Ref. 112) particularly helpful.

In a number of experiments, attempts have been made to find the actual physiological correlates of the learning or memory phenomenon. Notable among these are the experiments of Penfield (Ref. 68), who finds that electrical stimulation of selected points on the cortex may evoke long and vivid sequences of past experience, apparently with hallucinatory clarity. John (Ref. 39) has recently reviewed experiments in cortical conditioning, and reported a number of interesting results of his own, which suggest that memory may involve modification of the connections between the deep centers of the brain stem and the cerebral cortex, with the reticular formation playing a particularly significant role. The experiments of Olds (Refs. 64, 65, 66) on the reinforcing effects of electrical stimulation applied to certain points in the hypothalamus and adjacent structures suggest that these may be involved in the motivational aspect of learning. Such experiments, which have only recently become possible through the improvement of electrophysiological techniques, are likely to become increasingly valuable as guides to theory construction.

3.1.6 Field Phenomena in Perception

Early studies of perception were largely concerned with the absolute question of what perceptions are made of; such studies were concerned with range and sensitivity of sensory abilities, measurement of limits and thresholds, and the detailed dissection of sensory stimuli into fundamental components. Such studies form the main subject matter of classicial psychophysics. In psychology, they gave rise to an atomistic approach (reaching its utlimate expression in the work of Titchener) in which it was proposed that any phenomenon of perception could be accounted for by a proper compounding of sensory elements, each of which retains its own identity, like a piece of tile in a mosaic. During the last few decades, largely under the influence of the Gestalt psychologists, studies of perception have turned from the question of the constituents of perception to the question of the conditions under which a given perception occurs. It is now generally accepted that what is perceived depends not only upon the properties of the stimulus object, or image, which is recognized, but upon the organization of the entire sensory field in which it is embedded. This is true not only in vision, but in other sensory modalities as well.

The field phenomena which have been studied include the effects of contrast, figure-ground organization, frames of reference, depth perception, size constancy, and illusions. The reader is referred to Koffka (Ref. 44) and Gibson (Ref. 26) for detailed discussion of these topics. For present purposes, the most important implication of this work is that a physical model for a perceiving system must permit the interaction of all elements in a spatially organized field. It is not sufficient simply to detect sets of elements which represent a "pattern"; the perception of a pattern, and the

interpretation of it, depends in a fundamental way on metric relationships to other sense data from the same modality, and correlations with sensory data from entirely different modalities. The perception of a line as "upright", for example, depends on its observed angles relative to visual standards of "uprightness", such as the corners of a room, and also upon the gravity senses and kinesthetic data which provide a frame of reference for "up" and "down". The decision that two disjoint patches of illumination represent parts of the same object rather than different objects depends upon their contrast or resemblance to the field structure around them, as well as on their relationship to one another. It is possible (as Gibson has suggested) that recognition is never achieved, in biological systems, by the representation of a particular receptor configuration, but only by the representation of sets of relations (angles, ratios, etc.) as its elementary data. If this is the case, a suitable set of analyzing mechanisms, capable of measuring such variables must be included in the pre-recognition tracts of a brain model. As our models gain in sophistication, it is, in fact, becoming increasingly apparent that such analyzing mechanisms are essential for purposes of efficiency and economy of design.

The perceptrons to be considered initially will not possess intrinsic field-organization properties. With the introduction of cross-coupled systems, such properties begin to emerge. An evaluation of these systems by means of typical "Gestalt perception experiments" has barely begun at the present time, but represents one of the most important tasks to be undertaken.

3.1.7 Choice-Mechanisms in Perception and Behavior

Selective attention and "set" are fundamental phenomena in the control of psychological activity. They indicate mechanisms for choosing between alternative courses of action, or points of view, and play a logical role analogous to the selection of different branches in a "flow diagram" of a digital computing routine. Attention and psychological set are largely determined by the situational context in which behavior occurs, and by the current "goals" or "purposes" of the organism, which may be thought of as choices of a superordinate sort, under which sub-decisions are made to select particular modes of activity. For example, an individual who is set to look for a word in a dictionary will be most attentive to the sequence of letters in boldfaced type, while someone who is looking for torn pages will probably be unaware of the particular letter combinations, and someone who is simply scanning the volume to look for pictures is apt to notice neither the spelling nor the condition of the pages.

The importance of set, or attitude, for learning has been emphasized by Hebb (Ref. 33), but choice mechanisms of this type have rarely been incorporated in the detailed design of theoretical brain models. In purely logical models of behavior, they play a considerably more prominent role -- for example, in Tolman's learning theory, and in Newell and Simon's models for problem solving behavior (Refs. 62, 63), selective choice-mechanisms are specifically designated. In a brain model, it is clear that such phenomena must be closely related to the problem of "temporary memory", since the set under which the brain

is currently operating must be represented by a temporarily stable, but nonetheless readily altered, state of the system, capable of modifying processes which go on while it persists. It seems likely (although unsupported by any direct evidence) that pools of neurons connected by reverberating circuits may be important set-maintaining devices in the nervous system, exerting their influence on the brain as a whole by means of a widely distributed barrage of sub-threshold excitation or inhibition. The plausibility of such mechanisms will be considered in more detail in a later chapter.

3.1.8 Complex Behavioral Sequences

The discussion of psychological sets and choice mechanisms brings us to a consideration of even more highly organized behavior and thought patterns, such as the steps taken in performing an arithmetic computation, or driving to work, or performing a piece of research. All of these activities represent orderly sequences of decisions and action, and can be considered, as Newell and Simon have suggested, as programs to be performed. In some cases, these programs are highly stereotyped, and determined by rigid rules; in other cases, they employ chance mechanisms and heuristic procedures. Much of the classical psychological literature on problem solving and insight is relevant to this second class of programs, while a rat running a maze might be considered an example of the first type. As in the case of selective attention and set, these problems have not been dealt with in detail by any brain models proposed to date, but it seems likely that at this level the brain and the computer begin to approach a common meeting ground. Problems of memory span,

storage, and sequence control are present in both types of systems, and many of the logical problems confronted in "heuristic programming" (Refs. 60, 62, 63) seem to be direct translations from human problem-solving experience to the language of computing machines. This does not mean that the physical structure of a brain model must ultimately resemble that of digital devices, but rather that the same basic logical organization -- a memory for programs, a memory for data, and a mechanism for the sequential performance of a given program -- must be available. The "programs" themselves presumably take the form of sequences of selective sets, or bias states, arranged in a heirarchical manner, so that sub-operations are performed under the control of a "master set" or "master program" which determines the overall plan of activity. While the detailed properties of such systems must necessarily remain speculative at the present time, we shall see that such a concept is compatible with the organization of perceptrons not too far removed in complexity from those which we are now capable of analyzing.

3.2 Current Issues

While the discussion of the preceding section has attempted to stick to a relatively conservative and uncontroversial rendition of physiology and psychology as it applies to the brain model problem, it is clear that in the last pages we have been drawn into increasingly speculative and uncertain areas of discourse. In this section, an attempt will be made to highlight a number of issues which seem most salient in determining the fate of various brain models, and which are not answerable at the present time outside the realm of speculation.

Of necessity, a physical model will have to take a stand on most of these issues, and it is possible that by investigating the logical consequences of such a stand, a decision as to the plausibility of various alternatives might be made; the brain model approach has a chance, here, of providing answers which empirical studies have so far been unable to discover. In any event, the decision taken on these issues represent the points at which a brain model is most vulnerable to future attack, as new evidence is uncovered.

3.2.1 Elementary Memory Mechanisms:

The status of current information on basic memory mechanisms in the nervous system has been reviewed recently by Burns (Ref. 13). Most brain models employ some memory hypothesis, but evidence as to the nature of actual physiological mechanisms which might be involved is almost totally lacking. It is generally agreed, simply on the basis of definition, that whatever we call "memory" involves a modification of neural activity in the central nervous system or its output signals, as a function of exposure to previous events or "experience". In some models, this modification has been attributed to persistent activity in closed loops of neurons, but most theorists are now agreed that, while such a memory mechanism might account for "short term memory", and might play a significant role in the establishment of more permanent memory traces, there must also exist a non-volatile memory mechanism (e.g., a structural or chemical change) which can outlast periods of neural inactivity, and is relatively insensitive to transient activity in the nervous system (see Hebb, Ref. 33, pp. 12-16). The nature of this memory trace mechanism, it is generally agreed, must be such as to facilitate the use

or selection of neural pathways which have been active at the time of the "remembered" experience or behavior, and virtually all specific models assume that it takes the form of a facilitation of connections between sources of excitation and responding neurons in the motor system or CNS. In making such an assumption, the influence of the conditioned reflex model, which suggests that sensory neurons become coupled to association neurons, by which they are connected to motor neurons, is clearly evident. An alternative position, in which the preferred pathways "win out" by surviving deteriorative changes in unused pathways, rather than by active facilitation, has not been explored to any significant degree, but appears to be logically similar to its potentialities.

Granting that the memory mechanism takes the form of some means of selecting particular patterns of activity in preference to others, depending upon the input or current state of the nervous system, particular physiological models include: (1) mechanisms for reconstituting past activity states of the entire CNS or a major portion of it; (2) mechanisms for selecting particular output channels as a function of current activity or sensory inputs. The specific mechanisms proposed generally fall into one of the following four categories:

(1) Extracellular influences and modification of the neural medium: This has been proposed by Köhler (Ref. 45), Bok (Ref. 8), and others, who assume that, if a "structural trace" is present at all, it is not laid down in specific neurons, but in the surrounding medium, where it is capable of modifying activity in nearby neural tracts. The possible form that such a mechanism might take has never been specified in detail, and the approach is generally discounted by current theorists. The motivation for such a

hypothesis comes in part from attempts at preserving the isomorphism between a spatially distributed memory trace and spatially organized visual events, as in Köhler's system. While it is not implausible to assume that the surrounding medium participates in the memory trace structure, it seems likely that such interaction between medium and neurons would be highly localized, probably influencing only a single neuron or synaptic junction, rather than forming a widespread organized structure independent of the neurons themselves. If such a position is accepted, then whatever is left of this approach can be subsumed under one or another of the remaining neural modification mechanisms.

(2) Threshold Modification: The hypothesis that the threshold of an active neuron may be reduced as a consequence of the activity, thus making it more likely that this cell will respond to future stimuli, has frequently been proposed as a possible memory mechanism (c.f., Taylor, Ref. 99). If we take the "threshold", in its conventional sense, to mean the degree of membrane depolarization or the level of input excitation which will cause the neuron to discharge, regardless of the particular synapses involved in the transmission of excitation, then this model meets two main objections: first, the sensitivity which is acquired is non-specific, making it more likely that the cell will respond to any input, rather than just those which were effective at the time that the memory trace was established; second, after a long history of activity, we would expect the thresholds of all neurons to be reduced to a minimum level, unless some recovery mechanism exists. If such a recovery mechanism does exist, memory will tend to be lost as a consequence, and it must be shown that

the rate of forgetting would not vitiate the value of the system. Occasionally, the concept of "threshold reduction" seems to be used in the sense of an increase in specific sensitivity of a neuron to a particular afferent fiber. In this case, the threshold reduction mechanism becomes indistinguishable from a synaptic facilitation mechanism, which is considered below.

(3) Strengthening of active neurons: Eccles (Ref. 18), Uttley (Ref. 102), and Rosenblatt (Ref. 79) have proposed models in which the output signals of a frequently active neuron gain in strength or effectiveness, affecting all terminals alike. This model retains the specificity of response of a neuron (unlike the threshold reduction model) but increases its power to activate the neurons which follow it in series. If the output signal from a neuron goes to a single destination only, this is equivalent to a model which strengthens particular synaptic connections. If the output goes to a number of different locations, however, there is a lack of specificity in the channel-selection properties of this mechanism, which must generally be offset by auxiliary hypotheses. In Rosenblatt (Ref. 79) it is shown that by means of a suitably organized feedback mechanism, a particular output channel can be selected through a statistical bias. The feedback guarantees that these cells which are reinforced all have at least one "desirable" output connection, the other connections being distributed at random among a large number of alternative terminal neurons, each of which consequently receives only a fraction of the total reinforcement applied. While such a model is shown to be logically workable, the specific feedback connections required make it physiologically implausible, and it remains less efficient than a model in which specific synapses, rather than total neurons, are selected for modification.

(4) Modification of selected synapses: This model has been employed by Culbertson (Ref. 17), Hebb (Ref. 33), and others, and is employed in most current perceptron models. The mechanism takes account of the correlation of activity between an afferent synapse and the efferent neuron, augmenting the strength of the synaptic ending (or, equivalently, the sensitivity of the sub-synaptic membrane) if the correlation is positive, and, in some cases, diminishing it if the correlation is negative. The actual physiological process by which such a correlation might occur is obscure, but the logical advantages of such a mechanism are clear. Hebb has proposed that actual synaptic growth might occur, improving the contact between the transmitting and receiving neuron. While Eccles has considered possible synaptic growth mechanisms in some detail (Ref. 18) there is little evidence to support this conjecture. A possible biochemical mechanism has been proposed by this writer (Ref. 83), which assumes that large molecules used as catalysts for the production of transmitter substances in the endbulb must originate from the nucleoplasm of the post-synaptic cell, and that the exchange of these molecules is facilitated by membrane depolarization and periods of activity in both cells. An alternative possibility, in which the memory mechanism is entirely contained within the post-synaptic cell, is that a persistent sensitization of the subsynaptic membrane in the neighborhood of an active synapse occurs, given the hypermetabolic state which follows activity. The facilitation of a neuron's response to repeated subthreshold signals which has been reported by Bullock (Ref. 11) indicates that a localized persistent effect of the sort hypothecated does exist; it remains to be shown that the subsequent firing of the neuron may serve to "stamp in", or fix in a more permanent manner, the temporary sensitivity which has been observed.

The evaluation of a particular memory hypothesis must depend, at this stage, upon its logical power when employed in specific brain models, as well as its physiological plausibility. The mechanisms which are considered in this report have been selected for their simplicity and their demonstrated ability to yield interesting behavioral results. They suggest plausible directions in which to look for a physiological mechanism, but it remains possible that the actual mechanisms employed by the brain may be of a drastically different sort. It is fundamental to this approach, that any lasting change in the system, whatever its physical form, may act functionally as a memory trace. It seems likely that there is not a single memory mechanism, or even only two memory mechanisms at work in the brain, but rather a great number of dynamic processes, ranging from temporary facilitation and fatigue effects to permanent structural changes, all of which contribute in some way to the observed psychological phenomena called "memory". Among these processes, it is likely that one or two play an outstanding role, but likely candidates have not yet been found, and in the meantime, it seems wise to retain an open mind on the entire question.

3.2.2 Memory Localization

There is hardly any more agreement on the question of where memory traces are to be found (in the gross anatomy of the nervous system) than there is on the question of what they consist of. Lashley (Ref. 49) was largely responsible for the emphasis on "distributed memory" among many theorists over the last few decades, and Sperry (Ref. 95) has contributed a number of experiments which indicate that the residual

effects of learning must be widely dispersed throughout the brain. On the other hand, Penfield (Ref. 68) has shown that specific recall may be evoked by stimulation of specific selected points in the cerebral cortex. E. R. John, in a model which is supported by a certain amount of experimental evidence (Ref. 39), proposes that the memory traces are distributed between the thalamus and cortex, involving reverberating circuits and feedback loops between these two regions rather than being localized in one or the other of them.

The question of localization is of less importance for a functional model of the brain than is the question of mechanism; as long as we assume that it is the network topology, rather than the actual anatomical position of neurons, which is important in determining the brain's logical properties, there is no reason for requiring that a brain model resemble the biological system in its spatial organization. The indirect implications of the different theories of localization are of considerable importance, however. For one thing, the view that the brain contains its memories in a widely dispersed, intermingled form, suggests a mechanism in which the same cells participate in a great variety of different, and perhaps totaly unrelated, memory organizations. A model which can separate distinct memories from such a multiply overwritten system will be quite different in character from one in which each remembered event is stored in its own distinct location. For another thing, the apparent complexity of memory-sites which may interact in the recall of a single experience or association (as emphasized in John's work) impresses us with the possibility that human memory may be a product of a number of related processes and mechanisms, perhaps acting in a complex sequence of cause-and-effect, rather than a simple correlation of inputs and outputs.

Again, we are stuck with the necessity of simplifying for lack of detailed knowledge. While it is likely that memory and recall in the human nervous system involves the coordinated activity of several parts of a complex structure, we will attempt, at the outset, to see what psychological properties can be duplicated by a system in which memory is located in a single set of connections, with a minimum of structural differentiation. As perceptrons are elaborated into more highly structured models, the question of which connections should be allowed to participate in memory processes will be reconsidered, and alternative systems will be investigated.

3.2.3 Isomorphism and the Representation of Structured Information

Lashley, Köhler, Greene, MacKay, and others (Refs. 28, 45, 50, 55, 56, 110) have dealt with various aspects of the problem of isomorphism between the representation of an event in the central nervous system and the physical structure of the event in the outside world. In the naive isomorphism of Köhler, it is required that the representation in the brain should actually have a spatial structure resembling the thing that it represents; in the more sophisticated form advocated by Greene, it is sufficient that the representation should have a logical structure (not necessarily spatial in its physical manifestation) which permits it to be broken apart, dissected, and reassembled by suitable manipulations or attention-directing processes, in a way which is related to the parts, surfaces, or aspects of the real-world phenomenon. While some such structural representation seems to be inescapable in human perception, thinking, and imagery, the exact form that this might take is again almost totally unknown. This is essentially the problem of

determining the code employed by the brain in its representation of perceptual phenomena. We know that the code is one which enables us to recognize parts, relations, symmetries, and other organizational features which might be lost in a completely arbitrary representational system (such as a code which assigns binary symbols, in sequence, to all stimuli, and then lists all of those which are to be considered as "similar"). We also know that there are parts of the brain (the sensory projection areas) in which actual spatial organization of stimulus patterns is retained. We do not know, however, how far the representational code must go in the direction of spatial isomorphism in order to account for the organizational properties of experience. As usual, we shall begin with a simplification which assumes an unstructured coding, but it seems likely that this will have to be abandoned in order to deal with problems of figural representation, perception of relations, and other "gestalt problems". An attempt will be made in this report, however, to show that the required structuring for some of these problems may be acquired by adaptive processes and need not superficially resemble the phenomena which are represented.

3.2.4 Adaptive Processes in Perception

Much of the theoretical work on brain models (Hebb, Hayek, etc.) has been concerned with processes by which complex perceptual organizations can be "built up" out of sensory fragments, by a process of learning or association. Consequently, the question of adaptability, or modifiability, of perception is of paramount importance as a guide in model construction. The history of this problem has recently been reviewed by Hochberg (Ref. 34). Studies of "perceptual learning" have

been concerned (1) with the organization of given perceptual elements into "concepts", or "kinds of objects", and (2) with the modification of the perceptual elements or "impressions" themselves.

(1)The first type of experiment is concerned with the discrimination, rather than the "appearance" of stimuli. It is clear that much recognition and discrimination, as in the learning of speech sounds in a new language, is highly dependent upon learning. Such processes typically involve differentiation, rather than synthesis of complex patterns out of readily identified parts. Another, important part of perceptual concept formation is concerned with associating, or classifying readily discriminable patterns or symbols having the same significance (such as a Roman, italic, and script form for the letter "A"). (2) On the other hand, there are a number of studies concerned with attempts at modifying the seemingly intrinsic "appearance" of the stimulus itself. Such experiments are not concerned with refinements in discrimination or assignment of appropriate names to stimuli; they are concerned with re-structuring the sensory data at a considerably more "primitive" level. Such experiments include studies of figural aftereffects (Ref. 25), ambiguous figures (Ref. 107) the effect of memory upon color perception (Ref. 10), and the various experiments performed with inverting prisms to determine whether a human subject could learn to perceive normally with an inverted retinal field. Work with animals reared in darkness and exposed to the light for the first time in various test situations has been reported by Riesen (Ref. 75) and Gibson and Walk (Ref. 24) have conducted experiments with infants and newborn animals to determine whether depth perception is possible prior to learning. Other data have been collected by von Senden for congenitally blind human subjects to whom sight is restored by surgery (Ref. 106).

In general, the conclusions of this work seem to indicate that while recognition, in the sense of being able to discriminate and assign an appropriate name to an object, is largely dependent upon experience, the "subjective appearance" of a stimulus is relatively inflexible, and in some species, at least, may be innately given by the structure of the nervous system. Sperry's work with frogs, for example, in which the optic nerves are cut and then allowed to rejoin with the eyeballs inverted, suggests that no amount of relearning can compensate for so drastic a change (Ref. 94) and the Gibson-Walk experiments support the assumption of a highly developed sense of depth perception in many mammals from birth. To a much lesser degree, modification of visual images by experience is possible; generally, this takes the form of persistent field interactions (as in figural aftereffects) rather than a basic reorganization of perceptual experience. The extent to which perception might be organized by adaptive processes is currently unknown, and this is one of the main areas in which theoretical brain models may prove helpful to psychology.

3.2.5 Influence of Motivation on Memory

In psychological learning theories, it is commonly assumed that a "drive" or "motive" must be present in order for an animal to learn. Conditioned reflex experiments, on the other hand, frequently fail to show any relationship between the "motivation state" of the animal and the learning process. Speculation about the role of motivation in perceptual learning has also been quite extensive, and a number of experiments have been performed, to test the learning of perceptual discriminations or related tasks on the basis of "mere repetition" as opposed to directed learning. In these experiments, it is often hard to distinguish between

"attention" and "motivation", and the results are generally inconclusive. It seems that a certain amount of "incidental learning" does indeed occur, which is not directly relevant to the goal or task of the subject at the time; the actual degree of motivation, reward or punishment, or "reinforcement" that may have been involved, however, is impossible to ascertain in any absolute way. For the brain model problem, it is important to note that there are some learning situations, at least, in which "reward and punishment" can be used to control the acquisition of new responses; whether or not this is universally the case, and the actual physiological mechanisms involved, remain open questions at this time. It should be remembered, however, that any brain model which relies on the intervention of an outside agent or experimenter to direct the learning process is implicitly taking a stand on this issue. A possible compromise is found in the approach of Ashby (Ref. 3) where the brain is described as a complex homeostatic organization, in which particular "crucial variables" are capable of triggering random changes in organization if they exceed critical limits; stabilization of behavior, in such a system, is not a result of learning from reward, but is due to the cessation of disruptive changes which occur when the system makes a mistake. The main difficulty in making use of this approach is in guaranteeing that changes are sufficiently specific and well-directed so that the organism achieves its new behavior pattern in an economical and relatively direct fashion, rather than going on a random walk through all possible alternatives before arriving at the required solution.

3.2.6 The Nature of Awareness and Cognitive Systems

While it has been relegated by many theorists of the realm of philosophy or semantics rather than science, the question of the nature of consciousness or awareness keeps recurring in the literature. Current physiologists and psychologists represent the whole range of philosophical positions on this subject. For Eccles (Ref. 18) there is a conscious "mind" which controls the body by acting upon the nervous system. For Penfield and Jasper, awareness is a state of the nervous system involving heightened sensitivity and improved coordination, under the control of the centrencephalic system, and particularly the reticular formation (Ref. 38). John (Ref. 39) suggests that "awareness may be a property arising from the process of 'cortico-reticular resonance' ". For Culbertson (Ref. 17), consciousness is a property of trees of causal relations which tie together the events of the external physical world and the neural events in the brain. Lotka (Ref. 53) has suggested that we look to the world of molecular events for an explanation, and that consciousness involves particular unstable states of molecular or atomic particles.

To this writer, it seems likely that the question of the "nature of awareness" can be bypassed, in much the same way that we bypass the question of the "nature of perception", by concentrating on the experimental and psychological criteria which may be used to distinguish the actual phenomena in question. When a subject reports that he is "conscious" or that he was recently "unconscious", we are led to believe him or disbelieve him on the basis of his behavior, and what he is able to report about the content of his "experience" at the time in question. From an

operational point of view, the fact of "consciousness" is closely connected with the accessibility of information and its ability to influence overt behavior; it is, in fact, meaningless to say that an individual is "conscious" unless there is something that he is conscious of. The questions which can be asked concerning this phenomenon in a theoretical brain model (where we are not free to assume any intrinsic similarity of processes to those in the human brain) are questions of what can be discriminated, "seen", "attended to", or "remembered" under specified conditions. All that we can say, in the last analysis, is that the system acts as if it were conscious, leaving the question of the actual existence of consciousness in the system for metaphysicists to consider.

Systems which represent information internally, in such a way that it can be utilized for the control of certain kinds of responses (such as running, thinking, or talking) will be called cognitive with respect to the realm of information which is represented and the class of responses which this information controls. Note that this term is used in a relative, rather than an absolute sense. Thus the representation of information in the form of an image on the retina is not sufficient to permit us to say whether or not the organism is cognitive with respect to its visual environment; we must also demonstrate that this information is accessible to the organism for the control of some specified set of responses. We might say, for example, that a man who automatically stops for a red light, but is unable to state afterwards why he stopped is cognitive with respect to red signals at the level of overt motor-responses, but not at the level of verbal recall. Conversely, an unskilled pianist may be cognitive with

respect to errors in his performance at the verbal level, but not at the motor control level. We use the term cognitive, then, to indicate that knowledge of some realm of information is accessible for the control of some specified class of responses. This usage permits us to reserve judgement on the definition of such phenomena as perception and awareness, and still to recognize a class of psychological phenomena involving the accessibility of information, with which we shall be concerned.

3.3. Experimental Tests of Performance

The purpose of a theoretical brain model is to demonstrate how psychological phenomena can arise from a physical system of known structure and functional properties. In the preceding sections of this chapter, we have reviewed the physiological data which suggest the general form of the model, and the psychological data against which its performance must be measured. We now turn to a more specific consideration of the psychological tests which might be applied to a brain model in order to evaluate its performance, and to compare alternative systems with one another.

3.3.1 Discrimination Experiments

In the simplest type of experiment which can yield psychologically significant information about a system, two distinct stimuli are presented to the model, which is required to respond differentially to them. In the general case, it is not necessary to limit this experiment to two specific stimuli or sensory patterns; two or more classes of

patterns may be employed, each class consisting of "similar" patterns, such as squares, or triangles, or various sizes and styles of the letter "A". This experiment may be performed either to look for spontaneous discrimination by the system, in the absence of intervention or guidance by the experimenter, or to study forced discrimination in which the experimenter attempts to teach the system to make the required distinctions. In a learning experiment, a perceptron is typically exposed to a sequence of patterns containing representatives of each type or class which is to be distinguished, and the appropriate choice of response is "reinforced" according to some rule for memory modification. The perceptron is then presented with a test stimulus, and the probability of giving the appropriate response for the class of the stimulus is ascertained. Different results will be obtained, depending on whether or not the test stimulus is chosen to correspond identically to one of the patterns which were used in the training sequence. If the test stimulus is not identical to any of the training stimuli, the experiment is not testing "pure discrimination", but involves generalization as well. If the test stimulus activates a set of sensory elements which are entirely distinct from those which were activated in previous exposures to stimuli of the same class, the experiment is a test of "pure generalization". The simplest of perceptrons, which will be considered initially, have no capability for pure generalization, but can be shown to perform quite respectably in discrimination experiments particularly if the test stimulus is nearly identical to one of the patterns previously experienced.

3.3.2 Generalization Experiments

As indicated above, a pure generalization experiment is one in which the brain model, or perceptron, is required to transfer a selective response from one stimulus (say, a square on the left side of the retina) to a "similar" stimulus which activates none of the same sensory points (a square on the right side of the retina). Generalization of a weaker sort may be demonstrated if we simply require the system to transfer a response to members of a class of similar stimuli, which are not necessarily disjoint from the one which has been seen (or heard or felt) before. As in the case of discrimination experiments, it is possible to study either spontaneous generalization, in which the criteria for similarity are not supplied by an outside agency or experimenter, or forced generalization, in which the experimenter's concept of similarity is "taught" by means of a suitable training procedure. Some of the most significant problems in brain mechanisms concern generalization phenomena, and particularly the meaning of "similarity" for a particular kind of system. In common with a number of other theorists (e.g., Pitts and McCulloch, Ref. 71), this writer will assume that similarity is primarily determined by a group of transformations which stimuli may undergo in a particular physical environment. In the normal physical environment, for visual stimuli, this would include rigid motions, rotations, size changes, projective transformations, certain types of distortions or continuous deformations, and changes in color or contrast. A number of more subtle forms of similarity (as in styles of architecture, gestures and mannerisms, etc.) are presumably due to association of events into classes at a higher level of organization than we are concerned with at this point. It should be noted, however, that a perceptron which is taught

to form arbitrary classes of stimuli might be expected to generalize along completely arbitrary or abstract dimensions, "similarity of style" being as legitimate a candidate for a basis of classification as "similarity of shape". In the simple perceptrons, we will find that "pure generalization" does not occur, although an apparent generalization of responses to stimuli which share many sensory points with those previously experienced can be demonstrated. In this report, this weak form of generalization will be considered under "discrimination phenomena", the term "generalization" being reserved primarily for cases in which mechanism for recognizing actual similarity, rather than a rough approximation to identity, is involved.

3.3.3 Figure Detection Experiments

In the experiments considered above, two or more kinds of stimuli are always employed, in order to avoid the trivial case in which the desired response is automatically evoked by any stimulus that might occur. Since it is assumed that at each moment of time exactly one stimulus is present, these experiments represent a "forced choice" situation, in which the brain model is obliged to give one of several positive identifications in response to whatever it "sees". Such experiments have their counterparts in animal and human experimentation, and permit the study of an important class of psychological problems, involving simply structured situations. An alternative approach, which has been less studied to date, is to give the system the task of searching for a particular figure in a sensory field which may or may not contain it. In this case, the system is asked to discriminate between "figure present" and "figure absent", and is typically only instructed in the recognition of

one figure at a time. If the figure appears as a solitary object in an otherwise empty field, the task is a relatively trivial one. If the figure appears against a background, or as part of a complex of other patterns, the problem takes on a new aspect of complexity. In the most important case, this experiment permits us to study figure-ground organizing tendencies in a perceptron, by presenting it with embedded, or ambiguous figures which can be recognized as representing one thing if the field is appropriately structured, and a different thing if the field is structured differently. The Gestalt properties of "good figure" are supposed to determine the preference of a human observer to perceive one or another of the possible figures in such a field. Detection experiments permit us to compare the preferences and rules of "good figure" in a perceptron with those of human subjects, in controlled situations. Perceptrons considered to date show little resemblance to human subjects in their figure-detection capabilities, and gestalt-organizing tendencies. In Part IV of this report, some speculations concerning the development of such properties in more sophisticated perceptrons will be presented.

3.3.4 Quantitative Judgement Experiments

Another type of experiment with which little work has been done to date involves the estimation of quantitative properties of stimuli (size, distance, position, etc.) by perceptrons. It will be seen that simple perceptrons are capable of learning to represent stimuli by a continuously variable "analog" type of response. No work has been done to date, however, to investigate such questions as the generalization of quantitative judgement to new stimuli, or the accuracy which can be achieved in specific cases.

For more advanced systems, an important problem which must ultimately be faced is that of "perceptual constancies": the tendency in human subjects to perceive size, color, or other metric properties of a stimulus in terms of the "actual" physical properties of the object rather than its projection on the retina. A man, for example, is perceived to be about six feet tall regardless of whether his retinal image subtends one degree or fifteen degrees, and a dish appears to be circular in form regardless of whether its retinal image is a true circle or an elongated ellipse. It has been demonstrated in many psychological experiments that such phenomena are not based simply on familiarity with the particular objects involved; a completely unfamiliar form, seen in normal physical space, is perceived correctly, in terms of its "true" physical properties, except under exceptional circumstances (c.f. Gibson, Ref. 26).

3.3.5 Sequence Recognition Experiments

In the above experiments, it has been assumed that the stimuli are fixed, temporally invariant patterns. Analogous problems exist, involving discrimination, generalization, figure detection, and metric estimation for time-varying, or sequential patterns of all sorts. While static organization problems reach their greatest degree of complexity in the visual modality, temporal organization becomes comparably complex in the auditory field. Speech recognition is one particularly important case to be investigated. Problems include not only the recognition of particular movements, or sequences, but the segmentation of movement and sound patterns into figural units, words, or phrases as well. The recognition of sequences in rudimentary form is well within the capability of suitably organized perceptrons, but the problem of figural organization and segmentation presents problems which are just as serious here as in the case of static pattern perception.

3.3.6 Relation Recognition Experiments

In a simple perceptron, patterns are recognized before "relations"; indeed, abstract relations, such as " A is above B" or "the triangle is inside the circle" are never abstracted as such, but can only be acquired by means of a sort of exhaustive rote-learning procedure, in which every case in which the relation holds is taught to the perceptron individually. At the present time, the main hope for the abstraction of relations seems to lie in systems which are capable of executing a sequence of observations, according to a predetermined plan, in which first one member of the related pair is observed and then the other, the relationship between them being determined by the sequence of "experience" during the shift of attention from the first to the second. The problem of relation recognition is, at the outset, more complex than those previously considered, since it requires, by its very nature, the ability to recognize and attend selectively to at least two distinct "parts" of a total organization, specifying, for example, which part is larger and which smaller, or which part is "outside" and which "inside". The hypothesis that relation recognition involves a sequence, or program, of observation means that it must make use not only of figure organization capabilities (to separate the "parts" referred to) but of sequence recognition and sequential control capabilities as well. The actual experiments by which relation recognition can be detected must involve at least two components (such as square and triangle) which can be shown in such a way as to exemplify the relationship or not. In an ideal experiment, the system would be trained to recognize the relation by a number of examples with stimulus patterns or "parts" which do not resemble or intersect (in their retinal location) the test

patterns which are employed in evaluating the performance. If the perceptron can then indicate correctly, for entirely new stimuli, whether or not the relation holds, it will be considered that the relation has been abstracted by the system.

3.3.7 Program-Learning Experiments

The learning of sequences of behavior is the counterpart on the response side of the problem of sequence recognition. The problem has been discussed in detail by Lashley (Ref. 50). It requires, as a starting point, the ability to form "selective sets", which introduce a bias to give one of several alternative responses to a given stimulus. A capability of this sort has been shown to exist, to some degree, in relatively simple perceptrons, provided there is a feedback path from the response units to the association system (Ref. 79). To date, little has been done to study this capability in a quantitative fashion, but some of the heuristic arguments will be reviewed in Chapter 23. One of the most important applications of such a capability is in the control of the sequential activity involved in recognition of relations, and the "perceptual exploration" of a sensory field. Related phenomena, in which this capability plays a central part, are the sequential control of speech, thinking, and complex behavior patterns. The representation of problem solving activity in the human by heuristic programs has been studied by Newell, Shaw, and Simon (Refs. 62, 63), and it seems likely that many of their results might be transferred to a perceptron which is capable of program controlled activity.

3.3.8 Selective Recall Experiments

While most of the experiments described above involve "memory" in the sense of a change in behavior as a consequence of experience, they do not, in general, require substantive recall, of the sort which is displayed when we describe a person who we saw yesterday, or the location of furniture in a house where we lived last year. In selective recall experiments, the system is required to produce on demand information relevant to a particular time, place, or subject. This involves a particular case of "selective set" mechanisms, and can probably be demonstrated in most systems which are capable of program-controlled behavior.

3.3.9 Other Types of Experiments

In addition to the experiments considered above, we might ultimately wish to consider experiments in abstract concept formation, the formation and properties of a "self concept", creative imagery, and other higher-order psychological phenomena. At the present time, these problems seem sufficiently remote from the capabilities of present perceptrons that we need not consider them further here. Also relegated to the future is the consideration of such psychological phenomena as perceptual illusions, figural aftereffects, and related phenomena, even though these have been considered primary in some of the brain models hitherto advanced. It is this writer's belief that these phenomena are so likely to depend on inessential details of brain organization, at almost any level of complexity, that it would be a mistake to try to rest the case for or against a particular model on a demonstration that it can duplicate a

particular kinds of perceptual illusion. It seems more important, at this stage, to account for "veridical perception" than for its occasional failures, particularly since these are currently demonstrable in a single species only, and may lack any generality whatsoever.

3.3.10 Application of Experimental Designs to Perceptrons

The designs considered above have been discussed as if they were actual "flesh and blood" experiments, performed with real physical systems. In the study of perceptrons, it is not always practical or necessary to carry out such experiments in reality; the important thing is that an analysis of a given model should always be carried out in terms of an experimental design which is specified in sufficient detail so that it could be carried out if the system were actually constructed.

In practise, three main methods are employed in the study of perceptrons:

(1) Mathematical analysis, in which a stimulus environment, the rules for stimulus presentation and for the modification of the perceptron's memory state are clearly specified. The object of such analysis is, in general, to determine the probability of correct performance, or the probability of achieving a given performance criterion, for a specified class of systems.

(2) Digital simulation, in which the perceptron, its environment, and the memory modification rules are all represented in a digital computer program, which carries out the required operations of an experiment in

step-by-step fashion, calculating the response of every neuron and connection in the perceptron, and measures the performance of the system. Such a program, repeated for a sufficient sample of perceptrons in a class, yields much the same type of information as is obtained from a mathematical analysis. It has the advantage of being free from all approximations (which may be necessary in some analyses) but is less likely to yield important insights into the lawful relations which characterize a class of systems. Simulation programs are most valuable as an exploratory device, and for the study of systems of such complexity that an exact mathematical analysis is impossible.

(3) Study of physical models, involving the actual construction of a hardware device, and the performance of the indicated experiments. At present, little is to be gained from the study of actual physical models which cannot be learned from the other two methods, but as successive models grow in size and complexity, and as means are found for the inexpensive construction of electronic models, this method becomes increasingly important. Its main virtue is the flexibility and adaptability of a hardware perceptron to new types of learning experiments and procedures, and the ability to use ordinary physical objects and environments as stimuli, which would otherwise involve a great deal of time and expense in computer programming. The physical model itself, however, is apt to be less flexible than a simulated system, and is best suited for "case studies" of a single representative system, rather than statistical studies of a class of systems.

In most of the experiments considered in this report, (which are listed in Appendix D) human performance capabilities are sufficiently well known to permit us to draw conclusions about possible comparisons

between perceptrons and biological systems without further study. In some of the proposed experiments, however, (e.g., the figure organization experiments described in 3.3.3) additional data may be required on human performance in order to obtain a base-line for the quantitative evaluation of perceptrons. Thus it seems likely that in the near future, a program in experimental psychology with human and animal subjects may be a necessary adjunct to the evaluation of our brain models. When this occurs, the models are, in effect, being used as predictive devices, capable of generating data (probably grossly inaccurate at the outset) which have not yet been actually observed in human subjects. The ultimate test for a brain model, from the standpoint of psychological validity, is an experiment of this type, in which the model correctly predicts phenomena which have yet to be discovered in biological systems.

4. BASIC DEFINITIONS AND CONCEPTS

This chapter is devoted to basic definitions of terms which will be used throughout the report. It is recommended that the reader familiarize himself with this terminology in a general way, on first reading, and refer back to this chapter when the terms are reintroduced in the subsequent text. A list of standard symbols will also be found in Appendix A.

4.1 Signals and Signal Transmission Networks

The following definitions, which are not specific to perceptrons, are likely to be helpful:

DEFINITION 1: A signal may be any measurable variable, such as a voltage, current, light intensity, or chemical concentration. A signal is typically characterized by its amplitude, time, and location.

DEFINITION 2: A signal generating unit is any physical element, or device, capable of emitting a signal. The output signal of the unit a_i will be represented by the symbol a_i^*.

DEFINITION 3: A signal generating function is any function which defines the amplitude of the signal emitted by a signal generating unit.

DEFINITION 4: A connection is any channel (e.g., a wire or nerve fiber) by which a signal emitted by one signal generating unit (the origin) may be transmitted to another (the terminus). A connection c_{ij} is characterized by its origin and terminal units (u_i and u_j , respectively), and by a transmission function* which determines the amplitude of the signal induced at the terminus as a function of the amplitude and time of the signal generated by the origin unit. This signal will be symbolized by $c_{ij}^{*}(t)$.

DEFINITION 5: A signal transmission network is a system of signal generating units, linked by connections.

4.2 Elementary Units, Signals, and States in a Perceptron

A perceptron (which will be defined in the next section) is a signal transmission network containing three types of signal generating units: sensory units, association units, and response units. These units all have signal generating functions which depend on signals originating elsewhere in the network, or else externally, in an outside environment. The signals upon which the generating function of a unit depends are called

---

* In previous reports, the term "transfer function" has been used for this characteristic. Since "transfer function" has a somewhat different meaning in control system theory and elsewhere, it is avoided here, and the term "transmission function" is preferred.

the input signals to that unit. These units are defined here in a sufficiently general manner as to include biological neurons as a special case. We shall be chiefly concerned, however, with models which employ simplified versions of such neurons.

DEFINITION 6: A sensory unit (S-unit) is any transducer responding to physical energy (e.g., light, sound, pressure, heat, radio signals, etc.) by emitting a signal which is some function of the input energy. The input signal at time t to an S-unit $\mathcal{A}_i$ from the environment, W, is symbolized $\mathcal{C}^*_{Wi}(t)$. The signal which is generated by $\mathcal{A}_i$ at time t is symbolized $\mathcal{A}^*_i(t)$.

DEFINITION 7: A simple S-unit is an S-unit which generates an output signal $\mathcal{A}^*_i = +1$ if its input signal, $\mathcal{C}^*_{Wi}$ exceeds a given threshold, θ_i , and 0 otherwise.

DEFINITION 8: An association unit (A-unit) is a signal generating unit (typically a logical decision element) having input and output connections. An A-unit a_j responds to the sequence of previous signals $\mathcal{C}^*_{ij}$ received by way of input connections $\mathcal{C}_{ij}$, by emitting a signal $a^*_j(t)$.

DEFINITION 9: A simple A-unit is a logical decision element, which generates an output signal if the algebraic sum of its input signals, α_i , is equal or greater than a threshold quantity, $\theta > 0$. The output signal a^*_i is equal to $+1$ if $\alpha_i \geq \theta$ and 0 otherwise. If $a^*_i = +1$, the unit is said to be active.

DEFINITION 10: A response unit (R-unit) is a signal generating unit having input connections, and emitting a signal which is transmitted outside the network (i.e., to the environment, or external system). The emitted signal from unit r_i will be symbolized by r_i^*.

DEFINITION 11: A simple R-unit is an R-unit which emits the output $r^* = +1$ if the sum of its input signals is strictly positive, and $r^* = -1$ if the sum of its input signals is strictly negative. If the sum of the inputs is zero, the output can be considered to be equal to zero or indeterminate. (A physical unit which oscillates in response to a zero signal would have the required properties.)

DEFINITION 12: Transmission functions of connections in a perceptron depend on two parameters: the transmission time of the connection, τ_{ij}, and the coupling coefficient or value of the connection, v_{ij}. The transmission function of a connection c_{ij} from u_i to u_j is of the form: $c_{ij}^*(t) = f\left[v_{ij}(t), u_i^*(t-\tau_{ij})\right]$. Values may be fixed or variable (depending on time). In the latter case, the value is a memory function.

DEFINITION 13: The activity state of the network at time t is defined by the set of signals, u_i^*, emitted by all signal generating units at time t.

DEFINITION 14: The memory state of a network is the configuration of values associated with all (variable valued) connections at a specified time.

DEFINITION 15: The phase space of a network is the space of all possible memory states, for a given network. In general, if there are N variable-valued connections in the network, the phase space may be represented by a region in Euclidean N-space, each coordinate corresponding to the value of one connection. The memory state of the system at any specified time can be characterized by a point in this phase space, and the history of the system by a directed line, or path, followed by this point.

DEFINITION 16: The interaction matrix for a network of S, A, and R units is the matrix of coupling coefficients, $v_{i,j}$, for all pairs of units, u_i and u_j . If there is no connection from u_i to u_j , $v_{i,j}$ is defined as zero. Specifying an interaction matrix is equivalent to specifying a point in the phase space.

4.3 Definition and Classification of Perceptrons

DEFINITION 17: A perceptron is a network of S, A, and R units with a variable interaction matrix V which depends on the sequence of past activity states of the network.

DEFINITION 18: The logical distance from unit u_i to u_j is equal to the number of connections in the shortest path by which a signal can be transmitted from u_i to u_j.

DEFINITION 19: A series-coupled perceptron is a system in which all connections originating from units at logical distance d from the closest S -unit terminate on units at logical distance $d+1$ from the closest S -unit.

DEFINITION 20: A cross-coupled perceptron is a system in which some connections join units of the same type (S , A or R) which are at the same logical distance from S -units, all other connections being of the series-coupled type.

DEFINITION 21: A back-coupled perceptron is a system in which at least one A or R unit at a distance d_1 from the closest S -unit is the origin of a connection back to an S -unit or to an A -unit at a distance $d_2 < d_1$ from the closest S -unit; i.e., this is a system with feedback paths from units located near the output end of the system to units closer to the sensory end.

It should be noted that the above definitions are not exhaustive; they are intended to designate certain generic classes of perceptrons with which we shall be concerned. The initial models to be considered are of the type specified by the following definitions:

DEFINITION 22: A simple perceptron is any perceptron satisfying the following five conditions:

1. There is only one R -unit, with a connection from every A -unit.
2. The perceptron is series-coupled, with connections only from S -units to A -units, and from A -units to the R -unit.
3. The values of all sensory to A -unit connections are fixed (do not change with time).
4. The transmission time of every connection is either zero or equal to a fixed constant, τ .
5. All signal generating functions of S , A , and R units are of the form $u_i^*(t) = f(\alpha_i(t))$, where $\alpha_i(t)$ is the algebraic sum of all input signals arriving simultaneously at the unit u_i .

DEFINITION 23: An elementary perceptron is a simple perceptron with simple R- and A - units, and with transmission functions of the form $\rho_{ij}^*(t) = u_i^*(t-\tau)w_{ij}(t)$.

Perceptrons can be represented graphically in several different ways. In particular, frequent use is made of three types of diagrams, which will be called network diagrams, set diagrams, and symbolic diagrams. Depending upon the level of specificity required, any one of these diagrams may be used to represent the same system. The three types of diagrams

are illustrated in Figure 2. The network diagram shows each connection and signal unit individually; the arrows indicate the direction of signal transmission through the connections. The set diagram represents all S-units as a single set, connected to the set of A-units (or association system) which is represented by a Venn diagram, the subsets of which are connected to different R-units. Set diagrams of this general type are found to be particularly useful as an aid to analysis. The symbolic diagram for this same perceptron merely indicates the kinds of connections which exist, namely, S to A, A to R, and S to S. The perceptron illustrated would be called a three-layer perceptron, cross-coupled at the sensory layer.

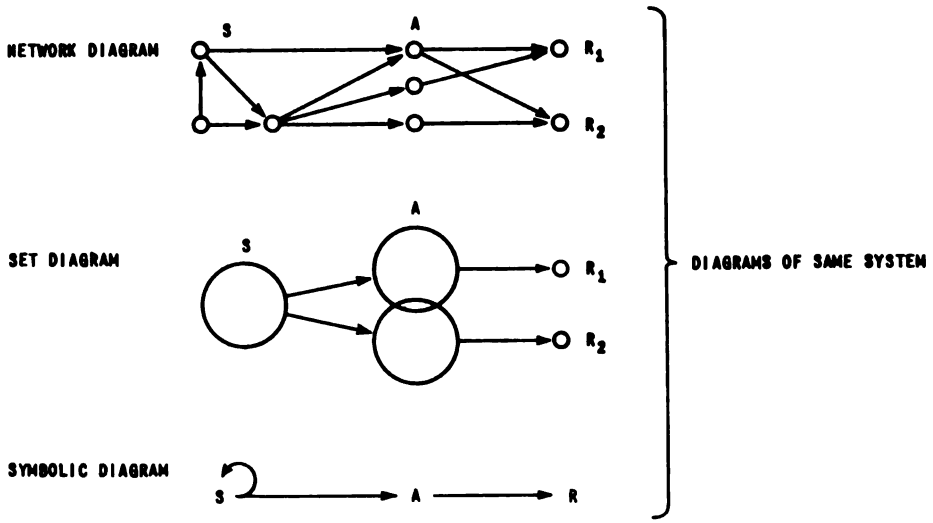


Figure 2 PERCEPTRON DIAGRAMS

4.4 Stimuli and Environments

DEFINITION 24: A stimulus is any non-zero set of input signals, $c^{*}_{wi}(t)$, to the S-units at time t. If there are N_A sensory units in the retina, then a stimulus can be characterized by a vector of N_A elements, representing the signal to each S-unit as an element of the vector. The condition in which all input signals are equal to zero is not considered a stimulus unless otherwise specified.

DEFINITION 25: A stimulus world (or environment) is any set of stimuli, defined for a specified S-unit set. The stimulus world will be symbolized by W. The number of different stimuli will usually be denoted by n.

DEFINITION 26: A stimulus-sequence world (or stimulus-sequence environment) is any set of stimulus sequences, each consisting of an ordered series of stimuli from the set W. (For example, if the image of a printed word is a stimulus, and W consists of all words in a dictionary, then the set of all English sentences would comprise a stimulus-sequence world.)

4.5 Response Functions and Solutions

DEFINITION 27: A response function is any assignment of R-unit output signals to stimuli in W. For a simple perceptron, the response function $R(W)$ is a vector of n elements,

($R_1, R_2, \cdots, R_n$) indicating the value of the response for each of the stimuli, $S_1, S_2, \cdots, S_n$ in the environment.

DEFINITION 28: A classification is an equivalence class of response functions. Two response functions are considered equivalent if their corresponding elements agree in sign. For any perceptron with one simple R-unit, a classification, $C(W)$, divides W into two classes: a positive class consisting of all stimuli for which $r^* = +1$, and a negative class, consisting of those stimuli for which $r^* = -1$.

DEFINITION 29: A response-sequence function is an assignment of sequences of R-unit output signals to stimulus sequences in a stimulus-sequence world. This is a generalization of the concept of a response function to include a time dimension.

DEFINITION 30: A solution to a response function (or classification) is said to exist for a given perceptron if there is a point in the phase space of the perceptron such that the response R_i (specified by the function) will occur if the stimulus S_i is shown, for all S_i in W.

4.6 Reinforcement Systems

DEFINITION 31: A reinforcement system is any set of rules by which the interaction matrix (or memory state) of a perceptron may be altered through time.

DEFINITION 32: A reinforcement control system is any system or mechanism external to a perceptron which is capable of altering the interaction matrix of the perceptron in accordance with the rules of a specified reinforcement system.

DEFINITION 33: Positive reinforcement is a reinforcement process in which a connection from an active unit u_i which terminates on a unit u_j has its value changed by a quantity $\Delta v_{ij}(t)$ (or at a rate dv_{ij}/dt) which agrees in sign with the signal $u_j^*(t)$.

DEFINITION 34: Negative reinforcement is a reinforcement process in which a connection from an active unit u_i which terminates on a unit u_j has its value changed by a quantity $\Delta v_{ij}(t)$ (or at a rate dv_{ij}/dt) which is opposite in sign from $u_j^*(t)$.

DEFINITION 35: A monopolar reinforcement system is a reinforcement system in which the values of all connections terminating on a unit u_j remain unchanged at time t unless $u_j^*(t)$ is strictly positive.

DEFINITION 36: A bipolar reinforcement system is a reinforcement system in which the values of connections are subject to change regardless of whether the output of the terminal unit is positive or negative.

DEFINITION 37: Alpha system reinforcement is a reinforcement system in which all active connections c_{ij} which terminate on some unit u_j (i.e., connections for which $u_i^*(t-\tau) \neq 0$) are changed by an equal quantity $\Delta v_{ij}(t) = \eta$ or at a constant rate while reinforcement is applied, and inactive connections $(u_j^*(t-\tau) = 0)$ are unchanged at time t. A perceptron in which α-system reinforcement is employed will be called an α-perceptron. The reinforcement will be called quantized if the change is a fixed quantity $(|\Delta v| = |\eta|)$ or non-quantized if the value may change by an arbitrary magnitude.

DEFINITION 38: Gamma system reinforcement is a rule for changing the values of the input connections to some unit, whereby all active connections are first changed by an equal quantity, and the total quantity added to the values of the active connections is then subtracted from the entire set of input connections, being divided equally among them. Such a system is said to be conservative in the values, since the total of all values can neither increase nor decrease. The change in v_{ij} is equal to

$$\Delta v_{ij}(t) = \left(\omega_{ij}(t) - \frac{\sum_i \omega_{ij}(t)}{N_j}\right)\eta$$

where $\omega_{ij}(t) = 1$ if $u_i^*(t-\tau) \neq 0$, 0 otherwise;

N_j = number of connections terminating on u_j

η = reinforcement quantity (typically ± 1 or 0).

Additional reinforcement rules, and variations of the above, will be presented as required. The above terminology has been standardized in previous work on perceptrons, and represents the systems on which most analysis has been done. In most of the cases to be considered, the reinforcement control system employs one of three training procedures, defined as follows:

DEFINITION 39: A response-controlled reinforcement system (R -controlled system) is a training procedure in which the magnitude of η is constant, and the sign of η is entirely determined by the current response, r^*, regardless of the current stimulus, S. In general, unless otherwise specified, this term implies that the reinforcement is always positive (i.e., the sign of η agrees with the sign of r^*, in a simple perceptron).

DEFINITION 40: A stimulus-controlled reinforcement system (S -controlled system) is a training procedure in which the magnitude of η is constant, and the sign of η is determined entirely by the current stimulus, S, and a predetermined classification, $C(W)$; the current response of the perceptron does not influence either the sign or magnitude of η.

DEFINITION 41: An error-corrective reinforcement system (error correction system) is a training procedure in which the magnitude of η is 0 unless the current response

of the perceptron is wrong, in which case, the sign of η is determined by the sign of the error. In this system, reinforcement is 0 for a correct response, and negative (see Definition 34) for an incorrect response, or, more generally, $\eta = f(R^* - r^*)$ where R^* is the required response, r^* is the obtained response, and f is a sign-preserving monotonic function, such that $f(0) = 0$.

In previous reports (Refs. 41, 82) the R-controlled system has been referred to as a "spontaneous learning system", since the perceptron evolves in an autonomous fashion, uninfluenced by the "correctness" of its outputs. The reinforcement control system requires no information from the environment in order to control the changes in the memory state of the perceptron. The S-controlled system has also been referred to as a "forced learning system", since the r.c.s. imposes a predetermined classification on the perceptron's responses, without taking the actual responses of the system into account at any time.

4.7 Experimental Systems

DEFINITION 42: An experimental system is a system consisting of a perceptron, a stimulus world, W, and a reinforcement control system. The reinforcement control system may be an automatic regulating device (e.g., a thermostat) or a human operator, capable of responding to the responses of the perceptron and the stimuli in the environment by applying the appropriate reinforcement rules, altering the memory state of the perceptron.

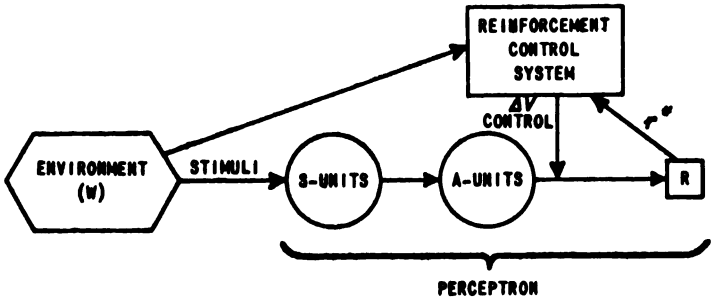


Figure 3 EXPERIMENTAL SYSTEM WITH A SIMPLE PERCEPTRON

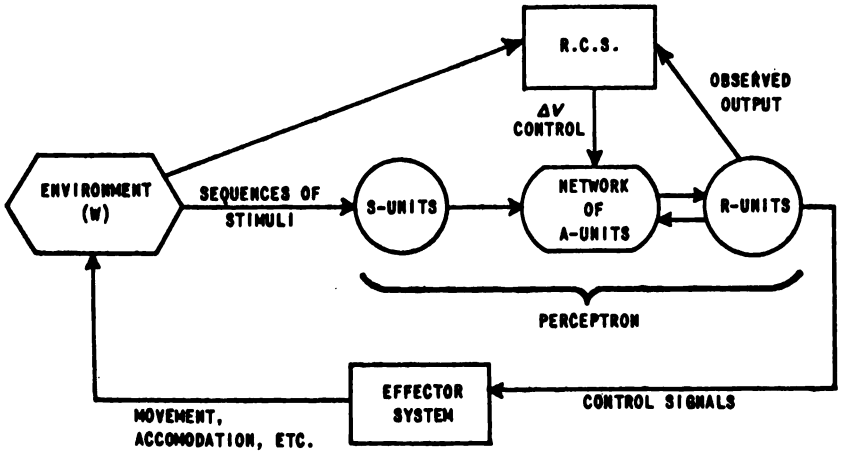


Figure 4 GENERAL EXPERIMENTAL SYSTEM

The basic organization of an experimental system with a simple perceptron is shown in Figure 3. A more general system, in which the perceptron may be of any variety, and where the output of the perceptron is capable of modifying its stimulus environment, is illustrated in Figure 4. A comparison with Figure 1 should indicate the basic similarity between the perceptron, in a general experimental system, and the biological nervous system. Analyses of perceptron performance always postulate an experimental system, involving, as a minimum, the components shown in Figure 3. The reinforcement control system can be considered a specialized part of the environment, in its relation to the perceptron, although it might actually be built into the same physical mechanism as the perceptron itself. In an R - controlled system, the information channel shown from *W* to the r.c.s. is non-functional, while in an S -controlled system the information channel from *W* to the r.c.s. is non-functional, and in an error-correction system, both channels are essential for reinforcement control. In digital simulation programs, the r.c.s. is the part of the program concerned with reinforcing the simulated perceptron, while in experiments with hardware systems it is generally a human operator.

An experiment involves an experimental system, a training procedure, and a procedure for testing the perceptron, or measuring its performance. A number of typical psychological experiments, which are of interest for perceptrons, were outlined in Chapter 3, and some of these will be analyzed in the following chapters.

PART II

THREE-LAYER SERIES-COUPLED PERCEPTRONS

5. THE EXISTENCE AND ATTAINABILITY OF SOLUTIONS IN ELEMENTARY PERCEPTRONS

The perceptrons to be considered in Part II all consist of three layers of units connected in series, with the topology S → A → R. In the following chapters, it will be seen that these perceptrons are capable of learning any set of responses which we might care to have them make to a universe of stimuli. Their main deficiencies are a lack of ability to generalize their performance to new stimuli or new situations where they have not been explicitly taught and a lack of ability to analyze complex environmental situations into simpler parts.

The first perceptron model to be considered in detail is the elementary α-perceptron. In this chapter, we shall examine the intrinsic ability of such systems to realize solutions to classification problems, including several theorems concerning the relationship of the size of the system to the existence of solutions, and the possibility of attaining such solutions by different training procedures. The term "solution" is used in the sense of Def. 30, in Chapter 4. Most of these results were first presented in Ref. 86.

5.1 Description of Elementary α-Perceptrons

Elementary α-perceptrons were defined in Chapter 4, as a subclass of simple perceptrons, in which S-units send connections to A-units, and the A-units all send connections to a single R-unit, no other connections being permitted, and all connections having equal trans-

mission times, τ . Without loss of generality, τ can be taken to be zero, and this assumption of instantaneous transmission will be made whenever we deal with simple perceptrons, unless otherwise stated. The A-units and R -unit in all elementary perceptrons are of the simple type, i.e., they have a threshold, θ , (equal to zero in the case of the R -unit) and emit a signal only if the input signal, α , is equal or greater than θ The connections from S to A -units have fixed values, and the connections from the A-units to the R -unit have variable values, which depend on the history of reinforcements applied to the perceptron. The connections, in an elementary perceptron, all have the transfer function (assuming τ to be zero).

$$c^*_{i,j}(t) = u^*_i(t)\, v_{i,j}(t)$$

In the α -system, which is to be considered initially, the reinforcement rule takes the form

$$\Delta v_{i,j}(t) = u^*_i(t) \cdot \eta = \begin{cases} \eta & \text{if } \alpha_i(t) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

In an elementary perceptron, where the only variable connections occur from A -units to the R -unit, the simplified notation v_i will generally be taken to mean the value of the connection from unit a_i to the R -unit. The basic parameters with which we shall be concerned in this chapter are the number of S -units, N_1 , and the number of A -units, N_2 . Without loss of generality, we can assume the N_1 sensory units to be situated at points in a two-dimensional field, or "retina", and regard the input stimuli as patterns of illumination on the retina. A typical system of this type is illustrated in Figure 5.

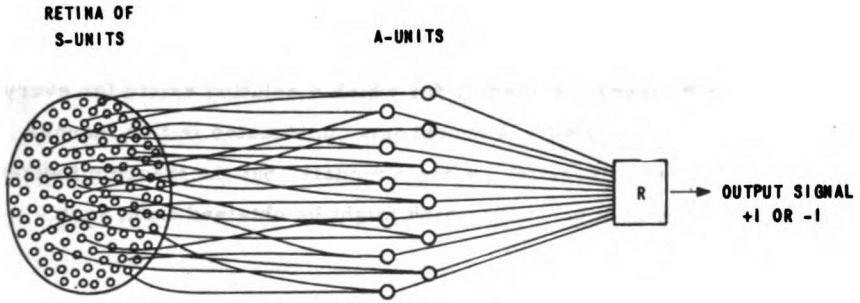


Figure 5 NETWORK ORGANIZATION OF A TYPICAL ELEMENTARY PERCEPTRON

5.2 The Existence of Universal Perceptrons

Most of the theoretical results obtained to date for elementary perceptrons are concerned with experiments in which a classification of an environment, $C(W)$, is taught to the perceptron by some training proce-

dure. The first theorems to be considered deal with the question of whether a solution to such a classification problem exists, or might exist, for a given perceptron. To begin with, the following theorem shows that the organization of an elementary perceptron is sufficient to permit the construction of a "universal system", for which a solution exists for every possible classification, $C(W)$. Perceptrons constructed in this manner are generally not very interesting as brain models, but the theorem indicates the wide range of possible behavior which might be obtained from such systems.

THEOREM 1: Given a retina with two-state (on or off) input signals, the class of elementary perceptrons for which a solution exists to every classification, $C(W)$, of possible environments W, is non-empty.

PROOF: Since it is sufficient to show the existence of such a perceptron, we proceed by construction. Let there be one A-unit for every possible stimulus configuration on the retina. Consider stimulus S_i and its corresponding A-unit, a_i. Let a_i have an excitatory connection (value equal to +1) originating from every "on" point in S_i, and an inhibitory connection from every "off" point in S_i, and let its threshold be equal to the number of excitatory connections. Then there will be one and only one A-unit responding to every possible stimulus, and no A-unit responds to more than one stimulus. (We say that a_i "responds" to S_i if $a_i \neq 0$.) Now consider any stimulus world, W, defined on the retina, and a corresponding classification, $C(W)$, which associates a positive or negative classification with each stimulus, S_i, in W.

In order to realize the classification, it is only necessary to set the value of the connection from a_i equal to + 1 if the class of S_i is positive, or - 1 if the class of S_i is negative. Q.E.D.

While this solution is clearly uneconomical and of little practical interest, it is sufficient to show that there are no "special cases" of classifications which have no solution, at least for a retina of binary elements. If the inputs to the S-units are capable of taking on more than two values, then a more elaborate construction (e.g., one which separates each combination of input values to a different set of A-units) would be required. It is left to the reader to satisfy himself that a system with less "depth" than an elementary perceptron (i.e., one in which S-units are connected directly to the R-unit, with no intervening A-units) is incapable of representing a solution to every $C(W)$, no matter how the values of the connections are distributed.

5.3 The G-matrix of an Elementary α-Perceptron

In practice, the cases of interest are those in which each stimulus activates some set of A-units, and each A-unit is likely to respond to a great many different stimuli in W. In order to deal with such systems, the concept of a G-matrix has been found to be particularly helpful, and this will now be defined. The definition given here is sufficient for elementary perceptrons, and will be generalized in a later chapter to permit us to deal with more complex systems.

DEFINITION: Consider a (simple) perceptron, and a stimulus world, W, consisting of n stimuli. Then the matrix

$$G = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix}$$

consists of elements g_{ij} called <u>generalization coefficients</u>. Each element, g_{ij}, is equal to the total change in value ($\sum \Delta v_k$) over all A-units in the set responding to S_i if the set of units responding to S_j are each reinforced with, η equal to $1/N_a$ (where N_a is equal to the number of A-units in the system).* For simple perceptrons and a given environment, G is fixed for all time.

If we are dealing with a particular α-perceptron, where $\Delta v_k = a_k^*(t) \cdot \eta$, we have

$$g_{ij} = \tilde{Q}_{ij}$$

where $\tilde{Q}_{ij}$ = the proportion of A-units which respond both to S_i and S_j.

If we are dealing with a randomly selected member of a class of perceptrons, $\tilde{Q}_{ij}$ is a random variable, and we have the equation for the expected value of g_{ij},

$$E\, g_{ij} = Q_{ij}$$

where Q_{ij} = the probability that an A-unit in a given class of perceptrons responds to both stimuli, S_i and S_j.

---

* With $\eta = 1/N_a$ we have a "normalized G-matrix". For some purposes it is convenient to take $\eta = 1$, in which case the "unormalized G-matrix" is equal to N_a times the normalized matrix defined above.

For the α-system, g_{ij} is simply a measure of the intersection of the sets of A-units responding to S_i and to S_j, and is equivalent to a "set intersection matrix". G is always symmetric for an alpha system. In any elementary perceptron (at a given time t) the net input signal to the R-unit from the set of A-units responding to stimulus S_i will be called u_i and is given by

$$u_i = \alpha_r(S_i) = g_{i1} x_1 + g_{i2} x_2 + \ldots + g_{in} x_n \qquad (5.1)$$

where x_j = the amount of reinforcement applied to the system, over all appearances of S_j prior to time t.* In matrix form, the vector u of signals u_i from all stimuli S_i in W is given by

$$u = Gx \qquad (5.2)$$

where x is a vector of elements x_j, defined as above.

5.4 Conditions for the Existence of Solutions

In general, if we are given the rules of organization of a perceptron and some classification, $C(W)$, it is by no means easy to say whether or not a solution to $C(W)$ exists for the perceptron in question. The following theorems deal with the existence of such solutions from several different points of view. We first define the bias ratio of an A-unit as follows:

DEFINITION: Given a classification, $C(W)$, the bias ratio of an A-unit, a_i, is defined for any set of stimuli in W as n_i^+/n_i^-, where n_i^+ = number of stimuli in the set which are members of the positive class C^+ and which activate a_i; n_i^- = number of stimuli in the set which are members of the negative class C^- and which activate a_i.

* It is assumed here that all initial $v_i = 0$.

THEOREM 2: Given an elementary perceptron and a classification $C(W)$, the following conditions are necessary but not sufficient for a solution to $C(W)$ to exist:

i) Every stimulus must activate at least one A-unit;

ii) There should be no subset of stimuli containing at least one member of each class, such that in the union of the responding A-unit sets, every A-unit has the same bias ratio (with respect to the stimuli of the subset).

PROOF: We first prove that the conditions are necessary. Condition i) is obvious. The proof that condition ii) is necessary is as follows:

Assume there is a subset violating this condition. Let a_j = input signal to R generated by stimulus S_j. Then summing the values of all such signals from stimuli of the positive class in this subset, we have (since violation of ii) requires that n_i^+/n_i^- is constant for A-units responding to stimuli in this subset).

$$\sum_{S_j \in C^+} a_j = \sum_i n_i^+ v_i = \frac{n_i^+}{n_i^-} \sum_i n_i^- v_i = \frac{n_i^+}{n_i^-} \sum_{S_j \in C^-} a_j$$

Thus the sum of the R-unit input signals for stimuli of the positive class must have the same sign as the sum of the R-unit input signals for stimuli of the second class. But then one of the sums must disagree in sign with the sign of the class, and therefore, one of its components (i.e., one of the a_j) must disagree in sign with the class, indicating that at least one stimulus must be classified incorrectly.

To show that these conditions are not generally sufficient, consider the following example: Let there be five stimuli, and four A -units. The A -units activated by each stimulus are:

S_1 activates a_1
S_2 activates a_2
S_3 activates a_3 and a_4
S_4 activates a_1, a_2, and a_3
S_5 activates a_1, a_2, and a_4

Let the positive class consist of S_1, S_2, and S_3, and the negative class consist of S_4 and S_5. Then the bias ratios for a_1 and a_2 are not the same as for a_3 and a_4. Also, there exists no subset with stimuli from each class, with equal bias ratios for all A -units. The values of a_1 and a_2 must be positive, and the sum of the values of a_3 and a_4 must also be positive, to obtain the required the required classification for the members of the first class. But then it is clear that either S_4 or S_5 must be classified incorrectly, which proves that conditions i) and ii) are not sufficient.*

In the next theorem we make use of the symbol u to denote a signal vector, such that the element u_i agrees in sign with the classficiation of S_i in $C(W)$. Such a signal vector will evoke the correct response for each stimulus in W. Two such vectors which agree in the signs of their elements are said to be in the same orthant (generalized quadrant, in n dimensions).

---

* In Theorem 9, a necessary and sufficient condition, closely related to the above, will be presented.

THEOREM 3: Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$; then in order for a solution to $C(W)$ to exist, it is necessary and sufficient that there exist some vector u in the same orthant as $C(W)$, and some vector v such that $Gv = u$.

PROOF: The proof would follow trivially from Equation (5.2) and the definition of u, were it not for the possibility that a solution might exist involving some unique assignment of values to the A-R connections, which could not be attained by any reinforcement vector, v, defined as in Equation (5.1). It will be shown, therefore, that if a solution exists, in the form of <u>any</u> assignment of values to A-R connections, an <u>equivalent solution</u> must exist corresponding to the reinforcement of each stimulus, S_i, by an amount v_i. For brevity, throughout the following discussion, we will speak of "the value of an A-unit" in place of "the value of the connection from an A-unit to the R-unit". The following definitions and notation will be used:

$$a_j^*(S_i) = \begin{cases} 1 \text{ if the A-unit } a_j \text{ responds to } S_i \\ 0 \text{ otherwise} \end{cases}$$

A is an n by N_a matrix, in which the element $a_{ij} = a_j^*(S_i)$. A solution to a classification problem is said to exist if there is some distribution of values over the A-units which enables the perceptron to perform the discrimination; i.e., there exist vectors v and u such that

$$Av = u$$

Consider the matrix AA'. The i,j^{th} element of this matrix (say A_{ij}) is

$$\sum_{\lambda} a^*_{\lambda}(s_i)\, a^*_{\lambda}(s_j) = A_{ij}$$

But the (un-normalized) G-matrix for an α-system, expressed in terms of the above functions, has elements,

$$g_{ij} = \sum_{\lambda} a^*_{\lambda}(s_i)\, a^*_{\lambda}(s_j)$$

so that the matrix $G = AA'$. Note that this shows that G is either positive definite or positive semidefinite.

We then have, for any vector x, such that $x'A = 0$

1) $x'A = 0 \Rightarrow x'AA' = x'G = 0$

2) $x'G = 0 \Rightarrow x'Gx = x'AA'x = (x'A, x'A) = 0 \Rightarrow x'A = 0$

Hence, the rank of G = rank of A, since any vector x which is in the left null space of G is also in the left null space of A; therefore the left null spaces of G and A are identical. Since the rank plus the dimension of the null space is equal to the dimension of the domain, G and A must be of the same rank.

But the columns of G are linear combinations of the columns of A, hence the space spanned by the columns of G is identical with the space spanned by the columns of A.

Since Av is a linear combination of the columns of A, the existence of v and u such that $Av = u$ implies the existence of a vector x such that $Gx = u$. Thus, if a solution exists, there is a solution to the equation $Gx = u$, so that the condition of the theorem is necessary. But it is also sufficient, since u by definition represents a solution vector. Q.E.D.

COROLLARY 1: Given an elementary perceptron and a stimulus world W, Then if G is singular, some $C(W)$ exists for which there is no solution.

PROOF: Each $C(W)$ requires a solution vector in a different orthant, and the set of all $C(W)$, for a given W, requires solutions in every possible orthant. But if G is singular, it maps the entire space into a hyperplane, and this plane must fail to intersect certain orthants. Consequently, the classifications $C(W)$ which are represented by vectors in these orthants have no solution.

COROLLARY 2: Given an elementary perceptron, if the number of stimuli in W is $n > N_a$, there is some $C(W)$ for which no solution exists.

PROOF: From Theorem 3 and Corollary 1, it is clear that there will be some $C(W)$ which has no solution if and only if G is singular. G has the same rank as the matrix A; but A is an n by N_a matrix, implying that A, and therefore G has rank $< n$.

COROLLARY 3: For any elementary perceptron, as the number n of stimuli in W increases, the probability that a randomly selected classification, $C(W)$, has a solution approaches zero (where $C(W)$ is chosen from a uniform distribution over the possible classifications of W).

PROOF: From Corollary 2, as n increases beyond the number of A-units in the perceptron, there must be some $C(W)$ without a solution. At the same time, increasing n increases the set of possible classifications in proportion to 2^n. But, owing to a theorem by R. D. Joseph and Louise Hay (Ref. 41, Appendix), the number $n(r)$ of classifications which have solutions is no greater than $2\left[\binom{n-1}{0}+\binom{n-1}{1}+\ldots+\binom{n-1}{r-1}\right]$ where $r \leq N_a$ is the rank of the G-matrix. Therefore, the upper bound of the probability of selecting at random one of the classifications which has a solution diminishes with $n(r)/2^n$ which goes to zero as n goes to infinity.

Several additional tests for the existence of solutions, which are of practical utility in diagnosing small systems, will be found in Theorems 9 and 10, at the end of this chapter.

5.5 The Principal Convergence Theorem

In the preceding section, the existence of solutions to classification problems in an elementary perceptron was considered, but nothing has been said about the ability to achieve such a solution by a training procedure. In this section, we consider the ability of an elementary α-perceptron to learn the solution to a classification $C(W)$ under an error correction procedure. The following theorem is fundamental to the theory of perceptrons.

A general definition of an error correction procedure was given in Definition 41, in Chapter 4. We now define in detail two specific forms of this procedure, as they apply to the elementary α-perceptron.

Consider some classification, $C(W)$. Let

$$\rho_i = \begin{cases} +1 & \text{if stimulus } S_i \text{ is to be in the positive class} \\ -1 & \text{if stimulus } S_i \text{ is to be in the negative class} \end{cases}$$

where $i = 1, \ldots, n$.

In order to obtain the most general conditions for the following theorem, a non-quantized error correction procedure is defined as follows: No response will be considered correct unless the magnitude of the input signal to the R-unit (u_i) is greater than θ, and the sign of u_i agrees with ρ_i for the current stimulus. (This corresponds to an R-unit with a threshold of θ, or for the special case where $\theta = 0$, it corresponds to a simple R-unit.) If no error occurs for stimulus S_i (i.e., $\rho_i u_i > \theta$) no reinforcement occurs; but if an error does occur a quantity $\eta = \rho_i \Delta x_i$ is added to the value of each active A-unit, Δx_i (the number of units of reinforcement) being just sufficient to bring the magnitude of the signal u_i past the threshold level, θ, to the level $\epsilon > \theta$. In a quantized correction procedure, the identical rules apply, except that $\eta = \rho_i \Delta x_i = \pm 1$, Δx_i representing a single unit of reinforcement.

THEOREM 4: Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists; let all stimuli in W occur in any sequence, provided that each stimulus must reoccur in finite time; then beginning from an arbitrary initial state, an error correction procedure (quantized or non-quantized)will always yield a solution to $C(W)$ in finite time, with all signals to the R-unit having magnitudes at least equal to an arbitrary quantity $\delta \geq 0$.

PROOF:* The matrix A is defined as in Theorem 3, so that $a_{i,j} = a_j^*(S_i)$. We recall that $AA' = G$. We also define the matrix B such that $b_{i,j} = \rho_i a_j^*(S_i)$; the matrix $H = BB'$; and the diagonal matrix D such that $d_{i,j} = \delta_{i,j} \rho_i$. Note that $DD = I$, $DA = B$, and $H = DGD$.

We first consider the non-quantized error correction procedure. In this case, no reinforcement is applied unless an error occurs; if an error does occur (when $\rho_i u_i \leq \delta$) the quantity $\rho_i \Delta x_i$ $(\Delta x_i > 0)$ is added to the value of each active A-unit, Δx_i being chosen so that the input to the response unit is exactly $\rho_i \epsilon$ $(\epsilon > \delta)$. It will be shown below that such a Δx_i exists.

---

* The proof of this theorem (which was first published by Rosenblatt in Ref. 86) has undergone a number of modifications. The original treatment was insufficient to prove the theorem in a rigorous fashion; subsequent forms have been due to Block, Joseph, Kesten, and others; and the present proof owes much to each of these. An interesting alternative approach, with a slightly modified reinforcement procedure, has recently been proposed by Papert (Ref. 67) who attempts to shorten the demonstration and avoids use of the G-matrix. Unfortunately, there are several logical errors in Papert's argument, the correction of which would tend to lengthen his demonstration.

It has been noted previously that the space spanned by the columns of G is the same as the space spanned by the columns of A (the rank of G being equal to the rank of A). Consequently, for any N_a -vector V , there is an n -vector Z such that $AV = GZ$.

An arbitrary initial state for the perceptron is represented by an N_a -vector V^0 of values for the A-units. Let Z^0 be a corresponding n -vector. Let Z be the n -vector whose i^{th} component, z_i , is equal to the total quantity of reinforcement given in all previous corrections for stimulus S_i , i.e.,

$$z_i = \sum \rho_i \Delta x_i \quad \text{(summing over all previous corrections).}$$

Let $U = GZ^0 + GZ = G(Z^0 + Z) = GD(X^0 + X)$ where $X^0 = DZ^0$ and $X = DZ$. The i^{th} component of U , u_i , would be the input to the R-unit if S_i were to occur at the present time. Let $W = DU$. This equation can be written

$$W = H(X^0 + X)$$

where a negative w_i (or more precisely, $w_i \leq \theta$) represents an error. The x_i are always non-negative, and this will be understood for the remainder of the proof. We now define M as the maximum diagonal element, h_{ii} , of H . We also define the function of the n -vector Z

$$K(Z) = Z'HZ - 2\epsilon \sum_{i=1}^{n} z_i$$

We then obtain the following results:

1) The existence of a solution means that there is an N_a vector V^* such that for all i

$$\sum_j a_j^*(s_i)\, v_j^* = \rho_i\, w_i^*$$

where $w_i^* > 0$. In matrix form $BV^* = W^*$.

2) Consider $X'HX$ for all X such that $\|X\| = 1$ (and of course $x_i \geq 0$). $X'HX = (X'B)(X'B)'$ so that $X'HX \geq 0$. Suppose $X'HX = 0$; then $X'B = 0$. Clearly $X'W^* > 0$, but $X'W^* = X'BV^* = 0$. This contradiction shows that $X'HX > 0$ on this closed, bounded set, so that there exists a minimum $\alpha > 0$ such that $X'HX \geq \alpha \|X\|^2$ for all X for which $x_i \geq 0$ for all i. Note that $M \geq \alpha > 0$ as a consequence. Note also that $g_{ii} = h_{ii} \geq \alpha > 0$.

3) $\sum x_i \leq \sqrt{n}\ \|X\|$ (Schwarz's inequality)
and $|X'HX^\circ| \leq \|HX^\circ\| \cdot \|X\| = k\ \|X\|$ (Schwarz's inequality)

4) $K(X^\circ + X) - K(X^\circ) = K(X) + 2X'HX^\circ$

$$\geq \alpha \|X\|^2 - 2\epsilon\sqrt{n}\|X\| - 2k\|X\|$$

$$\geq -\frac{(k + \epsilon\sqrt{n})^2}{\alpha}$$

5) $\dfrac{\partial K(X^\circ + X)}{\partial x_i} = 2w_i - 2\epsilon$

and $\dfrac{\partial w_i}{\partial x_i} = h_{ii} > 0$. This latter relation proves the contention at

the beginning of the proof that $\Delta x_i \geq 0$ exists. Specifically, we have $\Delta x_i = \frac{\epsilon - w_i}{h_{ii}}$.

6) A correction is made for S_i only if $w_i \leq \delta$. Denote the change in K when this is done by ΔK , and by subscript 0 the conditions before the correction.

$$\Delta K(X^0 + X_0) = 2\int_{x_{i0}}^{x_{i0}+\Delta x_i} (w_i - \epsilon)\, dx_i = 2\int_{w_{i0}}^{\epsilon} \frac{1}{h_{ii}} (w_i - \epsilon)\, dw_i$$

$$= \frac{1}{h_{ii}} (w_i - \epsilon)^2 \Big|_{w_{i0}}^{\epsilon}$$

$$= -\frac{(w_{i0} - \epsilon)^2}{h_{ii}}$$

$$\leq -\frac{(\epsilon - \delta)^2}{M}$$

7) From 4) and 6) we conclude that the maximum number of corrections is

$$N \leq \frac{M(k + \epsilon\sqrt{n})^2}{\alpha(\epsilon - \delta)^2}$$

8) In particular, if $X^0 = 0$ and $\delta = 0$ (corresponding to a perceptron with a simple R-unit and no initial reinforcement) then $\ell = \| HX^0 \| = 0$ and the bound becomes nM/α.

This proves the theorem for the case of the non-quantized correction procedure, since N is finite, implying that the process arrives at a solution in finite time. For the quantized case, we have the condition that Δx_i is always 1 when a correction occurs (the vector X representing the numbers of unit corrections for each of the n stimuli). For convenience, we take the case where $\delta = 0$ and $\epsilon = M = (g_{ii})_{max}$. Then in step 6) we have:

6a) $$\Delta K(X^0 + X_0) = 2\int_{x_{i0}}^{x_{i0}+1} (w_i - M)\, dx_i = 2\int_{x_{i0}}^{x_{i0}+1} \left[w_{i0} + h_{ii}(x_i - x_{i0}) - M\right] dx_i$$

$$= 2\left[w_{i0}\, x_i - M x_i + \frac{h_{ii}}{2}(x_i - x_{i0})^2\right]\Bigg|_{x_{i0}}^{x_{i0}+1}$$

$$= 2\left(w_{i0} - M + \frac{h_{ii}}{2}\right)$$

$$\leq -M$$

7a) From 4) and 6a) we have that the maximum number of corrections* is

$$N \leq \frac{(\ell + M\sqrt{n})^2}{\alpha M}$$

---

* An alternative bound, found by H. Kesten, is $\frac{n}{\alpha} \max_i (-2\rho_i w_i^0 + h_{ii})$. This under some circumstances represents a sharper bound; nonetheless, both bounds are generally quite poor, as estimates of the actual number of steps.

8a) This upper bound is again minimized when $X^0 = 0$ so that $k = \|HX^0\| = 0$. The bound is then nM/α .

This completes the proof of the theorem for the quantized case. Q.E.D.

COROLLARY : Given an elementary perceptron, a stimulus world W , and any classification $C(W)$; then if a solution to $C(W)$ exists, the set of possible solutions to $C(W)$ has positive measure over the phase space of the perceptron.

PROOF: From the proof of the theorem, we know that if a solution exists, there is a strictly positive vector X such that $HX = P$ (where P is a strictly positive vector). Let Y be any n -vector; then $\|HY\| \leq b\,\|Y\|$ where b is the absolute value of the maximum eigenvalue of H , or the norm of H . Let $\mu = \min_i p_i > 0$, and let $\epsilon = \mu/(b+1)$. Let U be in the ϵ -sphere around X , i.e., $U = P+Y$ where $\|Y\| \leq \epsilon$. Let $Z = HY$, and let $\xi = \max_i z_i \leq \|Z\| = \|HY\| \leq \frac{b\mu}{b+1} < \mu$. Then

$$p_i + z_i \geq \mu - \xi > 0$$

$$HU = H(U+Y) = P + Z$$

Therefore, HU is strictly positive, and U is an alternative solution.

This means that there is a cone of vectors including X which maps into the region which contains P , any such vector representing an equivalent solution. Since the volume of this cone has positive measure over the phase space, the corollary follows.

5.6 Additional Convergence Theorems

The theorem in the previous section deals with convergence to a solution state in an α-perceptron, trained by the error correction procedure. In this section, it will be shown, first, that a weaker form of correction procedure can also be guaranteed to yield a solution; secondly, that reinforcement procedures in which the magnitude of η does not depend on whether or not the current response is correct cannot, in general, be relied on to converge to a solution. If a solution state does occur in such a system, it will be shown that it is apt to be unstable except under special conditions.

DEFINITION: A random-sign correction procedure is one in which some quantity of reinforcement is applied to the perceptron when an error occurs, and zero reinforcement is applied when the response is correct. The sign of η is chosen at random, with an equal probability of being positive or negative, regardless of the response of the perceptron.

THEOREM 5: Given an elementary α-perceptron, with a finite number of memory states, a random-sequence stimulus world W, and any classification $C(W)$ for which a solution can be reached from the starting point by some reinforcement sequence, then a solution will be obtained in finite time with probability 1 by means of a random-sign correction procedure.

PROOF: The random-sign correction procedure consists of a random walk in which each step corresponds either to a step of the required correction process, or a step in the reverse direction. In the course of this process, the vector α (defined in connection with Theorem 4) will

eventually reach some attainable trapping state with probability 1. But the only trapping states are in the solution space. Consequently, a solution will be obtained in finite time.

In Chapter 4, (Definition 40) an S-controlled reinforcement system was defined as a training procedure in which the magnitude of η is constant, regardless of the current response of the system, the sign of η being chosen to agree with the sign of the classification of the current stimulus, S_i , in $C(W)$. Unlike the methods considered previously in this chapter, this is not a correction procedure; i.e., the magnitude of reinforcement does not depend on the occurrence of an error, and only the sign of the required response is taken into consideration in determining what reinforcement should be applied. In the following analysis, a solution will be called stable if, in a given experimental system, all future memory states will also satisfy the conditions of a solution, no matter how long the experiment continues. A system employing a correction procedure, since it receives no further reinforcement once a solution state is achieved, is inherently stable. The following theorem shows that this is not the case for an S -controlled system.

THEOREM 6: Given an elementary α -perceptron, a stimulus world W , and some classification $C(W)$ for which a solution exists, a solution can sometimes be achieved by an S -controlled reinforcement procedure. However, such a solution cannot be guaranteed for an arbitrary stimulus sequence; and may be unstable if it occurs.

PROOF: We will first consider a case in which a stable solution does occur, for the type of experimental system specified by the theorem. Let W consist of two stimuli, S_1 and S_2 . Let S_1 activate some set of A-units, A_1 ,

and let S_2 activate a disjoint set of A-units, A_2. Let $C(W)$ assign S_1 to the positive class and S_2 to the negative class. Regardless of the sequence and relative frequency of S_1 and S_2, it is clear that each occurrence of S_1 will augment u_1 in a positive direction, while each occurrence of S_2 will make u_2 increasingly negative. Since the intersection A_{12} is assumed to have zero measure, there will be no interference between the two stimuli, so that the acquired solution will remain stable no matter how long the process continues. This example proves the first part of the theorem. Let us now consider the case of intersecting A-unit sets. Suppose S_1 activates two units, a_1 and a_c, while S_2 activates units a_2 and a_c (the unit a_c responding to both stimuli). If the frequencies of S_1 and S_2 are equal, their effect on a_c will tend to cancel, and a solution with v_1 positive, v_2 negative, and v_0 equal to zero will tend to occur. As the sequence continues, the magnitudes of v_1 and v_2 will tend to increase without bound, so that the solution will become increasingly stable as time goes on. Suppose, on the other hand, that S_1 occurs with ten times the frequency of S_2. In this case, a_c will gain ten units of positive value for every unit of negative value received from S_2, so that v_c will tend to increase in a positive direction at nine times the rate that v_2 progresses in a negative direction. Thus the net signal, u_2, transmitted to the R-unit in response to S_2, which is equal to $v_2 + v_c$, will clearly become strongly positive as time goes on, resulting in an erroneous classification of S_2. Even if the initial state of the perceptron was a solution state (e.g., $v_1 = +1$, $v_2 = -1$, $v_c = 0$) it is clear that the S-controlled procedure will quickly destroy the existing solution, which is therefore unstable. Q.E.D.*

---

* H. D. Block has pointed out that, while a solution to $C(W)$ can not be guaranteed with a random stimulus sequence, nonetheless if a solution exists then there exists some S-sequence which <u>will</u> guarantee a solution with S-controlled reinforcement. In particular, if $Gz = u$ is a solution, then the occurrence of S_i with frequency $f_i = |z|$ (for all i) will guarantee a solution.

In the example considered above, it is clear that a frequency bias, in which the stimuli of one class are much more frequent than members of the other class, can strongly prejudice the perceptron to always give the response associated with the more frequent class, in an S-controlled system. Such a problem would exist, for example, in trying to teach a perceptron to distinguish the letters "E" and "X" occuring with their normal frequency in English text. Even if all stimuli occur with equal frequency, however, a similar effect exists if there is a size bias, in which the stimuli in one class activate more S-points (or illuminate a larger area of the retina) than the other class. As will be seen in the following chapter, larger stimuli generally tend to activate more A-units than smaller stimuli, and in the limiting case, the set of A-units responding to a smaller stimulus may be entirely contained within the set responding to a larger stimulus. Suppose for example, that S_1 activates units a_1 and a_2, while S_2 only activates a_2. A solution which classifies S_1 positively and S_2 negatively clearly exists (e.g., let $v_1 = +5$ and $v_2 = -1$) but if the stimuli occur alternately, u_1 will tend to become increasingly positive, while u_2 tends to oscillate about zero. The reader can satisfy himself that (starting with 0 values) a quantized error correction procedure yields a stable solution to this problem after five stimuli.

In the case of R-controlled reinforcement procedures (Definition 39 in Chapter 4) it makes no sense to talk about the probability of convergence to solution for an arbitrary classification, $C(W)$, since the required classification plays no part whatever in determining either the sign or the magnitude of the reinforcement. As will be shown later, it may happen that an R-controlled reinforcement system leads to the acquisition of an interesting stable response function by a perceptron, but this cannot generally be guaranteed, and any classification which is achieved is necessa-

rily one which is selected by the perceptron, rather than by the experimenter. The interesting questions concerning such systems deal with the types of classifications to which they converge, for different kinds of environments. In particular, we will be interested in any systems which tend to form classifications on the basis of some concept of stimulus "similarity". It will be shown in later chapters that elementary perceptrons do not, in general, tend to form classes on this basis except under special, and highly restrictive, environmental conditions, but that cross-coupled perceptrons appear to have a striking capability for such "spontaneous organization".

In the preceding theorems, only perceptrons employing alpha system reinforcement have been considered. The remaining two theorems consider two departures from this model. The first demonstrates that an even weaker form of reinforcement than that in the random-sign correction procedure can guarantee a solution in finite time, provided it is employed in a correction procedure, in which the application of reinforcement depends upon the occurrence of response errors. We define a <u>random perturbation correction procedure</u> as a reinforcement process in which, if an error occurs, reinforcement is applied to the active A-units, as in the α -system, except that the magnitude and sign of η are both chosen independently and separately for each reinforced connection in the system, according to some probability distribution.

<u>THEOREM 7</u>: Given an elementary perceptron with a finite number of memory states, a stimulus world W, and a classification $C(W)$ for which a solution can be reached from the starting point by some reinforcement sequence, then a solution can always be obtained in finite time by means of a random perturbation correction procedure.

PROOF: The reinforcement process is a random walk, which (for the given conditions) will eventually take the representative point of the system to every attainable point in phase space. Since the number of points is assumed to be finite, a solution must be reached in finite time.

Of the three reinforcement procedures which have been shown to guarantee solutions in elementary perceptrons (error correction, random-sign correction, and random perturbation correction procedures) the first is clearly the strongest, and can be expected to converge most rapidly. The random perturbation procedure will converge most slowly, since it must hunt through a large domain of the phase space of the system before achieving a satisfactory terminal state, and is not guided during this process by any directional constraints. In this respect, it shares many of the difficulties of Ashby's homeostat (Ref. 3); but it shares the virtue of the homeostat as well, that if the solution space is attainable, it will utlimately arrive at a solution no matter how complicated its functional representation may be. The random sign and random disturbance procedures may prove to be of interest in biological models, since the only information required for the control of reinforcement is whether or not an error has occurred.

In practice, it will be seen that a gamma system (Definition 38, Chapter 4) generally works at least as well and sometimes better than an alpha system. Nonetheless, the following theorem indicates that this system lacks the true universality of the alpha system.

THEOREM 8: Given an elementary γ-perceptron, a stimulus world W, and a classification $C(W)$, it is possible that a solution to $C(W)$ exists which cannot be achieved by the perceptron.

PROOF: Let each A-unit be activated for at least one stimulus in W, and let each stimulus activate a disjoint set of A-units. Let the classification function $C(W)$ be one which assigns every stimulus to the same class, either positive or negative, A solution clearly exists, if the values of all connections are positive (or negative, as required by the classification). But if the initial state of the system is one in which all values are zero, or of the wrong sign, a solution can never be achieved by the gamma system, since a solution requires that the total value of each set A_i of units responding to S_i, and consequently the total value over the entire A -set, should agree in sign with the classification. In the gamma system this is impossible, since the initial sum of the values is constant. The conservative property of the gamma system gives it one degree of freedom less than the alpha system, making it impossible to achieve a solution to such problems unless at least one surplus A-unit (which does not respond to any stimuli) exists.

The two remaining theorems were proposed by Joseph (Ref. 42), and establish useful diagnostic procedures for determining the existence of solutions in both alpha and gamma system perceptrons. As in Theorem 3, the activity function of the A -unit a_i is defined as

$$a_i^*(S_j) = \begin{cases} 1 & \text{if } a_i \text{ is active for } S_j \\ 0 & \text{otherwise} \end{cases}$$

For any n-vector, X, with components x_k, the bias number of a_i with respect to X is defined as

$$b_i(X) = \sum_{k=1}^{n} x_k \, a_i^*(S_k)$$

This quantity is clearly related to the bias ratio (defined in 5.4) if X is taken to be the class-assignment vector for the n stimuli. We will denote by X^* any n-vector X whose components x_i do not disagree in sign with the required classification, $C(W)$, i.e., $x_j \geq 0$ if S_j is in the positive class, and $x_j \leq 0$ if S_j is in the negative class. X^* will denote a vector in which the inequalities are strict (no zero components).

THEOREM 9: Given an α-perceptron, and a classification $C(W)$, a necessary and sufficient condition that the error correction procedure reach a solution (in finite time, with arbitrary starting point) is that there exists no non-zero X^* such that $b_i X^* = 0$ for all i.

PROOF: For conveneince, an un-normalized G-matrix will be assumed. For such a matrix,

$$g_{jk} = n_{jk} \equiv \sum_i a_i^*(S_j) \, a_i^*(S_k)$$

where n_{jk} is the number of A-units in the set responding to both S_j and S_k. Hence, for any n-vector X,

$$X'GX = \sum_{j,k} x_j \, x_k \, g_{jk} = \sum_{i,j,k} x_j \, x_k \, a_i^*(S_j) \, a_i^*(S_k)$$

But

$$\sum_i \left[b_i(X)\right]^2 = \sum_i \left[\sum_j x_j \, a_i^*(S_j)\right]^2 = \sum_{i,j,k} x_j \, x_k \, a_i^*(S_j) \, a_i^*(S_k)$$

Hence $X'GX = \sum_i \left[b_i(X)\right]^2$

If the condition of the theorem holds, then $X'GX \neq 0$ for $X = X^{\#}$, $X \neq 0$. But from the proof of Theorem 4, it can be shown that $X'GX \geq \alpha \|X\|^2$ for $X = X^{\#}$, where $\alpha > 0$. Then the proof of the correction procedure in Theorem 4 applies, and a solution exists, so that the stated condition must be sufficient.

If the condition does not hold, then there is a non-zero $X^{\#}$ such that $X'GX = 0$. Since G is positive semidefinite, this implies that $X'G = 0$. Thus, X is orthogonal to all the columns of G, and hence to any linear combination of the columns of G. Since for an arbitrary vector Z, GZ is a linear combination of the columns of G, GZ is orthogonal to X. $X^{\#}$ cannot be orthogonal to any vector U in which the signs of all u_i agree with $C(W)$, and hence it follows that there cannot exist vectors Z and U such that $GZ = U$. This means that there exists no solution to the classification problem, so the condition given must be necessary. Q.E.D.

COROLLARY: For an α-system, the condition that there exist no non-zero vector $X^{\#}$ such that $b_i X^{\#} = 0$ for all i is equivalent to the condition that there exist Z and U such that $GZ = U$ (where U is in the same orthant as $C(W)$). Alternatively, this condition is equivalent to $X'GX \neq 0$ for all non-zero $X^{\#}$.

THEOREM 10: Given a $\mathcal{T}$-perceptron, and a classification $C(W)$, a necessary and sufficient condition that the error correction procedure reach a solution (in finite time) is that there exists no non-zero X^* such that $b_i X^* = c$ for all i.

PROOF: For the $\mathcal{T}$-system, the normalized G matrix consists of elements

$$g_{jk} = n_{jk} - \frac{1}{N_a} n_j n_k = \sum_i a_i^*(S_j)\, a_i^*(S_k) - \frac{1}{N_a} \sum_{i,h} a_i^*(S_j)\, a_h^*(S_k)$$

It is readily seen that G is symmetric. For any n-vector X, $X'GX$ is given by

$$X'GX = \sum_{j,k} x_j\, x_k\, g_{jk}$$

$$= \sum_{i,j,k} x_j\, x_k\, a_i^*(S_j)\, a_i^*(S_k) - \frac{1}{N_a} \sum_{h,i,j,k} x_j\, x_k\, a_i^*(S_j)\, a_h^*(S_k)$$

We now define $b^*(X)$ as

$$b^*(X) = \frac{1}{N_a} \sum_i b_i(X)$$

From this, we see that

$$\sum_i \left[b_i(X) - b^*(X)\right]^2 = \sum_i \left[b_i(X)\right]^2 - N_a \left[\frac{1}{N_a} \sum_i b_i(X)\right]^2$$

$$= \sum_{i,j,k} x_j\, x_k\, a_i^*(S_j)\, a_i^*(S_k) - \frac{1}{N_a} \sum_{h,i,j,k} x_j\, x_k\, a_i^*(S_j)\, a_h^*(S_k)$$

Hence $X'GX = \sum_i \left[b_i(X) - b^*(X)\right]^2$.

From this it follows, first of all, that G is positive definite or positive semidefinite, as was the case for the α-system. Secondly, it is seen that $X'GX = 0$ if and only if $b_i(X) = c$ for all i . The proof now proceeds exactly as in Theorem 9.

COROLLARY: For a γ-system, the condition that there exists no non-zero vector X^* such that $b_i X^* = c$ for all i is equivalent to the condition that there exist Z and U such that $GZ = U$ (where U is in the same orthant as $C(W)$).

In practice, it is often possible to show that a given perceptron does not permit a solution to a given classification problem by substituting the classification vector itself, $C(W)$, for the vector X^* in the above theorems, and computing the b_i . If these turn out to be zero for all A-units, then no solution exists for either the alpha or gamma system. If they are a constant other than zero, a solution may exist for the alpha system, but not for the gamma system. If they are not all identical, then a solution may exist for either system. While it is sufficient to take the components of X^* to be integers, the vector with all components $x_i = \pm 1$ is not always sufficient. For example, if the $a_{ij}^*(S_i)$ matrix is $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ the b_i will all be anihilated by $X = (1, -2, 1)$, but not by $X = (1, -1, 1)$. The condition for the α-system is equivalent to the requirement that there should be no vector in the same orthant as $C(W)$ which is orthogonal to the linear manifold spanned by the activity vectors of the A-units.

6. Q-FUNCTIONS AND BIAS RATIOS IN ELEMENTARY PERCEPTRONS

Thus far, we have been mainly concerned with the general "qualitative" properties of elementary perceptrons. In the present chapter, the groundwork for a quantitative analysis of their performance will be presented. In the theorems of Chapter 5, it was shown that the existence and attainability of solutions, in an elementary perceptron, depends strongly on the properties of the G -matrix. Each element of this matrix, $g_{i,j}$, is a measure of the generalization of reinforcement from stimulus S_j to S_i This generalization coefficient, $g_{i,j}$, varies with the measure of the set of A-units which respond jointly to S_i and S_j . Until now, the actual quantitative measures of these sets have not been taken into consideration, and only the formal properties of the matrix G have been considered. The Q -functions, which are introduced in this chapter, represent the probabilities that an A-unit in a specified class of perceptrons will respond to a particular stimulus, or will respond jointly to a designated set of stimuli. These Q -functions not only determine the expected values of the generalization coefficients, $g_{i,j}$, but enter into the analysis of variability of perceptron performance as well, as will be seen in the following chapter.

6.1 Definitions and Notation

The Q -functions, defined below, are always specific to a particular class of perceptrons in which the origin point configurations of the A-units have been selected according to some designated set of rules from a specified S-set or retina. The functions Q are defined only for simple A-units, a_i , which are said to be active if the algebraic sums of their input signals, α_i , are equal to or greater than their thresholds,

θ_i . For such A-units, Q represents the probability of drawing an A-unit at random from the specified distribution which responds to each of a specified set of stimuli. The notation employed is as follows:

Q_i = probability that an A-unit in a specified class of perceptrons responds to stimulus S_i .

Q_{ij} = probability that an A-unit in a specified class of perceptrons responds to stimulus S_i and also to stimulus S_j .

$Q_{ij\ldots m}$ = probability that an A-unit in a specified class of perceptrons responds to each of the stimuli $S_i, S_j, \ldots, S_m$.

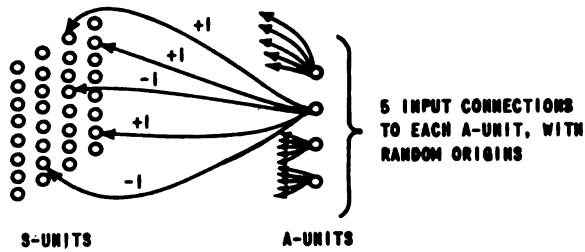
6.2 Models to be Analyzed

Three types of models will be considered which differ in the rules by which connections are made between S-units and A-units. It turns out that for the three cases, the distribution of input signals to the A-units is expressed in terms of binomial, Poisson, and normal random variables, respectively. These models are therefore named binomial, Poisson, and Gaussian models.

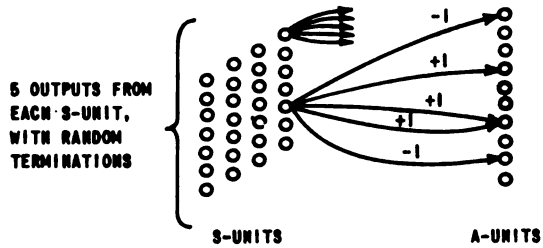
6.2.1 Binomial Models

In a binomial model the input signal, α_i , received by unit a_i , is distributed as the difference of two binomially distributed random variables. This model characterizes a type of perceptron in which each A-unit receives a fixed number of connections from the "retina",

(a) BINOMIAL MODEL, WITH $\alpha = 3$, $y = 2$



(b) POISSON MODEL, WITH CONSTRAINED ORIGINS



(c) POISSON MODEL, WITH RANDOM ORIGINS

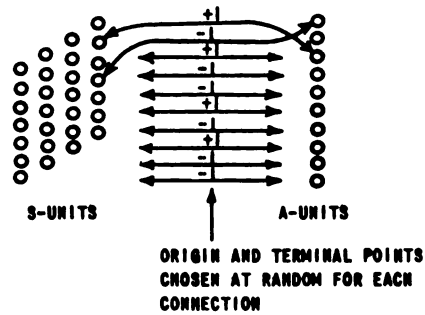


Figure 6 ILLUSTRATION OF TYPICAL S TO A-UNIT CONNECTIONS (ARROWHEADS INDICATE RANDOMLY SELECTED TERMINATIONS). IN GAUSSIAN MODELS, THE VALUES OF THE CONNECTIONS (SHOWN HERE AS ± 1) ARE NORMAL RANDOM VARIABLES.

consisting of exactly x "excitatory" and y "inhibitory" connections. Each of the excitatory connections has the value +1, and each inhibitory connection has the value -1. The threshold, θ, is assumed to be fixed for all A-units. The origins of the connections to an A-unit are selected independently, with uniform probability, from the entire set of S-units (or retinal points). Specifically, a set of equiprobable origin configurations can be constructed as follows: Let there be ν connections, numbered from 1 to ν. Let the S-units be numbered from 1 to N_A. Then the set of all possible sequences of ν integers, each having a value in the range $1 \leq n \leq N_A$ corresponds to the complete set of A-units. In this model, the number of distinguishable A-units possible for a retina of N_A points is $\binom{N_A + x - 1}{x}\binom{N_A + y - 1}{y}$*.

In the binomial model, Q functions do not depend on the number of sensory units, but on the fraction of them which are illuminated. A variation of this model has been analyzed in Ref. 79, where the additional constraint is introduced that no two connections to a single A-unit can originate from the same S-unit. It has been shown that for moderately large numbers of S-units, this model is practically indistinguishable from the true binomial model described above.

6.2.2 Poisson Models

In a Poisson model, α_i is distributed as the difference of two Poisson-distributed random variables. In this model, it is assumed that the number of input connections to an A-unit is not fixed, but is a random variable. The model corresponds to one of two situations, the equations for the Q-functions being identical for both:

---

* The derivation of this formula can be found in Feller, Ref. 21, page 52.

(1) In the constrained origin model, each S-unit emits a fixed number of output connections, consisting of ν_x excitatory, and ν_y inhibitory connections (with values +1 and -1, respectively). Terminal points are selected at random from a set of N_a A-units. For the model to hold exactly, N_A and N_a should both be infinite, the ratio N_A/N_a being a parameter of the system. For finite N_A and N_a, the model remains a close approximation.

(2) In the random origin model, a set of N_x excitatory and N_y inhibitory connections are each independently assigned an origin and a terminus at random, from a set of S-units and A-units, with uniform probabilities. In this case, for the model to hold exactly, the numbers N_x, N_y and N_a should all be infinite, with $\frac{N_x + N_y}{N_a}$ being a parameter of the system; as in the previous case, however, the model is a close approximation for finite systems.

In the Poisson model, for Case (1), the number of possible A-units is $(\nu_x + 1)^{N_A} (\nu_y + 1)^{N_A}$. For Case (2), the number of possible A-units is $(N_x + 1)^{N_A} (N_y + 1)^{N_A}$. The binomial model, the constrained-origin Poisson model, and the random-origin Poisson model yield increasingly large sets of possible A-units, for the same numbers of S-units, A-units, and connections.

6.2.3 Gaussian Models

In the Gaussian case, α_i is distributed as the difference of two normally distributed random variables, i.e., α_i is normally distributed. While both of the above cases converge to a Gaussian model as the number of input connections to an A-unit becomes large, we shall be concerned here with a model in which the number of connections remains finite, but the values of the connections are normally distributed.

6.3 Analysis of Q_i

For both the binomial and Poisson models, Q_i, the probability that an A-unit is activated by stimulus S_i, is given by the probability that the total input signal α is equal to or greater than the threshold, θ. Specifically,

(6.1)

$$Q_i = \sum_{\alpha \geq \theta} P(\alpha) = \sum_{E-I \geq \theta} P_x(E)\, P_y(I) = \sum_{E=\theta}^{E_{max}} \sum_{I=0}^{E-\theta} P_x(E)\, P_y(I)$$

where $E_{max} = \begin{cases} \chi & \text{for binomial model} \\ \infty & \text{for Poisson model} \end{cases}$

$P_x(E)$ = probability that exactly E of the excitatory connections to an A-unit originate from active S-points.

$P_y(I)$ = probability that exactly I of the inhibitory connections to an A-unit originate from active S-points.

For the binomial model,

$$P_x(E) = \binom{\chi}{E} R_i^{\,E} (1-R_i)^{\chi - E}$$
$$P_y(I) = \binom{y}{I} R_i^{\,I} (1-R_i)^{y-I} \qquad (6.2)$$

where R_i = fraction of retinal points (S-units) activated by stimulus S_i.

For the Poisson model,

$$P_x(E) = \frac{(R_i \bar{x})^E}{E!} \cdot e^{-R_i \bar{x}}$$
$$P_y(I) = \frac{(R_i \bar{y})^I}{I!} \cdot e^{-R_i \bar{y}} \qquad (6.3)$$

where $\bar{x} = N_x/N_a$ = expected number of excitatory input connections to an A-unit.

$\bar{y} = N_y/N_a$ = expected number of inhibitory input connections to an A-unit.

$P(\alpha)$ for the Poisson model can be expressed alternatively by the following identity (pointed out by Prof. H. D. Block):

$$P(\alpha) = P\left\{(e-i) = \alpha\right\} = e^{-R_i(\bar{x}+\bar{y})}\left(\frac{\bar{x}}{\bar{y}}\right)^{\alpha/2} I_\alpha\left(2R_i\sqrt{\bar{x}\,\bar{y}}\right)$$

Where $I_\rho(x)$ is a Bessel function of an imaginary argument, given by

$$I_\rho(x) = \sum_{y=0}^{\infty} \frac{\left(\frac{x}{2}\right)^{2y+\rho}}{y!(y+\rho)!} = i^{-\rho} J_\rho(ix)$$

The use of this equation makes it possible to compute Q-functions for the Poisson model by hand, with the aid of tables of Bessel functions (c.f., Ref. 37, pp. 224-233).

For the Gaussain model, equation (6.1) requires an additional factor representing the distribution of value for each of the connections. Specifically, if the absolute values of both excitatory and inhibitory connections are distributed with mean μ and standard deviation σ, we have

$$Q_i = \sum_{E=0}^{E_{max}} \sum_{I=0}^{I_{max}} P_x(E)\, P_y(I) \int_{D=\theta}^{\infty} \phi(D_{E,I})\, dD \qquad (6.4)$$

where

$$\phi(D_{E,I}) = \frac{1}{\sqrt{2\pi}\,\sigma_D}\, e^{-\frac{1}{2}\left(\frac{D-\mu_D}{\sigma_D}\right)^2}$$

$$\mu_D = E_\mu - I_\mu$$

$$\sigma_D^2 = (E+I)\,\sigma^2$$

$P_x(E)$ and $P_y(I)$, in equation (6.4) are given either by (6.2) or (6.3), depending on whether the number of input connections to an A-unit is fixed (as in the binomial model) or random (as in the Poisson model).

Figures 7 and 8 show representative families of curves for Q_i as a function of R_i, for the binomial and Poisson models, respectively. Note that both models are very similar in their basic characteristics. Specifically:

1. In all cases, for $R_i < .5$ and $x \geq y$, Q_i increases monotonically with R_i.

2. For purely excitatory models $(y = 0)$ Q_i goes to 1.0 as R_i approaches 1.0. (Figures 7a and 8a).

3. For models with $\theta > x - y$, Q_i goes to zero as R_i approaches 1.0. (Figures 7b and 8b).

4. For $x = y$, Q_i tends to remain invariant except for very small or very large values of R_i. The range over which Q_i tends to remain constant is increased if the number of connections becomes large (Figs. 7c and 8c). In the limit, with small θ and large x and y, Q_i approaches .5 for all values of R_i except 0 and 1.

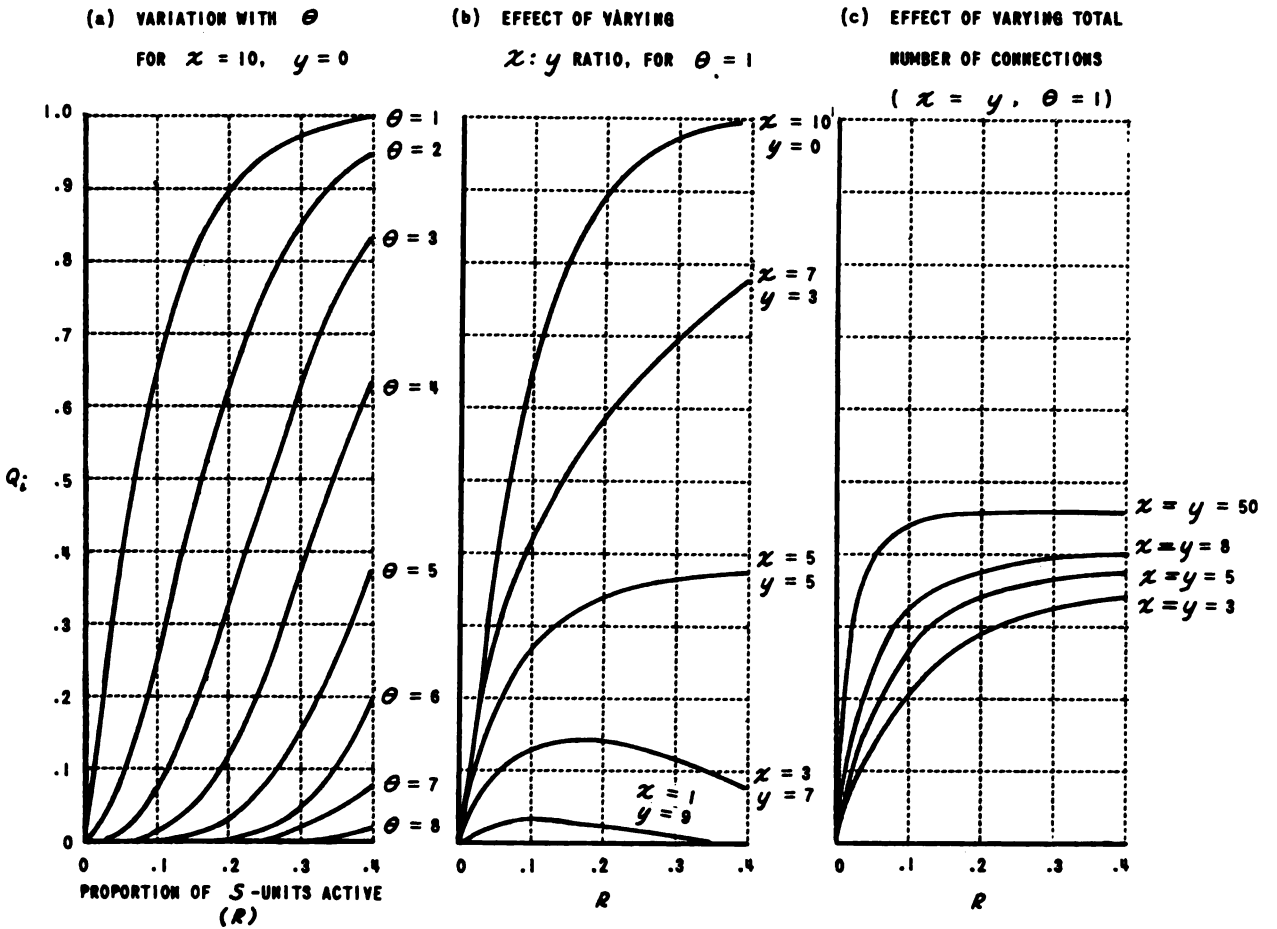


Figure 7 Q_i AS FUNCTION OF RETINAL AREA ILLUMINATED, FOR BINOMIAL MODEL

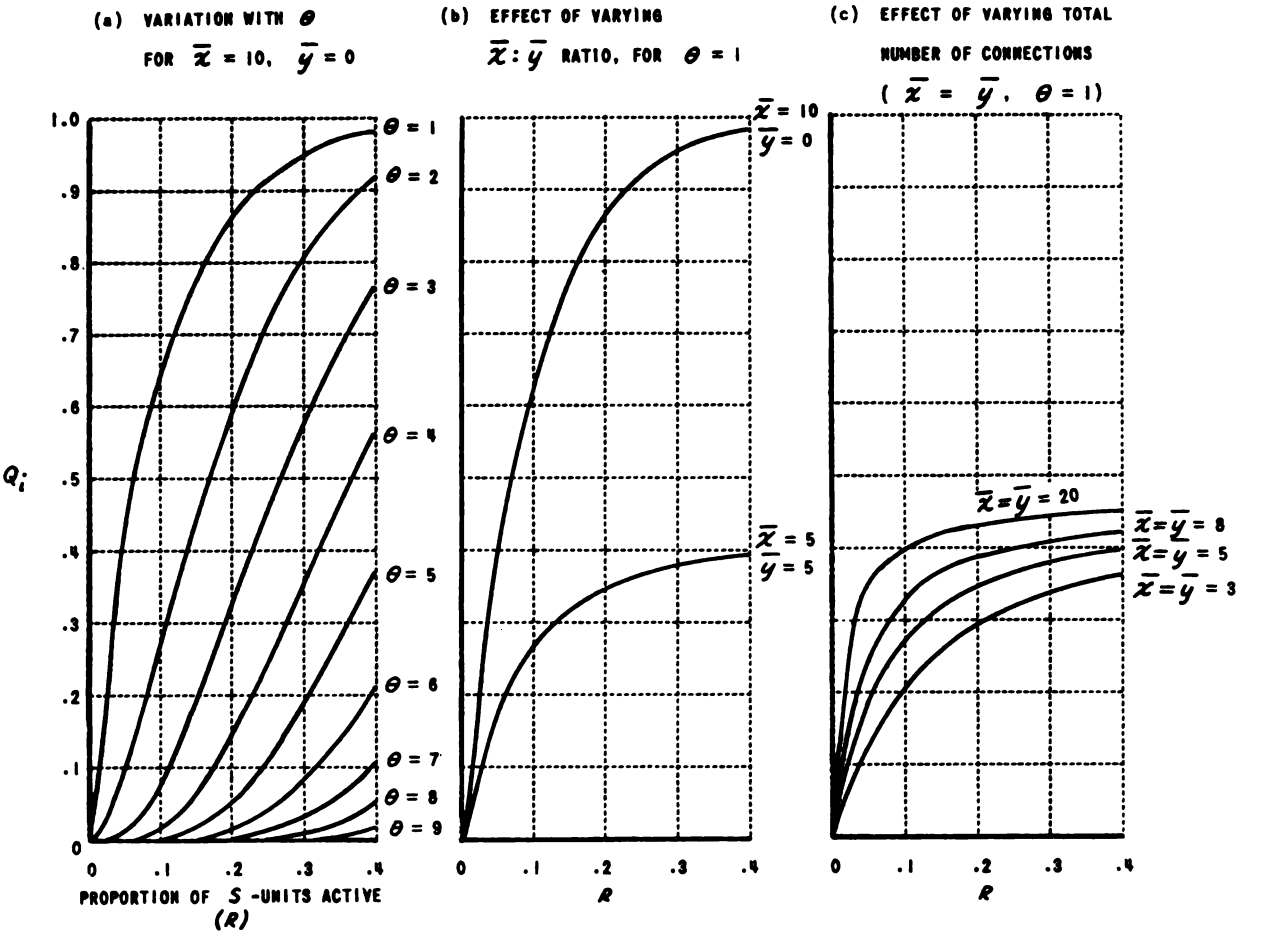


Figure 8 Q_i AS FUNCTION OF RETINAL AREA ILLUMINATED, FOR POISSON MODEL

5. Keeping x fixed, then for small θ, Q_i is generally greater for the binomial model than for the Poisson model. For large θ, Q_i is greater for the Poisson model.

6. For the binomial model, $Q_i = 0$ for $x < \theta$ while for the Poisson model, $Q_i = 0$ only if $x = 0$.

6.4 Analysis of Q_{ij}

Q_{ij} is the probability that an A-unit is activated by each of two stimuli, S_i and S_j. For both the binomial and Poisson models, Q_{ij} can be expressed by the equation:

$$Q_{ij} = \sum_{\substack{E_i + E_c - I_i - I_c \geq \theta \\ E_j + E_c - I_j - I_c \geq \theta}} P_x(E_i, E_j, E_c)\, P_y(I_i, I_j, I_c) \tag{6.5}$$

where θ = threshold of A-units

E_i = number of excitatory connections originating from points illuminated by S_i but not by S_j

E_j = number of excitatory connections originating from points illuminated by S_j but not by S_i

E_c = number of excitatory connections originating from points common to S_i and S_j

I_i = number of inhibitory connections originating from points illuminated by S_i but not by S_j

I_j = number of inhibitory connections originating from points illuminated by S_j but not by S_i

I_c = number of inhibitory connections originating from points common to S_i and S_j

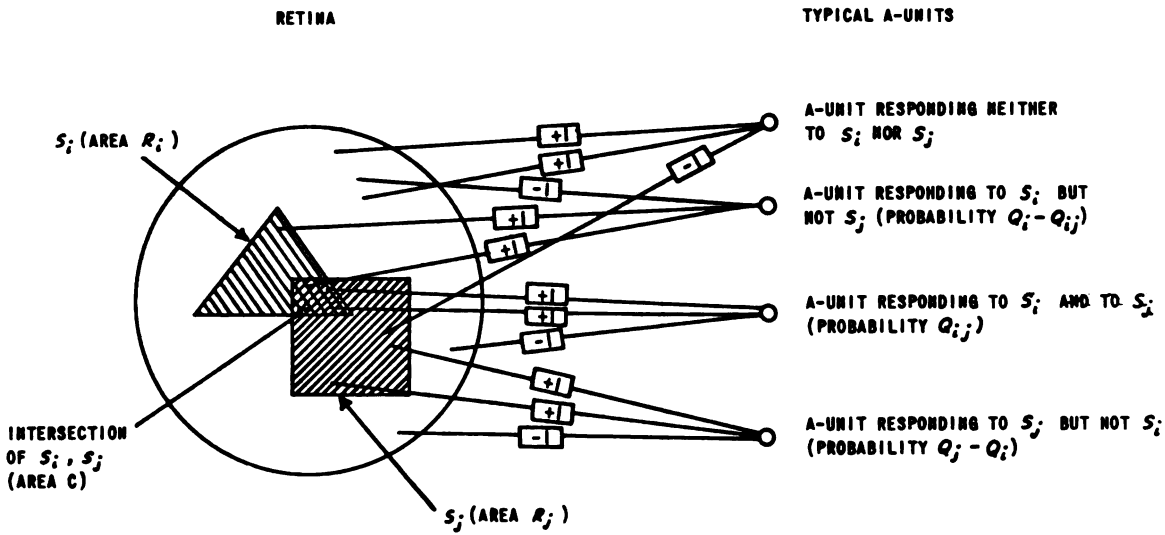


Figure 9 RETINAL SETS AND CONNECTIONS INVOLVED IN Q_{ij}. S_i IS TRIANGLE, S_j IS SQUARE. ASSUME $\theta = 2$, $\alpha = 2$, $y = 1$ FOR ALL A-UNITS

The point sets involved in the analysis of Q_{ij} are illustrated in Figure 9. For the binomial model, the required probabilities are given by the multinomial equations:

$$P_x(E_i, E_j, E_c) = \frac{x!}{E_i!\,E_j!\,E_c!\,(x-E_i-E_j-E_c)!}\, A_i^{E_i} A_j^{E_j} C^{E_c} (1-A_i-A_j-C)^{x-E_i-E_j-E_c} \quad (6.6)$$

$$P_y(I_i, I_j, I_c) = \frac{y!}{I_i!\,I_j!\,I_c!\,(y-I_i-I_j-I_c)!}\, A_i^{I_i} A_j^{I_j} C^{I_c} (1-A_i-A_j-C)^{y-I_i-I_j-I_c}$$

where C = proportion of retinal points illuminated both by S_i and S_j ;

$A_i = R_i - C$ where R_i is the proportion of retinal points illuminated by S_i ;

$A_j = R_j - C$ where R_j is the proportion of retinal points illuminated by S_j .

For the Poisson model (where $\bar{x}$ and $\bar{y}$ are the expected numbers of excitatory and inhibitory connections to an A-unit),

$$P_x(E_i, E_j, E_c) = (E_i!\,E_j!\,E_c!)^{-1} \cdot e^{-\bar{x}A_i}(\bar{x}A_i)^{E_i} \cdot e^{-\bar{x}A_j}(\bar{x}A_j)^{E_j} \cdot e^{-\bar{x}C}(\bar{x}C)^{E_c} \quad (6.7)$$

$$P_y(I_i, I_j, I_c) = (I_i!\,I_j!\,I_c!)^{-1} \cdot e^{-\bar{y}A_i}(\bar{y}A_i)^{I_i} \cdot e^{-\bar{y}A_j}(\bar{y}A_j)^{I_j} \cdot e^{-\bar{y}C}(\bar{y}C)^{I_c}$$

As in the case of Q_i , the Gaussian model for Q_{ij} requires an additional factor representing the normal distribution of connection values. The components of the input signal, α , which originate from the unique S-units in S_i , the unique points in S_j , and from the common retinal set are designated D_i , D_j , and D_c , respectively. By analogy to (6.4),

$$D_i = (E_i - I_i)$$

$$D_j = (E_j - I_j)$$

$$D_\ell = (E_\ell - I_\ell)$$

$$\mu_{D_\nu} = E_\nu \mu - I_\nu \mu$$

$$\sigma^2_{D_\nu} = (E_\nu + I_\nu)\sigma^2$$

$\phi(D_\nu) = \phi(D_{E_\nu, I_\nu})$, defined as in (6.4).

Then,

$$Q_{ij} = \sum_{\{E_i, E_j, E_\ell, I_i, I_j, I_\ell\}} P_x(E_i, E_j, E_\ell)\, P_y(I_i, I_j, I_\ell) \int_{D_\ell=-\infty}^{\infty} \int_{D_i=\theta-D_\ell}^{\infty} \int_{D_j=\theta-D_\ell}^{\infty} \phi(D_\ell)\,\phi(D_i)\,\phi(D_j)\, dD_\ell\, dD_i\, dD_j \tag{6.8}$$

For some purposes, the distribution of the input signals, α_i, and α_j, is of interest. The joint probability, $P(\alpha_i, \alpha_j)$, is given by

$$\sum_{\{E_i, E_j, E_\ell, I_i, I_j, I_\ell\}} P_x(E_i, E_j, E_\ell)\, P_y(I_i, I_j, I_\ell) \int_{D_\ell=-\infty}^{\infty} \phi(\alpha_i - D_\ell)\,\phi(\alpha_j - D_\ell)\, dD_\ell \tag{6.9}$$

It should be noted that $Q_i = Q_{ii}$ is a special case of these equations, for which $A_i = A_j = C$. Tables of Q_{ij} for binomial and Poisson models have been published in Ref. 87.

Figures 10 and 11 illustrate the quantitative properties of Q_{ij}, as a function of C, the measure of the intersection of stimuli S_i and S_j on the "retina". For convenience of representation, Q_{ij} is actually plotted

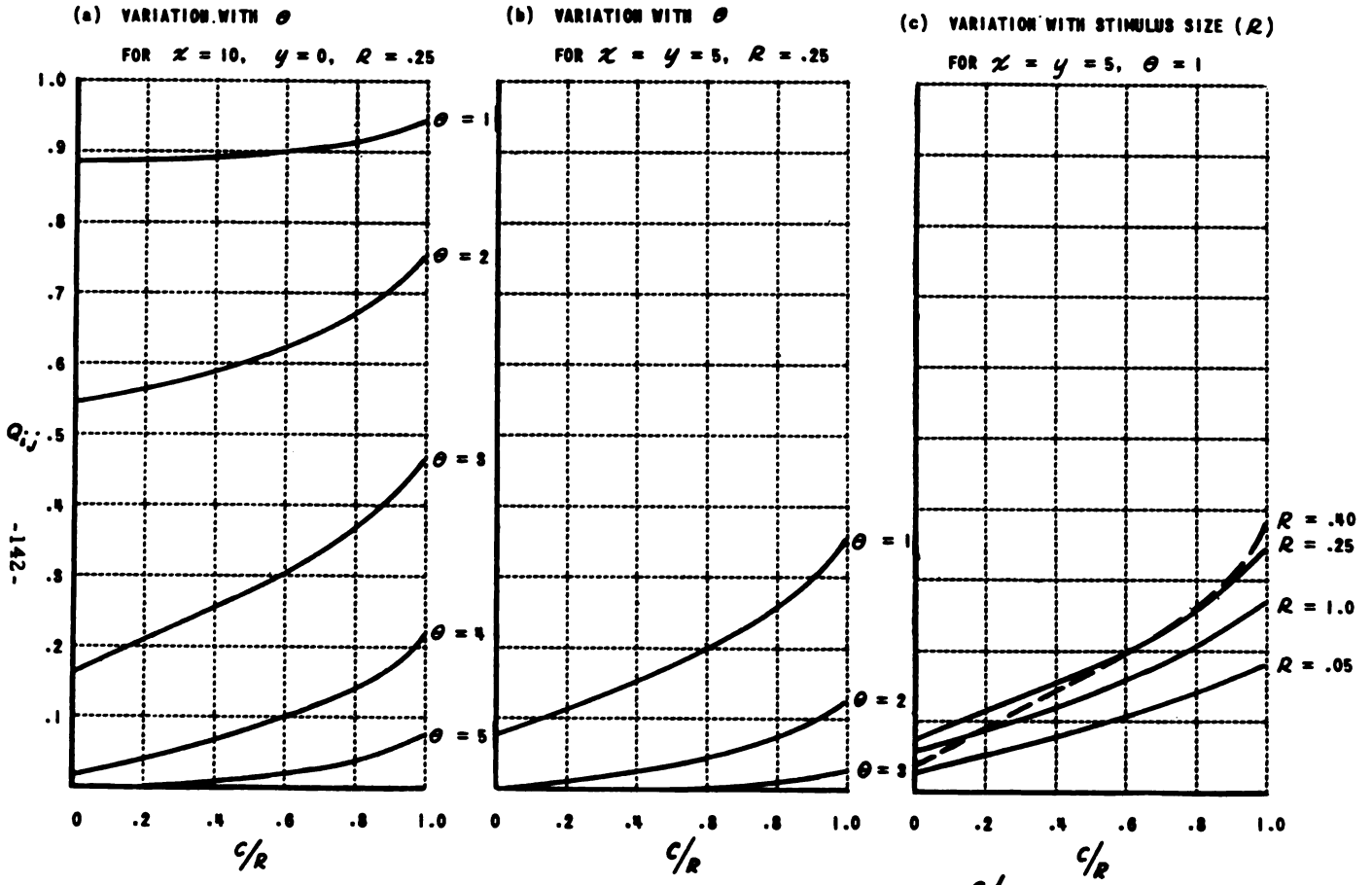


Figure 10 Q_{ij} AS A FUNCTION OF THE RELATIVE INTERSECTION, C/R FOR BINOMIAL MODEL. $R_i = R_j$ IN ALL CASES.

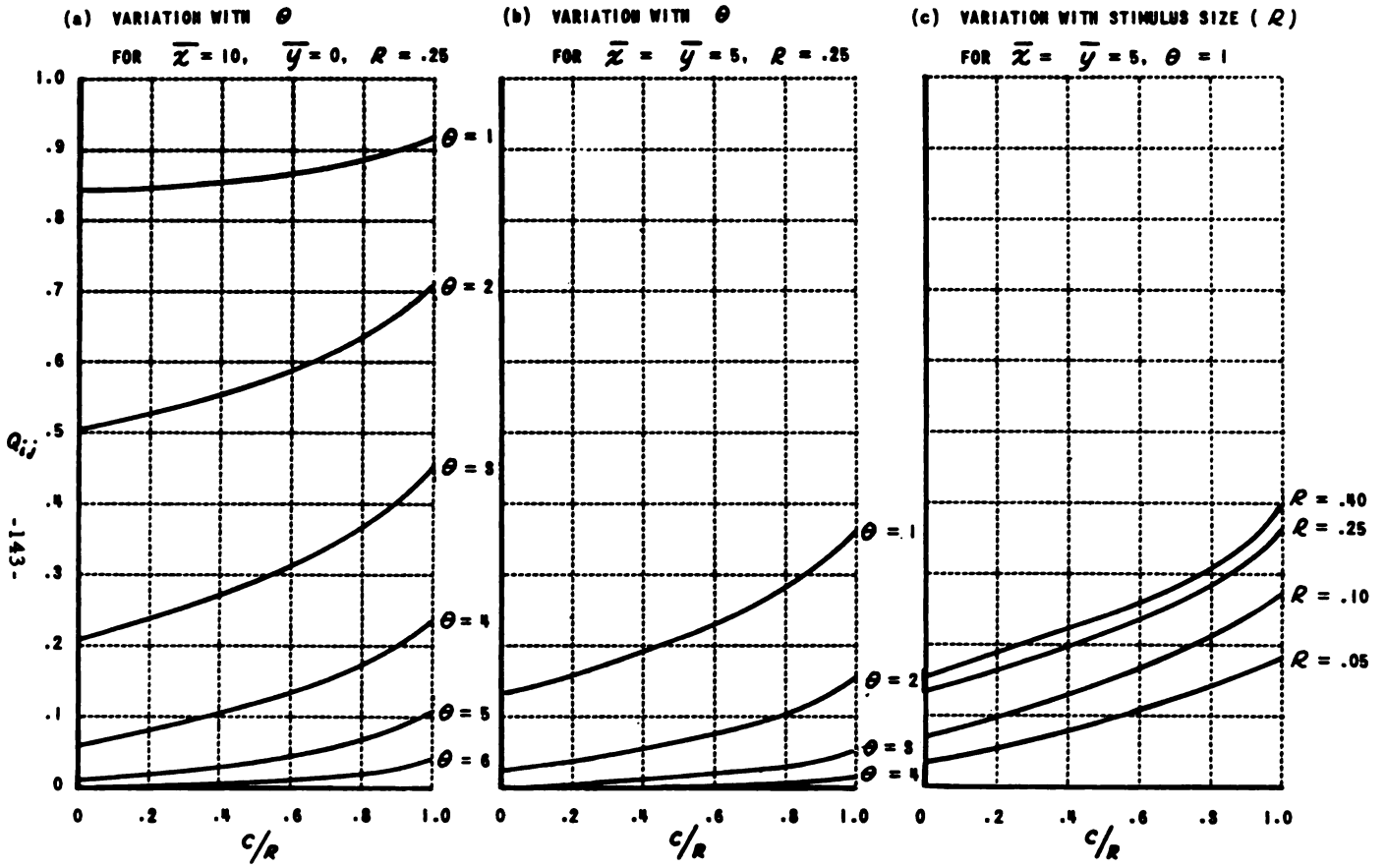


Figure 11 Q_{ij} AS FUNCTION OF THE RELATIVE INTERSECTION, c/R FOR POISSON MODEL. $R_i = R_j$ IN ALL CASES.

as a function of the relative intersection (or proportional intersection), C/R, R_i and R_j being equal for all cases shown. Note that for $C/R = 1$, $Q_{ij} = Q_{ii} = Q_i$. The main features of these curves are:

1. In all cases, Q_{ij} increases monotonically with C.

2. For large θ, Q_{ij} tends to remain close to zero, except for stimuli which approach perfect identity (C/R close to 1.0).

3. For large values of R, Q_{ij} tends to accelerate more rapidly as C approaches 1.

4. For the binomial model, Q_{ij} for disjoint or well separated stimuli ($C \approx 0$) may have a maximum with respect to R. This effect is not found in the Poisson model. (Figs. 10c and 11c.)

5. For equivalent parameters, Q_{ij} tends to show a sharper "shoulder" in the binomial model than the Poisson model.

The second of these properties is an important factor in determining the discriminative capability of a perceptron. It is shown best in terms of the conditional probability, $Q_{i|j}$, that an A-unit which responds to S_j also responds to S_i. $Q_{i|j}$ is equal to Q_{ij}/Q_j, and is shown for several typical cases in Fig. 12. Note that for large values of θ, the probability that an A-unit responding to S_j responds to a second stimulus, S_i, is virtually zero, unless the stimuli approach perfect identity. The difference between the binomial and Poisson models is shown most clearly in Figures 12(a) and 12(b). Figure 12(c) demonstrates that the conditional probability depends only slightly on stimulus size. Additional curves for

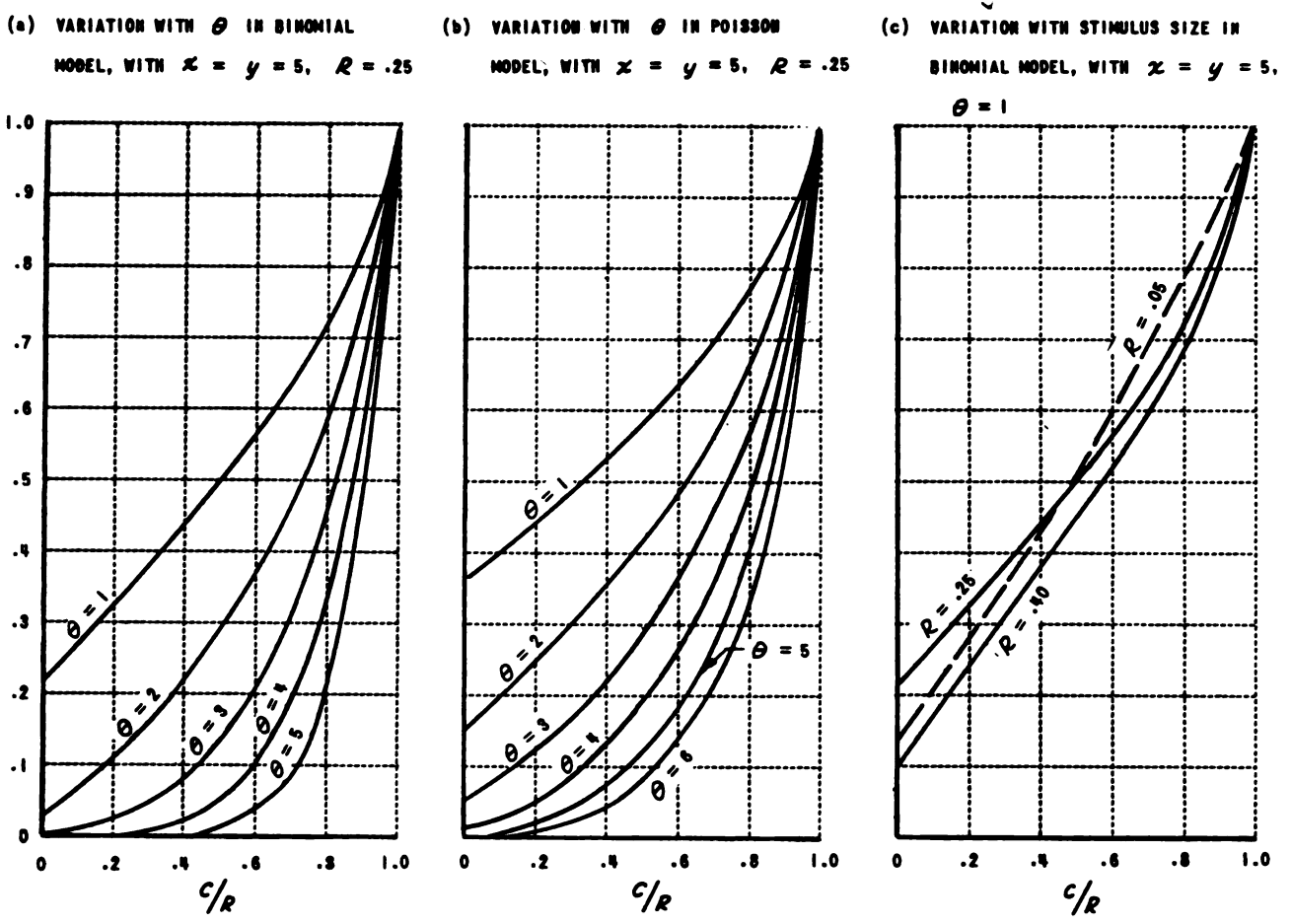


Figure 12 CONDITIONAL PROBABILITY ($Q_{i|j}$) THAT AN A-UNIT RESPONDING TO S_j ALSO RESPONDS TO S_i , SHOWN AS FUNCTION OF c/R

these functions can be found in Refs. 79 and 80.

In analyzing the gamma system, it will be seen that the conditions under which $Q_{i,j} = Q_i Q_j$ are of particular interest, since for the gamma system the expected value of $g_{i,j}$ is zero for such conditions. In the binomial model, $Q_{i,j} = Q_i Q_j$ if $C = R_i R_j$. This condition will tend to be met if the stimuli are randomly chosen sets of S-points, the expected intersection of any two such sets being equal to the product of the measures of the sets. It can readily be seen that under these conditions, the probability that an origin point which is in S_j is also in S_i is the same as the probability that an origin point which is not in S_j happens to be in S_i; in other words, the probability that the origin of a connection is in S_j does not depend on whether or not it is in S_i, and consequently the response to S_j is independent of the response to S_i, yielding $Q_{i,j} = Q_i Q_j$. In the Poisson model, however, $Q_{i,j} = Q_i Q_j$ only if $C = 0$ (i.e., for disjoint stimuli) since the connections received from any disjoint subset of S-units are independent of connections (or signals) from any other subset.

6.5 Analysis of $Q_{i,j,k}$

In the following chapter, it will be seen that the expected responses of a simple perceptron can generally be determined from the functions Q_i and $Q_{i,j}$. The variability of performance in a class of perceptrons, however, will be seen to depend on the joint probability, $Q_{i,j,k}$, that an A-unit responds to each of three stimuli, S_i, S_j, and S_k. The equations are a straightforward generalization of those employed in the last section for $Q_{i,j}$. Specifically, there are now seven excitatory and seven inhibitory signal components to be considered:

E_i = Excitatory signal from S-units responding to S_i but not to S_j or S_k

E_j = excitatory signal from S-units responding to S_j but not to S_i or S_k

E_k = excitatory signal from S-units responding to S_k but not to S_i or S_j

E_{ij} = excitatory signal from S-units responding to S_i and S_j but not S_k

E_{ik} = excitatory signal from S-points responding to S_i and S_k, but not S_j

E_{jk} = excitatory signal from S-points responding to S_j and S_k but not S_i

E_{ijk} = excitatory signal from S-points responding to all three stimuli.

Inhibitory components are defined analogously. This yields the equation:

$$Q_{ijk} = \sum_{\substack{\alpha_i \geq \theta \\ \alpha_j \geq \theta \\ \alpha_k \geq \theta}} P_x(E_i, E_j, E_k, E_{ij}, E_{ik}, E_{jk})\, P_y(I_i, I_j, I_k, I_{ij}, I_{ik}, I_{jk}, I_{ijk}) \quad (6.10)$$

where

$$\alpha_i = E_i + E_{ij} + E_{ik} + E_{ijk} - I_i - I_{ij} - I_{ik} - I_{ijk}$$

$$\alpha_j = E_j + E_{ij} + E_{jk} + E_{ijk} - I_j - I_{ij} - I_{jk} - I_{ijk}$$

$$\alpha_k = E_k + E_{ik} + E_{jk} + E_{ijk} - I_k - I_{ik} - I_{jk} - I_{ijk}$$

The multinomial and Poisson probabilities employed in (6.10) for the binomial and Poisson models, respectively, are obtained by extension of (6.6) and (6.7), with appropriate measures for the various double and triple intersections among the stimuli.

6.6 Bias Ratios of A-units

Bias ratios were defined in Section 5.4 as the ratio of the number of stimuli in the positive class to the number of stimuli in the negative class, which activate an A-unit. In Theorem 2, it was shown that there must be some variation in the bias ratios of the A-units in a perceptron, if a solution to a given classification is to exist, and Theorems 9 and 10 showed that the closely related "bias numbers" yield necessary and sufficient conditions for solutions. Clearly, the distribution of bias ratios depends on the probabilities $Q_{ij\cdots m}$, that the A-units will respond to various possible sets of stimuli, S_i, S_j, ..., S_m. Rather than undertake a detailed analysis of bias ratios, empirical data are presented for a typical case, to illustrate how we might expect the "responsiveness" of A-units to different classes of stimuli to be distributed. These data were obtained by a Monte Carlo procedure, in which 10,000 A-units were tested on a digital computer to determine to how many stimuli of each class they responded.*

---

* The program was written by A. Geoffrion, for the Burroughs 220 computer at Cornell University.

The "retina" consists of a 20 by 20 mosaic of S-units , and the stimuli consist of 4 by 20 bars, placed vertically or horizontally on the retina, in all possible positions. The retina is assumed to be toroidally connected, so that bars placed near one edge of the field may re-enter at the opposite edge. Thus, there are twenty possible horizontal bars (the positive class) and twenty possible vertical bars (the negative class). This universe will be used as a standard one in a number of learning experiments to be analyzed in the following chapters.** Table 1 shows the number of A-units out of 10,000 responding to each possible combination of N^+ horizontal bars and N^- vertical bars. An A-unit which responds to 4 horizontal and 6 vertical bars, for example, is tallied in the 5th row and 7th column of the table. Each A-unit had five excitatory and five inhibitory connections, and a threshold of 2.

For stimuli which are more similar to one another (in terms of possible intersection of S-sets) than horizontal and vertical bars, we would expect to find the A-units less well distributed, and a greater concentration around the diagonal. One would also expect that in a universe in which the stimulus classes are less symmetric in their properties, the distribution of A-units would be less symmetric than that shown in Table 1. Table 2 illustrates both of these features. In this case, the "positive" class consists of 4 by 20 horizontal bars, just as before; the "negative" class, however, consists of a set of 6 by 20 horizontal bars. Again, there are twenty members of each class, but the maximum intersection possible between stimuli of the positive and negative class is much greater than before, and the size difference introduces an asymmetry which was not previously present.

** The toroidal retina has the convenient property of being unbounded and isotropic, with a finite surface. Any relations which hold for a set of stimuli projected onto the retina hold equally well if all stimuli are displayed by any combination of horizontal and vertical translations. This model (with Born-von Kármán boundary conditions) is easier to analyze than a spherical retina which has similar properties.

TABLE 1

JOINT DISTRIBUTION OF 10,000 A-UNITS, WITH RESPECT TO NUMBERS OF HORIZONTAL BARS AND NUMBERS OF VERTICAL BARS TO WHICH THEY RESPOND

N^+ (HORIZONTAL BARS) \ N^- (VERTICAL BARS)	0	1	2	3	4	5	6	7	8
0	287	326	349	312	324	63	27	2	1
1	315	392	378	376	306	71	30	4	3
2	325	417	441	399	351	92	27	7	3
3	324	382	398	353	343	90	37	9	2
4	330	351	364	340	305	68	24	4	1
5	68	87	79	84	85	27	8	1	0
6	32	36	34	27	26	7	2	0	0
7	6	9	7	7	6	2	0	0	0
8	2	0	1	2	1	1	0	0	0

TABLE 2

JOINT DISTRIBUTION OF 10,000 A-UNITS, WITH RESPECT TO NUMBERS OF 4 x 20 AND 6 x 20 HORIZONTAL BARS TO WHICH THEY RESPOND

N^+ (4 x 20 BARS) \ N^- (6 x 20 BARS)	0	1	2	3	4	5	6	7	8	9	10	11
0	917	436	224	47	11	1	0	0	0	0	0	0
1	277	724	507	334	86	44	4	0	0	0	0	0
2	98	250	572	539	370	119	51	5	3	0	0	0
3	16	63	191	522	534	424	166	17	5	2	0	0
4	1	10	40	152	380	543	602	67	9	3	0	0
5	0	0	11	23	50	133	154	59	22	5	1	0
6	0	0	0	3	11	22	48	24	24	4	0	1
7	0	0	0	0	0	3	4	11	10	8	1	0
8	0	0	0	0	0	0	0	1	1	4	1	0

While the joint distributions illustrated here are not of great utility in analyzing perceptron performance, they provide considerable insight into what takes place within the association system when a perceptron learns a classification of stimuli. Units situated on the diagonal (i.e., units which respond equally to both classes of stimuli) are essentially "duds"; they contribute little to a discrimination, and are as likely to be reinforced positively as negatively. A-units which have a strong bias towards one class or the other, however, (those situated in the upper right or lower left corners of the tables) are useful "discriminators". In learning a classification, the perceptron relies on combinations of such units, transmitting large-valued signals, to establish a bias towards the proper class when a stimulus appears.

7. PERFORMANCE OF ELEMENTARY α-PERCEPTRONS IN PSYCHOLOGICAL EXPERIMENTS

So far, only the formal properties of elementary perceptrons have been analyzed, without regard to particular experimental situations or procedures. We are now ready to begin a quantitative analysis of the performance of these systems in "psychological" experiments, i.e., experiments in which the procedures and observations are analogous to those which might be performed on a biological organism. A number of such experiments were defined in Part I, Section 3.3. In this chapter, we shall be chiefly concerned with discrimination experiments (c.f., Section 3.3.1), since the capabilities of elementary perceptrons are largely limited to this category. Before going on to other types of systems, however, we will consider what kinds of behavior might be expected of an elementary system in generalization experiments, figure detection experiments, and other problems which were discussed in Chapter 3. The analysis of discrimination experiments which is reported here is basically similar to that which was originally presented in Ref. 79. The former models have been substantially simplified, however, and the analysis has been made more rigorous, thanks largely to the work of R. D. Joseph, (Ref. 41).

7.1 Discrimination Experiments with S-controlled Reinforcement

The first problem to be analyzed is that of a discrimination experiment in which the perceptron is presented with a sequence of stimuli from an environment, W, and is reinforced for each stimulus in the sequence in accordance with a predetermined classification, $C(W)$, with the reinforcement control constant, γ, taking the sign of the required

response. The perceptron is then shown a test stimulus (S_χ) and the response to this stimulus is determined. The measure of performance for a class of perceptrons (characterized by the parameters N_a, θ, χ, and y for a binomial model or by N_a/N_a, θ, $\bar{x}$, and $\bar{y}$ for a Poisson model) is the probability that a perceptron from the specified class will give the correct response to S_χ after having been "trained" with the specified sequence of stimuli.

7.1.1 Notation and Symbols

S_j = the j^{th} stimulus in the environment

ρ_j = $\begin{cases} +1 \text{ if } S_j \text{ is in the positive class} \\ -1 \text{ if } S_j \text{ is in the negative class} \end{cases}$

$a_i^*(jk..x)$ = $\begin{cases} 1 \text{ if the } i^{\text{th}} \text{ A-unit is active for } S_j, S_k, \ldots, \text{ and } S_\chi \\ 0 \text{ otherwise} \end{cases}$

$Q_{jk..x}$ = $Ea_i^*(jk..x)$ = probability that $a_i^*(jk..x) = 1$ (as defined in Chapter 6)

T = duration (number of stimuli) of the training sequence

$v_{ir}(T)$ = value of the connection from the i^{th} A-unit after the training sequence

$c_{ir}^*(x)$ = $c_{ir}^*(x,T) = a_i^*(x)\, v_{ir}(T)$ = signal received by the R-unit on connection c_{ir} when test stimulus S_χ is shown after the training sequence. The time T will be understood unless otherwise specified.

$u_\chi = \alpha_r^\chi(T) = \sum_i c_{ir}^* (\chi)$ = total input to the response unit when S_χ is shown after the training sequence. For present purposes, the symbol u_χ will be used, as in Chapter 5. Time T is understood unless otherwise specified.

In terms of these symbols, the reinforcement rule for a quantized α -system, with S -controlled reinforcement, can be represented by the following expression for the change in v_{ir} when stimulus S_j is shown:

$$\Delta v_{ir} = \rho_j(\alpha_i^*(j))$$

7.1.2 Fixed Sequence Experiments: Analysis

The first case to be considered is that of a fixed training sequence, in which a definite sequence of stimuli (S_1 , S_2 , . . . , S_T) is shown to the perceptron. In a later section, random training sequences will be considered. The fixed sequence consists of a fixed (though not necessarily equal) number of showings of each stimulus. For α -perceptrons, the order of occurrence of these stimuli does not affect the results. All values v_{ir} are assumed to be zero initially. The following analysis and theorem follow the treatment of Joseph (Ref. 41).

If a given perceptron is shown a training sequence, it will place a test stimulus S_χ in the positive class if u_χ is greater than zero, and in the negative class if u_χ if less than zero. For the given perceptron, training sequence, and test stimulus, u_χ is a determinate number.

Over the class of perceptrons, however, u_χ is a random variable. In order to determine the probability that a perceptron from the specified class will classify S_χ correctly, we must know the probability that u_χ has the correct sign. In order to obtain a conservative bound on the probability of correct response to S_χ, without making any assumptions about the distribution of u_χ, Joseph makes use of the Tchebysheff inequality, which states that for any random variable z with mean μ and variance σ^2,

$$\text{Prob}\left\{z > 0\right\} \geq 1 - \frac{1}{\mu^2/\sigma^2} \quad \text{if} \quad \mu > 0$$

$$\text{Prob}\left\{z < 0\right\} \geq 1 - \frac{1}{\mu^2/\sigma^2} \quad \text{if} \quad \mu < 0$$

Consequently, if the ratio $\mu^2(u_\chi)/\sigma^2(u_\chi)$ can be made arbitrarily large, the probability that u_χ for a randomly selected perceptron will agree in sign with its expected value over the class of perceptrons can be made arbitrarily close to 1*. It thus becomes important, first of all, to know whether or not the expected value of u_χ has the proper sign.

---

* Joseph, has pointed out that if the one-sided inequality $P_n\{z-\mu \geq 1\} \leq \frac{\sigma^2}{1+\sigma^2}$ is used in place of the two-sided inequality $P_n\{|z-\mu| \geq 1\} \leq \sigma^2$, slightly sharper bounds may be achieved, i.e.,

$$P_n\{z > 0\} \leq 1 - \frac{1}{1+\mu^2/\sigma^2} \quad \text{if} \quad \mu > 0$$

$$P_n\{z < 0\} \leq 1 - \frac{1}{1+\mu^2/\sigma^2} \quad \text{if} \quad \mu < 0$$

In the range of interest, this additional sharpness is insignificant..

DEFINITION: S_χ will be called a positive stimulus (with respect to a class of perceptrons, an environment, classification, and training sequence) if the expected value of u_χ agrees in sign with the assigned class of S_χ .

In terms of the symbols introduced above, S_χ is a positive stimulus if

$$\rho_\chi \cdot E(u_\chi) > 0$$

The expected value of u_χ for an α-perceptron (assuming that all A-R unit connections start out with zero value) is obtained as follows. Let P_j = the number of times stimulus S_j occurs in the training sequence, divided by T , the total number of stimuli in the sequence (i.e., the proportion of the training sequence which is S_j). Then the value of the connection from unit a_i at the end of the training sequence will be (since the magnitude of η is taken to be 1)

$$v_{ir} = T \sum_j \rho_j P_j \, a_i^*(j) \tag{7.1}$$

where the sum is over all stimuli in W . Consequently, summing over all A-units, the input signal to the response unit when the test stimulus S_χ occurs will be

$$u_\chi = T \sum_i \sum_j \rho_j P_j \, a_i^*(j\chi) = \sum_i c_{ir}^*(\chi) \tag{7.2}$$

The expected value of u_χ is therefore given by

$$\begin{aligned} E u_\chi &= E\left\{T \sum_i \sum_j \rho_j P_j \, a_i^*(j\chi)\right\} \\ &= T \sum_i \sum_j \rho_j P_j \, E\, a_i^*(j\chi) \\ &= T N_a \sum_j \rho_j P_j \, Q_{j\chi} \end{aligned} \tag{7.3}$$

From the above definition, it follows that S_χ is a positive stimulus (and will tend to be correctly classified) if

$$\sum_j \rho_j \rho_\chi P_j Q_{j\chi} > 0$$

From Equation (7.3) it is clear that $E u_\chi$ increases linearly with N_a. Let us now consider the variance of u_χ. This is obtained from the equation:

$$\sigma^2(u_\chi) = \sum_i \sigma^2(c^*_{ir}(\chi)) + \sum_i \sum_{i' \neq i} \text{cov.}\left[c^*_{ir}(\chi), c^*_{i'r}(\chi)\right] \qquad (7.4)$$

For the conditions currently being considered (an α-system with a predetermined training sequence) the only source of variability in $c^*_{ir}(\chi)$ is in the selection of the origin point configuration of the unit a_i. But if we assume (as in all models thus far considered) that the A-units are all chosen independently from a distribution of admissible origin configurations, the covariances will all be zero, and $\sigma^2(c^*_{ir}(\chi))$ does not depend on i. Therefore, the general equation (7.4) reduces to

$$\sigma^2(u_\chi) = N_a \sigma^2(c^*_{ir}(\chi)) = N_a\left[E c^{*2}_{ir}(\chi) - E^2 c^*_{ir}(\chi)\right] \qquad (7.5)$$

(See Rosenblatt, Ref. 79, pp. 82-83, for a more detailed algebraic discussion of this equality). Now, for an α-system,

$$c^*_{ir}(\chi) = \tau \sum_j \rho_j P_j a^*_i(j\chi)$$

and

$$c^{*2}_{ir}(\chi) = \tau^2 \sum_j \sum_k \rho_j \rho_k P_j P_k a^*_i(j k \chi)$$

This yields, for the required expected values in (7.5),

$$E\, e_{ir}^{*2}(x) = T^2 \sum_j \sum_k \rho_j \rho_k P_j P_k Q_{jkx}$$

and

$$E^2 e_{ir}^{*}(x) = T^2 \sum_j \sum_k \rho_j \rho_k P_j P_k Q_{jx} Q_{kx}$$

Substituting in (7.5) and simplifying, this yields

$$\sigma^2(u_x) = N_a T^2 \sum_j \sum_k \rho_j \rho_k P_j P_k (Q_{jkx} - Q_{jx} Q_{kx}) \tag{7.6}$$

Note that the variance depends on Q_{jkx}, while the expected value depends only on Q_{jx}. This variance, like the expected value, is of the order of N_a. We are now in a position to prove the following theorem (due to Joseph):

THEOREM: Given a class of elementary α-perceptrons, a finite stimulus world W, a classification $C(W)$, and a training sequence; then for every $\epsilon > 0$, there exists an $N_0(\epsilon)$ such that if $N_a > N_0(\epsilon)$, the probability of selecting a perceptron which will correctly identify the class of every positive stimulus will be greater than $1 - \epsilon$.

PROOF: From the Tchebyscheff inequality, we have seen that if $\mu^2(u_x)/\sigma^2(u_x)$ can be made arbitrarily large, the probability that u_x will agree in sign with its expected value over the class of perceptrons will approach unity.

It has also been demonstrated (Equations 7.3 and 7.6) that both $\mu(u_\chi)$ and $\sigma^2(u_\chi)$ are of the order of N_a; therefore, $\mu^2(u_\chi)/\sigma^2(u_\chi)$ will be of the order of N_a. Thus, for each positive stimulus, S_χ, the probability that u_χ agrees in sign with Eu_χ can be made arbitrarily clost to 1 by choosing N_a sufficiently large. Suppose there are n stimuli in W. Then, for the j^{th} positive stimulus there exists a quantity $N_j(\epsilon)$ such that if $N_a > N_j(\epsilon)$, the probability of selecting a perceptron which fails to correctly identify S_j will be less than ϵ/n. If we let $N_0(\epsilon) = \max_j N_j(\epsilon)$, the condition required by the theorem is satisfied. Q.E.D.

From Equations (7.3) and (7.6), it is seen that for a given set of stimulus frequencies p_j, the ratio μ^2/σ^2 does not depend on T. Thus any number of repetitions of the same training sequence can occur without affecting the performance of the system. Since μ^2/σ^2 varies linearly with N_a, the normalized ratio $\frac{1}{N_a}\mu^2/\sigma^2$ forms a convenient measure for the comparison of different perceptron models. Some numerical values for typical cases will be considered in the following section.

While the above analysis permits us to obtain a rigorous lower bound for the probability of correct identification of S_χ by a randomly selected perceptron, it does not actually yield an estimate of this probability. In order to estimate the probability of correct identification of S_χ, it will be assumed that u_χ is normally distributed. The justification for this assumption was discussed in Rosenblatt, Ref. 79, and subsequent analysis has shown that the approximation is very close, even for perceptrons with a

small number of A-units. Assuming a normal distribution, we have for the probability of a positive response to S_χ :

$$P = P(r_\chi^* | S_\chi) = \Phi\left(\frac{\mu(u_\chi)}{\sigma(u_\chi)}\right) \tag{7.7}$$

where $\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{\chi^2}{2}} d\chi$

Note that the above equations do not depend on whether the perceptron is constructed according to the binomial model, Poisson model, or any other other model, so long as the A-units are selected independently of one another. The performance does depend on the Q -functions, however, which will be different for different models. From equation 7.3 it is clear that any stimulus S_χ will tend to be classified correctly if the average value of $Q_{j\chi}$ for S_j in the same class as S_χ is greater than the average value of $Q_{j\chi}$ for S_j in the opposite class from S_χ . (If the frequencies P_j are not all equal, each $Q_{j\chi}$ must be multiplied by its appropriate frequency in obtaining these averages.) From the analysis of Q -functions in the preceding chapter, it is clear that this condition will generally be met if the stimuli of each class have large intersections with one another (on the retina) while stimuli from opposite classes have small intersections with one another. The ideal situation would consist of two disjoint clusters of stimuli, located in different parts of the retinal field, each cluster representing one class. In order to discriminate two stimuli reliably (i.e., to assign them to opposite classes) it is desirable that $Q_{i,j}$ for the two stimuli should be small, and particularly that the conditional probabilities $Q_{i|j}$ and $Q_{j|i}$ should be as small as possible. Figure 10,

in the last chapter, shows that this condition can readily be met if the stimuli have a small intersection with one another, but becomes increasingly difficult to meet as the intersection increases. This figure also shows that a binomial model is better suited to the discrimination of similar stimuli than a Poisson model, where $Q_{i|j}$ is apt to be relatively large even for disjoint stimuli.

7.1.3 Fixed Sequence Experiments: Examples

The environment which was considered in the last section of Chapter 6, involving twenty horizontal bars and twenty vertical bars on a 20 by 20 toroidally connected retina is a convenient one to use for a "calibration experiment", by which different classes of perceptrons can be compared. In particular, consider the following discrimination experiment:

EXPERIMENT 1: Given a perceptron with 400 sensory points arranged in a 20 by 20 toroidally connected array, or "retina", let W consist of the twenty possible 4 by 20 horizontal bars, and the twenty possible 4 by 20 horizontal bars. Let $C(W)$ be a classification which assigns every horizontal bar to the positive class, and every vertical bar to the negative class. Show every bar in W to the perceptron exactly once (or in a sequence with P_j equal for all stimuli). During this training sequence, the perceptron is reinforced with S -controlled reinforcement. Then select one of the bars, S_α, and determine whether the response is correct, according to $C(W)$.

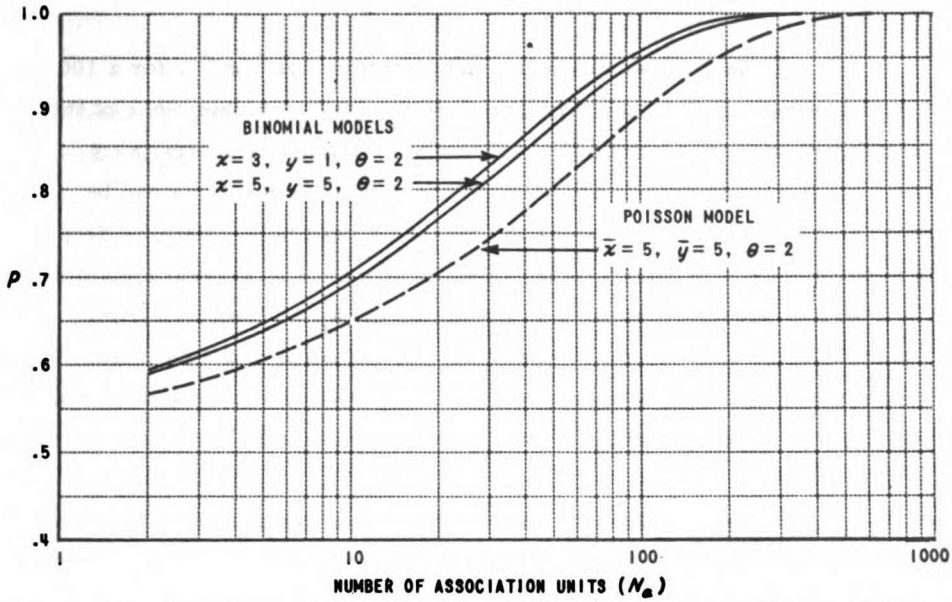


Figure 13 PROBABILITY OF CORRECT INDENTIFICATION OF A TEST STIMULUS BY AN ELEMENTARY α-PERCEPTRON, IN EXPERIMENT 1 (CURVES ALSO APPLY TO γ'-PERCEPTRONS; SEE CHAPT. 8)

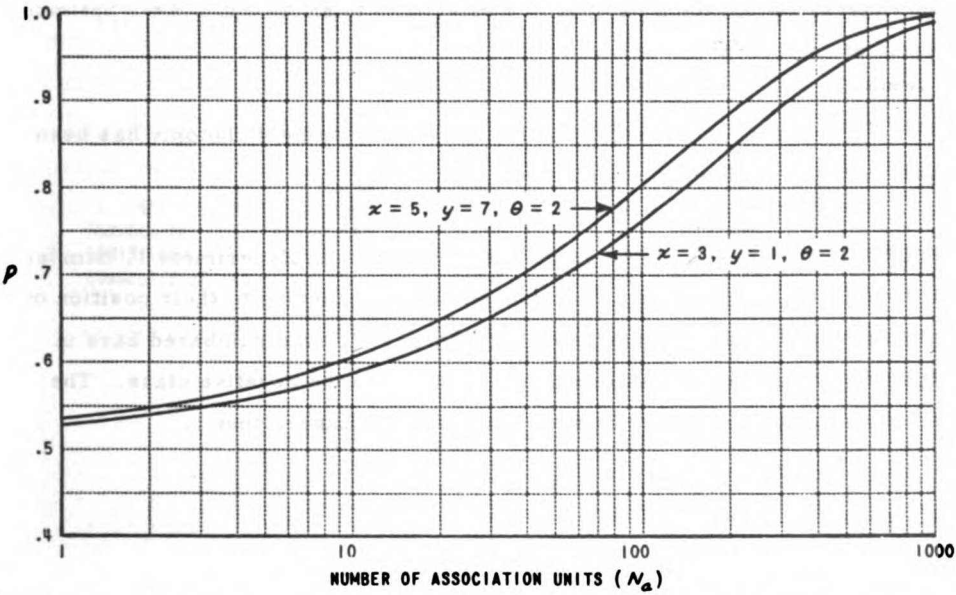


Figure 14 PROBABILITY OF CORRECT INDENTIFICATION OF A TEST STIMULUS BY AN ELEMENTARY α-PERCEPTRON, IN EXPERIMENT 2 (FOR TWO BINOMIAL MODELS). CURVES ALSO APPLY TO γ'-PERCEPTRONS (SEE CHAPT. 8)

Table 3 shows the performance ratios, μ^2/σ^2, for a 100 A-unit binomial model α-perceptron, with various combinations of the parameters x and y ($\theta = 2$ in all cases). The parameters $x = 3$, $y = 1$, $\theta = 2$, appear to be optimum for this experiment, as can be seen from the table. (Increasing the threshold results in a definite drop in performance.) Figure 13 shows the performance of several binomial and Poisson model perceptrons as a function of N_a, computed from Equation (7.7). The top curve shows the performance of the optimum (binomial) system. A comparison of the other two curves illustrates the relatively poor performance of the Poisson model on this particular problem.

It should be emphasized that the parameters found to be optimum in this experiment will not necessarily turn out to be optimum in other environments, or other classifications. In general, it appears that as the classes of patterns to be discriminated become more "similar", (i.e., as the maximum possible overlap between stimuli from opposite classes increases) the optimum number of connections to an A-unit and the optimum value of θ tend to increase.

A more difficult classification of the same dichotomy has been studied in the following experiment:

EXPERIMENT 2: With the same environment as in Experiment 1, number the horizontal and vertical bars consecutively according to their position on the retina. Let the classification $C(W)$ place all even numbered bars in the positive class, and all odd numbered bars in the negative class. The training and testing procedures are identical to Experiment 1.

TABLE 3

PERFORMANCE RATIOS $\left(\frac{\mu^2(u_x)}{\sigma^2(u_x)}\right)$ FOR 100-A-UNIT ELEMENTARY α-PERCEPTRONS (BINOMIAL MODEL) FOR EXPERIMENT 1 (HORIZONTAL/VERTICAL BAR DISCRIMINATION, FIXED SEQUENCE). $\theta = 2$ IN ALL CASES.

		x (NUMBER OF EXICITATORY CONNECTIONS PER A-UNIT)			
		2	3	4	5
y (NUMBER OF INHIBITORY CONNECTIONS PER A-UNIT)	0	2.474	2.831	1.540	.931
	1	2.063	2.912	2.104	1.349
	2	1.708	2.805	2.479	1.773
	3	1.406	2.592	2.670	2.140
	4	1.153	2.329	2.708	2.414
	5	.941	2.066	2.630	2.579
	6	.767	1.777	2.473	2.638
	7	.623	1.523	2.271	2.605

TABLE 4

PERFORMANCE RATIOS FOR 100-A-UNIT ELEMENTARY α-PERCEPTRONS (BINOMIAL MODEL) FOR EXPERIMENT 2. $\theta = 2$ IN ALL CASES.

		x (NUMBER OF EXCITATORY CONNECTIONS)			
		2	3	4	5
y (NUMBER OF INHIBITORY CONNECTIONS)	0	.358	.426	.328	.274
	1	.365	.502	.436	.363
	2	.362	.551	.526	.451
	3	.350	.578	.596	.533
	4	.333	.585	.646	.605
	5	.310	.578	.677	.664
	6	.285	.558	.690	.707
	7	.268	.529	.688	.736

In this case, the two most similar bars to any test bar (those which overlap it by 3/4 of its area on either side) are invariably in the opposite class. Nonetheless, all stimuli may be positive stimuli under these conditions, with a suitable choice of parameters. Table 4 shows the ratio μ^2/σ^2 for a 100 unit system in this experiment. Figure 14 shows the performance of a perceptron with the same parameters as before ($\alpha = 3$, $\gamma = 1$, $\theta = 2$) on this experiment, and also with the best parameters found to date ($\alpha = 5$, $\gamma = 7$, $\theta = 2$). These parameters are the best set for $\alpha \leq 5$ and $\gamma \leq 7$, but are probably not optimum, as it seems likely that a further increase in both α and γ would yield a further improvement in performance.

7.1.4 Random Sequence Experiments: Analysis

For the analysis of the performance of perceptrons trained with random stimulus sequences, it is convenient to make use of an unnormalized G-matrix (see footnote, page 75), where $\eta = 1$ instead of $1/N_a$. For such a matrix, in the α-system, g_{ij} = the number of units active for both S_i and S_j, or

$$g_{ij} = \sum_k a^*_k(i,j) \tag{7.8}$$

The mathematical properties of the unnormalized G-matrix are no different from those discovered for the normalized matrix, in Chapter 5.

In a random sequence experiment, the training sequence is assumed to consist of a series of T stimuli, in which each stimulus in the series is selected independently of the others. The probability of

selecting stimulus S_j for the t^{th} position in the sequence is p_j, for all t. We will let m_j = the number of times stimulus S_j occurs in the training sequence. The random vector $\vec{m} = (m_1, m_2 .. m_n)$ will have a multinomial distribution with T trials and probability vector $\vec{p} = (p_1, p_2, \ldots, p_n)$. The training sequence selected is assumed to be independent of the particular perceptron selected for a given experiment. At the end of the training sequence, the input to the R-unit in response to a test stimulus S_x will be

$$u_x = \sum_j p_j m_j g_{xj}$$

$$= \sum_i \sum_j p_j m_j a_i^*(jx)$$

Therefore, the expected value over perceptrons and training sequences is

$$E(u_x) = T N_a \sum_j p_j p_j Q_{jx} \tag{7.9}$$

which is of the order of TN_a. Note that this is identical to equation (7.3).

The variance over both perceptrons and training sequences is given by

$$\sigma^2(u_x) = \sum_j \sigma^2(m_j g_{xj}) + \sum_j \sum_{k \neq j} p_j p_k \text{ cov.}(m_j g_{xj}, m_k g_{xk})$$

$$= \sum_j \left[E(m_j^2) E(g_{xj}^2) - E^2(m_j) E^2(g_{xj}) \right]$$

$$+ \sum_j \sum_{k \neq j} p_j p_k \cdot \left[E(m_j m_k) E(g_{xj} g_{xk}) - E(m_j) E(m_k) E(g_{xj}) E(g_{xk}) \right]$$

(7.10)

For the components of the multinomially distributed vector $\vec{m}$ we have

$$E(m_j) = T p_j$$

$$E(m_j^2) = T(T-1) p_j^2 + T p_j$$

$$E(m_j m_k) = T(T-1) p_j p_k$$

Let $n_{ij..z}$ = number of A-units active for stimuli S_i, $S_j, \ldots, S_z$. The symbol $\sim$ over a subscript will be used to denote negation (e.g., $n_{j\tilde{k}}$ = the number of A-units active for stimulus S_j but <u>not</u> for S_k; $n_{j\tilde{k}} = n_j - n_{jk}$). From equation 7.8, it is clear that for the α-system, $n_{ij} = g_{ij}$. Now, any set of n's which is exhaustive (every A-unit counted in at least one $n_{ij..z}$), and such that each A-unit is counted in no more than one $n_{ij..z}$, will have a multinomial distribution. From this it follows that

$$E(g_{xj}) = N_a Q_{xt}$$

$$E(g_{xj}^2) = N_a(N_a-1) Q_{jx}^2 + N_a Q_{jx}$$

$$E(g_{xj}\, g_{xk}) = E\left[(n_{jkt} + n_{j\tilde{k}x})(n_{jkx} + n_{\tilde{j}kx})\right]$$

$$= E(n_{jkx}^2) + E(n_{jkx}\, n_{j\tilde{k}x}) + E(n_{jkx}\, n_{\tilde{j}kx}) + E(n_{j\tilde{k}x}\, n_{\tilde{j}kx})$$

$$= N_a Q_{jkx} + N_a(N_a-1)\left[Q_{jkx}^2 + Q_{jkx}(Q_{jx} - Q_{jkx})\right.$$

$$\left. + Q_{jkx}(Q_{kx} - Q_{jkx}) + (Q_{jx} - Q_{jkx})(Q_{kx} - Q_{jkx})\right]$$

$$= N_a Q_{jkx} + N_a(N_a-1) Q_{jx} Q_{kx}$$

Substituting in (7.10), this yields

$$\sigma^2(u_x) = TN_a \sum_j p_j Q_{jx} \left[(N_a - 1) Q_{jx} + 1 \right]$$

$$+ TN_a \sum_j \sum_k p_j p_k p_j p_k \left[(T-1) Q_{jkx} - (T + N_a - 1) Q_{jx} Q_{kx} \right] \qquad (7.11)$$

The variance of u_x is therefore on the order of $TN_a^2 + T^2 N_a$, at maximum. Since the square of the mean is on the order of $T^2 N_a^2$, the ratio μ^2/σ^2 becomes indefinitely large as N_a and T both increase, and the Theorem stated in Section 7.1.2 is seen to hold for random training sequences of sufficient length, as well as fixed sequences. As the length of the training sequence, T, increases, the relative frequencies m_i/T will approach the probabilities p_i, and the performance of the system will approach the performance in a fixed sequence experiment. As N_a goes to infinity, the ratio μ^2/σ^2 approaches

$$T \left[\sum_j p_j p_j Q_{jx} \right]^2 \Big/ \sum_j p_j Q_{jx}^2$$

7.1.5 Random Sequence Experiments: Examples

As a "calibration experiment" for comparing different systems, the horizontal vs. vertical bar discrimination problem is particularly convenient. The random sequence version of the experiment is as follows:

EXPERIMENT 3: For the same conditions and classification as Experiment 1, show the perceptron a random sequence of horizontal and vertical bars, each bar occurring with equal frequency ($p_j = 1/40$ for all bars). During this training sequence, S-controlled reinforcement is used, and the performance of the perceptron for an arbitrary bar, S_x , is then determined as before.

Figure 15 shows the performance of binomial model α -perceptrons of three different sizes on this problem, as a function of the length of the training sequence (T). The parameters x , y , and θ are the optimum values (3, 1, 2) found in Section 7.1.3. Further increases in N_a will not appreciably improve performance in this experiment.

The effect of a "frequency bias" on α -system perceptrons is illustrated in the following experiment:

EXPERIMENT 4: The conditions and classifications are the same as in Experiment 3, but the horizontal bars occur four times as frequently as the vertical bars; i.e., $p_j = .04$ for horizontal bars and $.01$ for vertical bars.

Figure 16 shows the performance of a 100 A-unit system on this experiment. The upper curve shows the probability of correctly identifying a horizontal bar, and the lower curve shows the probability of correctly identifying a vertical bar. The correct response to vertical bars is actually suppressed as training increases, due to the greater frequency of horizontal bars. The

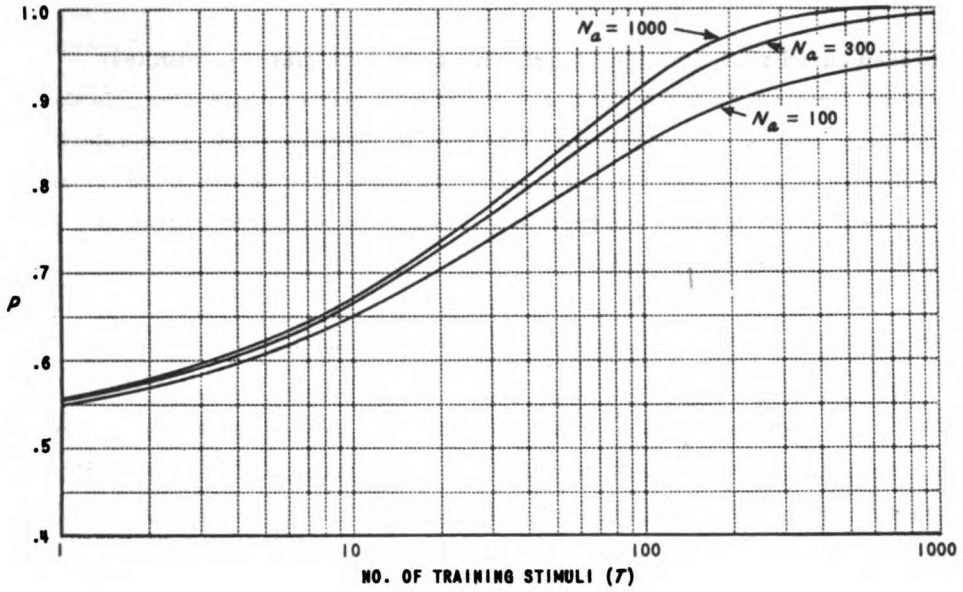


Figure 15 PROBABILITY OF CORRECT INDENTIFICATION OF TEST STIMULUS BY BINOMIAL α-PERCEPTRONS IN EXPT. 3 (RANDOM SEQUENCES) ($\chi = 3$, $y = 1$, $\theta = 2$)

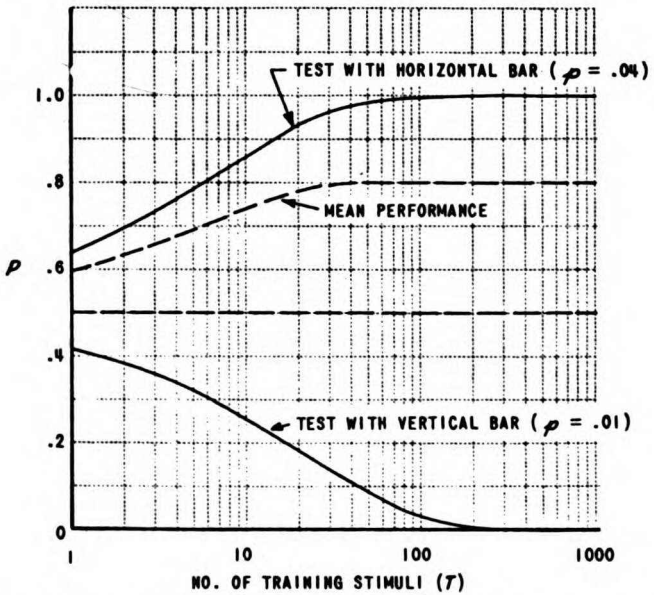


Figure 16 PROBABILITY OF CORRECT IDENTIFICATION OF TEST STIMULI IN EXPT. 4. BINOMIAL α-PERCEPTRON WITH N_a = 100, $\chi = 3$, $y = 1$, $\theta = 2$. P_i = .04 FOR HORIZONTAL BARS; .01 FOR VERTICAL BARS

broken curve shows the mean performance on both classes, with test stimuli drawn from each class with their appropriate frequencies. In the following chapter, it will be seen that this performance can be considerably improved in a $\mathcal{G}$ -system perceptron. It would also be improved for an α -perceptron if error correction training were employed instead of S-controlled reinforcement.

7.2 Discrimination Experiments with Error Correction Procedures

The analysis and experiments in the preceding section deal with S-controlled reinforcement experiments. In Chapter 5, Theorem 6, it was shown that this procedure cannot be guaranteed to yield a solution to a classification problem, even though a solution may exist, whereas an error correction procedure will always yield a solution if any solutions exist. The error correction procedure would therefore seem to be the method of choice in training a perceptron to discriminate between two classes of stimuli. Unfortunately, the type of analysis which was carried out for S-controlled experiments is not readily performed with error-correction experiments. Consequently, all data on learning curves for error correction procedures come from one of two sources: simulation on a digital computer*, and performance of actual experiments on the Mark I perceptron at the Cornell Aeronautical Laboratory (Refs. 29, 30, 31).

---

* Experiments performed by Carl Kesler on the Burroughs 220 computer at Cornell University.

Two main sets of experiments will be described here, the first with binomial model α-perceptrons, and the second with perceptrons having additional constraints imposed on their S to A-unit connections.

7.2.1 Experiments with Binomial Models

The following four experiments have been performed with binomial model perceptrons (having fixed numbers of sensory connections to each A-unit, with origins located at random in the sensory mosaic):

EXPERIMENT 5: The environment of horizontal and vertical bars used in Experiment 1 is employed, and the stimuli occur in fixed sequence, first showing all horizontal bars in fixed sequence, then all vertical bars, and repeating the sequence until perfect performance is achieved. The error correction procedure is employed, and the performance is tested at the end of each sequence.

EXPERIMENT 6: The same environment and training procedure is employed as above, but the stimuli occur in a random sequence, with $p_i = 1/40$ for each stimulus (as in Experiment 3).

EXPERIMENT 7: The environment consists of a set of triangles in all possible positions on a toroidally connected 20 by 20 retina, and a set of squares in all possible positions on the retina. The triangles and squares each cover 80 of the 400 retinal points. The sequence is random, as in Experiment 6, with $p_i = 1/800$ for each stimulus. (The set of possible stimuli is generated by translations of a standard image; rotations are not permitted.)

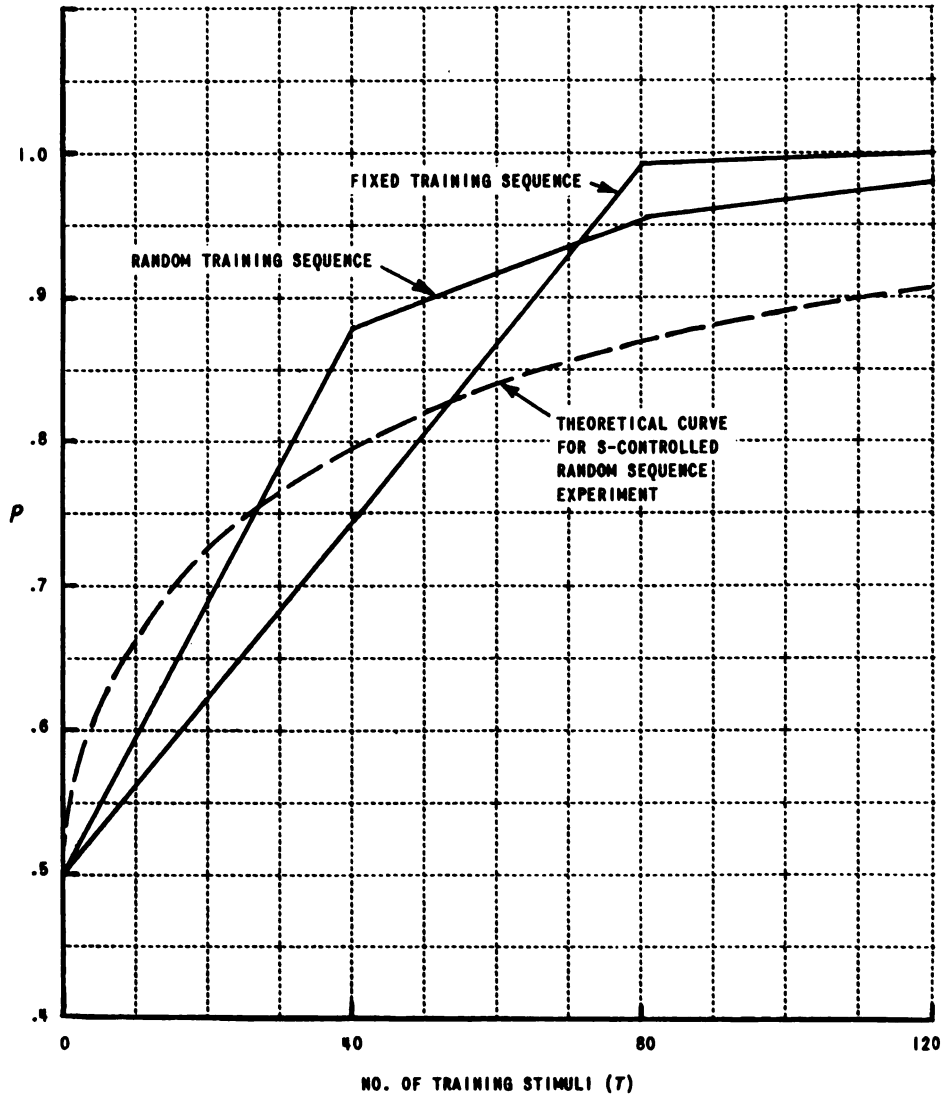


Figure 17 PERFORMANCE OF BINOMIAL α-PERCEPTRONS IN EXPERIMENTS 5 AND 6 (HORIZONTAL / VERTICAL BAR DISCRIMINATION WITH ERROR CORRECTION PROCEDURE). SOLID CURVES SHOW MEAN PERFORMANCE OF 25 PERCEPTRONS, WITH $N_a = 300$, $\alpha = 3$, $y = 1$, $\theta = 2$

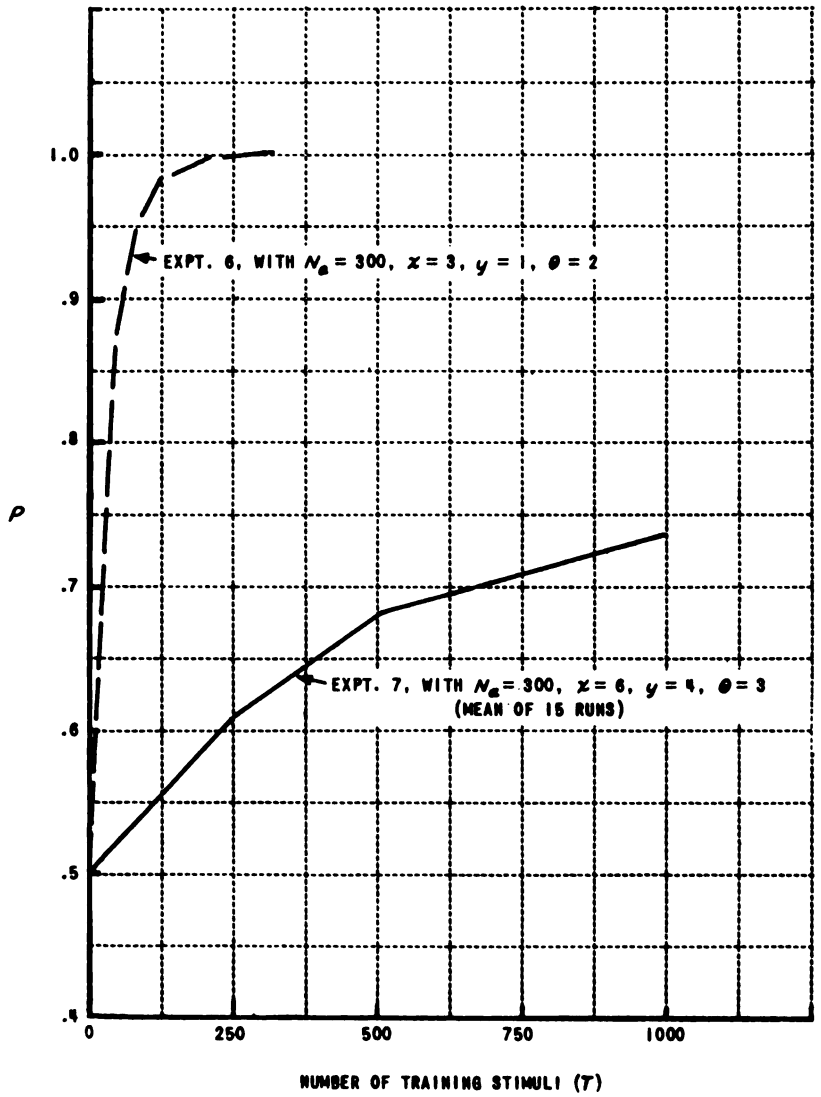


Figure 18 PERFORMANCE OF BINOMIAL α-PERCEPTRONS IN SQUARE / TRIANGLE DISCRIMINATION (EXPT. 7) COMPARED WITH HORIZONTAL / VERTICAL BAR DISCRIMINATION (EXPT. 6)

EXPERIMENT 8: The horizontal/vertical bar environment is employed, as in Experiment 6, with stimuli occurring in random sequence. A random sign correction procedure is employed for training the perceptron (see Definition, Section 5.6).

Figure 17 shows the results of Experiments 5 and 6, and includes a theoretical learning curve for an S-controlled experiment for comparison. The experimental curves show the mean performance for a set of 25 binomial perceptrons with 300 A-units, and the optimum parameters ($x = 3$, $y = 1$, $\theta = 2$) found in the preceding section. The same 25 perceptrons were employed in Experiments 5 and 6. It appears to be characteristic that a random training sequence leads to a more rapid learning rate initially, but is overtaken by the fixed sequence performance as the duration of training increases. Note that in both cases, the error correction method yields considerably better performance than the S-controlled method.

Figure 18 shows the mean performance of a set of 15 perceptrons on Experiment 7. The parameters are $N_a = 300$, $x = 6$, $y = 4$, $\theta = 3$. These were the best parameters tested, but are probably not optimum. The learning curve for the horizontal/vertical bar experiment (Experiment 6) is shown as a broken line for comparison. The slow learning rate in this experiment is largely due to the large number of distinct stimuli in the environment (800) compared to the number in the horizontal/vertical bar environment (40). The increased number of stimuli means that a much longer training sequence is required to guarantee a representative sample of all stimuli, with a reasonably uniform coverage of the retinal field. A further difficulty is introduced by the fact that the maximum overlap of a square and triangle is much greater than the maximum overlap of a horizontal and vertical bar, making the discrimination intrinsically more difficult.

Figure 19 shows a comparison of the performance of 10 perceptrons on Experiment 8 with the performance of the same 10 perceptrons on Experiment 6. In Experiment 8, the learning is not only much slower, but the variability between perceptrons is greatly increased. Of the ten perceptrons tested, two achieved perfect performance during the period of the experiment, which was discontinued after 2000 training stimuli. Nonetheless, each of the ten perceptrons would ultimately achieve perfect performance if the experiment were continued (due to Theorem 5, Section 5.6). With the directed error correction procedure, all ten perceptrons achieved perfect performance within 300 training stimuli.

While the performance of an elementary perceptron with the random sign procedure is clearly unsatisfactory for practical systems, it should be noted that the existence of a consistent bias in the proper direction still makes this a plausible component of a more reliable mechanism. If a "majority mechanism" is employed (e.g., a threshold device which responds to the difference of positive and negative signals from R-units) to determine the "majority vote" of n such elementary perceptrons, connected independently to the same retina, a highly reliable system would result. The error probability of this system would be:

$$P_\epsilon = \sum_{k=0}^{[n/2]} \binom{n}{k} p^k (1-p)^{n-k}$$

when p is the probability of correct response for a single perceptron (as shown in Figure 19).

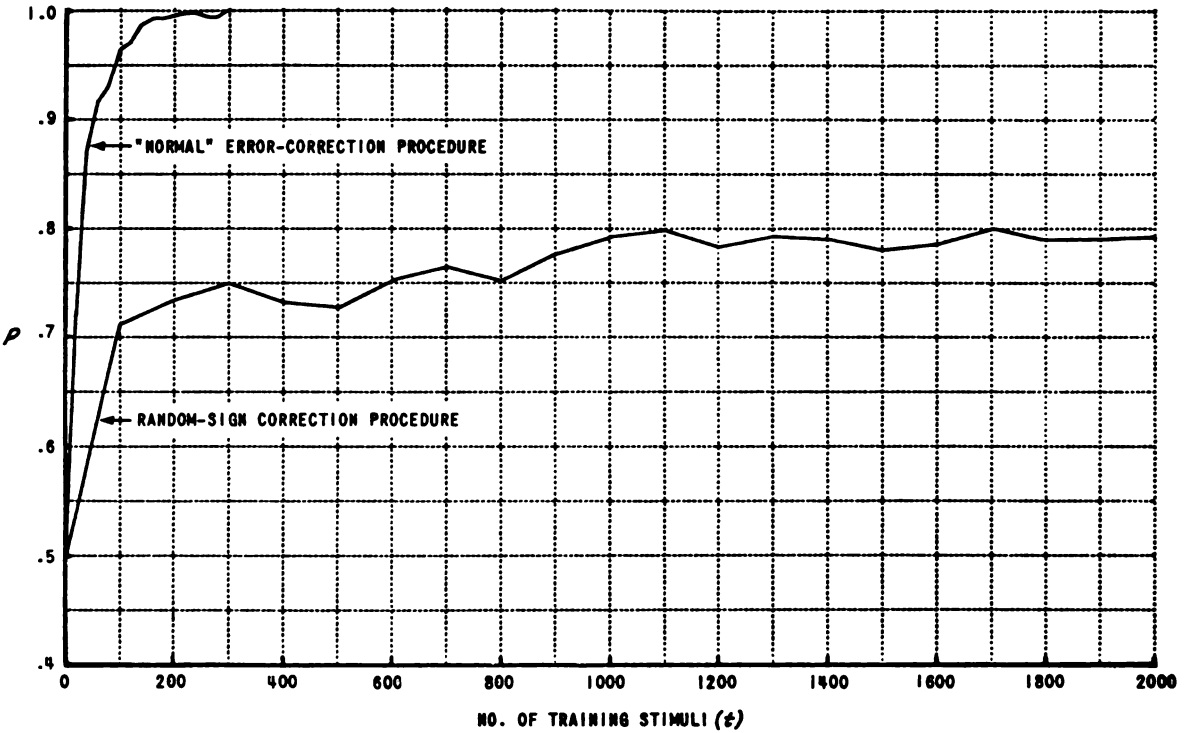


Figure 19 COMPARISON OF RANDOM-SIGN CORRECTION PROCEDURE (EXPT. 8) WITH "NORMAL" ERROR-CORRECTION PROCEDURE (EXPT. 6). SHOWS MEANS OF 10 BINOMIAL α-PERCEPTRONS WITH $N_a = 300$, $\chi = 3$, $y = 1$, $\theta = 2$

While the actual learning curve for error correction experiments cannot at present be stated analytically, R. D. Joseph has obtained an upper bound for the number of corrective reinforcements that must be applied, where a solution exists. In the proof of Theorem 4, Chapter 5, it was noted that an upper bound for the number of corrective reinforcements can be expressed in terms of the quantity α, as follows:

$$N_{max} \leq \frac{(\ell + M\sqrt{n})^2}{\alpha M} \tag{7.12}$$

where M = maximum diagonal element of the G-matrix,

α = minimum of the function $F(x) = x'Hx / \|x\|^2$ (as defined for Theorem 4, Chapter 5).

ℓ = $\|Hx^*\|$ (as in Theorem 4, Chapter 5).

For the case which is of primary interest here, the process starts from the origin, so that $\ell = \|Hx^*\| = 0$. In this case, (7.12) simplifies to

$$N_{max} \leq Mn/\alpha$$

7.2.2 Experiments with Constrained Sensory Connections

In all perceptrons considered thus far, connections from S-units to A-units have had their origins randomly chosen from the set of all sensory points, with equal probability. Such models will be called uniform input distribution models (u.i.d. models). It has occasionally been proposed that the performance of a perceptron might be considerably improved by the

introduction of special constraints on the admissible origin point connections. For example, the retinal connections could be made to resemble biological systems more closely by assigning a "retinal field" to each A-unit, and limiting its choice of origin points to S-units within this field. A similar procedure would be to construct a network of connections by assigning a center at random to each A-unit, somewhere on the retina, and selecting connections from a circular normal distribution about this center. Such systems will be called normal input distribution models (n.i.d. models). Further constraints might lead ultimately to specialized A-units, whose input configurations are specially designed to make them responsive to stimuli of particular shapes, or configuration properties. We will consider one further constraint in this section: the case in which the excitatory and inhibitory connections to an A-unit are assigned distinct centers on the retina, with origins selected from a circular normal distribution about these centers. This will be called the divided input distribution (d.i.d.) model. The n.i.d. model can be considered a special case of the d.i.d. model in which the excitatory and inhibitory centers and dispersions are identical.

In the general d.i.d. model, A-units are characterized by seven parameters: x , y and θ as before, the expected distance between excitatory and inhibitory centers (ED), the standard deviation of this distance (σD), and the standard deviations of the normal probability distributions about the excitatory and inhibitory centers (σx and σy). A number of experiments have been performed with such models in an attempt to discover what sort of improvement might be achieved by an optimum set of constraints on the sensory connections.

Experiments 6 and 7 have been used for the study of constrained input distributions. In the square/triangle discrimination experiment (Experiment 7) the performance of the d.i.d. models never showed any improvement over the original u.i.d. model. A large number of combinations of x , y , and θ were tested with various distribution parameters, in an attempt to find the optimum system for $x + y \leq 10$. The best performance was obtained for a set of 15 perceptrons with $x = 6$, $y = 4$, $\theta = 3$, $ED = 0$, $\sigma D = 0$, $\sigma x = 7$, and $\sigma y = 7$. This is equivalent to an n.i.d. model with the same centers for excitatory and inhibitory distributions, and $\sigma = 7$. The performance of this system did not differ from that of the equivalent u.i.d. model by more than 1% at any point on the learning curve, and was within 1/4% of the u.i.d. performance at most of the points tested. The same stimulus sequences were used for both models in order to make conditions as closely comparable as possible. These results suggest that for large but spatially concentrated stimulus patterns, little advantage is to be gained in an elementary perceptron by imposing radial constraints on the origin point configurations.

In the case of the horizontal/vertical bar discrimination (Experiment 6) a slight advantage was found for the d.i.d. model for the parameters $x = 1$, $y = 9$, $\theta = 1$, $ED = 12$, $\sigma D = 2$, $\sigma x = 2$, $\sigma y = 4$. On the basis of a number of simulation experiments, this appears to be close to an optimum configuration for the d.i.d. model for this experiment. Figure 20 shows the results obtained from 25 runs with these parameters, compared with 25 u.i.d. models with optimum parameters $(x = 3, y = 1, \theta = 2)$ using the identical training sequences. The difference, although slight, appears to be statistically significant.

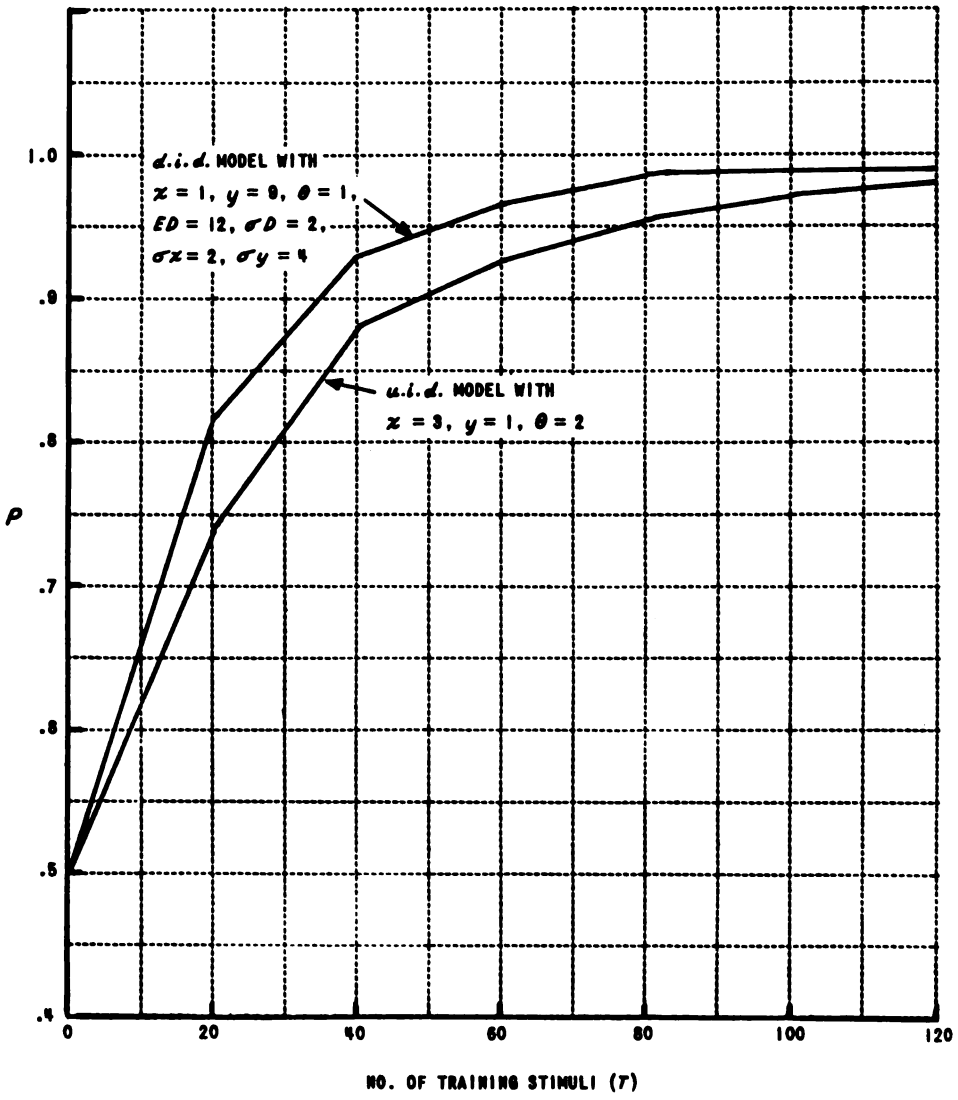


Figure 20 COMPARISON OF OPTIMUM *d.i.d.* AND *u.i.d.* MODELS IN HORIZONTAL / VERTICAL BAR DISCRIMINATION (EXPT. 6). CURVES SHOW MEANS OF 25 RUNS

The general conclusion from these experiments seems to be that (for large stimuli) little is to be gained from special constraints which affect only the dispersion, rather than the geometric form, of origin point patterns in elementary perceptrons. A further variation of the model, in which elliptical rather than circular distributions of origin points are employed might be more sensitive to contours and directions of elongation in the stimuli. No quantitative results are available on such a model at this time.*

7.3 Discrimination Experiments with R-controlled Reinforcement

In an experimental system with R-controlled reinforcement (Definition 39) the reinforcement control system receives information about the outputs of the perceptron, but receives no information directly from the environment. Such experiments are of interest in determining the "spontaneous organization" tendencies of perceptrons. It is readily seen, from theoretical considerations, that the performance of an elementary α-perceptron in such experiments is unlikely to be of psychological interest. In an α-perceptron, all $q_{i,j}$ are generally greater than zero, so that whatever response is associated to the first stimulus in a training sequence will tend to generalize to all other stimuli in the environment. Consequently, the perceptron, left to its own devices without any attempt to change its responses, will tend to form a classification $C(W)$ in which all stimuli in W are either in the positive class or else all in the negative class, with equal probability.**

---

* See Section 23.1.2 for a reconsideration of this problem from the standpoint of sensory analyzing mechanisms.

**In Ref. 82, such systems have been called "Class C perceptrons".

Two special cases are of interest, in which it is possible for a dichotomy to be formed with both classes non-empty. In the first case, some of the g_{ij} coefficients are zero. This might occur in a system with high thresholds on the A-units, so that some pairs of stimuli activate no A-units in common. If S_i and S_j are two such stimuli, then if S_i is the first stimulus and S_j is the second stimulus in the training sequence, it is perfectly possible that one will become associated to a positive response, and the other to a negative response. If these are the only two stimuli, or if there is no positive generalization from any of the stimuli which become associated to one class to the stimuli of the second class, this dichotomy may be stable. In general, however, one class is apt to become dominant, eventually pulling all stimuli into a single class as before. The second case in which a dichotomy might be formed is that in which the values are not initially all zero, but are distributed with some connections negative and some positive. In this case, the generalization from the first stimulus will not necessarily wipe out an initial bias in the opposite direction, and it is possible that a dichotomy will be formed.

While it is possible for dichotomies to be formed in the special cases mentioned above, there is little reason to suppose that such dichotomies would ever be of interest to a human observer. If the stimuli are uniformly distributed on the retina, or uniformly clustered about the center of the field, the g_{ij} coefficients which happen to be zero will generally be unrelated to possible "meaningful" classifications of the stimuli, so that any division into two classes will tend to be random, and unrelated to any concept of "intrinsic similarity" of the stimuli. Thus it is clear that in an elementary α-perceptron, psychologically meaning-

ful discriminations can be achieved only under the control of an experimenter, or r.c.s. which is capable of evaluating the correctness of the perceptron's responses according to some predetermined scheme. In the $\mathcal{T}$-systems, which are considered in the following chapter, somewhat more interesting performances in R-controlled experiments are likely to occur.

7.4 Detection Experiments

In discrimination experiments, such as those considered in the previous sections, the perceptron is required to give one of two responses to designate which of two well-defined classes of patterns is present. It is assumed that one of the two is always present, and that nothing else is present which might confuse the picture. In detection experiments, a single pattern, or class of patterns, is taught the perceptron as the "positive class", and anything else (such as noisy fields, arbitrary patterns, etc.) is considered to belong to the "negative class". Moreover, the positive pattern may appear with an admixture of background noise, irrelevant lines, or other sensory material. While such detection experiments differ considerably in their "psychological" character from discrimination experiments, from a theoretical standpoint they represent a special case of discrimination experiments in which the training and the two classes of stimuli are highly asymmetric, the positive class generally being smaller but more thoroughly trained than the negative class. Two cases are of interest: detection in noisy environments, and detection in organized environments. These are considered separately in the following sections.

7.4.1 Detection in Noisy Environments

A noisy environment will be defined as the product set of a set of well-defined stimulus patterns (including an empty field as a stimulus) and a set of "random noise patterns" superimposed on the members of the first set. The random noise patterns are generated by applying signals of random polarity (positive or negative with .5 probability) to a randomly selected set of S-units, chosen independently with probability P_n . P_n will be called the noise density of the environment, and represents the expected value of the proportion of S-points which emit random signals at any given moment of time.

Note that a noisy environment is, in its entirety, a well defined set of stimuli, with a probability ρ_j associated with each stimulus S_j . Such an environment consists of two classes: a positive class, in which one of the "positive stimuli" (e.g., a geometric form) is present in combination with one of the noise patterns, and a negative class, consisting of the noise patterns alone, or the "empty field" stimulus with a noise pattern superimposed. The task of the perceptron is to distinguish between positive and negative stimuli.

Let S_χ represent a test stimulus, selected from the positive class. Then the probability of correctly identifying S_χ as a positive stimulus in a random sequence experiment, with S-controlled reinforcement, is given by equation (7.7), with $E(u_\chi)$ defined by equation (7.9) and $\sigma^2(u_\chi)$ defined by equation (7.11), just as in an ordinary discri-

mination experiment. Similary, if S_n is a noise-stimulus, from the negative class, the probability of obtaining the correct (negative) response is given by the complement of the probability obtained from equation (7.7). Some special analytic features of this problem are worth noting.

For a binomial model, with a large retina and large association system (so that all Q -functions and retinal intersections of noise patterns can be assumed equal to their expected value) the intersection of a noise pattern with any other stimulus will be equal to the expected value of this intersection.* If we designate the noise patterns by S_n, $S_{n'}$, ..., and positive stimuli by S_χ, $S_{\chi'}$, ..., then (as explained on page 146),

$$Q_{nn'} = Q_n Q_{n'} \quad \text{and}$$
$$Q_{n\chi} = Q_n Q_\chi$$

Let S_χ and $S_{\chi'}$ represent the same positive stimulus pattern with different noise patterns superimposed. Then, if the noise density is low, $Q_{\chi\chi'} \approx Q_{\chi\chi} = Q_\chi$. But $Q_\chi >> Q_n Q_\chi$. Therefore, $Q_{\chi\chi'} >> Q_{\chi n}$, which means that the perceptron can be taught quite readily to give the proper positive response to a test stimulus, S_χ

---

* Actually, as noise patterns have been defined, the intersection of a pure noise pattern with a positive stimulus pattern will be slightly less than the expected value, since some of the points which normally are "on" for the positive stimulus will be turned "off" for the noise pattern. The conclusions above hold rigorously if the noise patterns are sets of positive signals only.

The same conclusion does not hold for the identification of a negative (noise) stimulus, however. In this case, the generalization from a previously trained noise stimulus, $S_{n'}$ to S_n is equal to $Q_{n'n} = Q_n^2$ (assuming all noise stimuli to be equal in area to their expected value). But the generalization from a positive stimulus is $Q_{xn} = Q_x Q_n$ which is generally greater than Q_n^2, since the area covered by the positive stimulus with noise superimposed is generally greater than the area of the noise stimulus alone. Consequently, we would expect the positive response to tend to generalize to the negative class as well, if both classes are represented with equal frequency in the training sequence.

A slight modification of the perceptron should improve its capability of distinguishing negative stimuli from positive ones. If the R-unit is given a threshold greater than zero, it will tend to remain "off" for the relatively weak signals coming from noise stimuli, but will go "on" (to its positive state) for the stronger signals coming from positive stimuli. With this modification, however, the system is no longer an elementary perceptron. An alternative procedure, which will improve the performance of an elementary perceptron, is to "overtrain" the negative stimuli, composing a stimulus sequence in which negative stimuli occur more frequently than positive ones. In an error correction experiment, it should be noted, this bias will be introduced automatically, regardless of the stimulus sequence, so that a detection problem should be solved much more readily than with an S-controlled system.

7.4.2 Detection in Organized Environments

In an "organized environment", where the background material may closely resemble the stimulus pattern in its characteristics, detection experiments take on some characteristics of special interest, psychologically. First of all, it should be noted that in attempting to distinguish a pattern such as the letter "X" against a background of lines occurring in random configurations, the environment may include stimuli which are fundamentally ambiguous in character, since patterns closely resembling the letter "X", or even identical to it, might arise by a chance superimposition of straight lines. In such a case, the only reasonable test of whether or not a pattern should be identified as an "X" would seem to be the human criterion of whether it looks more like an X or more like a random assemblage of line segments. While a similar problem might arise, in principle, in the case of detection experiments in noisy fields, it is less common there, except under extreme noise conditions. In the case of organized fields, ambiguous organizations are more the rule of the day, and the problem requires a different approach. In human perception, the properties of "good figure"are generally used to determine whether a particular set of line segments is seen as a letter, or some other known pattern, or simply as a random collection of unrelated components. Such judgements are not possible, however, for elementary perceptrons. We will return to the problem of figural organization in Part IV.

Treating the detection experiment simply as a special case of a discrimination experiment, the same conclusions apply as in the case of the noisy environment problem: it is possible, by exhaustively training

the perceptron with the product set of positive stimuli and irrelevant patterns to teach it to identify positive stimuli amidst extraneous material. The learning is apt to be slow, however, and will generally fall considerably short of what might be expected in a simpler discrimination experiment.

Most of the experimental work done to date on detection experiments has been carried out with the Mark I perceptron using a gamma system for the memory dynamics. This work will be reviewed in the following chapter, which deals with γ -perceptrons, but similar results might be expected with alpha systems.

7.5 Generalization Experiments

In the preceding experiments, it has been required that S_x should necessarily occur as one of the stimuli in the training sequence. When the perceptron is tested with a stimulus which has not been previously seen, a weak form of generalization is possible with elementary α-systems. Clearly, if the intersection of S_x with some other stimulus in the same class, $S_{x'}$, which did occur in the training sequence, is large enough, S_x will tend to evoke the same response as $S_{x'}$. In this case, S_x is correctly recognized only because, within the limits of tolerance of the perceptron, it appears to be identical, rather than merely similar to, the previously seen training stimulus. Thus, generalization, for an elementary α-perceptron, is based on an approximation to identity, rather than on similarity. In a "pure generalization" experiment, as defined in Chapter 3, the perceptron would be asked to recognize a pattern in a position where it does not overlap any previously seen patterns of the same class. If such an

experiment is performed with an α-system, with a single class of stimuli, the generalization will tend to be positive, due to the fact that Q_{ij} is never zero, for most systems, regardless of the relative positions of the stimuli. This result is trivial, however, and of no psychological interest, since any stimulus, whether it resembles the trained stimuli or not, will also tend to evoke the same response. To prevent such a tribial result, it is necessary to employ a discrimination test, training the system with two kinds of stimuli, and then testing it with similar stimuli in a disjoint portion of the retina to find out whether the appropriate responses have generalized for both kinds of stimuli. In this case, if the stimuli are of equal area, and equally trained, no generalization will be found, since the positive generalization from one class is exactly balanced by the negative generalization from the other class. Thus it is clear that an elementary α-system (and, in fact, any elementary perceptron) is incapable of abstracting similarity (in either the geometric or the psychological sense) but discriminates only by measuring a function of the overlaps of a test stimulus with representatives of both classes.

7.6 Summary of Capabilities of Elementary α-perceptrons

The elementary α-perceptrons, being the simplest class of perceptrons, provide a baseline of performance against which other systems can be compared. It has been demonstrated that the α-system, with both S-controlled and error correction reinforcement, is capable of discrimination learning, provided it sees a large representative sample of the stimuli which it is required to discriminate. It does not generalize well, to similar forms occurring in new positions in the retinal field, and

its performance in detection experiments, where a familiar figure appears against an unfamiliar background, is apt to be weak. More sophisticated psychological capabilities, which depend on the recognition of topological properties of the stimulus field, or on abstract relations between the components of a complex image, are lacking. The elementary perceptron has no capability of recognizing time sequences, since its responses are based on the momentary state of the system due to the current stimulus pattern alone, and are not influenced by the preceding sequence of events. Quantitative judgement might possibly be learned by an exhaustive training procedure, in which the system is required to give one response for stimuli above a certain area, or over a certain length, for example, and an opposite response if they fall short of the criterion. This is a rather crude approximation to quantitative estimation, however, and the problem can be handled much more satisfactorily with perceptrons with linearly responding R-units, as will be seen in Chapter 10. In R-controlled experiments, where the perceptron is required to form its own classification of stimuli, we have seen that the elementary α-perceptron tends either to classify everything identically (its most general tendency) or else to form a random dichotomy, which is of no psychological interest. It will be found that most of the weaknesses of elementary α-perceptrons are true of all simple perceptrons, and that it is necessary to go to topologically more complicated systems to find performances which are basically more satisfactory. In special cases, however, other types of simple perceptrons have advantages, as will be seen in the following chapters.

7.7 Functionally Equivalent Systems

It may be disturbing to some biologically oriented readers to think of an association unit that changes the sign of its output signal from excitatory to inhibitory as a function of its training. This is a conceptual simplification which makes analysis easier, but can be shown to be logically equivalent to an alternative model in which particular neurons, or A-units, are designated as excitatory, and others as inhibitory, with no change permitted in the sign of their outputs. The alternative model (which is analogous to the models originally presented in Refs. 79 and 80) is as follows:

Let the number of A-units be twice the number in the equivalent α-perceptron. Let half of the A-units be designated as excitatory units, and the other half be inhibitory units. All v_{ir} are initially assumed to be zero, or else to have positive signs if a_i is excitatory, negative signs if a_i is inhibitory. Each excitatory unit is paired with one of the inhibitory units, and the same origin point configuration is assigned to both members of the pair. Thus the responses of the inhibitory units exactly duplicate the responses of the excitatory units. The reinforcement rule is that a positive η from the r.c.s. affects only the excitatory units, while a negative η affects only the inhibitory units. With this rule, the signal u_i which goes to the R-unit in response to S_i is the sum of an excitatory component and an inhibitory component, the total being exactly equal to what it would be in the equivalent α-perceptron.

The exact pairing of the excitatory and inhibitory units is, of course, an inessential artifact, introduced only to guarantee that the two types of systems are truly identical in performance. If the origin configurations of all units are selected independently of one another, the expected values of the signals will be unaffected, but the variability will be somewhat increased, due to the greater number of independent A-units contributing to the signal. Such a system has been previously described as a "differentiated A-system" (Ref. 79).

8. PERFORMANCE OF ELEMENTARY γ-PERCEPTRONS IN PSYCHOLOGICAL EXPERIMENTS

It will be recalled that the reinforcement rule for a gamma system (defined in Chapter 4, Def. 38) is one which guarantees that the sum total of the value of all connections to any unit remains constant, even though the values of individual connections may change with time. In the notation of the last chapter, the change in the value of the connection c_{ir} due to the reinforcement of stimulus S_j was given by

$$\Delta v_{ir} = \rho_j \, a_i^*(j) \qquad \text{for an } \alpha\text{-system.} \qquad (8.1)$$

For a gamma system, the corresponding expression is

$$\Delta v_{ir} = \rho_j \left[a_i^*(j) - \frac{1}{N_a} \sum_k a_k^*(j) \right] \qquad (8.2)$$

A variation of the gamma system, which will be designated the γ'-system, is of interest chiefly because it is considerably easier to analyze. For this model,

$$\Delta v_{ir} = \rho_j \left[a_i^*(j) - Q_j \right] \qquad (8.3)$$

This is equal to the expected value of Δv_{ir} for the γ-system, and with large values of N_a the γ-system and γ'-system become indistinguishable.

The organization of this chapter will follow closely that of Chapter 7. The first section deals with the analysis of discrimination experiments with S-controlled reinforcement, and presents results of a number of experiments, including comparisons with the α-systems considered in the last chapter. Discrimination experiments with error correction, and discrimination experiments with R-controlled reinforcement are then presented, and the final sections deal with detection experiments, and other performances of γ-perceptrons.

8.1. Discrimination Experiments with S-controlled Reinforcement

8.1.1 Fixed Sequence Experiments: Analysis

As in the case of the alpha-system analysis, our object is to compute the ratio $\mu^2(u_\chi)/\sigma^2(u_\chi)$, for the class of perceptrons, test stimulus, and training sequence under consideration. The notation and definitions correspond to those employed in Chapter 7. The analysis again follows that of Joseph (Ref. 41). For the γ-system, the expected value of u_χ is obtained as follows: The value of the connection from the A-unit a_i at the end of the training sequence is given by:

$$v_{ir} = T\sum_j \rho_j \rho_j \left[a_i^*(j) - \frac{1}{N_a}\sum_k a_k^*(j)\right]$$

$$= T\sum_j \rho_j \rho_j \left[\frac{N_a - 1}{N_a}\, a_i^*(j) - \frac{1}{N_a}\sum_{k\neq i} a_k^*(j)\right]$$

Consequently, if the test stimulus S_χ is now shown, the input to the response unit will be

$$u_\chi = T \sum_i \sum_j \rho_j \rho_j \left[\frac{N_a - 1}{N_a} a_i^*(j\chi) - \frac{1}{N_a} a_i^*(\chi) \sum_{k \neq i}' a_k^*(j) \right]$$

yielding, for the expected value of the signal u_χ ,

$$E(u_\chi) = T \sum_i \sum_j \rho_j \rho_j \left[\frac{N_a - 1}{N_a} Q_{j\chi} - \frac{1}{N_a} Q_\chi \sum_{k \neq i} Q_j \right]$$

$$= T(N_a - 1) \sum_j \rho_j \rho_j (Q_{j\chi} - Q_j Q_\chi)$$

(8.4)

For a $\mathcal{F}'$-system, the analysis is considerably simplified. In this case, the value of the connection from unit a_i at the end of the training sequence is

$$v_{ir} = T \sum_j \rho_j \rho_j \left[a_i^*(j) - Q_j \right]$$

Collecting the signals from all active connections when S_χ occurs yields the input to the R-unit,

$$u_\chi = T \sum_i \sum_j \rho_j \rho_j \left[a_i^*(j\chi) - Q_j \, a_i^*(\chi) \right]$$

and the expected value of this signal is

$$E(u_\chi) = T N_a \sum_j \rho_j \rho_j (Q_{j\chi} - Q_j Q_\chi)$$

(8.5)

The variance of u_χ is again computed from the general equation (7.4), given in the last chapter. For a γ'-system, the same considerations apply as in the α-system, namely, that the only source of variability in the signals $c^*_{ir}(x)$ is due to the origin point configurations of the A-units, which are selected independently for the different A-units. Consequently, the equation (7.5) holds identically for a γ'-system. In a true γ-system, however, the signals $c^*_{ir}(x)$ are not independent. The value v_{ir} upon which c^*_{ir} depends is the result of a series of increments, ΔV_{lr}, each of which depends upon the particular set of A-units which are active at the time of reinforcement (as shown in Equation 8.2). Consequently, for a gamma system, the variance is

$$\sigma^2(u_x) = N_a \sigma^2 \left[c^*_{ir}(x)\right] + N_a(N_a-1)\, cov.\left[c^*_{ir}(x),\, c^*_{i'r}(x)\right]$$

$$= N_a\left[E c^{*2}_{ir}(x) - E^2 c^*_{ir}(x)\right] + N_a(N_a-1)\left[E c^*_{ir}(x)\, c^*_{i'r}(x)\right.$$

$$\left. - E c^*_{ir}(x)\, E c^*_{i'r}(x)\right] \qquad (8.6)$$

The reader who is interested in the detailed analysis of this expression will find a full algebraic expansion of its components in Ref. 41. The final equation which results is as follows:

$$\sigma^2(u_x) = \frac{T^2(N_a-1)}{N_a} \sum_j \sum_A p_j p_A p_j p_A \left\{(N_a-1)(1-2Q_x)\left[(Q_{jAx} - Q_j Q_A Q_x)\right.\right.$$

$$\left. - 2Q_A(Q_{jx} - Q_j Q_x)\right] - (N_a-2)\left[(Q_{jx} - Q_j Q_x)(Q_{Ax} - Q_A Q_x) - Q_x^2(Q_{jA} - Q_j Q_A)\right]$$

$$\left. + Q_x(Q_{jA} - Q_j Q_A)\right\} \qquad (8.7)$$

An analogous treatment for the $\mathcal{P}'$-system, based on Equation (7.5), yields the expression:

$$\sigma^2(u_x) = T^2 N_a \sum_j \sum_k \rho_j \rho_k p_j p_k \left[(Q_{jkx} - Q_j Q_k Q_x) - 2Q_k (Q_{jx} - Q_j Q_x) \right.$$

$$\left. - (Q_{jx} - Q_j Q_x)(Q_{kx} - Q_k Q_x) \right] \tag{8.8}$$

For both the $\mathcal{P}$-perceptron and the $\mathcal{P}'$-perceptron, the expectation of u_x and the variance of u_x are both on the order of N_a. Consequently, the ratio μ^2/σ^2 can be made arbitrarily large by increasing N_a. This means that the theorem stated in the last chapter (Page 159) holds for $\mathcal{P}$ and $\mathcal{P}'$-perceptrons as well as for α-systems. Equation (7.7) can again be used for a close approximation to the actual probability of correct response for a $\mathcal{P}$ or $\mathcal{P}'$-perceptron, substituting the appropriate expressions for the mean and variance in each case.

It is interesting to note that if the expected values of the generalization coefficients, g_{ij}, are substituted into equations (7.3), (8.4), and (8.5), identical expressions are obtained for the expectation of u_x for the α, $\mathcal{P}$, and $\mathcal{P}'$-systems. The expected value of the un-normalized coefficient, $\bar{g}_{ij}$, for a $\mathcal{P}$-perceptron is $(N_a - 1)(Q_{ij} - Q_i Q_j)$; for a $\mathcal{P}'$-perceptron it is $N_a(Q_{ij} - Q_i Q_j)$, while for an α-perceptron it is $N_a Q_{ij}$. Substituting these quantities, we obtain, for all three systems,

$$E(u_x) = T \sum_j \rho_j p_j \bar{g}_{xj} \tag{8.9}$$

$$a_x = T \sum_j \rho_j p_j g_{xj} \tag{8.10}$$

The special properties of the γ and γ'-perceptrons are due to the fact that their generalization coefficients for a binomial model tend to be negative for sufficiently well separated, or disjoint, stimuli, whereas in the case of an α-system, the generalization coefficients are all non-negative. In a Poisson model, while it is possible for negative generalization coefficients to occur due to random variability of individual perceptrons, the expected values of $g_{i,j}$ are always non-negative, since $Q_{i,j} = Q_i Q_j$ only if the stimuli are disjoint. These features are of interest for R-controlled experiments, as will be seen presently.

8.1.2 Fixed Sequence Experiments: Examples

Numerical analyses have been carried out mainly for the γ-perceptrons, since the equations are considerably simpler. For large values of N_a, the γ' and γ-systems will have identical performances. Tables 3 and 4 (in Chapter 7) apply identically to the γ'-system, for Experiments 1 and 2. The performance curves shown in Figures 13 and 14 are also applicable. Figure 21 shows a comparison of the γ and γ'-systems on Experiment 1 (horizontal vs. vertical bar discrimination), for the optimum parameters with a binomial model ($x = 3$, $y = 1$, $\theta = 2$). Figure 22 shows a similar comparison for the same parameters, with Experiment 2.

It is clear that under the conditions of Experiments 1 and 2, the γ-systems have no advantage over the α-perceptrons. The equivalence of the curves is due to the fact that in these experiments, all stimuli are equal in area (yielding equal Q_i for all stimuli), the number of stimuli in each class is equal, and all stimuli occur with equal frequency. If the sizes or frequencies are unequal, the γ-system may have a marked advantage, as will be seen in the analysis of Experiment 4, in Section 8.1.4.

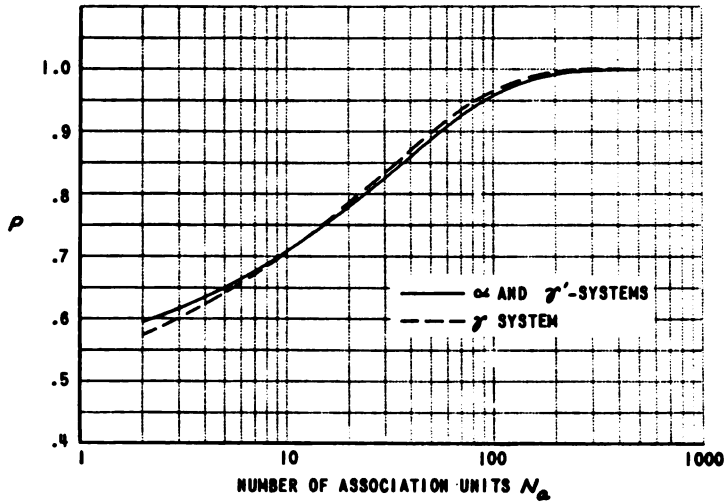


Figure 21 PROBABILITY OF THE CORRECT IDENTIFICATION OF ANY TEST BAR vs. THE NUMBER OF ASSOCIATION UNITS, IN EXPT. 1

HORIZONTAL 4 x 20 BARS vs. VERTICAL 20 x 4 BARS ON TOROIDAL 20 x 20 RETINA
$x = 3 \quad y = 1 \quad \theta = 2$

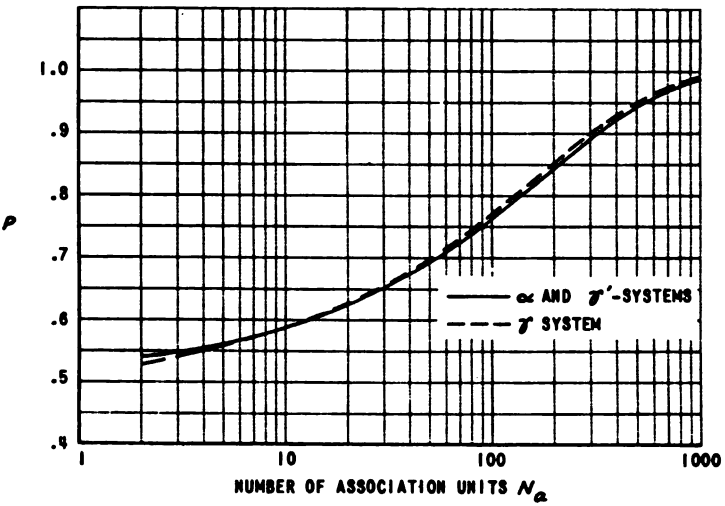


Figure 22 PROBABILITY OF THE CORRECT IDENTIFICATION OF ANY TEST BAR vs. THE NUMBER OF ASSOCIATION UNITS, IN EXPT. 2

$x = 3 \quad y = 1 \quad \theta = 2$ ALTERNATING DICHOTOMY

8.1.3 Random Sequence Experiments: Analysis

The un-normalized generalization coefficients for a $\mathcal{F}$ and $\mathcal{F}'$-system are given by

$$g_{ij} = n_{ij} - \frac{1}{N_a} n_i n_j \qquad \text{for a } \mathcal{F}\text{-system} \qquad (8.11)$$

$$g_{ij} = n_{ij} - Q_j n_i \qquad \text{for a } \mathcal{F}'\text{-system} \qquad (8.12)$$

where n_{ij} = the number of A-units responding both to S_i and to S_j.

As in the α-system analysis (Section 7.1.4) the training sequence is assumed to consist of T stimuli, where each stimulus, S_j, has a probability p_j of being selected at any step of the training sequence. The analysis has been carried out only for the $\mathcal{F}'$-perceptron, since the true $\mathcal{F}$-system leads to excessively cumbersome expressions for the variance. For large N_a, as observed in the preceding section, the two systems should be virtually indistinguishable in performance.

For the $\mathcal{F}'$-system, the input to the response unit when S_x occurs after the training sequence is

$$u_x = \sum_j p_j m_j (n_{xj} - Q_j n_x)$$

where m_j, as before, is the number of times that S_j occurs in the training sequence. Taking the expected value of this expression, we obtain

$$E(u_x) = T N_a \sum_j p_j p_j (Q_{jx} - Q_j Q_x) \qquad (8.13)$$

The variance of u_χ over both perceptrons and training sequences is again given by equation (7.10). In the present case, this yields:

$$\sigma^2(u_\chi) = TN_a \sum_j p_j \left[Q_{j\chi} - 2Q_j Q_{j\chi} + Q_j^2 Q_\chi + (N_a - 1)(Q_{j\chi} - Q_j Q_\chi)^2 \right]$$

$$+ TN_a \sum_j \sum_k p_j p_k p_j p_k \left[(T-1)(Q_{jk\chi} - Q_j Q_{k\chi} - Q_k Q_{j\chi} + Q_j Q_k Q_\chi) \right.$$

$$\left. - (T + N_a - 1)(Q_{j\chi} - Q_j Q_\chi)(Q_{k\chi} - Q_k Q_\chi) \right]$$

(8.14)

The detailed derivation of this expression can be found in Ref. 41. It can readily be seen that the theorem of Section 7.1.2 continues to hold for this system. Actual performances can again be calculated by using Equation (7.7).

8.1.4 Random Sequence Experiments: Examples

A comparison of binomial α and γ'-perceptrons on the random sequence version of the horizontal/vertical bar experiment (Experiment 3) is shown in Figure 23. A curve obtained from the simulation of a true γ-system with the same parameters is included for comparison. The simulation curve shows the average of 100 runs. Figure 24 compares the performance of the binomial model with that of a Poisson model, on the same experiment.

In Figure 25, the performance of a γ'-system in the "frequency bias" experiment (Experiment 4) is shown, with the mean performance curve of the equivalent α-system, from Figure 14, included for comparison. A comparison with Figure 16 shows that under

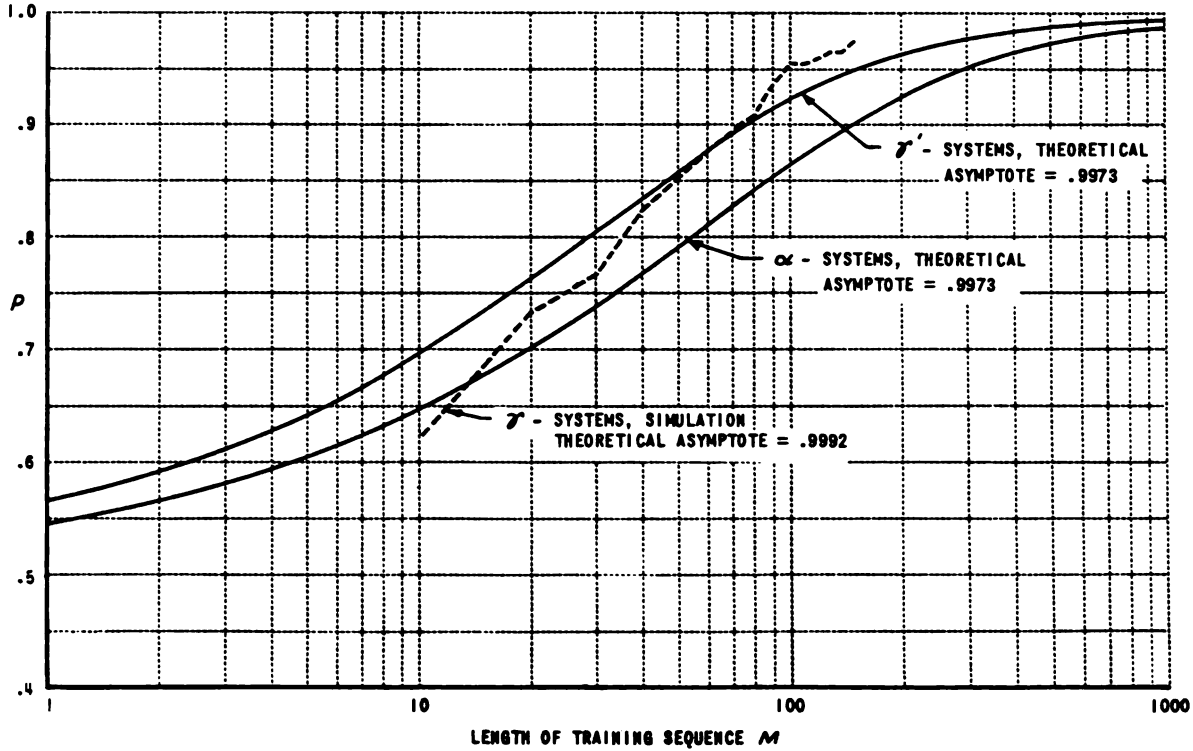


Figure 23 COMPARISON OF BINOMIAL α AND γ-PERCEPTRONS ON RANDOMIZED TRAINING SEQUENCE (EXPT. 3). PROBABILITY OF THE CORRECT IDENTIFICATION OF A TEST BAR vs THE LENGTH OF THE TRAINING SEQUENCE

HORIZONTAL 4 x 20 BARS vs VERTICAL 20 x 4 BARS ON TOROIDAL 20 x 20 RETINA
$N_a = 300$ $X = 5$ $Y = 5$ $\theta = 2$

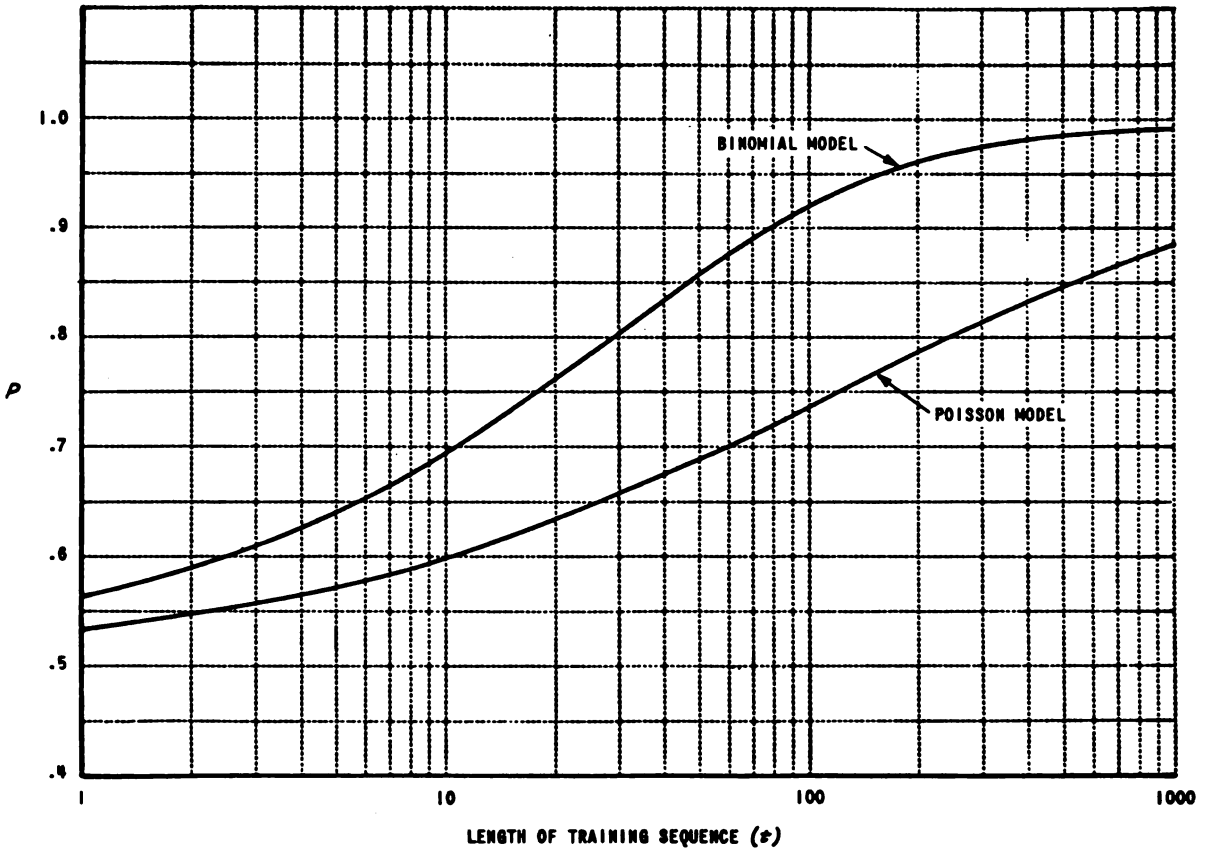


Figure 24 . RANDOM TRAINING SEQUENCE, $\mathcal{S}'$-SYSTEM (EXPT. 3). PROBABILITY OF CORRECT IDENTIFICATION OF A TEST BAR vs. LENGTH OF TRAINING SEQUENCE. $N_a = 300$, $x = 5$, $y = 5$, $\theta = 2$

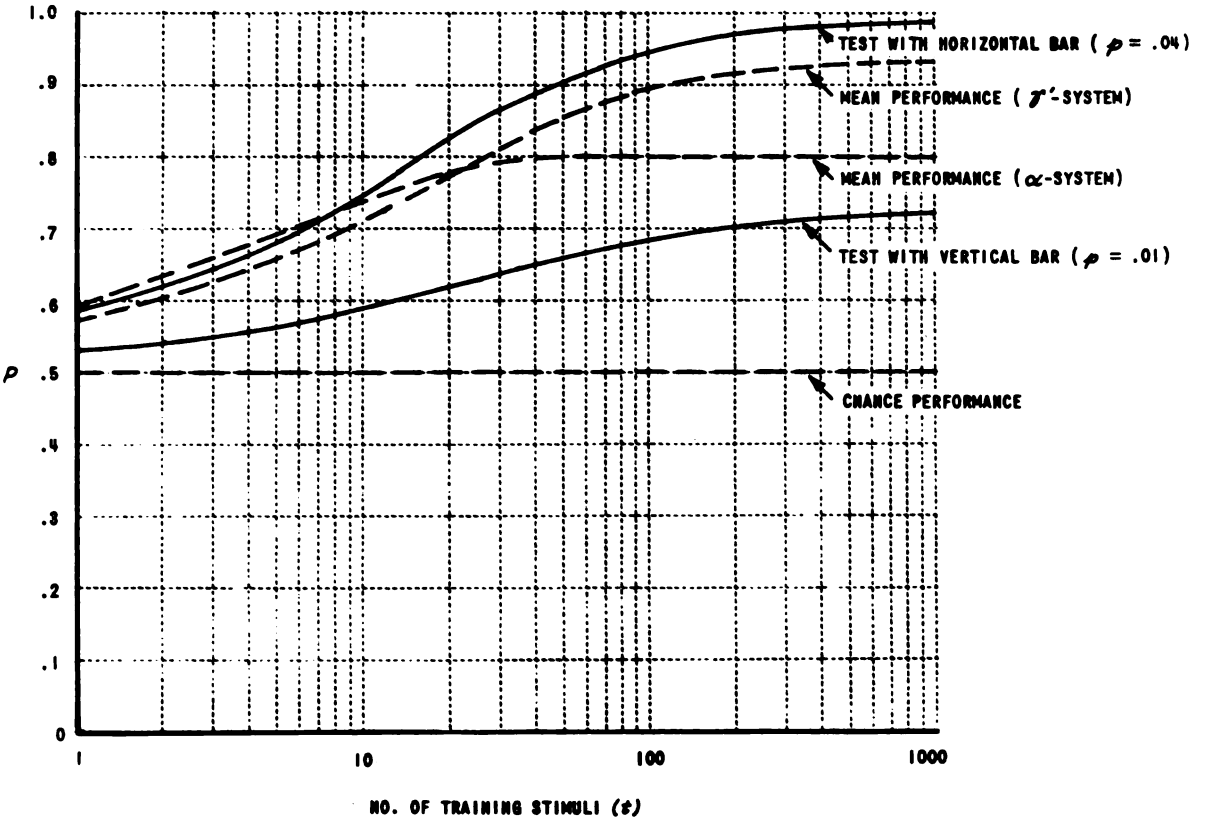


Figure 25 PROBABILITY OF CORRECT IDENTIFICATION OF TEST STIMULI IN FREQUENCY BIAS EXPERIMENT (EXPT. 4) WITH BINOMIAL γ'-PERCEPTRON $N_a = 100$, $x = 3$, $y = 1$, $\theta = 2$. (COMPARE FIG. 16 FOR α-SYSTEM)

conditions of unequal frequency for the two classes to be discriminated, the $\mathcal{F}$ -system may have a marked advantage. The effect of frequency bias on a $\mathcal{F}$ -system is also shown in a number of simulation experiments with the IBM 704 computer, which have been described previously (Ref. 84). The horizontal/vertical bar discrimination problem happens to show up the $\mathcal{F}$ -system to its best advantage, since, with a binomial perceptron, the expected value of the generalization coefficient, g_{ij} , where S_i and S_j are in opposite classes, is zero for this particular problem. A Poisson model, where the interaction between the horizontal and vertical bar classes is non-zero, would not perform as well in this experiment, and the binomial model would also perform less well in experiments with classes of stimuli which could achieve greater intersections.

Figures 26, 27 and 28 show some typical experiments performed with a digital simulation program, for binomial $\mathcal{F}$ -perceptrons of sizes up to $N_a = 1000$, and a 72 by 72 retina. The stimuli are kept within the retinal field in these experiments by requiring that their centers remain within a 13 by 13 field, so that there are 169 possible positions for each stimulus. In Figure 26(b), the effect of allowing rotations up to 30 degrees and up to 359 degrees (inclusive), in addition to displacements within the retinal field, is illustrated. Figure 28 shows the effect of size bias where one class of stimuli (the letter "F") can be considered as subsets (on the retina) of stimuli of the other class (the letter "E"). With purely excitatory connections from the retina, the situation is clearly much worse than with both excitatory and inhibitory connections, as shown in Figures 28(a) and (b).

From the equations for the expected value of the signal (Equation 8.13, for example) it can be seen that a bias in the correct direction may exist even when the perceptron is occasionally reinforced

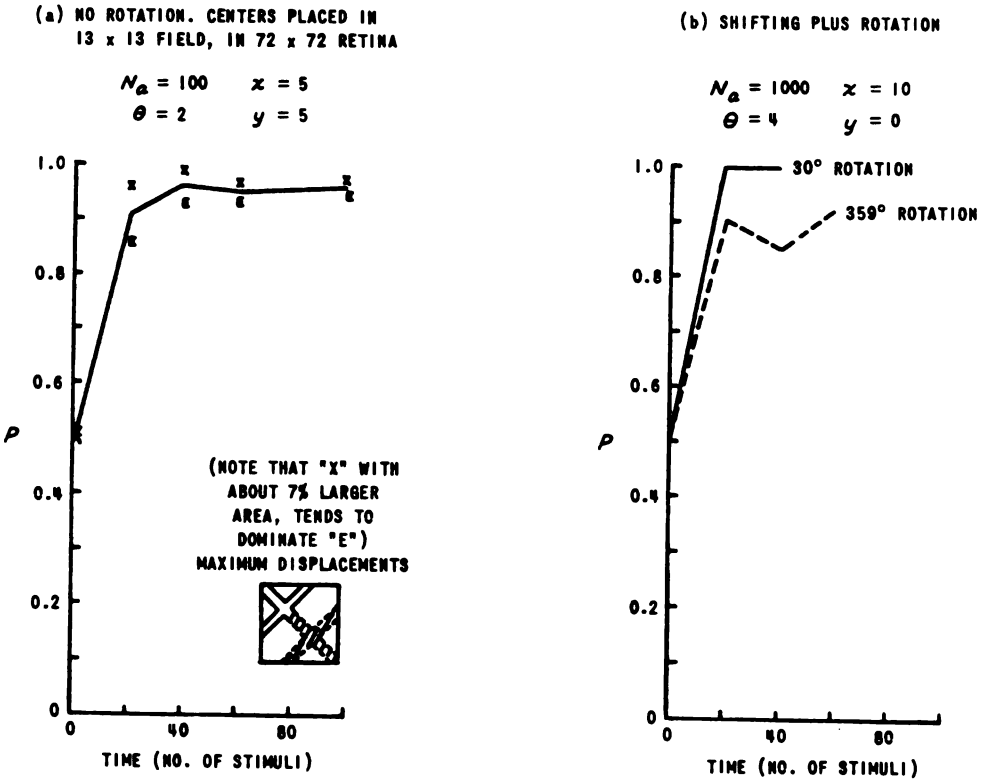


Figure 26 "E" vs. "X" DISCRIMINATION EXPERIMENTS

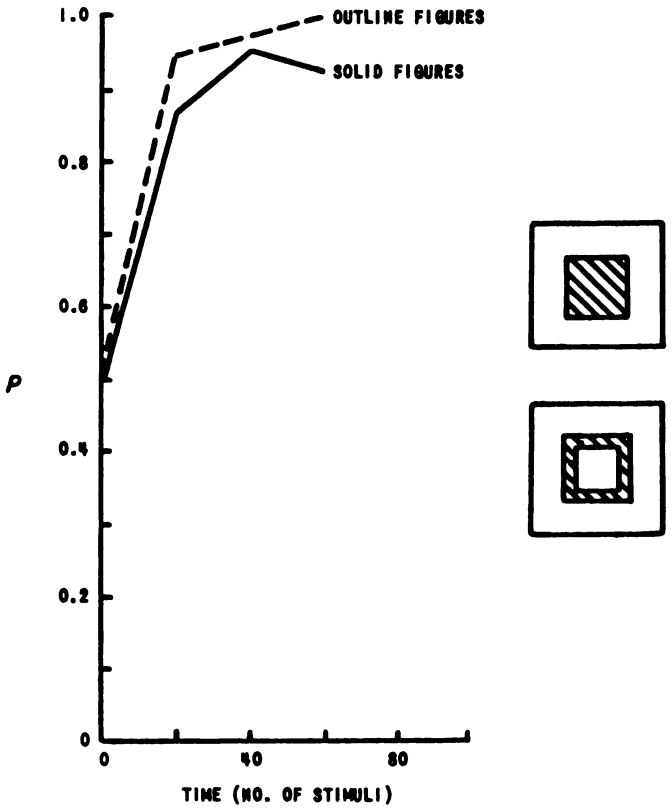


Figure 27 SQUARE-DIAMOND DISCRIMINATION. $N_a = 1000$, $x = 10$, $y = 0$, $\theta = 4$ CENTERS PLACED IN 13 x 13 FIELD

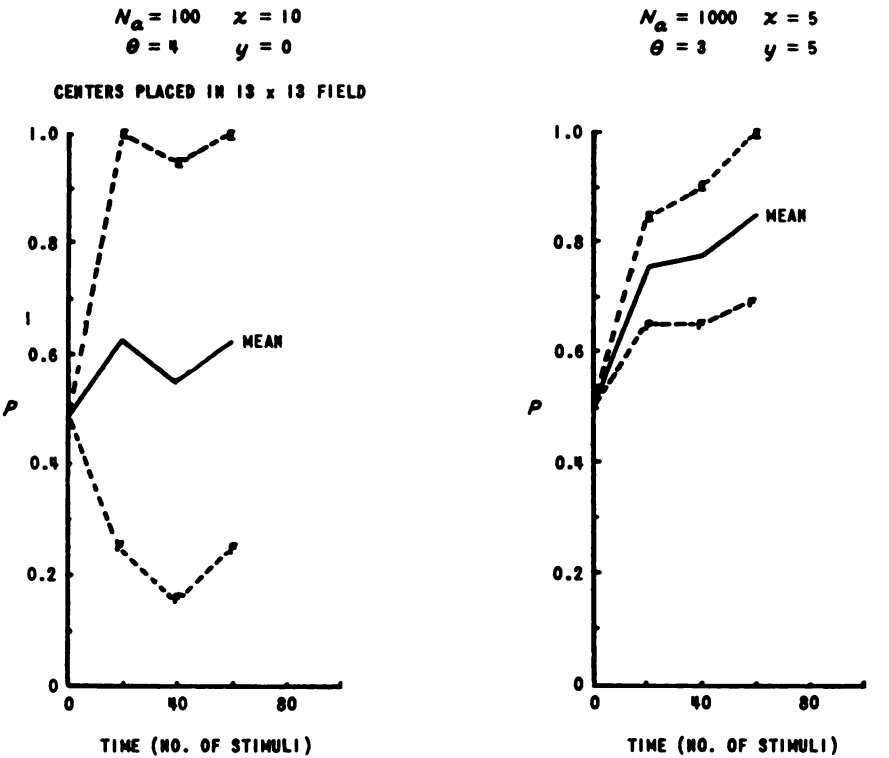


Figure 28 "E" vs. "F" DISCRIMINATION

in the wrong direction. Several experiments have been carried out by Hay using the Mark I perceptron at CAL, to study the effect of "random errors" by the experimenter training the machine (Ref. 30). In an experiment on the discrimination of the letters "E" and "X" with a $\mathcal{F}$-perceptron employing S-controlled learning, it was found that the perceptron learned to discriminate the letters with 100% accuracy despite the introduction of 30% misidentifications by the experimenter (i.e., by the r.c.s.). This experiment emphasizes the fact that the perceptron can exceed the level of performance of its "teacher" or reinforcement control system.

8.2 Discrimination Experiments with Error-Corrective Reinforcement

While it has been demonstrated in Chapter 5 (Theorem 8) that the error correction procedure will not always lead to a solution with the $\mathcal{F}$-system, practical systems seem to work about as well as α-systems, and may actually learn somewhat faster in some cases. Figures 29 and 30 illustrate two sets of experiments on $\mathcal{F}$-perceptrons, using the Burroughs 220 computer at Cornell University, in which performance is compared with perceptrons having the same topological organizations, but employing an α-system memory rule. Since the error correction procedure will lead to a solution regardless of sequence or relative frequency of stimuli in the classes being discriminated, and regardless of relative sizes of stimuli, the special advantages of the $\mathcal{F}$-system in overcoming frequency bias and size bias are relatively unimportant here. In most experiments with error-corrective reinforcement, therefore, the simpler α-rule is generally employed.

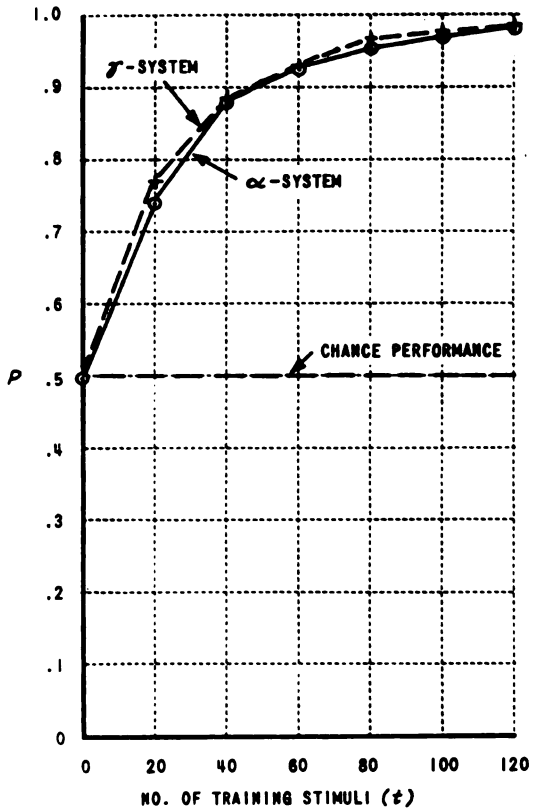


Figure 29 COMPARISON OF BINOMIAL α AND γ SYSTEMS IN HORIZONTAL/VERTICAL BAR DISCRIMINATION WITH ERROR CORRECTION (EXPT. 6). MEAN OF 25 CASES WITH $N_a = 300$, $x = 3$, $y = 1$, $\theta = 2$

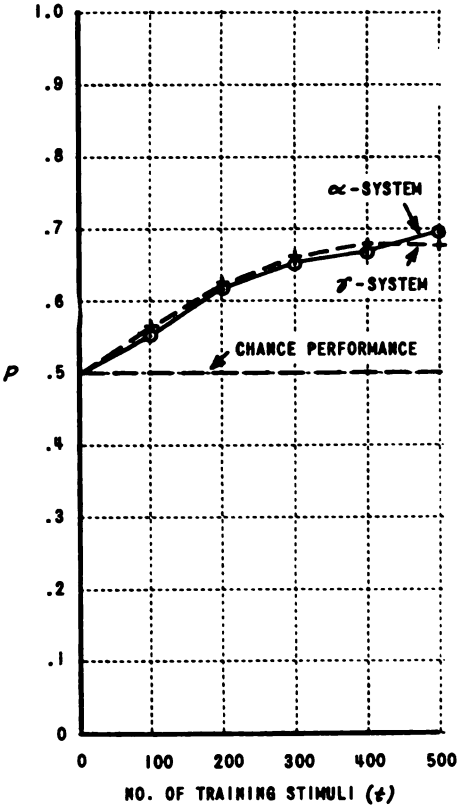


Figure 30 COMPARISON OF α AND γ SYSTEMS IN SQUARE TRIANGLE DISCRIMINATION WITH ERROR CORRECTION (EXPT. 7). MEAN OF 5 CASES, WITH $N_a = 300$, $x = 6$, $y = 4$, $\theta = 3$

8.3 Discrimination Experiments with R-controlled Reinforcement

The performance of a γ -perceptron in R-controlled experiments (where the r.c.s. is entirely isolated from the environment and reinforces the perceptron positively at all times, regardless of what its current response happens to be) is somewhat more interesting than that of the α -perceptron. Since it is possible to have negative generalization coefficients for the γ -model, two distinct possibilities suggest themselves which were not present before: (1) The system may form an unstable classification of the environment, with individual stimuli continually shifting membership from one class to the other, due to negative interaction between successive reinforcements; (2) the system may form a stable dichotomy with some stimuli in the positive class and some in the negative class. The third possibility corresponds to the expected situation with an α -system, namely: (3) The system may form a stable classification with every stimulus in the same class, the alternative class being empty.

An unpublished theorem by H. Kesten* proves that (for a γ -system in which the values are allowed to grow without bound) the first alternative is impossible. Every perceptron will ultimately form a "stable" classification, in which every stimulus is assigned to one of the two classes and will remain in the same class with probability 1 at any future time. The remaining two alternatives both remain possible, however.

At the present time, a fully satisfactory analysis of the classification tendencies of γ-perceptrons which are "left on their own" in an R-controlled experiment is not available. A number of special cases can

---

* Personal communication.

be analyzed heuristically, however, and some of these are illuminating. Moreover, a series of simulation experiments has been completed which illustrates performance on some typical problems.

The basic feature of this system in an R-controlled experiment is a tendency to classify stimuli on the basis of retinal location, rather than geometrical similarity. If two stimuli occur in the same location on the retina, covering largely the same set of sensory points, g_{ij} will tend to be positive, so that the reinforcement of one stimulus will tend to generalize automatically to the other. A "cluster" of such stimuli, projected onto a limited region of the field, will tend to be classified the same way, either all positive or all negative. On the other hand, two stimuli which cover disjoint sensory sets will (in a binomial model) tend to have a negative g_{ij}*. In this case, reinforcing S_i with η positive will automatically assign S_j, to the negative class, if its value was previously zero. Thus, clusters of stimuli which are "well separated" will tend to go into opposite classes, with a binomial γ-perceptron. The following experiment illustrates this tendency quite clearly:

EXPERIMENT 9: For the same retina and environment of horizontal and vertical bars described in Experiment 1, let the stimuli occur in a random sequence, as in Experiment 3. During the training sequence, R-controlled reinforcement is employed. The response to each of the 40 bars is then determined, to establish the classification which has been developed by the perceptron.

---

* In a Poisson model, the expectation of g_{ij} for disjoint stimuli is zero, in the γ-system, and all stimuli will tend to go into the same class unless they form completely disjoint clusters, in which case the class assignment will be random for each cluster.

In a number of repetitions of this experiment (which was simulated with a 704 computer for a very large, or infinite N_a, binomial perceptron, it was found in every case that the perceptron placed ten adjacently located horizontal bars and ten adjacent vertical bars in the positive class, and the other ten bars of each type in the negative class. The dynamics of the process can be readily followed in a heuristic fashion. The first bar to be seen -- say a vertical bar -- may evoke a positive or negative response at random. If $r^* = +1$, then the connections from the responding A-units will each gain a positive increment of value, and connections from inactive A-units will become slightly negative, so that the total value is conserved. For two disjoint bars in the "same" class (i.e., both horizontal or both vertical) g_{ij} will be negative, but for the two closest neighbors on either side, g_{ij} will be positive. The generalization, g_{ij}, to members of the "opposite" class (i.e., one horizontal and one vertical) will be zero, since the intersection between any horizontal and vertical bar, in this environment, is equal to its expected value, yielding zero generalization for a binomial γ-system (see Page 146). Consequently, the horizontal and vertical bars will never interact, regardless of the sequence in which they occur, and each of these two sets of stimuli will organize independently. Consider, therefore, the development of a classification for the vertical bars, after the first has been associated to $r^* = +1$. If the second vertical bar in the training sequence should happen to be one of the two close neighbors on either side of the original bar, this will immediately evoke the response $r^* = +1$, and will be reinforced in the same direction as the previous bar, extending the net positive generalization to at least one additional member of the vertical set. At the same time, vertical bars which are more than two positions removed from both of the bars already seen will now have twice the negative reinforcement that they received before, due to the summation

of the negative g_{ij} . If one of these bars should occur, the response will be -1 and η will be negative. This will not only spread negative value to the adjacent stimuli, but will add to the positive value of the stimuli which were previously placed in the positive class. Thus two mutually supporting "nuclei" of stimuli are formed, one in the positive class and one in the negative class, which tend to spread their domain to neighboring stimuli, but tend to "repel" remote stimuli, supporting their adhesion to the opposite class. Under these conditions, it is plausible that the most stable balance between classes will be found when the classes are evenly divided, each tending to attract marginal stimuli from the other to the same degree.

Simulation experiments with this procedure show that a stable dichotomy tends to be formed after the first few hundred stimuli of the training sequence, the probability of a change in class membership being very small thereafter. The terminal condition is of the type indicated above, with 10 horizontal and 10 vertical bars in each class of the dichotomy.

8.4 Detection Experiments

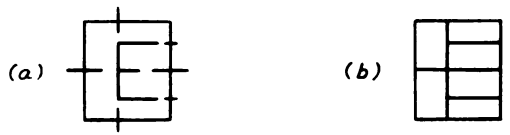
In detection experiments, the same general conclusions hold true as in the case of α -systems (Section 7.4). In the case of noisy environments with a large retina, it was noted that the intersection of a noise pattern with any other stimulus will be equal to the expected value of the intersection, i.e., to the product of the measures of the active S-sets. For the binomial γ -system, this implies zero generalization from a reinforced "positive" stimulus to a noise pattern, and zero generalization from one noise pattern to another. This means that a class of positive stimuli can be learned without any generalization to noise

patterns, but that negative training on a limited sample of noise patterns does not generalize effectively to new noise patterns. As in the case of the α-system, the use of a threshold greater than zero on the R-units should effectively separate positive stimuli from noise patterns. It is worth noting that for discriminating a single class of positive stimuli from noise, a monopolar reinforcement system (Defintion 35, Chapter 4) will work as effectively as a bipolar system, since reinforcement given for negative responses has little or no effect on future performance (except for those noise patterns actually seen, or nearly identical to those seen).

Several experiments have been performed with the Mark I perceptron at CAL to evaluate the performance of γ-perceptrons in noisy environments, and in problems in which positive stimuli such as letters of the alphabet have been mixed with extraneous, but similarly organized stimuli (geometric patterns, other letters, etc.). Performance on the discrimination of the letters "E" and "X" with various amounts of noise present has been reported by Hay in Ref. 30. Two 240 A-unit perceptrons were tested, both learning to perfection in the absence of noise. With noise present, one perceptron learned as well as before, the second falling to about 75% accuracy. The amount of noise introduced was not carefully quantified in these experiments, but it is clear that the perceptron can perform appreciably better than chance as long as a human observer can still detect the original letters embedded in the image. In the experiments with superimposed images of irrelevant patterns, a poorer level of performance is obtained. A perceptron trained to respond positively to the letter X, with monopolar γ-reinforcement, will generally give the proper response whenever an "X" is present, but tends to give the

positive response quite frequently to triangles, squares, or other letters as well. The introduction of a high response threshold improves performance considerably, but a system capable of responding in terms of figure-ground organization would clearly have a great advantage in such experiments. As the quantity of background material is increased, the performance of an elementary perceptron in detection experiments deteriorates rapidly.

A striking difference between an elementary perceptron and a human observer in detection experiments is that the human will show vast differences in performance depending upon organizational properties of the background and its relationship to the figure. For example, the human observer will readily recognize the letter "E" in Figure (a), but will find it hard to segregate the "E" from the extraneous lines in Figure (b). An elementary perceptron would show little or no difference between these two situations.



Typical test patterns for detection experiments

8.5 Generalization and Other Capabilities

In "pure" generalization experiments, where the test stimuli are disjoint from the training stimuli, the γ-system has no advantages over the α-system. In fact, the binomial γ-system, due to its negative $g_{i,j}$ for disjoint stimuli, will actually tend to place a disjoint stimulus in the opposite class from the reinforced stimulus, unless members of the opposite class have also been reinforced, in which case the effects tend to cancel.

Where the training stimuli cover the retina in a representative sample of locations, the gamma system has the possible advantage of low or negative generalization to patterns which have small intersections with the trained patterns. This shows best in such experiments as the horizontal/vertical bar discrimination experiment, where generalization from horizontal to vertical bars is zero. As was noted in the case of R-controlled discrimination experiments, generalization in γ-systems, as with all elementary perceptrons, tends to be based on the location rather than the similarity of the stimuli, in any more fundamental sense. Ideally, we would hope to find a system in which $g_{i,j}$ is large for all pairs of stimuli, S_i and S_j, which are "similar" or "equivalent" under some group of spatial transformations, such as rigid motions, dilatations, or projective transformations, and small or negative otherwise. Except in exceptional and highly restrictive environmental conditions, this condition is not to be found in elementary perceptrons. Highly artifactual organizations which have the required property can be designed in the case of four-layer series coupled perceptrons,

as will be seen in Chapter 15. Systems which spontaneously acquire the required organizational properties are found chiefly among the cross-coupled perceptrons, however, and will be discussed in Part III of this volume.

In general, it is seen that γ-perceptrons have much the same properties as α-systems. In S-controlled experiments, especially with frequency and size bias present, they perform somewhat better, but in error correction experiments there is little to be gained from the gamma rule, and there is the possibility that the γ-system may fail to work where an α-system would have succeeded, as proven in Chapter 5. The performance in R-controlled experiments is somewhat more interesting than that of α-systems, but the classifications which are formed spontaneously tend to form on a basis of classification related to position of stimuli on the retina, rather than similarity, and are consequently of minimum psychological interest.

The γ-system may be somewhat more plausible as a biological memory mechanism, due to its fundamental conservative property. If biological memory is due to a physical process which maintains some overall equilibrium, such as a chemical substance the total amount of which remains invariant, or a competition among afferent processes for "Lebensraum" in the neighborhood of an efferent neuron, this property would certainly be indicated. It should be emphasized, however, that the conservation of the total value, as in the systems considered in this chapter, is insufficient to keep individual coupling coefficients, v_{ij}, from becoming indefinitely great, since they may be balanced by negative values of equal magnitude. Such a condition is quite implausible in any real physical system. In the next chapter, elementary perceptrons with memory dynamics which limit the growth of the values are considered.

9. ELEMENTARY PERCEPTRONS WITH LIMITED VALUES

Two basically different mechanisms for limiting the growth of values, v_{ij}, will be considered in this chapter. The first mechanism is a simple upper and lower bound, such that the value may grow up to the designated limit but no further. Systems employing this mechanism show "saturation properties" as the connections attain their limits. The second mechanism is an exponential decay, which determines an equilibrium point for each v_{ij} depending upon the frequency with which it is reinforced. If the decay rate is very small, such systems tend to approach a terminal state resembling the performance characteristics of a perceptron with unlimited values after a long training sequence. Systems with strictly bounded values will be considered first.

9.1 Analysis of Systems with Bounded Values

Two types of analysis have been carried out for systems having upper and lower bounds for v_{ij}. The first deals with the terminal distribution of the values after a long period of exposure to a random sequence of stimuli, with S-controlled reinforcement. The second deals with the actual performance of a bounded-value perceptron. In both cases, we will follow the method of analysis originally employed by Joseph, in connection with bounded γ'-perceptrons (Ref. 41)*. All of these analytic results apply to experimental systems using S-controlled reinforcement procedures.

---

* Bounded γ'-systems have been called λ-systems in Ref. 41.

9.1.1 Terminal Value Distribution in a Bounded α-system

Suppose an α-perceptron has upper and lower limits L and ℓ for the values v_{ir}. Suppose a particular connection, c_{ir}, receives a reinforcement of +1 with probability p, -1 with probability q, and 0 with probability $1-p-q$. If all stimuli are equiprobable, and the perceptron is trained by an S-controlled procedure, this would correspond to a connection from an A-unit with bias ratio p/q (see Definition, Page 77). It is assumed in the following analysis that the reinforcements occurring at different times are statistically independent. For convenience, L and ℓ are taken to be integers. Then the value, v_{ij}, may assume any one of $L-\ell+1$ distinct states ($\ell, \ell+1, \ldots, L$). Clearly, if unit a_i responds more often to stimuli of the positive class than to stimuli of the negative class, v_{ir} will tend to grow in a positive direction. Eventually it will arrive at the limit L. At this point, a run of "negative" stimuli may bring it down again, but it can never exceed L. If the unit has a negative bias, v_{ir} will similarly tend to remain in the neighborhood of the lower limit, ℓ. The problem is to find the terminal probability distribution (if one exists) for the value v_{ir}, as the duration T of the training sequence goes to infinity.

In the following analysis, it will first be assumed that a stable terminal probability distribution for v_{ir} exists, which will not be altered by the addition of more stimuli to the training sequence. On the basis of this assumption, an equation for the distribution can be found. It will then be proven by induction that the proposed distribution is, in fact, a stable probability distribution.

Let $\pi(x)$ = probability that $v_{ir} = x$, in the terminal probability distribution. Let $\pi(\ell) = c$. This will be equal to the probability of v_{ir} arriving at ℓ from above, plus the probability that v_{ir} remains in state ℓ if it is already there. Thus,

$$\pi(\ell) = c = q\left[\pi(\ell) + \pi(\ell+1)\right] + (1-p-q)\,\pi(\ell)$$

Hence

$$\pi(\ell+1) = \frac{p\pi(\ell)}{q} = \frac{pc}{q} \tag{9.1}$$

For any integer $2 \le i \le L - \ell$,

$$\pi(\ell+i-1) = q\pi(\ell+i) + p\pi(\ell+i-2) + (1-p-q)\,\pi(\ell+i-1)$$

Hence,

$$\pi(\ell+i) = \frac{p+q}{q}\,\pi(\ell+i-1) - \frac{p}{q}\,\pi(\ell+i-2) \tag{9.2}$$

Thus, all values of $\pi(x)$ can be computed if the probability c of v_{ir} being at the lower limit is known. Since the sum of π for all possible values of v_{ir} must be 1, the value of c can be obtained from the equation:

$$\sum_{i=0}^{L-\ell} \pi(\ell+i) = 1 \tag{9.3}$$

For the distribution to be stable, it is sufficient that the probability of v_{ir} being at its upper limit satisfies the equation.

$$\pi(L) = p\pi(L-1) + (1-q)\pi(L) \tag{9.4}$$

By induction on i, it will be shown that

$$\begin{aligned}\pi(\ell+i) &= p\left[\pi(\ell+i-1)\right] + (1-q)\pi(\ell+i) \\ &= \frac{p}{q}\pi(\ell+i-1)\end{aligned} \tag{9.5}$$

for $1 \le i \le L-\ell$. (9.4) is only a special case of (9.5). To begin with, for $i = 1$, we have $\pi(\ell) = c$ and from (9.1), $\pi(\ell+1) = \frac{pc}{q}$. This clearly agrees with (9.5). Now assume (9.5) is true for $i = r$ $\left[1 \le r \le L-\ell-1\right]$. That is

$$\pi(\ell+r+1) = \frac{p}{q}\pi(\ell+r-1)$$

But by (9.2), letting $i = r+1$, we then obtain

$$\begin{aligned}\pi(\ell+r+1) &= \frac{p+q}{q}\left[\frac{p}{q}\pi(\ell+r-1)\right] - \frac{p}{q}\pi(\ell+r-1) \\ &= \frac{p}{q}\pi(\ell+r)\end{aligned}$$

Thus, having assumed (9.5) to be true for $i = r$, we find that it is also true for $i = r+1$; consequently it is true for all i, and (9.5) must be true. From (9.5) it is also clear that the quantities π will all be non-negative, so that the function $\pi(x)$ meets the requirements for a probability distribution.

Equation (9.5) can be used to compute $\pi(x)$ by assuming an arbitrary value for c, and then normalizing the distribution as in (9.3). The equation can be simplified by taking the lower limit, ℓ, equal to zero, and setting $c = 1$ for the unnormalized distribution. Then $\pi(x) = \left(\frac{p}{q}\right)^x$ prior to normalization. For the normalized distribution, $c = \left[\sum_{v=\ell}^{L} \left(\frac{p}{q}\right)^{v-\ell}\right]^{-1} = \frac{q-p}{q-p\left(\frac{p}{q}\right)^{L-\ell}} = \frac{1-\left(\frac{p}{q}\right)}{1-\left(\frac{p}{q}\right)^{L-\ell+1}}$. This completes the proof of the following theorem:

THEOREM: In a bounded α-perceptron, with S-controlled reinforcement, the probability distirubtion $\pi(v)$ (for the value of a particular connection) approaches a stable terminal distribution of the form $\pi(v) = c\left(\frac{p}{q}\right)^{v-\ell}$ where c is a normalization constant equal to $\frac{1-\left(\frac{p}{q}\right)}{1-\left(\frac{p}{q}\right)^{L-\ell+1}}$.

Figure 31 shows the probability distribution for v_{ip} for several values of $\frac{p}{q}$ and for 40 increments between the upper and lower limits. (The distributions are symmetric for equivalent values of $\frac{q}{p}$, with upper and lower limits reversed.) Note that with even a slight bias ($\frac{p}{q} \neq 1$) there is a very low probability that v_{ip} will have a sign opposite to the bias. For $\frac{p}{q} = .9$, for example (and taking $\ell = -20$ $L = +20$, as in the figure) the probability of a positive v_{ip} in the terminal distribution is only .0097. If the range were half as great (20 increments instead of 40) the probability of positive v_{ip} for the same conditions would be increased to .2295.

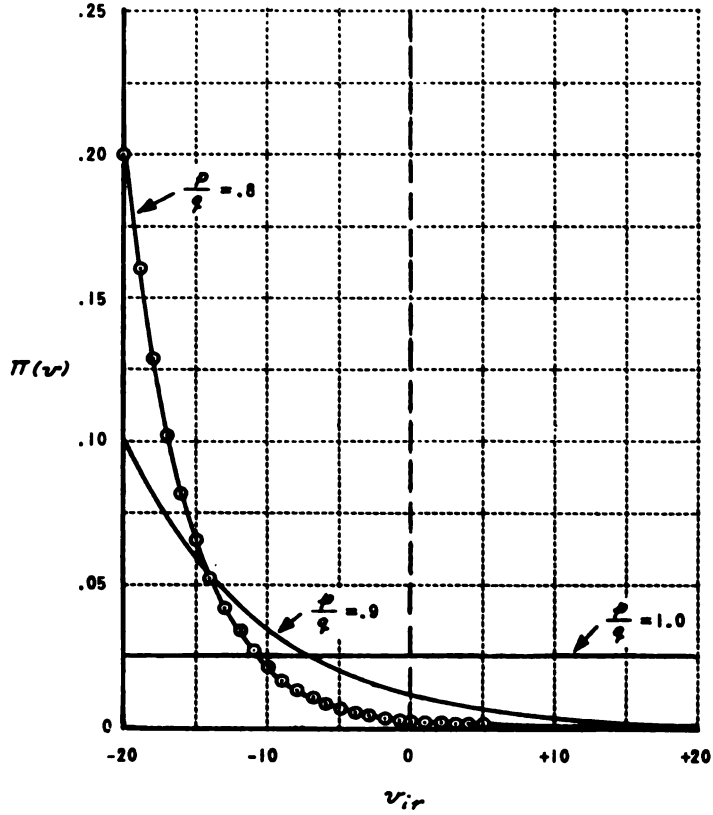


Figure 31 TERMINAL PROBABILITY DISTRIBUTION OF v_{ir} IN BOUNDED α-SYSTEM
$\ell = -20$, $L = +20$

The frequency of possible ratios $\frac{p}{q}$ for A-units responding to horizontal and vertical bars can be determined from Table 1. From this, it is clear that the majority of units have a pronounced bias towards one class or the other, so that one might expect fo find the majority of active connections having values in the neighborhood of the appropriate limit, L or ℓ. This heuristic argument supports the conjecture that the bounded system should still be capable of learning discrimination tasks in S-controlled experiments, even though the system tends to "saturate", with all values in the neighborhood of the upper or lower limit. The quantitative performance of such systems will be taken up in Section 9.1.3.

9.1.2 Terminal Value Distribution in Bounded γ-systems

In a bounded γ-perceptron, the analysis of the terminal distribution for v_{ij} is complicated by two considerations. First, there are at least four possible values of Δv, namely $1-Q_i$, $-1+Q_i$, $-Q_i$, and $+Q_i$, each with its own probability. If Q_i is not equal for all stimuli, the number of possible values for Δv is increased in proportion to the number of different values for Q_i. The second consideration is that the conservation rule, which requires the sum of all values to remain constant, makes the admissible increment for one connection dependent on how many of the other connections are currently free to move. For example, if all of the "active" connections have values equal to L, the expected decrement, $-Q_i$, for the inactive connections due to the application of a positive Δv cannot occur.

Due to these complications, an analysis for a true γ-system has never been carried out. An analysis has been completed by Joseph for a γ'-system with monopolar reinforcement (i.e., reinforcement is applied only for stimuli of the positive class, and $\eta = 0$ for stimuli of the negative class). In this case there are only two non-zero changes which might occur, $1 - Q_i$ for active connections and $-Q_i$ for inactive connections, and the reinforcement of a given connection does not depend on the state of any other parallel connections, as it does in the γ-system. The analysis is a somewhat more complicated form of that presented in the preceding section (due to the inequality of positive and negative changes in v_{ir}). Since the equations are of limited interest aside from the specific model considered, they will not be repeated here, but they can be found, together with typical distribution curves, in Ref. 41.

9.1.3 Performance of Bounded α-systems in S-controlled Experiments

From the preceding analysis, it is clear that with a large number of increments between the upper and lower limits of v_{ir} , the value will ultimately tend to remain in the neighborhood of the upper or lower bound, depending upon the bias ratio of a_i . In the following analysis, the problem is simplified by assuming that the limits are actually trapping, so that once a connection has arrived at value L or l , it remains there permanently, regardless of future reinforcement.

Consider a basic training sequence of m stimuli, $S_1 \cdots S_m$, which is then repeated a sufficient number of times to "saturate" the system, i.e., to drive all biased values to their limits. If the value of a connection is v after the first m stimuli, then after r repetitions of the training sequence, the value will be

$$\begin{aligned} &\min (L, rv) && \text{if } v > 0 \\ &\min (\ell, rv) && \text{if } v < 0 \\ &0 && \text{if } v = 0 \end{aligned}$$

for a bounded α-system. An unbounded α-system will have the same performance after r repetitions of the training sequence as after a single repetition. The following analysis compares the performance of the "saturated" bounded α-system with that of the unbounded α-system at the end of the training sequence. The analysis will be accurate for the assumption of a large range between L and ℓ, so that after the first m stimuli none of the values have reached their limits.

Let P_χ be the probability that $R = +1$ for test stimulus S_χ, for the unbounded α-system, and P'_χ be the corresponding probability for the bounded α-system. Then the conditional probability $(P'_\chi | P_\chi)$ gives the performance of the bounded system as a function of the performance of the unbounded system (which is known from Chapter 7).

Suppose N_a^* A-units are activated by the test stimulus, S_χ. Then for the unbounded system, $(P_\chi | N_a^*) = \Phi(z)$ where Φ is the cumulative distribution function defined by equation (7.7) and

$$z = \frac{\sqrt{N_a^*}\, E(v_{ir})}{\sigma(v_{ir})}$$

where $E(v_{ir})$ = expected value of a connection activated by S_χ, and $\sigma(v_{ir})$ = standard deviation of such a connection. The bounded α-system,

on the other hand, will give response +1 if the proportion of the N_a^* active connections having value L is greater than $-\ell/(L-\ell)$. If $\ell = -L$, then this reduces to a requirement that the number of active connections having value L should be greater than the number having value ℓ. The connections having value 0 may be ignored. As with the unbounded system, it is assumed that after the first m stimuli, v_{ir} is normally distributed with expected value $E(v_{ir})$ and variance $\sigma^2(v_{ir})$. This assumption is reasonable if the range of v_{ir}, $(L-\ell)$ is greater than $2m$ and m is fairly large. If the range of v_{ir} is less than $2m$, the analysis can be considered only an approximation, which becomes increasingly poor as the range diminishes.

Under these conditions, in the bounded system, the probability that the terminal value of a connection is L is equal to the probability that v_{ir} is positive after the first m stimuli. This is equal to $\Phi\left(\frac{z}{\sqrt{N_a^*}}\right)$. Since Φ is a cumulative probability distribution it is a one-to-one function from its domain to its range, and is therefore invertible. Thus, given P_χ and N_a^*, the probability P_L that a connection activated by S_χ goes to value L will be:

$$(P_L | P_\chi, N_a^*) = P_L = \Phi\left(\frac{\Phi^{-1}(P_\chi)}{\sqrt{N_a^*}}\right) \tag{9.6}$$

and this yields

$$(P_\chi' | P_\chi, N_a^*) = \sum_{y=r}^{N_a^*} \binom{N_a^*}{y} P_L^{\,y} (1-P_L)^{N_a^*-y} \tag{9.7}$$

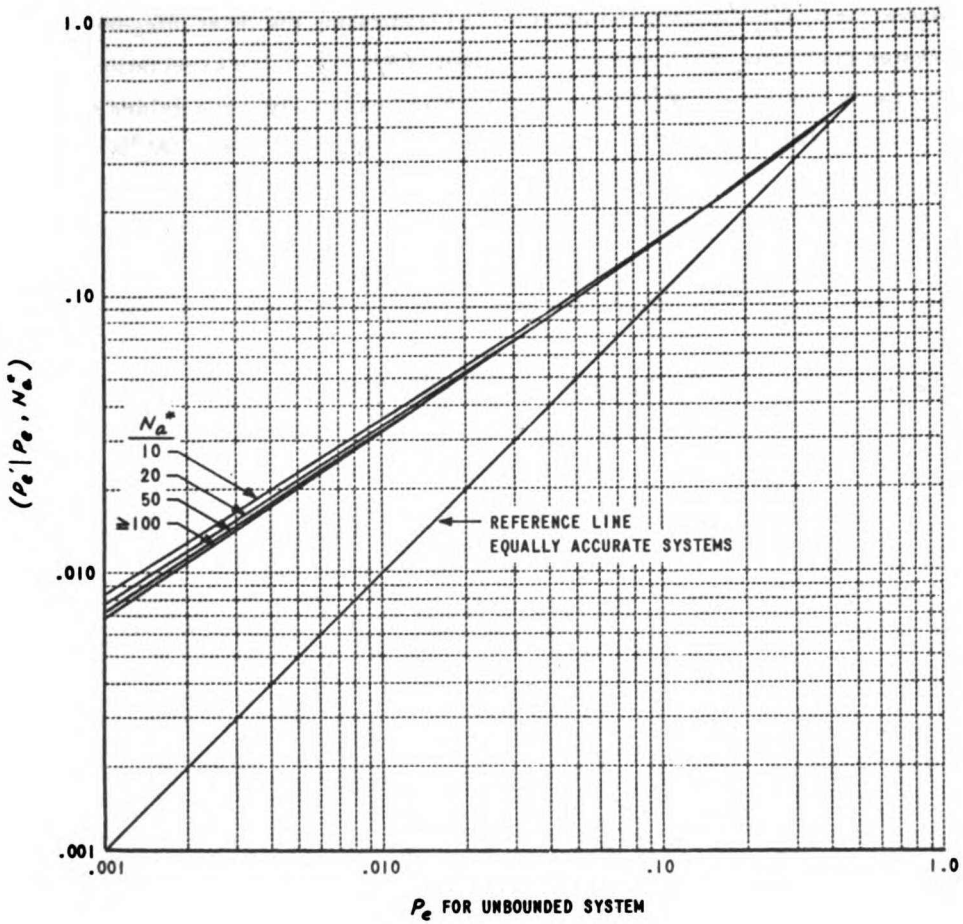


Figure 32 CONDITIONAL ERROR PROBABILITY FOR BOUNDED α-SYSTEM vs. ERROR PROBABILITY FOR UNBOUNDED SYSTEM. ($\ell = -L$)

where $r = \left[\frac{N_a^* |\ell|}{L + |\ell|}\right]$, the notation $[n]$ indicating the least integer greater than or equal to n.* To obtain $(P_\chi' | P_\chi)$, the expectation of (9.7) with respect to N_a^* is required. For reasonably large values of N_a, $(P_\chi' | P_\chi) \approx (P_\chi' | P_\chi, E(N_a^*))$. Substituting $Q_\chi N_a$ for $E(N_a^*)$ this finally yields:

$$(P_\chi' | P_\chi) \approx \sum_{y=r}^{Q_\chi N_a} \binom{Q_\chi N_a}{y} P_L^y (1-P_L)^{Q_\chi N_a - y} \tag{9.8}$$

where $P_L = \Phi\left(\frac{\Phi^{-1}(P_\chi)}{\sqrt{Q_\chi N_a}}\right)$, $r = \left[\frac{Q_\chi N_a |\ell|}{L + |\ell|}\right]$.

In Figure 32, the conditional probability of error in a bounded α-perceptron is shown as a function of the error probability $(1-P_\chi)$ for the unbounded system, for several values of N_a^*. $\frac{|\ell|}{L+|\ell|}$ is taken to be 1/2. Curves of this function for cases where upper and lower limits are not symmetric can be found in Joseph, Ref. 41 (Figures 10-14).

9.1.4 Performance of Bounded γ-systems in S-controlled Experiments

The analysis in the preceding section, and the curves shown in Fig. 32, can be applied without modification to bounded γ'-perceptrons. The true γ-system, however, may perform somewhat better than the γ'-system, since not all values can "saturate" independently. If more than half of the connections have a positive bias, for example, not all of the positively biased connections can go to the limit L, since this would

---

* It is assumed here that $L > 0$, $\ell < 0$.

require that the remaining connections take on values less than ℓ, in order to satisfy the conservation rule. In the γ-system, therefore, we would expect a greater number of connections to remain at intermediate values, rather than going to the limits, and this should result in a "compromise" between the performance of an unbounded and a bounded value system. An exact analysis of the γ-system has not been carried out.

9.2 Analysis of Systems with Decaying Values

The bounded value systems have two disadvantages relative to the "ideal" unbounded systems. First, they permit a smaller number of memory states, and second, in S-controlled experiments they tend to arrive at a saturation condition in which their performance is actually poorer than that obtained during the transient learning phase; that is, their performance curve first increases to a maximum, and then declines to a terminal asymptote as the system saturates. The first disadvantage is not serious, if the range of $v_{i,r}$ is reasonably large. The second may be more critical, since it means that units with a low "utility" for a given discrimination are pulling as much weight in the saturated system as units with high utility (as measured by their bias ratios). In the cross-coupled perceptrons considered in Part III, this latter consideration is more salient than in elementary perceptrons.

An alternative value-limiting mechanism, which is also of interest due to its apparent biological plausibility, is obtained by allowing the values to decay exponentially towards a resting state (generally taken to be zero). This mechanism is relatively free from the difficulties

encountered in the bounded value system. In this model, v_{ir} will continue to grow in the direction determined by the bias ratio of a_i, until the expected rate of reinforcement is exactly balanced by the rate of decay. At this point a dynamic equilibrium will occur, with v_{ir} tending to fluctuate about the equilibrium level. This means that connections which are frequently reinforced, in a consistent direction, will attain higher values, in the limit, than infrequently reinforced connections, or connections with low bias.

Consider an α-system with decaying values. Let the decay rate be equal to $\delta (\delta << 1)$. Let the probabilities of positive and negative increments to v_{ir} be p and q, as in the analysis of bounded α-systems. As long as δ is small, v_{ir} will tend to approach an expected asymptotic value equal to $(p-q)/\delta$. At this point, the expected rate of gain, per unit time, is $p-q$, and the expected rate of loss is $\delta v_{ir} = p-q$. If the value of δ is very small, and the relaxation time correspondingly long relative to the expected recurrence rate of stimuli from the environment, this system should approach as a limit the same performance as the unbounded α-system, where v_{ir} tends to grow in proportion to $p-q$. If δ is somewhat larger, however, we find that the most recent stimuli in the training sequence will have the most pronounced effect, progressively earlier stimuli exerting a progressively dimishing effect due to the decay of v_{ir}. Such a perceptron tends to forget its remote experience in favor of more recent experience.

The dependence of these systems on the sequence as well as the identity of training stimuli makes them difficult to analyze when the relaxation time, or "half-life" of v_{ir} is on the same order as, or

shorter than, the training sequence. If δ is sufficiently small, performance can be assumed identical to the unbounded system. An absolute bound on the maximum attainable magnitude of $v_{i,r}$ for a decaying value perceptron will be $1/\delta$, corresponding to a situation in which $c_{i,r}$ is reinforced continuously in the same direction.

9.3 Experiments with Decaying Value Perceptrons

9.3.1 S-controlled Discrimination Experiments

The essential features of S-controlled discrimination experiments with decaying value perceptrons have already been noted in the preceding section. If the decay rate is small, the decaying value system approaches the performance of the corresponding "ideal" or unbounded system. If the decay rate is relatively large, forgetting occurs, which is greatest for temporally remote events and negligible for recent events in the training sequence.

9.3.2 Error-correction Experiments

In discrimination experiments with error corrective reinforcement, a more complicated situation exists than in the case of S-controlled experiments. In the error correction system, once the perceptron has learned a task, reinforcement ceases, and the values of a decaying system would be expected to decay back towards zero. In a perfectly noise-free system, the values would all decay in proportion

to their magnitudes, however, and consequently their ratios would never change as long as no further reinforcement was applied. Thus once perfect performance is achieved, it will not be lost as long as the values remain above the noise-level of the system, despite the decay effect. This also means that if a "run" of correct responses occurs during training, the ratios of v_{ir} for different connections will be unaltered, so that the next error to occur will be no different in the decaying value model than in the unbounded model. Consequently, the application of reinforcement just sufficient to correct this error will bring the ratios of the values to precisely the state that they would have in the unbounded model, and ability to achieve a solution to a classification problem should be unaffected, in principle.. In actuality, however, the continuously decaying values clearly present a problem, since any physical system will ultimately forget, when the values become small enough to be undetectable.

A variation of the decaying value model is capable of eliminating the problem caused by the diminution of the values in an unreinforced system. If v_{ir} is held constant so long as no reinforcement signal is received from the reinforcement control system, but decays exponentially in the presence of such a signal, the learning ability of the perceptron will still be unaltered (by the same argument as above), and no change will occur once the task has been properly learned. This means that the increment to the value of v_{ir} at time t will be

$$\Delta v_{ir}(t) = \left[a_i^*(t) - \alpha v_{ir}(t)\right] \cdot \eta(t)$$

where $\eta(t)$ may be $+\epsilon$, $-\epsilon$, or 0.

It should be noted that in the error-correction procedure, the loss of temporally remote experience with large values of δ does not occur, in an ideally functioning (noise-free) system. Unlike the S-controlled system, where the magnitude of new reinforcements remains unchanged as the values decay, the error correction procedure will require smaller or less frequent increments in order to correct an error, and earlier experience tends to be retained about as well as in the unbounded, or non-decaying system. A loss of early experience does occur, in such systems, but it is due to "writing over" earlier memory traces with more recent reinforcement, rather than to a passive decay, as in the case of the S-controlled system. This observation would seem to indicate a closer correspondence of the error-corrective system with what is known of forgetting in biological systems.

The mean performance curves for eight simulated perceptrons with $\delta = 0$, $\delta = .001$, and $\delta = .01$ are shown in Fig. 33. Note that for these actual systems, there is a progressive deterioration of performance as the decay rate is increased.

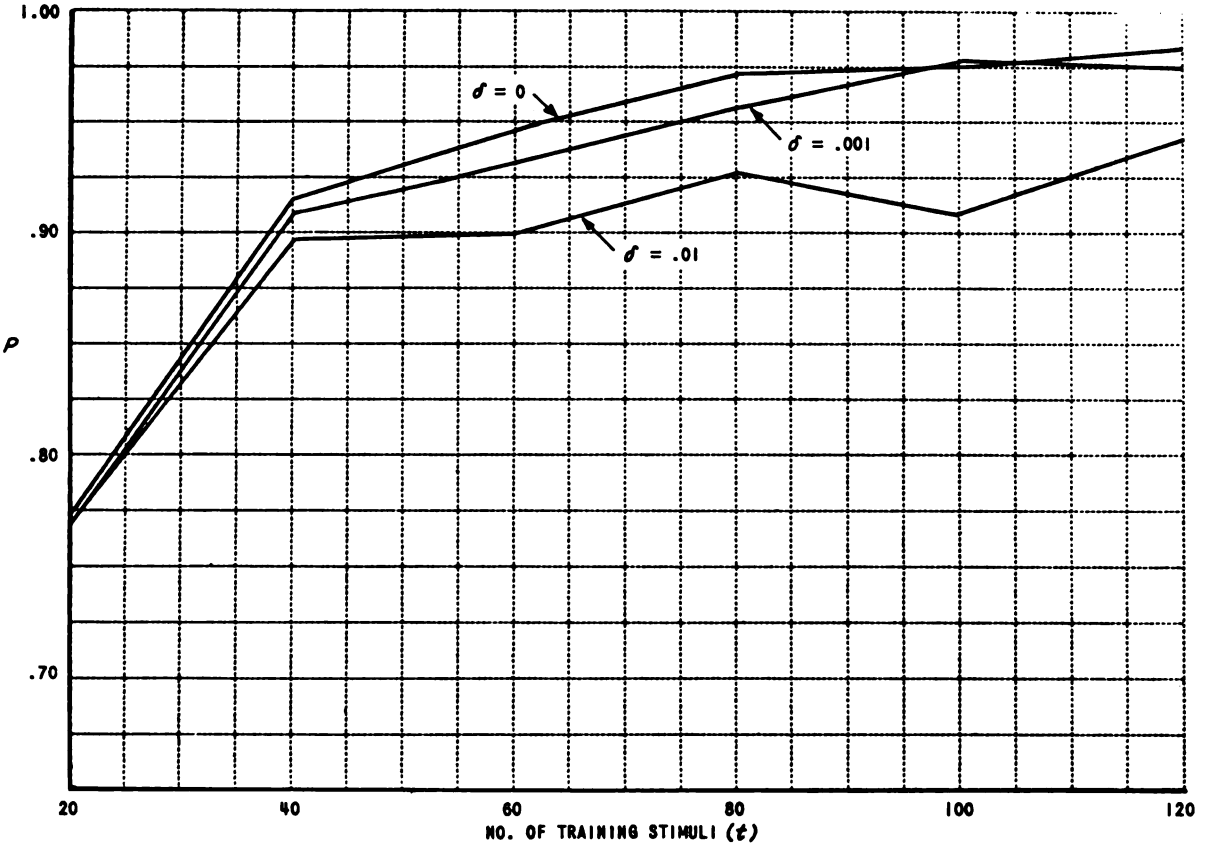


Figure 33 PERFORMANCE OF 8 DECAYING α-PERCEPTRONS, WITH $x = 3$, $y = 1$, $\theta = 2$, ON HORIZONTAL/VERTICAL BAR. DISCRIMINATION EXPERIMENT WITH ERROR CORRECTIVE REINFORCEMENT

9.3.3 R-controlled Experiments

The most interesting experimental results obtained to date with decaying value perceptrons deal with the performance of decaying $\mathcal{T}$-systems in R-controlled experiments. Experiment 9 has been studied most extensively, by means of simulation experiments representing a very large, or infinite N_a, perceptron. Unlike the previous experiments (discussed in Section 8.3) monopolar reinforcement was employed, i.e., the perceptron was reinforced positively for $r^* = +1$, and was not reinforced at all for $r^* = -1$. The system was further modified by assuming a slight negative quantity to be added to $\Delta v_{ir}(t)$ for all i; that is, an invariant negative reinforcement component was added uniformly to all connections, regardless of what stimulus occurred, and regardless of the activity state of the connection. In the absence of any other components, this would cause a progressive downward drift of all v_{ir} until they achieved an equilibrium with the decay rate. It was assumed that this negative component was sufficient to add a quantity equal to -0.0001 to the set of connections activated by a single stimulus. Thus, apart from the decay, the change in values for each reinforcement could be expressed by the equation:

$$g_{ij} = Q_{ij} - Q_i Q_j - 0.0001$$

The effect of the fixed negative component in these experiments is to create a negative generalization from the first stimulus to occur (say a horizontal bar) to all members of the opposite class (vertical bars) in place of the zero generalization which would otherwise occur with a

$\mathcal{C}'$-system. The result is that after having seen a single stimulus which activates a positive response, all members of the opposite class are thenceforth permanently classified in the negative class, as no further events can occur which will make one of them positive. If the initial stimulus is a horizontal bar, then, with monopolar reinforcement, no vertical bar will be reinforced, since all vertical bars evoke a -1 response. The next stimulus which can possibly be reinforced is, in fact, another horizontal bar which happens to be close enough to the previous one to have received positive generalization from the first reinforcement, i.e., the first or second neighbor on either side. The result is a gradual growth of the positive stimulus set, by accretion of near neighbors which have received positive generalization from those bars already classified as "positive". Thus, having started out by randomly placing a horizontal bar in the positive class, the system has no choice but to include only horizontal bars in the positive class, and, with sufficient time, all horizontal bars are so classified.

While this phenomenon occurs even if the decay rate is zero, it is markedly accelerated by a non-zero decay rate. With $\theta = 0$, the perceptron shows a high degree of "rigidity" in its early classification, in which some horizontal bars are positive, and the remainder still negative (as in Section 8.3). This is due to the continually increasing magnitude of the negative values evoked by the "incorrectly" classified stimuli, which must be overcome in order to change their classification. Thus, as time progresses, it becomes harder and harder to switch each additional horizontal bar into the positive class, since an increasingly large number of

"marginal" positive stimuli must be reinforced in order to obtain the required amount of positive generalization. Moreover, as the positive class expands, the stimuli which are centrally located within the "positive band" all contribute further negative generalization to the remaining stimuli, rather than helping to make them positive. These combined effects lead to a convex, negatively accelerating learning curve, as illustrated in Figure 33. The addition of a non-zero decay rate limits the negative value which must be overcome in order to change the classification of an "incorrect" stimulus, and thus makes the system more flexible.

If the decay rate is increased progressively, it is found that there is an optimum at about $\delta = 0.01$. If the decay rate is increased further, instability occurs, due to the loss of stimuli which were previously classified correctly, but whose positive values have decayed to such an extent as to be overcome by negative generalization from other stimuli. These effects are shown both in the learning curves of Fig. 34(a) and in Fig. 34(b), which shows the expected learning time to perfect performance (i.e., perfect dichotomization of horizontal and vertical bars), obtained from a sample of 10 runs.

It might seem, from these results, that perceptrons organized in the manner indicated could be expected to form "meaningful" classifications of stimuli, on some basis other than retinal position. Unfortunately, the results, while illuminating, are highly restricted in generality. The proposed dynamics are too contrived to be biologically plausible, and it is found that in any environment in which classes of stimuli to be

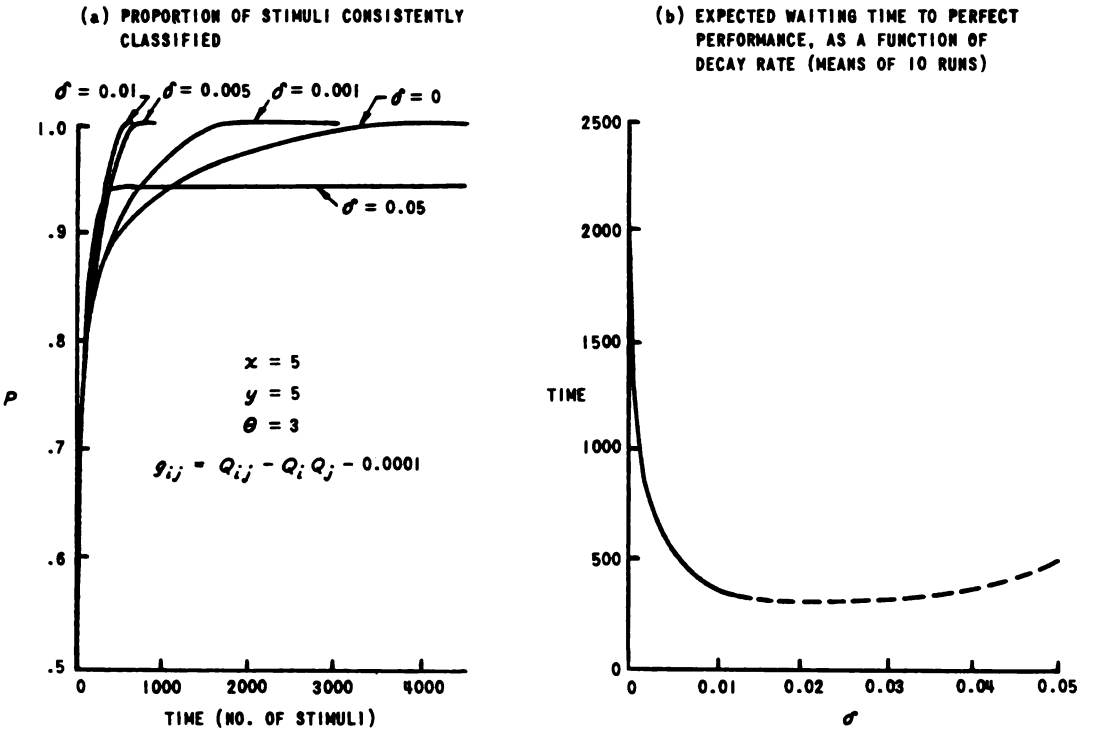


Figure 34 SPONTANEOUS ORGANIZATION OF INFINITE PERCEPTRON IN ENVIRONMENT OF 4 x 20 VERTICAL AND HORIZONTAL BARS

differentiated permit positive generalization between members of different classes (a much more usual situation) the mechanism which yields good separation in the above example breaks down. If g_{ij} between a single horizontal bar and any of the vertical bars were positive, for example, the spread of generalization would not stop with the members of the horizontal class, in the above case, but would invade the opposite class as well. If, instead of 4 by 20 horzontal and vertical bars, the perceptron is confronted with an environment consisting of the twenty horizontal bars and a set of twenty pairs of parallel 2 by 20 horizontal bars, separated by a space of 3 units on the retina, the perceptron will not spontaneously learn to distinguish single bars from double bars (although this task presents no difficulty in an S-controlled experiment).

Another shortcoming of the spontaneous organization phenomenon which has been demonstrated here is the basically unbiological character of the learning curves. It has already been noted that these curves are convex, or decelerating. A human subject, or even an animal subject, confronted with the problem of distinguishing horizontal from vertical bars might make many mistakes initially, but would soon accelerate his learning as he began to generalize to new stimuli. If he had a hundred bars, in different retinal positions, to classify, the hundredth bar would certainly not present the almost insurmountable obstacle that it represents for the elementary perceptron. Thus it is clear that the most sophisticated generalization phenomena which have yet been found in elementary perceptrons are still far short of what one should expect from an adequate brain model, if biological standards are employed. This problem will be re-examined at greater length in Part III, where it will be seen that multi-layer and cross-coupled perceptrons perform such tasks in a much more suitable fashion than those systems which have been considered thus far.

This completes the presentation of elementary perceptrons. In the following chapters, some other types of minimal (S-A-R) perceptrons will be considered, but it will be seen that none of these have capabilities for generalization appreciably beyond those discovered in the elementary systems.

10. SIMPLE PERCEPTRONS WITH NON-SIMPLE A AND R-UNITS

In Chapter 4, a simple perceptron was defined as one which satisfies the following five conditions:

1. There is a single R-unit, with a connection from every A-unit.
2. The perceptron is series coupled, with an S-A-R topology.
3. The values of all S-A connections are invariant.
4. Transmission times of all connections are equal (τ generally taken as 0).
5. All signals generated by S, A, and R-units are functions of the algebraic sum of input signals arriving simultaneously at the unit.

In the preceding chapters, we have considered elementary perceptrons, which are characterized by the additional constraints that all A and R-units are "simple" units, and that the transmission function of the connection c_{ij} takes the form: $c^{\#}_{ij}(t) = a^{\#}_{i}(t-\tau)\, v_{ij}(t)$. A simple A-unit is a signal generating unit which emits an output signal $a^{\#}_{i} = +1$ if the algebraic sum of the input signals, α_i , is equal or greater than the threshold θ , and 0 otherwise. A simple R-unit emits a +1 signal if the sum of its input signals is strictly positive, and -1 if the sum of its inputs is strictly negative. In this chapter, we shall consider the properties of simple perceptrons in which these contraints are dropped. This will include a brief consideration of linear networks

in which all signals are transmitted in proportion to their value; the properties of perceptrons with linear R-units but non-linear A-units will then be considered, and finally the question of optimum transmission functions will be discussed. In later chapters, the remaining constraints of simple perceptrons will be modified, and a number of non-simple systems will be analyzed.

10.1 Completely Linear Perceptrons

A completely linear perceptron is one in which all signal functions and transmission functions are linear, i.e., the output of unit u_i is of the form $u_i^* = c_i a_i$, and the signal transmitted by a connection c_{ij} is of the form $c_{ij}^* = u_i^* v_{ij}$. We will consider linear perceptrons in environments such that the inputs to an S-unit are either 1 or 0 (so that the conclusions apply equally well to perceptrons which are linear everywhere except in the S-units). By analogy to Section 5.4, we define the bias ratio of an S-unit as n^+/n^-, where n^+ is the number of positive stimuli, and n^- the number or negative stimuli which activate the S-unit. For such systems, the following theorem holds:

THEOREM 1: Given a completely linear perceptron, a stimulus world, W, and a classification $C(W)$ such that the bias ratio of every S-unit is equal (and non-zero), no solution to $C(W)$ can exist.

PROOF: Let k^+ = index of any stimulus in positive class (S_{k^+}).
k^- = index of any stimulus in negative class (S_{k^-}).
λ = index of λ^{th} sensory unit
$c_{\lambda i}^*(k)$ = signal transmitted from the λ^{th} sensory unit to the i^{th} A-unit in response to stimulus S_k.

When stimulus S_k occurs, unit a_i transmits a signal equal to $\alpha_i(k) v_{ir}$ to the R-unit, where

$$\alpha_i(k) = \sum_\lambda c^*_{\lambda i}(k)$$

The total signal, u_k, received by the R-unit from S_k is therefore:

$$u_k = \sum_i \alpha_i(k)\, v_{ir} = \sum_i \sum_\lambda c^*_{\lambda i}(k)\, v_{ir}$$

Since every signal u_k must agree in sign with the classification of S_k for a solution to exist, we require that the following inequalities be satisfied:

$$\sum_i \sum_\lambda \sum_{k^+} c^*_{\lambda i}(k^+)\, v_{ir} > 0 \qquad (10.1)$$

$$\sum_i \sum_\lambda \sum_{k^-} c^*_{\lambda i}(k^-)\, v_{ir} < 0 \qquad (10.2)$$

But it has been stipulated that the bias ratio of each S-point is equal to a constant, $r > 0$. This means that, for any i and λ,

$$\sum_{k^+} c^*_{\lambda i}(k^+) = r \sum_{k^-} c^*_{\lambda i}(k^-) \qquad (r > 0)$$

or, summing over S-units,

$$\sum_\lambda \sum_{k^+} c^*_{\lambda i}(k^+) = r \sum_\lambda \sum_{k^-} c^*_{\lambda i}(k^-) = c$$

Substituting in the expressions (10.1) and (10.2) we get the contradiction

$$\begin{cases} \sum_i c\, v_{ir} > 0 \\ r \sum_i c\, v_{ir} < 0 \end{cases}$$

which proves that a solution cannot exist.

This means that if two stimulus patterns are placed in all possible positions on a retina, the resulting classes of stimuli cannot be correctly discriminated by a linear perceptron. As a consequence, such systems are relatively uninteresting, even though they may successfully discriminate a moderate number of patterns which are restricted to limited positions on the retina. In all systems considered from here on, there will be at least one set of non-linear components subsequent to the S-units in the perceptron network.

10.2 Perceptrons with Continuous R-units

The next type of perceptron to be considered has simple A-units, but continuous R-units, such that the response $r_i^* = \phi(u_i)$, with ϕ an arbitrary monotonic function of u_i. This includes the case of linear R-units, where $\phi(u_i) = c\,u_i$. An important theorem which is analogous to Theorem 4 of Chapter 5 deals with the ability of such systems to learn arbitrary response functions (Definition 27, Chapter 4) under the error correction procedure. A response function assigns an arbitrary output signal (rather than just ± 1) to every stimulus in W. We first prove the following Lemma:

LEMMA 1: Given a symmetric positive definite or positive semidefinite matrix, H, and any vector z, then $(z, Hz) = 0$ only if $Hz = 0$.

PROOF: Since H is positive definite or semidefinite, there exists a matrix B such that $H = B'B$.

$$0 = (z, Hz) = (Bz, Bz)$$

$$\Rightarrow Bz = 0 \Rightarrow 0 = B'Bz = Hz$$

THEOREM 2: Given a simple α -perceptron with simple A-units, an R-unit with a continuous monotonic sign-preserving signal generating function, a stimulus world W (in which each stimulus ultimately reoccurs) and any response function $R(W)$ for which a solution exists, then by means of the error-corrective reinforcement procedure, the given response function can always be approximated in finite time by an output vector $R(W) + \epsilon$, where ϵ is a vector of elements $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, $|\epsilon_i| < \epsilon'$, where ϵ' may be an arbitrarily small quantity greater than zero.

PROOF: The following proof was suggested by R. D. Joseph. From Theorem 3 of Chapter 5, we know that under the conditions of the theorem, a solution v to the equation $Gv = u$ exists. Suppose the system is currently in the state x , represented by $Gy = x$. From the definition of the G-matrix, and the fact that every stimulus must activate at least one A-unit for a solution to exist, we have

$$1 \geq g_{ii} \geq (g_{ii})_{min} > 0$$

The difference between the solution vector u and the present state x is given by

$$G(v-y) = u - x$$

Let $v - y = z$ and

$u - x = w$

Then $Gz = w$.

We wish to show that by applying an error correction method to one component at a time of the vector z, w must ultimately go to a point within the ϵ' cube about 0. (The method will apply a correction of the proper size until a response $r^* = R^*$ is obtained.) We know that $u_i = \sum_j g_{ij} x_j$. Therefore, for the difference, w, we have

$$w_i = \sum_j g_{ij} z_j$$

$$\frac{\partial w_i}{\partial z_i} = g_{ii} > 0$$

Since G is non-negative definite, we know that $F = (z, Gz) \geq 0$, $\frac{\partial F}{\partial z_i} = w_i$, and from Lemma 1 we know that if $w \neq 0$, $F > 0$. Therefore, if $w_i > 0$ decreases as a result of decreasing z_i, F decreases; also, if $w_i < 0$ increases by increasing z_i, F decreases (see Proof of Theorem 4, Chapter 5). To prove the theorem, it is sufficient to show that this implies that w must ultimately enter the ϵ'' cube about zero.

Let w_i' = initial value of w_i at start of a correction step

z_i' = initial value of z_i at start of a correction step

Then for the correction, we have

$$\Delta w_i = -w_i'$$

$$\frac{\partial w_i}{\partial z_i} = g_{ii}$$

$$\Delta z_i = -\frac{w_i'}{g_{ii}}$$

$$\frac{\partial F}{\partial z_i} = 2 w_i = \left[w_i' + g_{ii} (z_i - z_i') \right]$$

$$\int_{z_i'}^{z_i' - \frac{w_i'}{g_{ii}}} \frac{\partial F}{\partial z_i} \, dz_i = 2 \int_{z_i'}^{z_i' - \frac{w_i'}{g_{ii}}} \left[w_i' + g_{ii} (z_i - z_i') \right] dz_i$$

$$\Delta F = \frac{1}{g_{ii}} \left[w_i' + g_{ii} (z_i - z_i') \right]^2 \Bigg|_{z_i'}^{z_i' - \frac{w_i'}{g_{ii}}}$$

$$= -\frac{w_i'^2}{g_{ii}}$$

Therefore, $\Delta F < -w_i'^2 < -\epsilon''^2$

Hence, there can be only a finite number of corrections, since $F \geq 0$, and the vector $w = u - x$ must converge to a point within the ϵ'' cube about zero. But u is the input to the R-unit. Since $r^*(u)$ is continuous, there exists an ϵ'' such that $|r^*(u+\sigma) - r^*(u)| < \epsilon'$ if $|\sigma| \leq \epsilon''$. Therefore the response function coverges together with the vector w. Q.E.D.

The following Lemma and Corollaries establish that the various weaker forms of correction procedures are also capable of yielding a solution to $R(W)$.

LEMMA 2: For the same conditions as Theorem 2, given that a solution exists, the set of all solutions forms a hyperplane of dimension equal to the nullity of G .

PROOF: Let $Gx = u$ be a solution. Of necessity $u_i = r_i$. Let $Gy = u$ be another solution. Then $G(x-y) = 0$, consequently $x-y$ is in the null space of G . Conversely, if $z-x$ is in the null space of G, then $G(z-x) = 0$. Therefore, $Gz = u$, so that z is a solution. Q.E.D.

COROLLARY 1: For the conditions of Theorem 2, and a phase space which is unbounded in all dimensions, the probability of convergence to an arbitrarily close approximation to $R(W)$ by means of a random-sign correction procedure or a random-perturbation correction procedure may be less than 1.

PROOF: The random-sign and random-perturbation procedures were defined in Section 5.6. $R(W)$ is taken to be any response function, obtainable by an R-unit with a monotonic signal generating function. For convergence to occur, it would be necessary that a series of steps by increments of fixed magnitude, $|\eta|$, but of random sign, should carry the system from its initial state to an arbitrarily small distance, ϵ , from its required state. From Lemma 2, the solution states form a hyper-

plane of dimension equal to the nullity of G, which has zero measure over the phase space of the system. But a random walk of the type described may carry the system arbitrarily far from its starting point, in a random direction, and the probability that a vertex of this path will fall within a distance ϵ of the solution hyperplane may be less than unity.

COROLLARY 2: Given the conditions of Theorem 2, and a phase space bounded in all dimensions, then (given that a solution to $R(W)$ exists in this bounded space) the response function can always be approximated by means of the random-sign correction procedure, the system converging in finite time to an approximation $R(W) + \epsilon$, ϵ a vector, where $|\epsilon_i| < \epsilon'$ for arbitrarily small $\epsilon' > 0$.

PROOF: Since the phase space is finite, the set of solution points within the bounds defined above has positive measure. The random-sign correction procedure cannot carry any of the A-unit outputs beyond the limit set for its value; therefore, if the values approach their limit in any direction, a random walk in the opposite direction will follow. This procedure will ultimately take the representative point of the system into every set with positive measure, provided η is sufficiently small. Consequently, a solution within the bounds stated by the theorem will be obtained in finite time.

COROLLARY 3: Given the same conditions as Corollary 2, the response function can always be approximated by the random-perturbation correction procedure, the

system converging in finite time to an approximation $R(W) + \epsilon$, ϵ having elements of magnitude $|\epsilon_i| \leq |\eta|$ if the reinforcement is quantized, or $|\epsilon_i| \leq \epsilon' > 0$, if η is chosen from a continuous distribution around zero.

PROOF: The proof follows the same line as that of Corollary 2. Since each connection can be set to an independent value, in the quantized case the total error over the set of all connections need not be greater than η , while in the continuous case it may be made arbitrarily small.

Theorem 2 and its corollaries indicate that it is possible to teach a simple perceptron to produce responses which are proportional to some metric feature of the input stimuli, such as their size, or coordinates of their center of gravity on the retina. In the latter case, the output of such an R-unit can be fed back to the optical system to control the centering of a stimulus in the field.

10.3 Perceptrons with Non-linear Transmission Functions

In all perceptrons considered thus far, the transmission functions of connections from A-units to the R-unit have been of the form

$$c_{ir}^* = a_i^* v_{ir}$$

We will now consider functions of the more general form:

$$c^*_{ir} = f(a^*_i, v_{ir})$$

Where time is not specified, this is understood to mean

$$c^*_{ir}(t) = f\left(a^*_i(t-\tau), v_{ir}(t)\right)$$

Since a^*_i is a function of the input signal, α_i, the transmission function can be written in a still more general form (allowing for various types of signal-generating functions in the A-units),

$$c^*_{ir}(t) = F(\alpha_i, v_{ir})$$

This form will be employed in the following theorems.

THEOREM 3: Given a simple perceptron with a simple R-unit, and with transmission functions for all A-R connections of the form $f(\alpha_i)\, v_{ir}$, where f is any function, and given the existence of a solution to a classification function $C(W)$ for this perceptron, then if $p(v)$ is any polynomial of odd degree in v, there also exists a solution if the transmission function is changed to $f(\alpha_i)\, p(v_{ir})$.

PROOF: A polynomial of odd degree can assume all possible values. Therefore if v_{ir} is the original value of the connection c_{ir}, there exists a solution to $p(x) = v_{ir}$ yielding a new value, x, for the connection c_{ir} which will cause it to transmit an identical signal under the new transmission function.

THEOREM 4: Given the perceptron of Theorem 3, if a solution exists for some transmission function $f(\alpha_i)\,v_{ir}$, a solution does not necessarily exist for the transmission function $g(\alpha_i)\,v_{ir}$, $g \neq f$.

PROOF: Suppose the number of A-units is equal to the number of stimuli in W. Let B = matrix of elements b_{ij} representing the value of the function $f(\alpha_i(j))$ which is the coefficient of v_{ir} for stimulus S_j. Then for a solution to exist, there must be some vector V and some vector U in the orthant required by $C(W)$, such that $B'V = U$. But if B is singular, there must be some $C(W)$ for which no solution exists. This can be demonstrated by noting that each $C(W)$ requires a solution vector in a different orthant, the set of all $C(W)$ requiring solutions in every possible orthant. But if B is singular, it maps the entire space into a hyperplane, and this plane must fail to intersect certain orthants. Consequently, the functions $C(W)$ which are represented by vectors in those orthants have no solution. Now consider the following cases:

CASE 1: For the transmission function αv, let the matrix

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{pmatrix}$$

This is singular, and consequently there are some insoluble classifications. Now change the transmission function to $\alpha^2 v$, yielding

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 9 \\ 4 & 9 & 16 \end{pmatrix}$$

This matrix is non-singular, so that with the non-linear transfer function, all classifications are soluble.

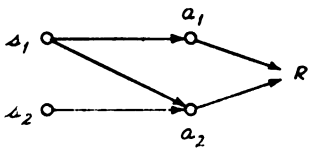
CASE 2: In this case it is shown, conversely, that there may be situations in which a linear transmission function will yield solutions which are unobtainable with a particular non-linear function. Let the transmission function be αv, with the matrix $B = \begin{pmatrix} 3 & 5 & 8 \\ 4 & 12 & 15 \\ 5 & 13 & 17 \end{pmatrix}$ This matrix is non-singular, so there is a solution for every $C(W)$. But now let the transmission function be $\alpha^2 v$. Then $B = \begin{pmatrix} 9 & 25 & 64 \\ 16 & 144 & 225 \\ 25 & 169 & 289 \end{pmatrix}$ which is singular, implying that there is some $C(W)$ with no solution.

THEOREM 5: Given a simple perceptron with A-R connections which differ in their transmission functions (or with uniform transmission functions but non-simple A-units) a response function $R(W)$ may have a solution which is unattainable by either the error correction procedure or the random-sign correction procedure.

PROOF: Consider a perceptron with a single sensory unit and two A-units. Let the R-unit be a linear amplifier with gain of 1. Let the sensory unit emit signals 0, 1, or 2 depending upon the intensity of the stimulus. The required response function is $R(W) = (0, +1, -1)$ corresponding to a null stimulus, a low-intensity stimulus, and a high-intensity stimulus, respectively. Let the transmission function of $\mathcal{L}_{1r}$ be $|\alpha| v$, and the transmission function of $\mathcal{L}_{2r}$ be $\alpha^2 v$. The response function $R(W)$ then has a solution if we

set $v_{1r} = 2.5$ and $v_{2r} = -1.5$. But this is the only possible solution, and is unattainable by the error correction or random-sign procedures, since both connections are always activated together and consequently must always be equal in value under these procedures (assuming that their initial values are equal). This example is sufficient to prove the theorem for the case of non-uniform transmission functions.

For the second case, in which all transmission functions are uniform, but the perceptron has non-simple A-units, consider the following perceptron:



The values of all S-A connections are +1, and the A-units are both linear, with transmission function αv. Let the environment consist of the two stimuli $S_1 = A_1$ and $S_2 = (A_1, A_2)$. Then a solution exists to the response function $R = (+1, -2)$, namely $v_{1r} = +3$, $v_{2r} = -2$. However, the error-correction or random-sign correction procedures will not work, since both A-units are always active (where "active" means that they emit a non-zero signal). Note that a solution also exists to the classification $(+1, -1)$ for this perceptron, and that this is also unattainable by the methods indicated.

The sixth theorem was proposed by R. D. Joseph.

THEOREM 6: Given a simple perceptron with any mixture of transmission functions $f_j(\alpha_j, v_{jr})$ for the connections c_{jr}, and a response function $R(W)$ for which a solution exists; then there exists some transmission function $g(\alpha, v)$ which is uniform for all connections, such that a solution to $R(W)$ exists.

PROOF: Let $f_j(\alpha_j, v_{jr})$ = signal from unit a_j when stimulus S_i occurs. Then we can fit a polynomial

$$f_j(\alpha_j(i), v_{jr}) = \sum_{k=0}^{n-1} c_{jk}\, \alpha_j^k(i)$$

for each stimulus S_i. The coefficients, c_{jk}, (which depend on the A-unit, a_j) can be replaced by polynomials

$$c_{jk} = c_k(j) = \sum_{l=0}^{N_a-1} b_{lk}\, j^l$$

Thus we have, for all values of j,

$$f_j(\alpha_j(i), v_{jr}) = \sum_{k=0}^{n-1} \sum_{l=0}^{N_a-1} b_{lk}\, j^l \alpha_j^k(i) = g(\alpha, j)$$

which satisfies the conditions required by the theorem for $g(\alpha, v)$ if we set $v_{jr} = j$.

It should be noted that this theorem applies only to a given response function for which a solution exists; if a different response function also has a solution, then there will again be a uniform transmission function for all A-units which will solve the problem, but this transmission function may differ from the one obtained for the original response function.

We have seen in Theorem 5 that if the connections differ in transmission functions, or the A-units differ in signal generating functions, response functions may have solutions which cannot be obtained by the more systematic correction procedures. The following theorem proves that in this case the weakest of the correction procedures (the random perturbation method) can still be used successfully.

THEOREM 7: Given a simple perceptron with an R-unit which is either simple or has a continuous signal generating function, and with any combination of transmission functions from its A-units (all continuous functions of $v_{i,r}$, equal to zero if $\alpha_i = 0$), and given a bounded phase space within which a solution exists for $R(W)$; then, if each stimulus in W ultimately reoccurs, an approximate solution $R(W) + \epsilon$ is always attainable in finite time by the random-perturbation correction procedure.

PROOF: For an R-unit of the specified type, and a bounded phase space, the solution set has positive measure, over the region defined by $R(W) + \epsilon$ (where ϵ consists of arbitrarily small elements, $\epsilon_i \leq \epsilon'$). To achieve an approximate solution within this set, it is only necessary to adjust the values of the active A-units for each stimulus. Since, under the random

perturbation procedure, each active connection will independently tend to assume a value in every admissible range with positive measure, the active set of connections as a whole will ultimately attain a value configuration within the solution set.

10.4 Optimum Transmission Functions

The general conclusions of the preceding pages are that while a completely linear perceptron does not work satisfactorily, there are many possible transmission functions which seem to work quite well. For many of these, there is no choice to be made from the standpoint of ability to achieve a solution, for they all seem to be capable of solving the same problems equally well. From the standpoint of efficiency of discrimination and speed of learning, however, the various transmission functions might differ considerably from one another. In this section, making use of an analysis due to Joseph, it will be shown that with some fairly weak constraints on the system under consideration, an optimum transmission function exists, and that this takes the form of a quadratic function of v_{ir} rather than a linear function.

The constraints on the system to be analyzed are as follows:

1. The analysis deals with S-controlled discrimination experiments, with a fixed training sequence.

2. The conditional distribution of v_{ir} for connections activated by a test stimulus of the positive class, S_x, is assumed to be independent of the choice of S_x. Similarly, the distribution of v_{ir} for active

connections is assumed to be independent of the exact choice of S_χ when the test stimulus is selected from the negative class.

3. It is assumed that the conditional distribution of v_{ir} for the connections activated by S_χ is a normal distribution, and that either the distributions are different or the probabilities Q_i are different, for test stimuli in the positive and negative classes. These constraints will generally be met satisfactorily if the positive class consits of all possible positions on the retina of a large stimulus, and the negative class consists of all possible positions of a small stimulus. The main requirement is one of equivalence of stimuli within each class, and dissimilarity between classes, with respect to the distribution or number of signals transmitted from A-units to the R-unit.

The discrimination problem can be stated as one of testing a hypothesis about the test stimulus, S_χ. The response unit is required to test the hypothesis that S_χ is a member of the positive class against the possibility that it is a member of the negative class. If the test stimulus is a member of the positive class, the output of an A-unit (subject to the above assumptions about the system being analyzed) will have the distribution

$$\begin{cases} 0 & \text{with probability } 1 - Q_\chi(+) \\ v & \text{with density function } \dfrac{Q_\chi(+)}{\sqrt{2\pi}\,\sigma_{(+)}} \exp.\left\{-\dfrac{1}{2\sigma^2_{(+)}}(v-\mu_{(+)})^2\right\} \end{cases} \tag{10.3}$$

where $Q_\chi(+)$, $\sigma_{(+)}$, and $\mu_{(+)}$ are the parameters characterizing stimuli of the positive class. Similarly, if the test stimulus is a member of the negative class, the output of an A-unit will have the distribution

$$\begin{cases} 0 & \text{with probability } 1 - Q_\chi(-) \\ v & \text{with density function } \frac{Q_\chi(-)}{\sqrt{2\pi}\,\sigma_{(-)}} \exp.\left\{-\frac{1}{2\sigma_{(-)}^2}(v-\mu_{(-)})^2\right\} \end{cases} \tag{10.4}$$

where $Q_\chi(-)$, $\sigma_{(-)}$, and $\mu_{(-)}$ are the parameters characterizing stimuli of the negative class. Thus, the problem can be restated as one of testing whether the output of an A-unit has the distribution (10.3) or the distribution (10.4).

There is thus a simple hypothesis (dealing with a single distribution) and a simple alternative. As Joseph has observed, under these conditions, for any significance level, the likelihood ratio test is most powerful. In performing this test, we would make N independent observations of v (corresponding to a sample of N A-units with independent origin point configurations), and obtain the likelihood ratio:

$$L = \left(\frac{1-Q_\chi(+)}{1-Q_\chi(-)}\right)^{N-N^*}\left(\frac{Q_\chi(+)\sigma_{(-)}}{Q_\chi(-)\sigma_{(+)}}\right)^{N^*} \exp.\left\{-\frac{1}{2\sigma_{(+)}^2}\sum_i (v_i-\mu_{(+)})^2 + \frac{1}{2\sigma_{(-)}^2}\sum_i (v_i-\mu_{(-)})^2\right\} -$$

where N^* is the number of active A-units, and the summation on i is over active units only. If L is greater than a preassigned constant L_0, we accept the hypothesis that S_X is a member of the positive class; if L is less than L_0, we accept the alternative, that S_X is a member of the negative class. The constant L_0, corresponding to the threshold of the R-unit in a perceptron employing this procedure, determines the power and significance of the test. (The "significance" is measured by the probability of erroneously rejecting a positive stimulus, and the "power" is the probability of correctly classifying a negative stimulus.) In logarithmic form, the condition $L \geq L_0$ becomes

$$\sum_i \left\{ \left(\frac{1}{2\sigma_{(-)}^2} - \frac{1}{2\sigma_{(+)}^2} \right) v_i^2 - \left(\frac{\mu_{(-)}}{\sigma_{(-)}^2} - \frac{\mu_{(+)}}{\sigma_{(+)}^2} \right) v_i + \frac{\mu_{(-)}^2}{2\sigma_{(-)}^2} - \frac{\mu_{(+)}^2}{2\sigma_{(+)}^2} + \ln \frac{Q_X(+)(1-Q_X(-))\sigma_{(-)}}{Q_X(-)(1-Q_X(+))\sigma_{(+)}} \right\}$$
$$\geq \frac{L_0 (1-Q_X(-))^N}{(1-Q_X(+))^N}$$

Thus, the required test is effectively performed if the perceptron is designed with R-units having a threshold $\ln L_0 + N \ln \frac{1-Q_X(-)}{1-Q_X(+)}$ and the transmission functions from A to R-units are of the form

$$f(\alpha, v) = \begin{cases} 0 \text{ if } \alpha < \theta \\ \left(\frac{1}{2\sigma_{(-)}^2} - \frac{1}{2\sigma_{(+)}^2} \right) v^2 - \left(\frac{\mu_{(-)}}{\sigma_{(-)}^2} - \frac{\mu_{(+)}}{\sigma_{(+)}^2} \right) v + \frac{\mu_{(-)}^2}{2\sigma_{(-)}^2} - \frac{\mu_{(+)}^2}{2\sigma_{(+)}^2} + \ln \frac{Q_X(+)(1-Q_X(-))\sigma_{(-)}}{Q_X(-)(1-Q_X(+))\sigma_{(+)}} \text{ if } \alpha \geq \theta \end{cases}$$

The actual savings that might be obtained by the use of such a quadratic form have not been investigated numerically. In practise, they are probably slight. A further discussion of the optimization problem, including the optimization of the upper and lower bounds in a bounded value perceptron, can be found in Joseph, Ref. 41.*

---

* Prof. A. Gamba, in a related paper, has observed that not only the transmission functions but the reinforcement rule might be profitably modified in order to optimize the overall decision function of the system (Ref. 23).

11. PERCEPTRONS WITH DISTRIBUTED TRANSMISSION TIMES

One of the requirements for a simple perceptron is that the transmission time, $\tau_{i,j}$, should be equal for all connections, $c_{i,j}$. In this chapter, we consider the consequences of allowing a distribution of transmission times. It is obvious that under these conditions the set of A-units active at time t will depend not on the single momentary stimulus occurring at time $t - \tau$, but rather on the entire sequence of stimuli occurring between $t - \tau_{min}$ and $t - \tau_{max}$. We shall first consider the cases of binomial and Poisson models where $\tau_{i,j}$ is distributed with a discrete spectrum, $\tau_{i,j}$ always being an integer equal to or greater than 1. We shall then consider the case of a continuous Gaussian distribution for $\tau_{i,j}$.

11.1 Binomial Models with Discrete Spectrum of $\tau_{i,j}$

For the binomial case, we shall consider only the case where each A-unit receives a fixed number of connections of each type (excitatory and inhibitory) with $\tau_{i,j} = 1$, and a fixed number with $\tau_{i,j} = 2$. Specifically, the parameters of an A-unit are:

θ = threshold (defined as usual)

x_1 = number of excitatory connections with $\tau_{i,j} = 1$

y_1 = number of inhibitory connections with $\tau_{i,j} = 1$

x_2 = number of excitatory connections with $\tau_{i,j} = 2$

y_2 = number of inhibitory connections with $\tau_{i,j} = 2$

Models with a greater number of possible values for τ_{ij} can be analyzed by extensions of the method applied here. The object of the analysis is to find Q_i and Q_{ij} at time t, as functions of the two-step sequences of stimuli:

$$\mathcal{J}_i = S_i'(t-2),\ S_i(t-1)$$

$$\mathcal{J}_j = S_j'(t-2),\ S_j(t-1)$$

The notation S_i' will be used consistently to denote the stimulus preceding the terminal stimulus in sequence $\mathcal{J}_i$. Similarly, in sequences of more than two stimuli, S_i'' will be used to denote the third stimulus from the end, etc. In the present model, sequences of length greater than 2 need not be considered. If it is assumed that A to R-unit connections all have equal transmission times, the analysis of performance in terms of the Q-functions will be identical with the analysis for simple perceptrons, the important difference being that the perceptron is now learning to recognize sequences of stimuli, rather than isolated momentary events.

The total input signal to an A-unit at time t, $\alpha(t)$, is now a sum of four components, namely,

$$\alpha(t) = E_1 + E_2 - I_1 - I_2$$

where E_1 = number of excitatory connections with $\tau = 1$, having origins active at $t-1$

I_1 = number of inhibitory connections with $\tau = 1$, having origins active at $t-1$

E_2 = number of excitatory connections with $\tau = 2$, having origins active at $t-2$

I_2 = number of inhibitory connections with $\tau = 2$, having origins active at $t-2$.

As usual, $a_i^*(t) = 1$ if $\alpha_i(t) \geq \Theta$, and 0 otherwise. Q_i is then given by the following equation, which is analogous to (6.1):

$$Q_i = \sum_{E_1+E_2-I_1-I_2 \geq \theta} P_{x_1}(E_1)\, P_{x_2}(E_2)\, P_{y_1}(I_1)\, P_{y_2}(I_2) \qquad (11.1)$$

where the probabilities P_{x_1}, P_{x_2}, P_{y_1} and P_{y_2} are defined as in (6.2), with the substitution of the appropriate parameters, and the stimulus measures R_i in the expressions for P_{x_1} and P_{y_1} and $R_{i'}$ in the expressions for P_{x_2} and P_{y_2}.

In a similar manner, the expression for Q_{ij} can be obtained by the extension of the treatment employed in Equations 6.5 and 6.6. However, there are now eight components to be considered for α for each stimulus sequence. Specifically,

$$\alpha(i) = E_i + E_c + E_{i'} + E_{c'} - I_i - I_c - I_{i'} - I_{c'}$$

$$\alpha(j) = E_j + E_c + E_{j'} + E_{c'} - I_j - I_c - I_{j'} - I_{c'}$$

where E_i and I_i are defined, as before, as the excitatory and inhibitory components originating from the set of retinal points situated in S_i and not in S_j, $E_{i'}$ and $I_{i'}$ are the corresponding components originating from the set of retinal points situated in $S_{i'}$ but not in $S_{j'}$, and E_j, I_j, $E_{j'}$, and $I_{j'}$ are similarly defined. Likewise, E_c and I_c are the

excitatory and inhibitory components coming from the retinal set common to S_i and S_j, and $E_{\mathcal{L}'}$ and $I_{\mathcal{L}'}$, are the components from the set common to $S_{i'}$ and $S_{j'}$. Thus we have the equation

$$Q_{ij} = \sum_{\substack{\alpha(i) \geq \theta \\ \alpha(j) \geq \theta}} P_{x_1}(E_i, E_j, E_{\mathcal{L}}) P_{y_1}(I_i, I_j, I_{\mathcal{L}}) P_{x_2}(E_{i'}, E_{j'}, E_{\mathcal{L}'}) P_{y_2}(I_{i'}, I_{j'}, I_{\mathcal{L}'}) \qquad (11.2)$$

The required multinomial probabilities being computed from equations (6.6) with an obvious extension of the above notation to the quantities A_i, A_j, C, $A_{i'}$, $A_{j'}$, and C'.

Since the Poisson model is much easier to compute, and has properties which are similar in all essentials to the binomial model, no numerical examples are given for the binomial model, but examples for the Poisson model can be found in the following section.

11.2 Poisson Models with Discrete Spectrum of τ_{ij}

The Poisson model to be considered again has two values of τ, namely $\tau = 1$ and $\tau = 2$, the parameters $\bar{x}_1$, $\bar{x}_2$, $\bar{y}_1$, and $\bar{y}_2$ being defined analogously to $\bar{x}$ and $\bar{y}$ in the Poisson model considered in Chapter 6. The equations for Q_i and Q_{ij} can, of course, be developed by extension of the equations of Chapter 6, as has just been done for the binomial model. A considerably simpler approach is possible in the Poisson model, however, if the corresponding stimulus areas at times $t-1$ and $t-2$ are also equal, i.e., $A_i = A_{i'}$, $A_j = A_{j'}$, and $C = C'$. In this

case, the previous equations (6.1, 6.3, 6.5, and 6.7) hold without modification, except that $x = x_1 + x_2$ and $y = y_1 + y_2$. More generally, the previous equations can always be employed by making the appropriate substitutions:

$$\begin{aligned}
\bar{x} R_i &= \bar{x}_1 R_i + \bar{x}_2 R_i' \\
\bar{x} R_j &= \bar{x}_1 R_j + \bar{x}_2 R_j' \\
\bar{x} A_i &= \bar{x}_1 A_i + \bar{x}_2 A_i' \\
\bar{x} A_j &= \bar{x}_1 A_j + \bar{x}_2 A_j' \\
\bar{x} C &= \bar{x}_1 C + \bar{x}_2 C'
\end{aligned}$$

and similarly, for the inhibitory components. If $\bar{x}_1 = \bar{x}_2$ and $\bar{y}_1 = \bar{y}_2$, the equations for Q_i and Q_{ij} again become identical with the equations of Chapter 6 where $R_i = \frac{1}{2}(R_i + R_i')$, $A_i = \frac{1}{2}(A_i + A_i')$, etc. By an obvious extension to a spectrum with three or more values of τ, where $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_n$, and $\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_n$, we can apply the same equations, substituting the parameters

$$\begin{aligned}
R_i &= \frac{1}{n}(R_i + R_i' + R_i'' + \ldots) \\
A_i &= \frac{1}{n}(A_i + A_i' + A_i'' + \ldots) \\
C &= \frac{1}{n}(C + C' + C'' + \ldots)
\end{aligned}$$

and similarly for R_j and A_j.

As an example of the performance of such a system, consider a Poisson model perceptron with an expected value of 6 excitatory and 6 inhibitory connections to each A-unit and $\theta = 2$. Let the environment consits of a set of 4 by 20 vertical bars, such as were employed in the experiments of the preceding chapters. The object will be to discriminate a bar arriving at a certain fixed location by movement from the left from a bar which arrives at the same location by movements from the right. Clearly, if a single value of $\tau_{i,j}$ is permitted, this task is impossible. Consider first the case in which half of the excitatory and half of the inhibitory connections have $\tau = 1$ and the remaining half have $\tau = 2$, so that $\bar{x}_1 = \bar{x}_2 = \bar{y}_1 = \bar{y}_2 = 3$. Let sequence $\mathcal{A}_i$ denote $(S_a(t-3), S_b(t-2), S_c(t-1))$ and $\mathcal{A}_j$ denote $(S_e(t-3), S_d(t-2), S_c(t-1))$, where $S_a, \ldots, S_e$ represent successive adjacent positions of the vertical bar on the retina. Then $Q_i = Q_{ii} = .153$, and $Q_{ij} = .094$. Next, suppose one third of the excitatory connections and one third of the inhibitory connections have delays $\tau = 3$, one third have $\tau = 2$, and one third have $\tau = 1$, so that $\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{y}_1 = \bar{y}_2 = \bar{y}_3 = 2$. In this case, $Q_{ii} = .153$, as before, but Q_{ij} is reduced to .063. Further increasing the spread of the τ distributuion will have the effect of further reducing Q_{ij} (for correspondingly lengthened stimulus sequences) while keeping Q_{ii} constant. Thus, the greater the spread of the τ distribution, the more readily can such "divergent" time sequences be distinguished. Conversely, two sequences which are identical save for a momentary divergence in recent time (say at $t-1$) can be distinguished most readily by a perceptron with $\tau_{i,j}$ concentrated at small values, and increasing the spread of the τ distribution will only increase the difficulty of discrimination.

It should be emphasized that the set of active A-units depends on the order and not merely on the constituents of a stimulus sequence. Thus the sequence (S_1 , S_2 , S_3) will generally activate a different set of A-units from the sequence (S_1 , S_2 , S_3) in which the first two members have been inverted. In principle, a perceptron of this type which receives sequences of sound spectra from a set of audio-filters (instead of visual patterns) should be capable of distinguishing spoken words, or other characteristic sound sequences, such as progressions of chords or melodic fragments.

11.3 Models with Normal Distribution of $\tau_{i,j}$

A somewhat more "natural" model than the discrete spectrum models considered above is one where the transmission time of each connection is an independent random variable drawn from a normal distribution, with parameters $\mu(\tau)$ and $\sigma(\tau)$. If an A-unit is to have a non-zero probability of being active at time t in such a model, the dynamics must be modified by the introduction of an "integration period", Δt , such that

$$\alpha(t) = \sum_{\tau=t-\Delta t}^{t} E(\tau) - I(\tau) \qquad (11.3)$$

summing over all values of τ for which E or I (the numbers of excitatory or inhibitory impulses arriving at the A-unit) are non-zero.

The qualitative properties of such a system are clear without further analysis. If Δt is short compared to $\sigma(\tau)$, the presentation of a "momentary" or transient stimulus will lead to a gradual increase in the

proportion of responding A-units (or the value of Q_i) followed by a gradual decrease. If Δt is greater than $\sigma(\tau)$, the system will respond with a momentary burst of activity, maintained for a period equal to Δt, and will then immediately relapse to inactivity. We are chiefly concerned with the case where Δt is less than $\sigma(\tau)$. In this case, the performance of the system in discriminating sequences will be close to that of the Poisson or binomial models, with an appropriate discrete spectrum of τ_{ij}, to approximate the normal distribution. There will be a maximum sensitivity to differences between the two sequences $\mathcal{A}_i$ and $\mathcal{A}_j$ occurring at time $t - \mu(\tau)$, with less sensitivity to more recent or more remote differences between the sequences.

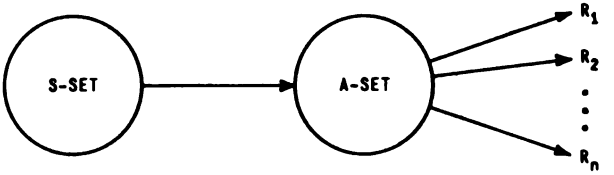
12. PERCEPTRONS WITH MULTIPLE R-UNITS

Up to now, the simple "three-layer" topology (S-A-R) with a single R-unit has been the only one considered. In this chapter, we will still consider only three-layer perceptrons, but more than one R-unit will be permitted. The performance of such systems, it will be seen, does not differ significantly from that of perceptrons which have been considered in previous chapters, except for the fact that it is now possible to form classifications with more than two classes, with simple R-units, or to have perceptrons respond simultaneously to several different attributes of a stimulus pattern. The most interesting analytic problems for such systems are concerned with the optimum coding of the classes of patterns to be recognized, in order to optimize performance.

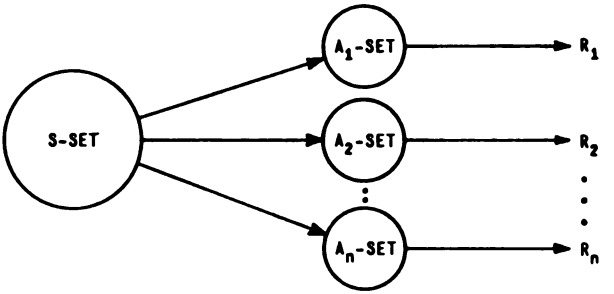
12.1 Performance Analysis for Multiple R-unit Perceptrons

Several types of topological organization which are possible for networks with more than one R-unit are illustrated in Figure 35. The set of A-units which are connected to a given R-unit will be called the source-set of that R-unit. The organization which is most economical in the number of A-units employed is that shown in Fig. 35(a), where every A-unit is connected to every R-unit. This is logically equivalent to the disjoint source-set model shown in Fig. 35(b), if every source set is required to have the same composition of origin point configurations for its A-units. Unless otherwise specified, it will be assumed that each R-unit receives the same number of input connections; however, if the R-set is large, and the terminus of each connection from an A-unit is selected at random, the total number of inputs to each R-unit

(a) EVERY A-UNIT CONNECTED TO $\mathcal{L}$ R-UNITS. (IN FULLY COUPLED CASE, $\mathcal{L} = N_R$)



(b) DISJOINT SOURCE-SET FOR EACH R-UNIT. (SPECIAL CASE OF (a) WHERE $\mathcal{L} = 1$)



(c) EACH R-UNIT HAS SOURCE SET OF N RANDOMLY SELECTED A-UNITS. (EQUIVALENT TO (a) IF $N = N_a$)

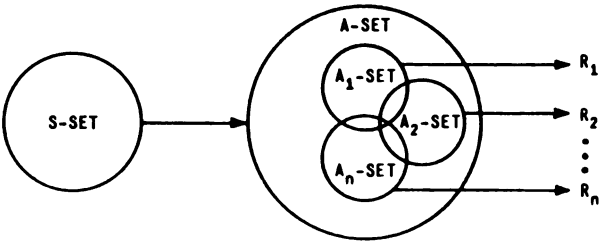


Figure 35 TYPES OF TOPOLOGICAL ORGANIZATION FOR PERCEPTRONS WITH MULTIPLE R-UNITS

(i.e., the size of its source set) will be a binomially distributed random variable. An inversion of this connection procedure is shown in Fig. 35(c). In this case, each R-unit receives exactly N connections, but the origins are assigned at random among the A-units. Here the number of output connections from an A-unit will be a Poisson distributed random variable.

It can be readily seen that as N_a becomes large, the various topological connection schemes illustrated in Fig. 35 all become logically equivalent in their performance characteristics, since it does not matter to the performance of the perceptron whether two R-units are connected to the identical A-unit or to two different A-units with equivalent origin point configurations. For the sake of specificity, the following discussion will assume the organization illustrated in Fig. 35(b), with a disjoint source-set for each R-unit.

In S-controlled discrimination experiments, it is obvious that performance of such a system in equivalent to that of N_R simple perceptrons (where N_R is the number of R-units) each of which is exposed to the same training sequence, but trained on its own independent dichotomy of the environment. For example, if $N_R = 2$, one R-unit might be trained to discriminate between stimuli in the upper and lower halves of the field, while the second R-unit is taught to discriminate between right and left halves. The probability that both responses are correct, at the end of the training sequence, will be the product of the probability that R_1 is correct on its dichotomy, and the probability that R_2 is correct on its dichotomy. In the present case, assuming that stimuli occur with equal frequency in all parts of the field, we would expect the two dichotomies to be equally difficult, so that the probability of correct performance on the joint response would be the square of the probability of correct response for either dichotomy considered separately.

In an error correction procedure, a more interesting problem arises. Clearly, if each R-unit and its set of input connections are corrected on an assigned binary classification or response function independently of the other R-units, the same situation exists as in S-controlled experiments, and the probability of correct response on the entire set of N_R R-units after a given training sequence will be the product of the probabilities for each of the response functions considered separately. More generally, if we let

$P_\chi(R_i(W), N_i)$ = probability of correct response on test stimulus S_χ for the i^{th} response function, given a source-set with N_i members connected to the R-unit, we have

$$P_\chi(R_1, \ldots, R_n) = \prod_i P_\chi(R_i(W), N_i) \qquad (12.1)$$

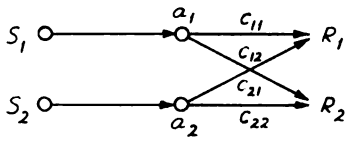
for the probability that the joint response to S_χ is correct on all R-units.

Suppose, however, the reinforcement control system is only capable of recognizing that the total response (on all R-units jointly) is right or wrong, and cannot tell which individual R-units are contributing to the error. In this case, it might be supposed that the system would eventually learn the correct joint response by assuming that every R-unit is wrong whenever an error in the composite response occurs, and correcting the perceptron accordingly. This supposition, unfortunately, is not true, as proven by the following theorem.

THEOREM: Given a perceptron with more than one R-unit, and a response function $R(W)$ or a classification $C(W)$ for which a solution exists, it may be impossible to achieve

this solution by an error correction procedure which applies negative reinforcement jointly to all R-units based on errors in the joint response.

PROOF: The theorem can be proven by a simple example. Consider the perceptron illustrated below, which has two sensory units, two A-units, and two R-units. (The topology corresponds, in this case, to Fig. 35(a).



Assume all v_{ir} initially = + 1. Let W consist of two stimuli: S_1 illuminates sensory point a_1 alone, and S_2 illuminates a_2 alone. Let the required joint classification function be:

$$(r_1^*, r_2^*) = (+1, -1) \quad \text{for} \quad S_1$$

$$(r_1^*, r_2^*) = (-1, +1) \quad \text{for} \quad S_2$$

A solution clearly exists, e.g., by making v_{11} and v_{22} positive, and v_{12} and v_{21} negative. Since all v_{ir} are initially positive, whichever stimulus occurs first (say S_1) will elicit a positive output from both R-units, which is wrong. The error correction procedure would then apply negative reinforcement to both R-units, having the effect (if S_1 is the stimulus) of making both connections from a_1 negative. But this now makes both

R-units negative, which is still wrong. Clearly, the error cannot be corrected by reinforcement in the presence of S_1, since the signals to both R-units are coupled, and must rise or fall together. If the second stimulus should occur, the situation is not improved, and the same oscillatory behavior will continue, with the perceptron switching from $(r_1^*, r_2^*) = (+1, +1)$ to $(-1, -1)$ alternately. Thus a solution will never be achieved, which proves the theorem.

Note that if, instead of administering negative reinforcement to all R-units (which assumes that each one is currently wrong) the error correction procedure were to be modified to apply a correction to each response unit according to the rule

$$\eta_i = (R_i^* - r_i^*) \tag{12.2}$$

where η_i = value of η employed in reinforcement of the R_i connections, and R_i^* and r_i^* are the required and obtained responses, respectively, for i^{th} R-unit, we then have the same conditions as in the case of independent correction of each R-unit (see Definition 41, Chapter 5). Thus, if we let $\vec{\eta} = \vec{R^*} - \vec{r^*}$ be a vector of N_R components, the i^{th} component being given by (12.2), the system will always converge if a solution exists. This implies, however, that the r.c.s. must not only be able to recognize the existence of an error in some R-component, but must be able to determine the magnitude (or at least the sign) of the error for each R-unit independently, and control an appropriate value of η_i for each section of the network. A logically similar procedure, which also yields a solution, is to allow the r.c.s. to scan the R-units sequentially, checking

the correctness of each one in turn, and applying a correction only to the R-unit currently being examined by applying negative reinforcement when it is wrong. This requires a longer training process, but requires the r.c.s. to act on only one component at a time, just as in a simple perceptron.

12.2 Coding and Code-Optimization in Multiple Response Perceptrons

A perceptron with a large number of R-units can clearly be used to identify many more than two alternative kinds of stimuli. A number of possible schemes for the representation of information in such systems have been suggested. As a first possibility, each response may be used to identify an independent trait, or property of the stimulus, such as left/right location, size, horizontal or vertical elongation, etc. The combination of responses occurring when a test stimulus is presented should then serve as a description of the stimulus in terms of its traits. An alternative scheme is to assign a distinct response unit to each kind of stimulus, and train the perceptron to emit a +1 response only if that type of stimulus is present. In this case, only one R-unit at a time would be active, the active unit identifying the stimulus class. Unlike the first scheme, where some response must be made for every binary trait whether applicable or not, the second scheme has the possibility of rejecting a stimulus altogether as "unknown", in which case all R-unit outputs would be negative. On the other hand, the second scheme lacks the economy of which the first is capable, and requires that every combination of traits which is to be distinguished must be assigned a special category and taught to the perceptron before it can be recognized. In the "trait discrimination" approach, a new configuration may still be correctly described, in terms of the characteristics present, even though it

has not been seen before. (This last feature is only weakly present in the perceptrons considered thus far, since it depends strongly on generalization. Some of the perceptrons to be considered in later chapters, which generalize more effectively, can make optimum use of "descriptive codes".)

The above examples illustrate two types of response-codes, which will be called configuration codes and position codes, respectively. A configuration code employs the R-units independently of one another, assigning an arbitrary dichotomy to each. This results in the assignment of a binary number (if the R-units are two-state devices) to each stimulus. The total number of stimulus types which can be encoded in this fashion, for a perceptron with N_R R-units, is 2^{N_R}. A position code, on the other hand, permits only one R-unit to be "on" (or in the positive state) for any one stimulus; the code takes the form of a binary number of N_R bits all but one of which are zeros. The position of the non-zero bit indicates the class of the stimulus identified. With this system, only N_R types of stimuli can be recognized. The position code can be considered a special case of a configuration code in which the positive classes of all dichotomies are disjoint, and the negative classes are almost completely intersecting. A compromise between the two approaches (which permits a descriptive statement to be obtained about a stimulus without forcing a decision on inapplicable characteristics) would assign n response units to each set of n mutually exclusive traits (for example, 2 R-units would be assigned to left/right description, 3 to horizontal, vertical, or diagonal specification, etc.). Each R-unit would then be made to discriminate between "trait present" and "trait absent", permitting any combination to occur. Such a system will be classed under configuration codes.

The problem of finding an optimum code for a particular task can be specified for a given value of N_R, an environment, W, and a classification, $C(W)$, into N types of stimuli. Clearly, if N is greater than N_R, a configuration code must be used, or the problem is insoluble. If N is commensurate with N_R, however, we have a choice of either assigning a position code, in which each R-unit identifies the presence or absence of a single type of stimulus, or assigning a configuration code, in which each R-unit is assigned an arbitrary dichotomy. In general, the problem is to find the optimum set of dichotomies to be assigned to the R-units, so as to obtain the greatest probability of correct identification for an arbitrarily selected test stimulus. Let us assume all stimuli equally likely to occur, and all classes of equal size (i.e., an equal number of stimuli in each). The number of A-units connected to each R-unit is also assumed to be constant.

Let the vector $R^* = (r_1^*, r_2^*, \ldots, r_{N_R}^*)$ = the correct response vector for a given test stimulus. Then, from equation (12.1) we are required to maximize

$$P_\chi(R^*) = \prod_i P_\chi(r_i^*)$$

Since we further assume that S_χ is chosen arbitrarily, and that every stimulus is equally likely to be chosen as a stimulus, we require the expected value

$$E(P_\chi) = \frac{1}{n}\sum_\chi P_\chi = \frac{1}{n}\sum_\chi \prod_i P_\chi(r_i^*) \qquad (12.3)$$

to be maximal. The choice of dichotomies which maximizes (12.3) would be considered an optimum code for the environment and perceptron in question.

At present, no general solution to this problem has been found. Several heuristic cues as to the organization of optimal codes are worth noting, however.

(1) If a given stimulus class has members which are disjoint from the stimuli of all other classes, while the remaining classes have large retinal intersections, it will clearly be advantageous to employ a single R-unit for the recognition of the stimulus class in question, with a highly assymmetric dichotomy which does not attempt to divide the remaining stimuli into two sub-sets, but takes advantage of the "natural" dichotomy formed on the basis of location.

(2) If the relationships of all stimulus classes are symmetric, so that no two classes tend to "stick together" more than any other two classes, and no pair of classes are easier to discriminate than any others, and if S-controlled reinforcement is to be used, it will probably be best to use equal dichotomies for all R-units, ($n/2$ stimuli in each positive set) so as to avoid asymmetric generalizations from the larger set to the smaller one. The results of the frequency bias experiments, illustrated in Figs. 16 and 25, appear to support this conjecture. Where an error correction method is used, however, empirical results suggest that asymmetric dichotomies are preferable.

(3) There exist classifications which cannot be achieved by means of a position code, which can be achieved with a configuration code. For example, consider the following case: Let there be three stimuli in W, such that S_1 activates a_1, S_2 activates a_2 and S_3 activates

both a_1 and a_2. Let there be three simple R-units, each connected to both a_1 and a_2. It is required to assign a unique code number to each of the three stimuli. With a position code, the R-unit assigned to identify S_3 must give a positive response when both a_1 and a_2 are active, but a negative response when either a_1 or a_2 alone is active. This is clearly impossible, with simple R-units. However, if a configuration code is employed, we can assign the R-function $(r_1^*, r_2^*, r_3^*) =$

(+1, -1, -1) for S_1
(-1, +1, -1) for S_2
(+1, +1, -1) for S_3

which is readily soluble, by an error correction procedure. R_3 is obviously redundant here, and is arbitrarily set to -1 for all stimuli.*

(4) A general rule, proposed by Joseph, is the following: The smallest possible number of R-units should be required to distinguish between very similar stimuli. The more dissimilar two stimuli are, the more R-units may be allowed to place the two in opposite classes.

---

* Note that in this example, it is possible to assign an arbitrary classification to an environment of 3 stimuli with only 2 A-units. This could not be done with a simple perceptron (as proven in Corollary 2 of Theorem 3, Chapter 5). The addition of a second R-unit in this model substitutes for the missing A-unit which would otherwise be required.

In empirical tests with the Mark I perceptron (such as the experiments described in the following section) it has been found that the choice of a code, even with binary numbers of a fixed length, can easily determine whether or not a particular task is within the perceptron's capability.

12.3 Experiments with Multiple Response Systems

The Mark I perceptron at C.A.L. is equipped with eight binary R-units, and 512 A-units, which can be employed in any combination. The network topology is of the type shown in Fig. 35(b). A number of experiments have been performed (Ref. 30) dealing with the recognition of letters of the alphabet and sets of geometrical patterns where multiple classifications are required. Two such experiments are illustrated in Figures 36 and 37.

In Fig. 36, learning curves are shown for an S-controlled experiment on the left, and for an error-correction experiment on the right. In each case, the perceptron was taught to identify eight letters of the alphabet, presented in the form of large block letters in random locations, over a considerable part of the retinal field. In the error correction procedure, each of the erroneous R-units is corrected simultaneously.

Figure 37 shows the learning curve for the entire alphabet, presented in fixed position. A partially optimized binary code employing five R-units was used here. This represents about the limit of the capacity of the Mark I system. Attempts at teaching the Mark I to recognize all 26 letters in two type faces simultaneously have been unsuccessful, the

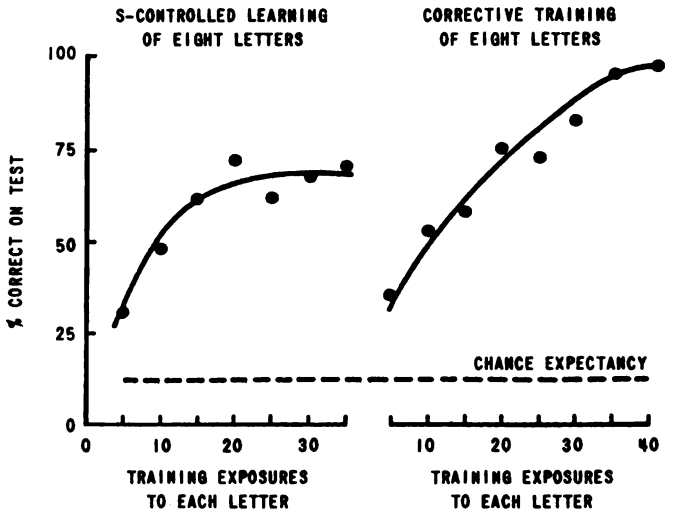


Figure 36 LEARNING CURVES FOR EIGHT LETTER IDENTIFICATION TASK

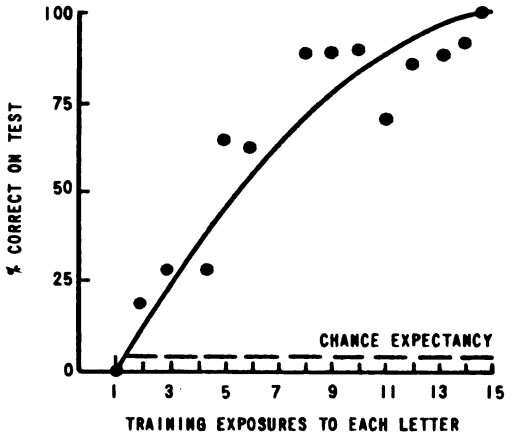


Figure 37 LEARNING CURVE FOR 26 LETTERS: CORRECTIVE TRAINING

maximum performance being about 85% on the combined alphabets, With a discrimination task of this difficulty, any displacement of the patterns from the position where they have been learned is likely to abolish the correct response.

On easier problems, such as a four-letter discrimination task, the choice of code is found to make little difference in system performance. The code becomes critical only when the discrimination capability is marginal, as in the 26 letter identification task. Given the choice between a position code and a configuration code with the number of A-units in a source-set held constant, the position code generally seems preferable with the kinds of stimulus material employed in these experiments. If the same total number of A-units must be divided among the source sets of the additional R-units used for the position code, however, better performance is obtained with the more economical configuration code, which uses binary numbers for identification, with larger source sets.

13. THREE-LAYER SYSTEMS WITH VARIABLE S-A CONNECTIONS

In the foregoing chapters, we have almost exhausted the possible ramifications of minimal three-layer perceptrons, having an S→A→R topology. Only one constraint remains to be dropped, in order to obtain the most general system of this class: this is the requirement that S to A-unit connections must have fixed values, only the A to R connections being time-dependent. In this chapter, variable S-A connections will be introduced, and the application of an error-correction procedure to these connections will be analyzed. It would seem that considerable improvement in performance might be obtained if the values of the S to A connections could somehow be optimized by a learning process, rather than accepting the arbitrary or pre-designed network with which the perceptron starts out. It will be seen that this is indeed the case, provided certain pitfalls in the design of a reinforcement procedure are avoided.

13.1 Assigned Error, and the Local Information Rule

In order to apply an error correction procedure to all connections of a perceptron, including the S - A connections, we must first re-examine the concept of "error" which has been employed so far as a criterion for reinforcement. In the theorem of Section 12.1, it was shown that it will not do to assume that all units of the perceptron are equally in error when a mistake in the total response occurs. It was seen that if all connections are corrected, on the assumption that both R-units are wrong (in the two R-unit case employed for demonstration) a solution may never be achieved.

The alternative was to assign an error independently to each R-unit, by a suitable criterion, and correct the connections leading to each R-unit in accordance with the corresponding error indication. In the present case, where A-units as well as R-units are to have their input-connections modified, it becomes necessary to assign an error indication to each A-unit, as well as to each R-unit.

In preceding chapters, the assigned error for an R-unit, E_r, was taken to be equal to $(R^* - r^*)$, where R^* is the desired response, and r^* is the obtained response. A positive error meant that the R-unit was to be turned to its positive state, and a negative error meant that it was to be turned to its negative state, in the case of simple R-units. Similarly, for an A-unit a_i, we might use a positive assigned error, E_i, to indicate that the unit is to be turned "on", and a negative E_i to indicate that it is to be turned "off", or made inactive, in response to the current stimulus. The difficulty is that whereas R^*, the desired response, is postulated at the outset, the desired state of the A-unit is unknown. We can only say that we desire the A-unit to assume some state in which its activity will aid, rather than hinder, the perceptron in learning the assigned classification or response function.

One possible way of obtaining the required activity states of the A-units would be to examine each possible state of the system, with its corresponding G-matrix, and determine whether or not a solution to the assigned problem exists. If a state is found in which a solution does exist, then the appropriate responses can be taught to each A-unit, by means of a standard error-correction procedure, operating on the A-units in the same

manner as on the R-units. Such an approach, however, evades the real issue of finding a procedure which will guarantee convergence to a solution without requiring that the reinforcement control system know the solution state ahead of time. Specifically, in assigning an error-indication to an A-unit, we wish to base the assignment only on the state of the network at the time and locality where the error occurs. The following rule will therefore be accepted as a working premise for all models to be considered:

LOCAL INFORMATION RULE: For any A-unit, a_i, the assignment of an error $E_i(t)$ can depend only on information concerning the activity or signals received by a_i, the value of its output connections, and the error assignment at their terminal points at time t.

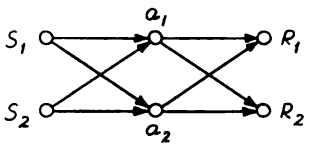
In other words, only a_i itself and the points to which it is directly connected can determine the error assignment.

13.2 Necessity of Non-deterministic Correction Procedures

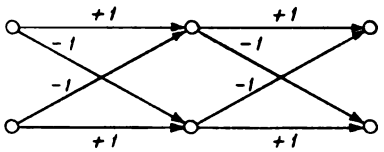
By a "deterministic reinforcement procedure" we mean that if the same state of the system should occur repeatedly with all signals and values unchanged, an identical reinforcement will be applied; and that if two similar subnetworks are in the same state of activity, value, and error assignment, they will be modified identically. Up to this point, no problem has been found for which a solution exists, where a suitably defined deterministic reinforcement procedure could not find a solution. The first exception to this is stated in the following theorem.

THEOREM 1: Given a three-layer series-coupled perceptron with simple A and R-units, and variable-valued S-A connections, and a classification $C(W)$ for which a solution exists, it may be impossible to achieve a solution by any deterministic correction procedure which obeys the local information rule.

PROOF: The proof is by example. Consider the following network:



Let a_1 and a_2 have thresholds of 1, and let the stimuli of W consists of A_1 alone (stimulus S_1) or A_2 alone (stimulus S_2). Let the required classification be $(R_1^*, R_2^*) = (+1, -1)$ for S_1 and $(-1, +1)$ for S_2 A solution clearly exists; for example, the following assignment of values would be satisfactory:



In this problem , a solution clearly requires an asymmetric assignment of values for "parallel" and "crossed" connections from each sensory unit and from each A-unit. If we assume that all values are initially equal, then either a_1 and a_2 are both on, or else both are off. In either case, one of the R-units is wrong, and whichever one is wrong will induce a symmetric correction of the values from both A-units. Moreover, since both a_1 and a_2 are in indistinguishable states (whichever R-unit happens to be wrong) under the local information rule both units must receive an identical error indication. But then the connections from whichever S-unit is active will both be modified identically, and the result is that the members of each value-pair (from each S-unit and from each A-unit) are still identical. The required asymmetry between "parallel" and "crossed" connections can therefore never arise, and the same response must always occur for S_1 and S_2 . Q.E.D.

While this theorem shows that a deterministic procedure cannot be guaranteed to work, it remains to be shown that a non-deterministic procedure will work. In the most extreme case, we could employ a procedure which randomly varies the value of every connection, independently of the others, as long as errors continue to occur. In this case, if the phase space of the system is bounded, a solution will certainly occur in finite time, but we have already seen the devastating consequences of a much less drastic randomization of the reinforcement process on learning time (c.f., Figure 19). In the following section, a more systematically directed procedure is presented, which can be shown to lead to a solution with probability 1, as in the case of error correction procedures considered for elementary perceptrons.

13.3 Back-Propagating Error Correction Procedures

The procedure to be described here is called the "back-propagating error correction procedure" since it takes its cue from the error of the R-units, propagating corrections back towards the sensory end of the network if it fails to make a satisfactory correction quickly at the response end. The actual correction procedure for the connections to a given unit, regardless of whether it is an A-unit or an R-unit, is perfectly identical to the correction procedure employed for an elementary perceptron, based on the error-indication assigned to the terminal unit. Thus, if the error E_i is positive, a correction is applied to the values of the active connections terminating on a_i which would tend to increase the signal to a_i algebraically, eventually turning it "on"; if E_i is negative, a correction, η, of the opposite sign is applied to all active connections terminating on a_i. The essential feature of the method is a probabilistic procedure for assigning the errors, E_i.

The rules for the back-propagating correction procedure are as follows:

1. For each R-unit, set $E_r = R^* - r^*$, where R^* = required response and r^* = obtained response.

2. For each association unit, a_i, E_i is computed as follows, for each stimulus: Begin with $E_i = 0$.

   a) If a_i is active, and the connection c_{ir} terminates on an R-unit with a non-zero error E_r which differs in sign from v_{ir}, add -1 to E_i with probability P_1.

b) If a_i is inactive, and the connection c_{ir} terminates on an R-unit with an error E_r which agrees in sign with v_{ir}, add +1 to E_i with probability p_2.

c) If a_i is inactive, and the connection c_{ir} terminates on an R-unit with an error E_r which does not agree in sign with v_{ir} (or if v_{ir} is zero) add +1 to E_i with probability p_3.

For all other conditions, E_i is not changed.

3. If $E_j \neq 0$, add η to all active connections terminating on the A or R-unit u_j, taking the sign of η to agree with the sign of E_j. In symbols,

$$\Delta v_{ij} = a_i^* \, sgn(E_j)\,\epsilon$$

where ϵ is the magnitude of η.

In general, p_1 and p_2 are taken large relative to p_3. The effect of these rules is to try to turn off any A-units (with probability p_1) whose output is currently contributing to an error in an R-unit, and to try to turn on any A-units (with probability p_2) which are currently off, but whose output signals would help correct an error in one or more R-units if they were on. The purpose of the third probability, p_3, is twofold; first, if no A-units respond to a stimulus, and all of the values have the wrong sign or are zero (as in typical initial conditions) it guarantees that some A-units will come on; second, it prevents the permanent loss of A-units which might be necessary for the proper response to some stimulus,

even though their values may have the wrong sign at some time during the training procedure. If ρ_1 and ρ_2 are larger than ρ_3, the main changes in the network will clearly all tend to go in the direction of a solution. The following theorem proves that the procedure is sufficient to guarantee a solution, if a solution exists, in the form of some assignment of values to the network.

THEOREM 2: Given a three-layer series-coupled perceptron, with simple A and R-units, variable-valued S-A connections, bounded A-R values, and a classification $C(W)$ for which a solution exists, then a solution to $C(W)$ can be obtained in finite time with probability 1 by means of a back-propagating error-correction procedure, given that each stimulus in W always reoccurs in finite time, and that probabilities ρ_1, ρ_2, and ρ_3 are all greater than 0 and less than 1.

PROOF: The state of the S-A network can be characterized, for present purposes, by an N_a by n matrix, A^*, which consists of the N_a row vectors:

$$A_i^* = (a_{i1}^*, a_{i2}^*, \ldots, a_{in}^*)$$

where $a_{ij}^* = 1, 0 =$ signal generated by unit a_i in response to stimulus S_j. Two assignments of values to S-A connections which yield the same A^*-matrix will be called equivalent S-A states. To each such matrix, A^*, there corresponds a G-matrix for the perceptron. We will say that a given S-A state permits a solution if the corresponding G-matrix is one for which a solution to $C(W)$ exists.

First, suppose the system is initially in a state which permits a solution. Then if it remains in this state sufficiently long, a solution must occur with probability 1, due to Theorem 4, of Chapter 5. Since S-A connections only change in value if the errors E_i are assigned magnitudes other than zero, and since the probabilities p_1 , p_2 , and p_3 of assigning non-zero E_i are all less than 1, there is a probability $p > 0$ that the perceptron will remain in its initial state for any given finite time. Thus, there is a probability greater than zero that a solution will be achieved before any change in the A^*-matrix occurs.

Next, suppose the A^*-matrix changes to some different state before a solution is achieved, or suppose that the system starts out in a state which does not permit a solution. Then it is sufficient to show that the system will always return to a state which does permit a solution in finite time with probability 1, and that the probability P of obtaining a solution for a given S-A state does not approach zero with successive returns to the same state. If it does always return to such a state, then each time it arrives at such a state, there will be a probability greater than zero (and bounded away from zero) that it finds a solution before the state is destroyed. Thus, with sufficiently many returns to states which permit solutions, a solution will be found with probability 1.

It is now necessary to show that from an arbitrary starting state, the system will always achieve an A^*-matrix which permits a solution in finite time with probability 1.

If the current A^*-matrix does not permit a solution, then either or both of the following conditions must be present:

(a) Some c^*_{ij} which should be 1 for a solution to be possible is actually 0;

(b) For some c^*_{ij} which should be 0 and is actually 1, there must be a v_{ir} the sign of which disagrees with R^* for stimulus S_j .

The second condition follows from the fact that if every active connection from A to R-units has a v_{ir} with proper sign for every S_j , and if condition (a) is not present, then a solution already exists. Now suppose, for an arbitrary A^*-matrix, Stimulus S_j occurs. Then condition (a) may exist for some A-units, and condition (b) for others. For each A-unit which is currently off (including all of those to which condition (a) applies) Rule 2b or 2c of the correction procedure becomes operative, and there is some probability that each such unit will receive an error indication. Since we have assumed the activity of these units to be necessary for a solution, and have postulated that a solution exists, there must be some assignment of S-A values for each such unit which will turn it "on" for S_j . Since S_j is postulated to reoccur infinitely many times, then it follows from Theorem 4 of Chapter 5 (treating the A-unit and its input connections as equivalent to an R-unit) that the required c^*_{ij} will ultimately be obtained. Since each A-unit is corrected independently of the others, a state will ultimately occur in which all of the A-units which were wrong by condition (a) have been corrected. Next consider those A-units for which condition (b) applies. For these units Rule 2a of the error correction procedure is

applicable, and by the same argument as above, the c^*_{ij} will ultimately all be corrected. But in that case, we have arrived in a state which permits a solution. Since there is nothing in the above argument which depends on states prior to the arbitrary starting state, the system can arrive at states permitting solutions indefinitely often, and a solution must therefore occur with probability 1, provided the probability P of finding a solution while in such a state does not approach zero. This last assumption, though plausible, still remains to be rigorously proven for the general case.

For the special case in which the values v_{ir} are bounded, the remaining assumption can be proven without difficulty. In the proof of Theorem 4, in Chapter 5, it was shown that the number of corrections necessary to find a solution is at most equal to

$$\frac{M(\ell + \epsilon\sqrt{n})^2}{\alpha(\epsilon - \sigma)^2}$$

where M and α are constants depending only on the G-matrix (and therefore on A^*), and ℓ is the length of the vector Hx^0. Thus the number of corrections required to find a solution can incrase only as a result of an increase in the magnitude of some components of the starting vector, x^0, upon successive returns to the same S-A state. But if all values v_{ir} are bounded, the components of x^0 are also bounded. Consequently, ℓ has an upper bound for any given H (or for any given A^*). This means that there is a maximum number of corrections that might possibly be required (assuming that a solution exists) and that the probability p of arriving at a solution before destruction of the A^* state is not only greater than zero but must be bounded away from zero. Q.E.D.

13.4 Simulation Experiments

At the present time, no quantitative theory of the performance of systems with variable S-A connections is available. A number of simulation experiments have been carried out by Kesler, however, which illustrate the performance of such systems in several typical cases, shown in the accompanying figures.* In order to show the performance of the variable S-A system to its best advantage, small perceptrons were used, for which the learning of a horizontal/vertical bar discrimination (Experiment 6) falls short of what might be obtained with an optimum S-A organization.

Figure 38 illustrates the effect of various combinations of the probabilities ρ_1 , ρ_2 , and ρ_3 (including the 0,0,0 case where all S-A connections remain fixed, for comparison). The curves show the mean performance for 20 perceptrons, with 50 A-units, having 10 input connections to each. The initial values of all S-A connections are set equal to +10, and the threshold is 50. The same set of 20 networks and training sequences was used for each probability combination.

It is found that if the probabilities of changing the S-A connections are large, and the threshold is sufficiently small, the system becomes unstable, and the rate of learning is hindered rather than helped by the variable S-A network. Under such conditions, the S-A connections are apt to change into some new configuration while the system is still trying to adjust its values to a solution which might be perfectly possible with the old configuration. Better performance is obtained if the rate of change in the S-A network is sufficiently small to permit an attempt at solving the problem before drastic changes occur. To improve the stability

* The experiments were carried out with the Burroughs 220 computer at Cornell University, and the IBM 704 at the A.E.C. Applied Mathematics Center at New York University.

of the network, in all experiments shown here, the A-R connections are reinforced, for each stimulus, before determining whether a correction should be propagated back to the S-A network. Thus, S-A connections are changed only if the system fails to correct an error at the A-R level.

In Figure 39, mean performances of a number of 20 A-unit perceptrons are shown, in one case with 4 connections, and in a second case with 50 connections to each A-unit. These perceptrons are small enough so that in many cases we would expect no solution to exist to the horizontal/vertical bar problem (which requires the classification of 40 stimuli with only 20 A-units) were it not for the modifiable S-A network. Initial values of S-A connections are again equal to 10, and thresholds are $2m$, where m = number of connections to each A-unit. Note that with 50 fixed connections to each A-unit the performance is poorer than with only 4 connections, but that with $P_1 = .9$, $P_2 = .3$ and $P_3 = .1$, the performance overtakes the 4-connection model. This is because with large numbers of S-A connections, the perceptron can effectively take its pick of whatever organization might be most helpful, and can always reduce excess connections to zero value, while with only a small number of connections at its disposal it is seriously limited in its potentialities. With only 4 connections, variable S-A connections have little effect on performance.

These experiments suggest that the best performance will generally be obtained by taking $P_1 > P_2 > P_3$.

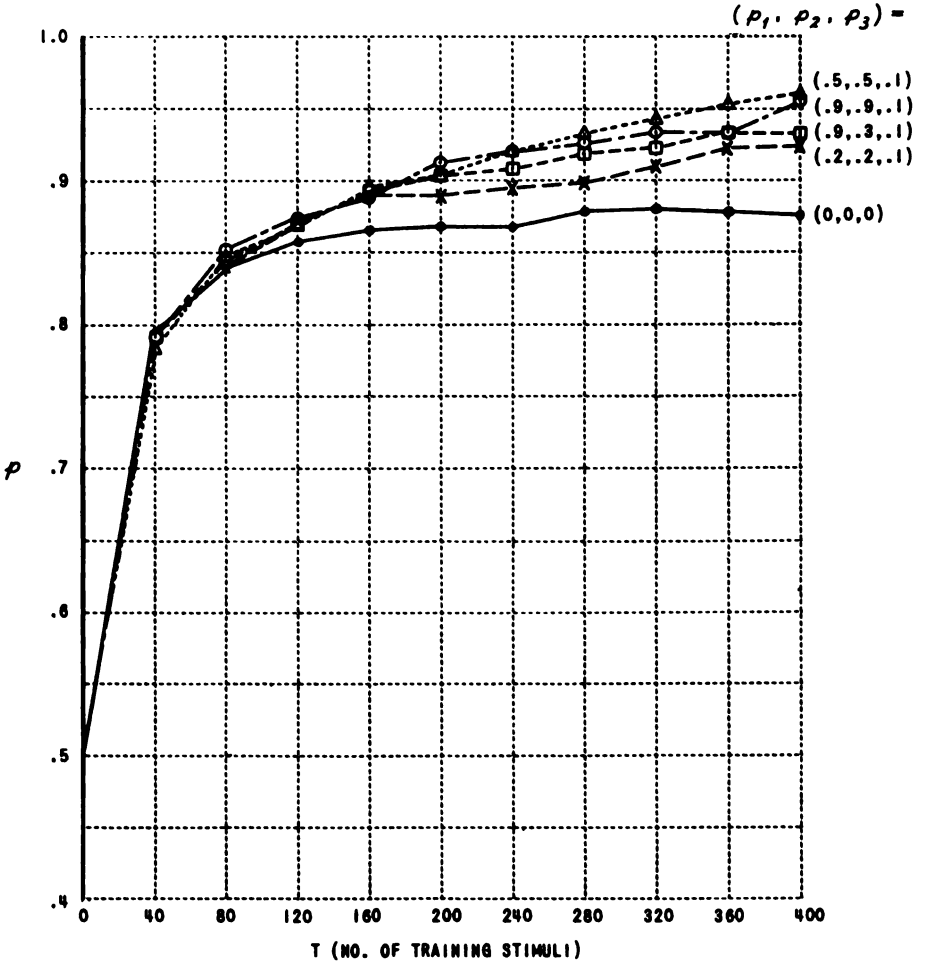


Figure 38 BACK-PROPAGATING ERROR-CORRECTION EXPERIMENTS: MEANS OF 20 PERCEPTRONS $N_a = 50$, $\theta = 50$, 10 CONNECTIONS TO EACH A-UNIT. HORIZONTAL/VERTICAL BAR DISCRIMINATION (EXPT. 6). (0-SIGNALS COUNTED CORRECT WITH .5 PROBABILITY).

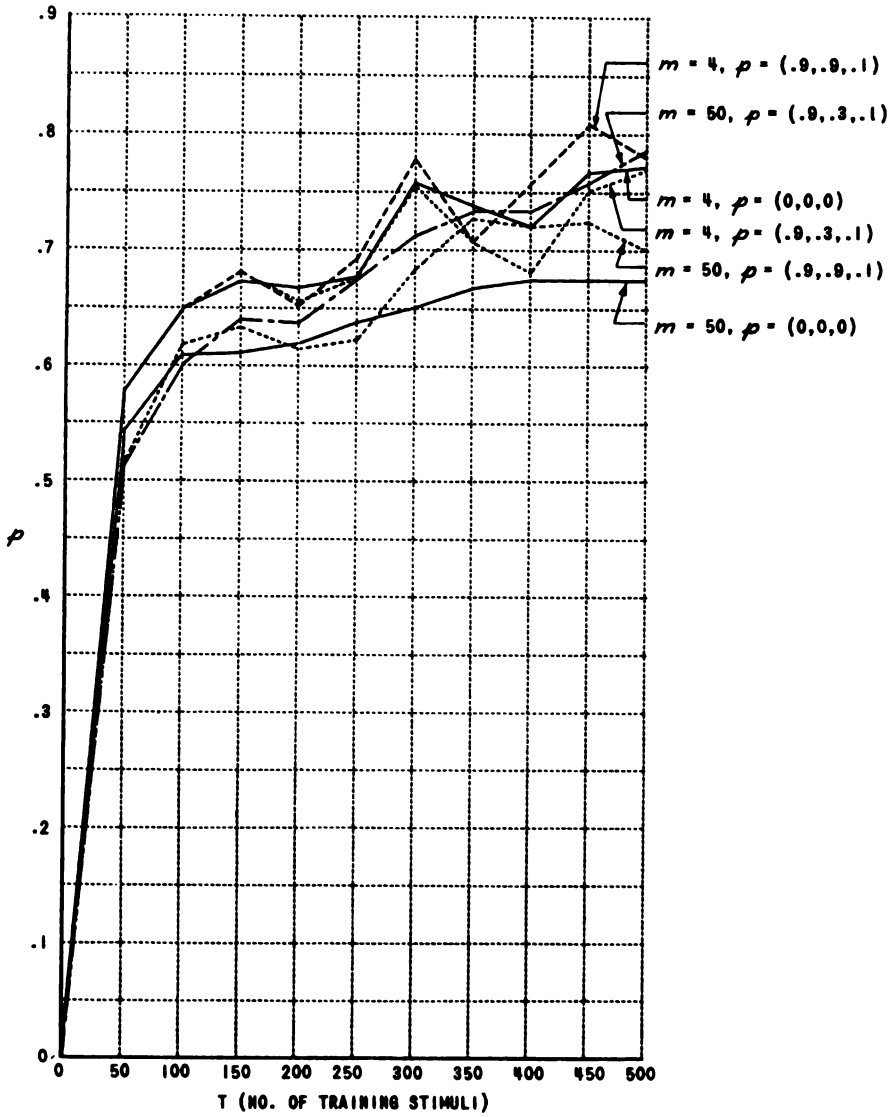


Figure 39 BACK-PROPAGATING ERROR-CORRECTION EXPERIMENTS: MEANS OF 10 PERCEPTRONS WITH $N_a = 20$, m CONNECTIONS TO EACH A UNIT, $\theta = 2m$. HORIZONTAL/VERTICAL BAR DISCRIMINATION (EXPT. 6). 0-SIGNALS COUNTED AS ERRORS.

An interesting application of the variable S-A system is in pre-conditioning a perceptron for stimuli of a particular type (such as line figures, or blob patterns) by giving it a number of discrimination tasks to perform on typical material of the given type, and then trying to teach it a new discrimination on the same kind of stimuli. Due to the prior adaptation of the S-A system, it is to be expected that the learning curve for the final discrimination task should show faster learning after the period of pre-conditioning than if the same discrimination task had been attempted with the original randomly organized S-A network. In other words, the S-A network should become adapted to the stimuli of a particular kind of universe, performing better on typical discrimination tasks involving "familiar" kinds of stimuli than on tasks involving radically different or "unfamiliar" kinds of stimuli.

14. SUMMARY OF THREE-LAYER SERIES-COUPLED SYSTEMS: CAPABILITIES AND DEFICIENCIES

The three-layer series-coupled perceptron (S→A→R perceptron) is the least complicated topological organization which yields fully general response-capabilities. The analysis presented in the preceding chapters leads, in effect, to the following conclusion: With a suitable design and training procedure, a three-layer series-coupled perceptron can be taught to duplicate the performance of any finite automaton. This means that if we have a finite universe of potential input sequences ($\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_n$) and a finite set of possible response sequences ($\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_m$), then it is possible to construct a minimal perceptron such that any response sequence, $\mathcal{R}_i$, can be associated with each possible input sequence, $\mathcal{J}_k$. In order to do this with full generality, of course, a suitable spectrum of time delays, τ_{ij} , must be present, as indicated in Chapter 11.

Both the generality and the practical limitations of the above statement should be emphasized. It is perfectly possible, in principle, to teach a minimal perceptron to duplicte the performance of an arbitrary digital computer. To do this, every possible sequence of coded instructions and data must be represented as a stimulus sequence (one of the $\mathcal{J}_i$) and the set of output numbers generated by the computer as a response sequence (one of the $\mathcal{R}_j$). If the perceptron is large enough, it can then be trained, with an error correction procedure, to make the appropriate association of input and output sequences. But what the perceptron learns by this process is to simulate the behavior of the digital computer; it does not acquire the

computer's logic. If any one of the trillions of possible programs were omitted from the training sequence, the perceptron would probably fail to perform correctly if tested on the omitted sequence. The failure to generalize, or to learn logical rules, in such a problem makes such an application of these minimal perceptrons totally impractical.

For practical purposes, we will limit our remarks to the performance of these perceptrons in recognizing and reporting environmental events. In this connection, the following capabilities have been established:

(1) A three-layer series-coupled perceptron can be taught to associate an arbitrary coded output, or sequence of outputs, $\mathcal{R}_i$, to each stimulus, or stimulus sequence, $\mathcal{S}_i$, in a finite environment.

(2) The perceptron need not be explicitly designed for the task which it is required to learn. The same network may be taught a variety of alternative outputs, or codifications, of the same environment.

(3) The required training can be accomplished by means of an arbitrary sequence of events from the specified environment, regardless of the order or frequency with which they occur, provided each event ultimately reoccurs in finite time.

(4) The training can be accomplished regardless of the initial state of the perceptron's memory, and without specifying in detail the changes which must take place in the state of the system (i.e., general dynamic laws are sufficient to bring about the required adaptation).

(5) A perceptron will tend to assign the same response to any two stimuli or stimulus sequences, $\mathcal{S}_i$ and $\mathcal{S}_j$, which are close to identity under temporal translation. By means of discrimination training, however, it can be made to associate a different response to each such stimulus.

With this kind of universality in the performance of the system, we obviously cannot hope to find any new kinds of response capabilities in more complex or sophisticated networks, which cannot be realized by minimal perceptrons after suitable training. Nonetheless, the three-layer series-coupled perceptron clearly falls far short of biological systems in some respects. The differences lie not in what the system can learn to do, but rather in the speed, efficiency, economy, and reliability of learning or adaptation. An S→A→R perceptron can be taught to play a game, such as checkers, only by teaching it what response to make in every conceivable situation; a biological system can anticipate most of this training by learning the rules of the game. Or, similarly, an S→A→R perceptron can distinguish a circle from a triangle in the lower half of its retina only if it has previously been trained with triangles and circles in the lower half of its retina; it will not generalize from experience with similar forms in the upper half of the field. In Nature, the enormous number of sensory situations which comprise the potential universe (each situation, individually, having exceedingly low probability of occurrence) makes the capabilities of generalization, analysis, and abstraction absolutely essential for an advanced organism, or recognition device, to function properly. Two main ingredients of such performance are recognition of similarity and recognition of functional parts, or entities. The first of these is basic to generali-

zation and induction, while the second is basic to analysis, the abstraction of relations, and the reduction of complex situations to familiar terms. Seen in this light, the principal deficiencies of these minimal-topology perceptrons are:

(1) An excessively large system may be required.

(2) The learning time may be excessive.

(3) The system may be excessively dependent on external evaluation (by an independent r.c.s.) during learning.

(4) The generalizing ability (inductive ability) is insufficient.

(5) Ability to separate essential parts in a complex sensory field (analytic ability) is insufficient.

Point (1) is largely attributable to (5); the excessive size of the perceptrons necessary to deal with complex environmental situations is due largely to the necessity of having a characteristic set of A-units representing every possible sensory field or sequence in its entirety. A preliminary coding of the field in terms of its parts and relations would greatly reduce the size of the system required to describe a given universe of situations. To take an extreme case, if a three-layer series-coupled perceptron is required to produce as an output the coded representation of the sum of a sequence of a million digits, it must be capable of representing in its association system every possible sequence of a million digits

(presented either serially or simultaneously): 10^{10^6} possibilities in all. On the other hand, a perceptron which could attend selectively to each digit, form a partial sum, and then go on to the next digit, requires only 10^8 possible states: 10^7 to represent the possible values of the partial sum, multiplied by a factor of ten to allow for each of the possible incoming digits. The second method is the one employed by a digital computer, or a man adding a sequence of numbers. In the field of sensory pattern recognition, similar conditions occur. The recognition of a sentence is made much easier by breaking it into words, and the recognition of a scene is made easier by analyzing it into objects and relations.

Similarly, the excessive learning time (point 2) can be largely attributed to (4), the insufficient generalizing ability of the system. With improved generalization, several examples should be sufficient to teach the perceptron to recognize all members of a class of similar events, whereas at present an unduly large sample is required in order to extend the response over the class. The insufficient generalizing capability has been frequently pointed out in the preceding chapters, and is common to all of the S→A→R perceptrons. Thus points (3), (4) and (5) appear to be the primary deficiencies.

In connection with point (3), we note the failure of minimal perceptrons to reach "useful" terminal states under R-controlled reinforcement procedures, except under exceptional environmental and organizational conditions. This means that the reinforcement control system must itself have a great deal of information about the environment, and must generally know, or have built into it, the precise discrimination

or response functions which the perceptron is supposed to learn. Thus the r.c.s. must either be a free agent (e.g., a human trainer) or else some kind of homunculus within the same physical system as the perceptron. It has been noted that a perceptron can improve over the performance of the r.c.s. in some cases (Section 8.1.4) but the functioning of the r.c.s. still seems to be rather remote from what might be expected of a biological motivating system. By using a random-sign correction procedure, the information required from the r.c.s. is minimized; with such a procedure, the possible outputs of the r.c.s. can be interpreted to mean "hold steady" or "change", while with a directed correction procedure the three alternatives "hold steady", "increase values", or "decrease values" are all required. But the efficiency of a system employing the randomized procedure is greatly reduced (c.f., Figure 19) and the only hope for such systems seems to be in a "majority rule" procedure, which increases the size and complexity of the total organization.

If a system could be contrived which would guarantee generalization of a response from one stimulus of a class to all other stimuli of that class, an r.c.s. which employs the "trial-and-error" process of the random-sign procedure might become practical, and a simple motivation system which senses only the suitability or unsuitability of the present response or state of the organism might be substituted for the more complicated r.c.s. assumed for most of the preceding experiments. In Part III, it will be shown that multi-layer and cross-coupled perceptrons are capable of providing just this sort of generalizing capability, and, moreover, that this capability may be "self-organizing" under reasonable environmental conditions. That is to say, R- controlled systems

can learn to form reasonable classes on the basis of a similarity criterion, provided there is some support for this organization from the environment. The required support takes the form of a "continuity constraint", which says, in effect, that stimuli do not occur as momentary flashes, but are more likely to persist for a time, during which they undergo a series of movements or transformations. It will be seen that such a sequential organization provides sufficient information to enable a multi-layer or cross-coupled perceptron to abstract a concept of similarity which can then be employed to obtain immediate generalization in later situations.

The improvements which have been demonstrated to date in multi-layer and cross-coupled perceptrons will be seen to be primarily in the field of generalization phenomena, and their main virtue is in reducing the learning time of a perceptron. Some reductions in size requirements have also been demonstrated, and the dependence on external evaluation of performance is largely eliminated. Thus points (1) through (4), in the list of criticisms of minimal perceptrons can be largely or entirely eliminated with a multi-layer or cross-coupled topology. Point (5), however, remains the least understood of the current problems. While there is some indication that perceptrons of the types to be considered in Part III may have some analyzing ability (for example, they can isolate contours from solid figures, and may possibly learn to suppress the partial response of the association system to irrelevant aspects of the stimulus field) it is not yet possible to say whether such systems are really sufficient to meet the challenge of point (5), or not. The psychological problems of figure-ground organization, recognition of relations, and

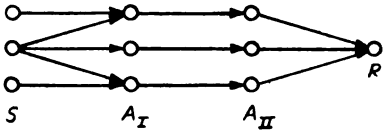
"cognitive set" are all involved here. It is likely that "back-coupled perceptrons", in which R-units or deep association layers feed back to more superficial layers, may be necessary to deal with these problems. Several possible approaches will be considered in Part IV, which deals with current problems, and attempts to establish directions for future study.

PART III

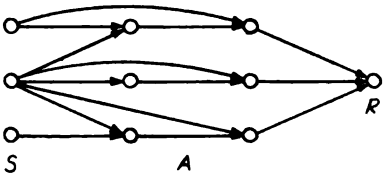
MULTI-LAYER AND CROSS-COUPLED PERCEPTRONS

15. MULTI-LAYER PERCEPTRONS WITH FIXED PRETERMINAL NETWORKS

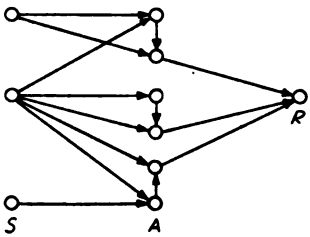
The perceptrons considered in Part II have all consisted of three "layers" of signal generating elements: a sensory layer, a single layer of association units, and a layer of R-units (containing only a single unit in the case of simple perceptrons). A perceptron with additional layers of A-units between S and R-units will be called a multi-layer system. Thus the network diagram:



represents a four-layer series-coupled system, whereas the diagram



represents a three-layer cross coupled system, since all A-units are at least the same logical distance from the sensory units (see Definition 18, Chapter 4). The three-layer structure of the second diagram can be made clearer if it is drawn in the form:



which is topologically identical to the preceding network. Cross-coupled systems will be considered in detail in the following chapters.

It has been demonstrated that three-layer, series coupled perceptrons are capable of learning any type of classification, or associating any responses to stimuli or to sequences of stimuli, that might possibly be required. Therefore, if a multi-layer topology is to offer any functional advantages, it will not be in the form of new kinds of responses to stimuli (since any such response can be achieved with a three-layer system) but rather in increased efficiency in the acquisition of such responses. It can, in fact, be demonstrated that the adaptability, or ease of acquisition of responses, may be greatly improved with a suitable multi-layer topology. The most striking improvements are to be found in the generalizing ability of such networks -- an ability to give appropriate responses to stimuli for which they have not been taught. It has been seen that this "inductive" or generalizing capability is present only in rudimentary form in three-layer series-coupled systems. Some multi-layer systems also show improvements in sensitivity to differences between highly similar stimuli, making such discriminations easier to learn, as will be seen in Section 15.1.

In the following sections, we will first consider systems in which all connections other than connections to R-units have fixed values, only the R-unit input connections being reinforced. The connections to the R-units will be called terminal connections, all other connections (from S to A-units, and A-units to other A-units) being called preterminal connections. It will be seen in Section 15.2 that the most interesting effects which can be obtained by such systems depend on special constraints in the organization of the preterminal network. The following chapter will therefore be devoted to the examination of dynamic rules by which the preterminal connections between layers of A-units can be modified, so as to yield the required organizations as a result of the system's adaptive functioning, in a suitably organized environment.

The analysis of multi-layer systems is of interest not only in its own right, but also because it introduces many of the problems and formal techniques of analysis which will be encountered in the following chapters on cross-coupled systems, with feed-back loops within the network. In fact, it is found that with a suitable transformation, many "closed-loop" cross-coupled systems can be represented by an equivalent "open-loop" multi-layer system.

15.1 Multi-layer Binomial and Poisson Models

The most straightforward extension of our previous models to a multi-layer topology is to assume that each A-unit in the first association layer is assigned an origin point configuration in the retina, or sensory layer, chosen independently for each A-unit, as before. Each A-unit in the second layer (designated $A^{(2)}$) is similarly assigned an origin point configuration in the $A^{(1)}$ layer, independently for each such A-unit. In general, every A-unit in the $A^{(k)}$ layer is independently assigned an origin point configuration from an appropriate distribution (binomial or Poisson model), the connections originating from the $A^{(k-1)}$ layer. All connections from one A-layer to the next are assumed to be fixed in value, the final A-layer sending variable-valued connections to the R-units. In order to analyze the performance of such a perceptron, it is sufficient to determine the Q-functions for the A-units of the last layer, before the R-unit, since, given these Q-functions, we can then apply the same equations and analysis which were employed in Part II, for three-layer perceptrons. The notation $Q^{(1)}_{ij..n}$ will be used to denote the Q-functions for A-units in the first layer (which are identical with the Q-functions discussed in Chapter 6), and $Q^{(k)}_{ij..n}$ to denote Q-functions for units in the k^{th} layer.

Even in the simplest case, of a four layer perceptron, the combinatorial analysis required for a rigorous statement of $Q^{(2)}$ functions is awe-inspiring. A special case, in which all inter-layer connections are inhibitory, and the thresholds of all $A^{(2)}$ units are zero, has been

analyzed by Joseph (Ref. 41), and the reader is referred to his contribution for the detailed considerations. The basic difficulty stems from the fact that a second layer Q-function, such as $Q_{ij}^{(2)}$ depends on the distribution of the numbers of A-units in the first layer which respond to S_i alone, S_j alone, and jointly to S_i and S_j. The expected values of these numbers are obtainable from the $Q^{(1)}$ functions in a straightforward manner, but the non-central moments of the distributions enter into the analysis in such a way that it becomes unduly complicated.

A practical solution is obtained by assuming that the numbers of A-units in the 1st, 2nd,...i-1th layers (designated by $N_a^{(1)}, N_a^{(2)}, \ldots, N_a^{(i-1)}$)are all very large, or infinite. In this case, the proportion of active units in each layer in response to S_i will be equal to Q_i, and the expected values of all set-intersections can be employed in the analysis. In this case, the equations of Chapter 6 can be employed without modification to compute $Q_{ij\ldots n}^{(i)}$ by using $Q_i^{(i-1)}$ in place of the stimulus area, R_i, $Q_{ij}^{(i-1)}$ in place of the intersection C, etc. The error introduced by assuming infinite N_a for the preterminal layers will be slight, as long as the actual N_a is reasonably large.

The addition of extra A-unit layers can have one of several interesting effects, depending upon the parameters α, γ, and θ (or $\bar{\alpha}$, $\bar{\gamma}$, and θ in a Poisson model) for each layer. The special case of inhibitory connections and zero thresholds was investigated by Joseph (Ref. 41), who finds that by optimizing the number of input connections to each layer, so as to achieve highest probability of correct recognition, Q_i approaches a constant as the number of layers increases,

regardless of the size of the stimuli or the dichotomy which the perceptron is required to learn. At the same time, $Q_{i,j}$ approaches Q_i^2, $Q_{i,j,k}$ approaches Q_i^3, etc. In effect, this represents a condition in which, in the terminal association layer, a statistically independent set of A-units responds to each stimulus in the environment. The consequence is that all discriminations become equally easy. Specifically, it was found that the ratio $\left(\frac{\mu^2(u_\chi)}{\sigma^2(u_\chi)}\right)$ for 100 A-units in the terminal layer approaches 1.941 as the number of layers is increased, with an environment of 40 stimuli. A comparison with Table 3, in Chapter 7, shows that this performance is less than would be achieved with a three-layer perceptron for the task of discriminating horizontal from vertical bars, but it is considerably better than the performance of a three-layer perceptron on a more difficult task, such as the odd-even bar discrimination illustrated in Table 4. Thus the addition of extra association layers can be used to improve discrimination in difficult problems, but only at the cost of reduced generalizing ability, since two adjacent stimuli with a large intersection are now no more closely related (in the $A^{(i)}$ layer) than two totally disjoint stimuli.

In Joseph's model, with all inhibitory connections, the above results are obtained only by optimizing the number of connections to each new layer of A-units. If, instead of carrying out this optimization, a fixed number of connections is assumed for all A-units in the system, the perceptron will be unstable, and will tend to develop oscillations such that alternate A-layers are totally "on" or totally "off", making all discrimination impossible. Moreover, it is to be expected that a model which has

been optimized for one environment, with a given size of stimuli, will be unstable in a different environment, with a slightly different size of stimuli. In more practical cases, a mixture of excitatory and inhibitory connections must be used, with thresholds greater than zero, in order to guarantee stability and convergence of Q_i for a range of environmental variations. Clearly, if $x < y + \theta$, $Q_i^{(k)}$ will not go to 1 as k increases. If $x = y$, a suitable choice of $\theta > 0$ will generally guarantee, as well, that Q_i will not go to zero. From Figure 7(b), for example, it is clear that if $x = y = 5$, and $\theta = 1$, an equilibrium should occur at about $Q_i = .37$, since at this point $Q_i^{(k)} = Q_i^{(k-1)}$. If $Q_i^{(k-1)}$ should rise above .37, we will have $Q_i^{(k)} < Q_i^{(k-1)}$, while if $Q_i^{(k-1)}$ falls below .37 we will have $Q_i^{(k)} > Q_i^{(k-1)}$.* If we increase the amount of inhibition by making $x = 3$, $y = 7$, then (from the same Figure) we find that the equilibrium value of Q_i is reduced to .14. If the inhibition is increased still further (e.g., to $x = 1$, $y = 9$, as in the bottom curve of Fig. 7b) the equilibrium value of Q_i is zero, and no matter how large a stimulus is presented, activity will die away entirely in the "deeper" association layers.

---

* This observation will generally not be valid for a small perceptron, where the actual level of activity may go to zero in one of the layers, due to random variations in the network. In this case, $Q_i^{(k)}$ will be zero for all subsequent layers. Thus, for a finite system, $Q_i^{(k)} \xrightarrow[k \to \infty]{} 0$.

15.2 The Concept of Similarity-Generalization

So far, the addition of extra association layers has had no important effect beyond the sharpening of the discriminative acuity of the perceptron, generally counterbalanced by a loss in the generalizing capability of the system. In the next section, we will consider a four-layer perceptron with special constraints in the organization of the connections to the A-units, such that the system tends, spontaneously, to generalize a response associated to a given stimulus pattern to all "similar" stimuli, regardless of their location in the retinal field. In the following chapter, it will be shown that such constraints need not be built into the system ab initio, but can arise through a spontaneous adaptation process (without any intervention by the r.c.s.) if some simple dynamic laws are introduced. In all of these systems, the concept of "similarity" is of fundamental importance.

The term "similarity" has been used in a number of different ways, some of them well-defined, as in "two triangles are similar", some relatively vague and ambiguous, as in "two faces are similar" or "two ideas are similar". For present purposes, we have need of a concept which will cover the range of relationships which might make two objects appear "similar" to a perceiving observer, but which will still permit exact definition for purposes of analysis. We must also distinguish between the "objective similarity" of objects in space, the similarity of stimuli on the retina, and the "subjective similarity" which the observer recognizes and reports. While the concepts proposed here do not cover all of the possible meanings of "similarity" in psychology, they are sufficient to permit the design of a number of perceptual experiments related to the similarity problem.

15.2.1 Similarity Classes

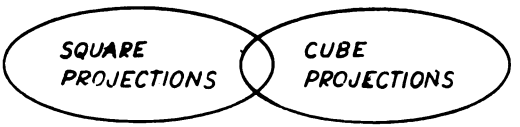
We will first consider a definition of similarity which is applicable to the classification of stimuli. From this point of view, two stimuli either are similar or they are not; there are no intermediate degrees of similarity. In the following section, a quantitative definition which permits a multidimensional ordering of objects or stimuli according to their similarity will be considered.

For present purposes, the only constraints which will be placed on the logical nature of the similarity relation are that it should be symmetric and reflexive; that is, if A sim B, then B sim A, and A is always similar to itself. It is not required that the relation of similarity should be transitive; that is, A sim B and B sim C does not imply A sim C, except under very special conditions, as will be seen below. There are clearly a large number of possible relations which meet the logical conditions for a similarity relation. For example, equality, geometrical congruence, equality of area, and topological equivalence are all admissible possibilities. Thus, in specifying the similarity of two stimuli the notation A sim $B|\mathcal{R}$ will be used, where $\mathcal{R}$ is a particular relation, meeting the conditions of symmetry and reflexivity.

The set of stimuli which are similar under a given relation will be said to form a similarity class under that relation. For example, if $\mathcal{R}$ is defined as the relation of similarity under a rotation group, then A sim $B|\mathcal{R}$ means that A is a rotated image of B, and B is a rotated image of A.

In perceptual problems, a particular kind of similarity class is of particular importance. This will be called a projective similarity class, and is defined as follows. Let the sensory points of a perceptron be embedded in an r-dimensional sensory manifold, $\mathcal{S}$. Let $\mathcal{S}$ be embedded in an $r+k$ dimensional world manifold, $\mathcal{M}$. An object in $\mathcal{M}$ is defined as any set of points in $\mathcal{M}$.* Let Ω be a set of admissible objects in $\mathcal{M}$. Let $\mathcal{G}$ be any transformation group in $\mathcal{M}$. Let a projection π be defined as an operation which maps every point in $\mathcal{M}$ into at most one point in $\mathcal{S}$. Then A sim B$|\mathcal{G}, \mathcal{S}, \Omega, \pi$ means that stimuli A and B are both π-projections onto the sensory points in $\mathcal{S}$ of transforms under $\mathcal{G}$ of the same object in Ω.

A few moments reflection should show that this encompasses most of the cases in which we say that two stimuli are perceptually "equivalent"; for example, any group of rigid movements of an object in 3-space will yield a projective similarity class on a two-dimensional retina. Note that this similarity relation is not generally transitive. For example, if we let $\mathcal{G}$ be the group of rigid motions in 3-space, and let $r = 2$, then the similarity classes generated by a flat cut-out of a square in $\mathcal{M}$, and by a cube in $\mathcal{M}$ (with orthogonal projection onto the retina) are related by the Venn diagram:



* The term "object" is used in much the same sense as "distal stimulus" in psychology. Our use of the term "stimulus" always signifies a "proximal stimulus" unless otherwise specified.

where the intersection includes all cases where the square and a face of the cube are both parallel to the retinal surface (assuming $\mathcal{S}$ to be Euclidean, which it is not in a vertebrate eye). A tilted square will be projected as a parallelogram, whereas a tilted cube is projected either as a rectangle, pentagon, or hexagon, so that the classes, although they intersect, are not equivalent.

For the special case in which the points of an object and all of its transforms in $\mathcal{M}$ can be placed in one-to-one correspondence with the S-points in $\mathcal{S}$, the relation of projective similarity will be transitive. This includes the case in which $\mathcal{M}$ and $\mathcal{S}$ are of the same dimensionality and coextensive, objects and transforms consisting only of sensory points in $\mathcal{M}$. Most stimulus classes considered in experiments up to this point have been interpretable in this fashion. Alternatively, $\mathcal{M}$ might have a higher dimensionality than $\mathcal{S}$, but the group $\mathcal{G}$ may be limited to motions parallel to the surface of $\mathcal{S}$. Here again, with a suitable choice of $\mathcal{G}$, a transitive similarity relation can be obtained.

The case of greatest psychological interest is that of a three-dimensional world-manifold, $\mathcal{M}$, and a two-dimensional sensory manifold, $\mathcal{S}$, where $\mathcal{G}$ is the group of rigid motions and dilatations in $\mathcal{M}$. A perceptron which generalizes strongly between any two members of a similarity class defined by such a relation, and generalizes weakly between stimuli which are not in the same similarity class, will duplicate a large fraction of the perceptual behavior of a biological organism, in the visual domain.*

---

* A consideration of some of the projection operations which apply to this problem can be found in Gibson, Olum, and Rosenblatt, Ref. 27.

15.2.2 Measurement of Similarity, Objective and Subjective

Let $\mathcal{G}$ be a Lie-group* (of dimension r) of transformations of the manifold $\mathcal{M}$. Let B be a canonical system of coordinates defined in the Euclidean r-space, E^r, such that every system of equations $g_i(t) = a_i t$ (where g_i is the i^{th} coordinate of g in B) gives a one-parameter subgroup $g(t)$. Then the distance $d(O, g)$ for any $g \in (\mathcal{G})$ $(g = (g_1, g_2, \ldots, g_r))$ is given by

$$d(O, g) = \sqrt{\sum g_i^2}$$

We then define the similarity measure $\mu(X, Y) | \mathcal{G}, B$ for the objects X and Y with respect to $\mathcal{G}$ and B as

$$\mu(X, Y) | \mathcal{G}, B = \inf_{g \in \Gamma} d(O, g) \qquad (15.1)$$

where $\Gamma = \{ g : X = gY \}$, $g \in \mathcal{G}$ (That is, Γ is the set of all transformations in $\mathcal{G}$ which will transform the object Y into the object X.)

Note that this measure is applicable only to objects in $\mathcal{M}$ which are similar under $\mathcal{G}$; it is not applicable to stimuli unless $\mathcal{S}$ is coextensive with $\mathcal{M}$. Consequently, the measure μ will be called the objective similarity measure with respect to $\mathcal{G}$ and B. This measure represents the length of a sort of "shortest path" by which Y

---

* Readers who are unfamiliar with the theory of Lie-groups will find a useful discussion of this subject in Pontrjagin (Ref. 111).

can be continuously transformed into X, by means of transformations of the group $\mathcal{G}$. The choice of the basis, B, determines the relative weighting attached to various subgroups of $\mathcal{G}$. For example, if $\mathcal{G}$ is the group of translations in $\mathcal{M}$, then μ can be made proportional to the length of the displacement vector which would carry Y into X.

Let us also define the subjective similarity measure with respect to a perceptron, $\mathcal{P}$, a response unit, R, and a projection operator Π, by

$$\mu^*(X,Y) \Big| \Pi, \mathcal{P}, R = Q_{XY}(R) \Big/ \sqrt{Q_X(R)\, Q_Y(R)} \leq 1 \qquad (15.2)$$

where $Q_{XY}(R)$ is the value of Q_{ij} for the stimuli corresponding to the objects X and Y (under the projection Π) measured in the source set of the response unit R. For an α-system, and stimuli of fixed size, $\mu^*(X,Y)$ is proportional to the generalization coefficient g_{XY}, for the response R. For two identical stimuli, $\mu^*(X,Y) = 1$. If the value of $\mu^*(X,Y)$ is a monotonic function of the objective similarity of the objects X and Y, we would expect the response r^* to generalize most strongly to highly "similar" objects, and most weakly to dissimilar objects. Over any given subgroup of transformations of an object in $\mathcal{M}$, this induces a "generalization gradient" equivalent to the use of the term in experimental psychology.

A perceptron which is to simulate perceptual performance must have or acquire a close correlation between the subjective and objective similarities of objects in physical space, under the group of

rigid motions and some kinds of cintinuous deformation. A perceptron in which such a correlation exists is said to be capable of similarity generalization. Similarity generalization implies that the perceptron not only tends to generalize to similar objects, but retains its ability to respond differentially to dissimilar objects. The demonstration of such a capability will be our main concern for the remainder of this chapter and the following four chapters.

15.3 Four-Layer Systems with Intrinsic Similarity Generalization

15.3.1 Perceptron Organization

The four-layer perceptrons to be analyzed have fixed connections except for the terminal A to R-unit connections, and a topology which is illustrated in Figure 40. S, A, and R-units are all assumed to be of the simple variety, resembling those of an elementary perceptron. The special features of this system (which might be called a "similarity-constrained perceptron") are the following:

(1) Each $A^{(1)}$ unit has a threshold θ, x excitatory and y inhibitory input connections, and a single output connection to one of the $A^{(2)}$ units.

(2) Each $A^{(2)}$ unit receives connections from a source set of $mA^{(1)}$ units, and has a threshold equal to 1.

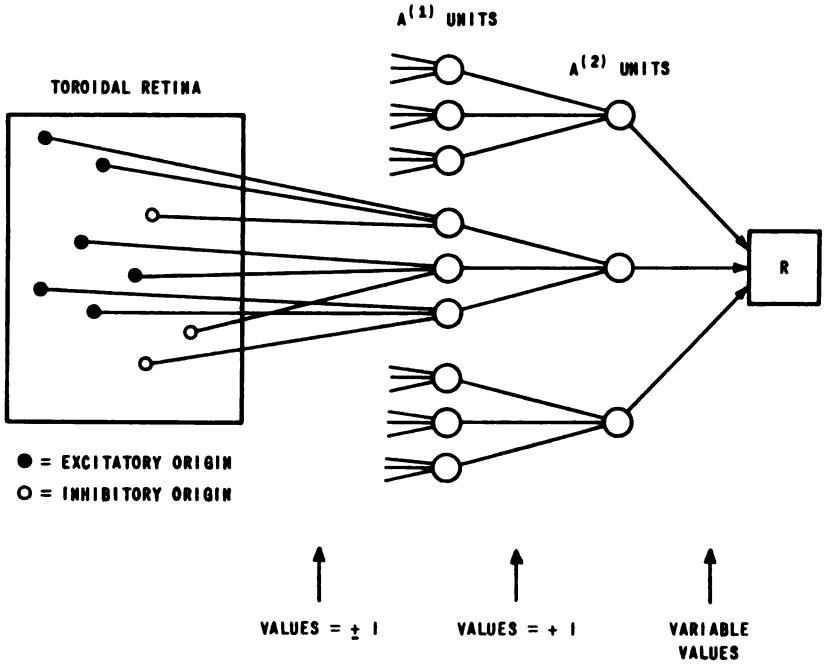


Figure 40 ORGANIZATION OF A SIMILARITY-CONSTRAINED PERCEPTRON ($x = 2$, $y = 1$, $m = 3$). $\mathcal{D}$ = TRANSLATION GROUP IN TOROIDAL RETINA.

(3) The values of all connections from $A^{(1)}$ to $A^{(2)}$ units are equal to +1.

(4) All $A^{(1)}$ units in the source set of a given $A^{(2)}$ unit have origin point configurations which are members of a similarity class, under some similarity relation $\mathcal{R}$.

The subsequent discussion will be limited to the special case in which the similarity relation $\mathcal{R}$ is equivalent to similarity under a transformation group, $\mathcal{G}$, in the sensory space of the perceptron.* This means that, when an origin configuration has been picked for one of the $A^{(1)}$ units connected to a given $A^{(2)}$ unit, the remaining $\mathcal{M}-1$ $A^{(1)}$ units connected to the same $A^{(2)}$ unit must have origin configurations which are transforms under $\mathcal{G}$ of the first configuration selected. This is illustrated in Fig. 40 for a case in which $\mathcal{M} = 3$, and the transformation group is the group of horizontal and vertical translations on the retina. In the model to be analyzed, it is assumed that a single template configuration is chosen at random for each $A^{(2)}$ unit, and the m origin configurations actually assigned to the $A^{(1)}$ units are obtained by selecting m transformations at random, without replacement, from the group $\mathcal{G}$. This yields the auxiliary condition that no two $A^{(1)}$ units in the same source set have identical origin point configurations.

---

* In the case considered here, the world manifold $\mathcal{M}$ and the sensory space $\mathcal{S}$ are taken to be coextensive, with a one-to-one correspondence between objects in $\mathcal{M}$ and stimuli in $\mathcal{S}$.

15.3.2 Analysis

To begin with, we will attempt to provide an intuitive basis for understanding the functioning of the similarity-constrained perceptron. At one extreme, if $\mathcal{m} = 1$, note that the system becomes functionally equivalent to an elementary perceptron of the binomial variety, with A-units having the same parameters as the $A^{(1)}$ units in the 4-layer model. At the other extreme, where m is equal to the order of the transformation group, there is one $A^{(1)}$ unit in each source set for every possible transform of the "template configuration". Now if one of the $A^{(1)}$ units whose origin configuration is ω responds to a stimulus S_χ, any transform $T(S_\chi)$ will necessarily activate the $A^{(1)}$ unit whose origin configuration is the transform $T(\omega)$. Since both of these $A^{(1)}$ units are connected to the same $A^{(2)}$ unit, this unit will respond both to S_χ and $T(S_\chi)$, since its threshold is 1, and the values of the connections from $A^{(1)}$ to $A^{(2)}$ units are fixed at 1. Thus we have the rule that any $A^{(2)}$ unit which responds to a stimulus S_χ will also respond to all transforms $T(S_\chi)$ under the group $\mathcal{T}$. Alternatively, we could state that if S_χ sim $S_y | \mathcal{T}$, and an $A^{(2)}$ unit a_i responds to S_χ, then this unit will also respond to S_y. Next suppose that in addition to making m equal to the order of the group, the threshold of the $A^{(1)}$ units is θ = number of excitatory origins = area of the stimuli, and the number of inhibitory origins is equal to the complementary area, so that an $A^{(1)}$ unit will respond to only one stimulus. We then have an ideal situation, in which an $A^{(2)}$ unit responds to all the members of a given similarity class, and only to members of that similarity class. Under these conditions, if we show the

perceptron a stimulus, say a square, and associate a response to that square, this response will immediately generalize perfectly to all transforms of the square under the group $\mathcal{G}$, and will not generalize at all to any stimulus which is not a transform of the square under $\mathcal{G}$.

The conditions considered above, where m is equal to the order of the group, and each $A^{(1)}$ unit responds to only one possible stimulus, are impractical in the extreme, for a retina of reasonable size. It should be clear from the above arguments, however, that even with smaller values of m (so long as $m > 1$) and lower thresholds, a bias will exist for an $A^{(2)}$ unit to respond to similar stimuli, rather than dissimilar stimuli, under the group $\mathcal{G}$. We now pass on to a quantitative analysis of the performance of this system, first for an environment of random "salt-and-pepper" stimuli, and then for an environment of square stimuli.

The performance of a four-layer perceptron of the type under consideration can be obtained from preceding analyses of elementary perceptrons if we know the G-matrix or the Q-functions of the $A^{(2)}$ units. The expected performance of the system (or the actual performance of a very large system) is entirely determined by the functions $Q_{ij}^{(2)}$, i.e., the probability that a second-layer A-unit will respond both to S_i and to S_j. We will consider the case of a perceptron with N_A sensory points, and a universe of random dot-stimuli, each consisting of $RN_A = n_A$ sensory points chosen at random from a uniform distribution. Let T be any transformation in $\mathcal{G}$, such that the measure of the set of fixed points

under the transformation is zero. We will use the notation S_i' to denote the transform $T(S_i)$, and S_i^* to denote some other transform $T^*(S_i)$, $(T^* \neq T)$. With this notation, $Q_{ii'}^{(2)}$ is the probability that an $A^{(2)}$ unit responds to S_i and to $T(S_i)$, and $Q_{ii^*}^{(2)}$ is the probability that it responds to S_i and to $T^*(S_i)$.

First of all, we have

$$Q_{ii'}^{(2)} = Q_i^{(2)} Q_{i'|i}^{(2)} \tag{15.3}$$

where $Q_{i'|i}^{(2)}$ = conditional probability that an $A^{(2)}$ unit responds to S_i' given that it responds to S_i. For the first factor of this expression, we have the close approximation

$$Q_i^{(2)} \approx 1 - (1 - Q_i^{(1)})^m \tag{15.4}$$

This approximation assumes that the m $A^{(1)}$ units connected to an $A^{(2)}$ unit all have an independent chance of responding to stimulus S_i. This will be approximately true if $\theta << n_\blacktriangle$ for the $A^{(1)}$ units. In this case, since the stimuli consist of random point configurations, the knowledge that an origin point of the first $A^{(1)}$ unit falls on an active S-point still leaves $n_\blacktriangle - 1$ possible S-points in the same stimulus, any one of which might coincide with the transform of the origin point for one of the other $A^{(1)}$ units. In the range of parametric conditions with which we are generally concerned, equation (15.4) approaches a perfect equality.

For the second factor in (15.3) we have the approximation (which is accurate for small $Q_i^{(1)}$)

$$Q_{i'|i}^{(2)} \approx \frac{m-1}{\omega-1} + \left(1 - \frac{m-1}{\omega-1}\right)\left[1 - \left(1 - Q_{i'|i^*}^{(1)}\right)^m\right] \tag{15.5}$$

where ω is the order of the group $\mathcal{G}$. The first term of this expression, $\frac{m-1}{\omega-1}$, is the probability that one of the $m-1$ $A^{(1)}$ units, other than the one which is known to have responded to S_i , has an origin configuration which is a T -transform of the configuration of the "known" A-unit. There are $m-1$ non-identical possibilities that this transform is present, and $\omega-1$ transforms from which they are chosen. If this condition is met, then the $A^{(2)}$ unit must certainly respond to $T(S_i)$. If this condition is not met, with probability $1 - \frac{m-1}{\omega-1}$, it is still possible that one of the $A^{(1)}$ units responds to $T(S_i)$, and this probability is given by the last term of the above expression. Here $Q_{i'|i^*}^{(1)}$ is the probability that an $A^{(1)}$ unit, which is known to respond to some transform $T^*(S_i)$ will also respond to $S_{i'}$. Since T^* may be any transformation (including the identity) so long as it is not equal to T , all of the m $A^{(1)}$ units are equally good candidates for such a response. Specifically, for the case under consideration,

$$Q_{i'|i^*}^{(1)} = \sum_{n_c=0}^{n_A} P(n_c)\, Q_{i'|i^*}^{(1)}(n_c) = \underset{n_c}{E}\, Q_{i'|i^*}^{(1)} \tag{15.6}$$

where n_C = the number of common sensory points in S_i' and S_i^*, with probability

$$P(n_C) = \binom{n_A}{n_C} p^{n_C} (1-p)^{n_A - n_C}$$

$$p = \frac{n_A - 1}{N_A - 1} \tag{15.7}$$

Note that the probability p that a point in S_i^* is in the common area is based on $N_A - 1$ possible locations, since it cannot occupy the location of its transform in S_i'; however, there are $n_A - 1$ other points in S_i' whose locations it might occupy. The only quantity which we still lack is $Q^{(1)}_{i'|i^*}(n_C)$ which is given by

$$Q^{(1)}_{i'|i^*}(n_C) = \frac{Q^{(1)}_{i'i^*}(n_C)}{Q^{(1)}_{i^*}} = \frac{Q_{ij}(C)}{Q_i}$$

where $Q_{ij}(C)$ is computed from Equation (6.5) with $C = n_C/N_A$. Substituting, we have

$$Q^{(1)}_{i'|i^*} = \frac{1}{Q_i} \sum_{n_C=0}^{n_A} \binom{n_A}{n_C} \left(\frac{n_A - 1}{N_A - 1}\right)^{n_C} \left(1 - \frac{n_A - 1}{N_A - 1}\right)^{n_A - n_C} \left[Q_{ij}\left(\frac{n_C}{n_A}\right)\right]$$

$$= \underset{C}{E}\, Q_{i|j}(C) \tag{15.8}$$

Note that as N_Δ, the number of retinal points, goes to infinity, (with n_Δ / N_Δ constant) this quantity approaches

$$\frac{Q_{ij}(R^2)}{Q_i}$$

which is equal to Q_i for the binomial model. At the same time, the first term of (15.5) goes to zero if m remains finite and the order of the group increases with the number of possible retinal locations of the stimulus. Thus, for an infinite retina and a transformation group of infinite order, we have

$$Q^{(2)}_{i'|i} = 1-\left(1-Q_i^{(1)}\right)^m \tag{15.9}$$

and

$$Q^{(2)}_{ii'} = \left[1-\left(1-Q_i^{(1)}\right)^m\right]^2 \tag{15.10}$$

which is identical to the expression for $Q^{(2)}_{ij}$ for a pair of random, unrelated stimuli. Thus, with an infinite retina, no additional generalization is to be expected from a random stimulus to its transform under the conditions assumed above. For a finite retina, however, (or for a finite group $\mathcal{G}$) we have the inequality

$$Q^{(2)}_{ii'} > Q^{(2)}_{ij}$$

due to the effect of the first term in equation (15.5).

Let us now turn to a modification of the above problem, in which the environment consists of square patterns with edges alligned in a square (toroidal) retina, and the group $\mathcal{G}$ consists of all possible translations. In particular, we will take the transformation T to be a lateral translation by half the width of the retina. The notation $S_{i'}$ will be used for $T(S_i)$, and T^* will be taken to mean any transformation in $\mathcal{G}$ not equal to T and not equal to the identify transformation. For convenience, we restrict the area of the stimuli so that $R \leq .25$. This guarantees that S_i and $S_{i'}$ are always disjoint patterns. $Q_i^{(1)}$ is again assumed to be small. In this case we have, in place of (15.5)

$$Q_{i'|i}^{(2)} \approx \frac{m-1}{\omega-1} + \left(1-\frac{m-1}{\omega-1}\right)\left[1-\left(1-Q_{i'|i}^{(1)}\right) E\left(1-Q_{i'|i_1}^{(1)}\right)\left(1-Q_{i'|i_2}^{(1)}\right)\cdots\left(1-Q_{i'|i_{m-1}}^{(1)}\right)\right]$$

where the expectation is with respect to selections of transformations such that $T_j(S_i) = S_{i_j}$.

To avoid the computation of this expectation, we make the further approximation that the expectation of the product of the above sequence of Q-functions is equal to the product of the expected values of the Q-functions. Now it can be shown that for any distribution of $Q_{i'|i_j}$,

$$E\,\Pi(1-Q) \leq \Pi E(1-Q) = \Pi(1-EQ)$$

It follows from this that the approximation which we now propose to make will be a conservative one, yielding values of $Q_{i'|i}^{(2)}$ which are slightly smaller than they should be. With this approximation, we now have:

$$Q^{(2)}_{i'|i} \approx \frac{m-1}{\omega-1} + \left(1-\frac{m-1}{\omega-1}\right)\left[1-\left(1-Q^{(1)}_{i'|i^*}\right)^{m-1}\left(1-Q^{(1)}_{i'|i}(0)\right)\right] \tag{15.11}$$

since the "known" $A^{(1)}$ unit which responds to S_i has the conditional probability

$$Q^{(1)}_{i'|i}(0) = \frac{Q_{i,j}(0)}{Q_i}$$

of responding to the disjoint transform $S_{i'}$. The expression for $Q^{(1)}_{i'|i^*}$ is again given by (15.6), only the probability $P(n_c)$ is different from the random stimulus case. A general equation for $P(n_c)$ will not be developed here, for a finite retina; in particular cases, it is obtained by counting all of the possible ways in which a square and its translate can intersect to yield n_c common points. Some numerical examples will be considered in the following section. Note that the modification from Equation (15.5) to (15.11) will have the effect of tending to diminish the value of $Q^{(2)}_{ii'}$ for small values of m, so that for $m = 1$ the generalization to a disjoint square will always be less than the generalization from a square to a random stimulus of the same area, which is still given by

$$Q^{(2)}_{ij} = \left[1-\left(1-Q^{(1)}_{i}\right)^{m}\right]^2 \tag{15.12}$$

If we go to the limit of an infinite retina, (and infinite transformation group) with the environment of square stimuli just considered, the results differ considerably from the random stimulus case. The difference is due to the distribution of the common area, C, which, in the case of the random stimuli, went to R^2 with probability 1. In the case of randomly placed square stimuli, the probability of a zero intersection in an infinite retina is given by

$$P(C = 0) = 1 - \frac{4k^2}{r^2} \tag{15.13}$$

where k = length of edge of square,
r = width of retina $(r \geq 2k)$.

The probability of $0 < C \leq q$ will be $4/r^2$ times the area under the hyperbola $y = q/x$ from $y = 0$ to k. Specifically,

$$P(0 < C \leq q) = \frac{4}{r^2}\left[\int_{q/k}^{k} \frac{q}{x}\,dx + q\right]$$

$$= \frac{4q}{r^2}\left[1 + 2\ln k - \ln q\right] = \frac{4q}{r^2}\left[1 + \ln\left(\frac{k^2}{q}\right)\right]$$

Differentiating,

$$P(c = q) = \frac{4}{r^2}\left[\ln\left(\frac{k^2}{q}\right)\right] \tag{15.14}$$

Thus, for a square stimulus of area R in a retina of area 1 ($R \leq \frac{1}{4}$) we have

$$Q^{(1)}_{i'|i^*} = \frac{1}{Q_i}\left[\lim_{\epsilon \to 0}\int_{C=\epsilon}^{R} P(C)\, Q^{(1)}_{i;j}(C)\, dC + (1-4R)\, Q^{(1)}_{i;j}(O)\right]$$

$$= \frac{4}{Q_i}\left[\lim_{\epsilon \to 0}\int_{\epsilon}^{R} \left[\ln \frac{R}{C}\right] Q^{(1)}_{i;j}(C)\, dC + \left(\frac{1}{4} - R\right) Q^{(1)}_{i;j}(O)\right] \qquad (15.15)$$

Substituting this in (15.11) yields an expression for $Q^{(2)}_{i'|i}$ for the infinite retina, and $Q^{(2)}_{i';i}$ can be computed by (15.3), as usual.

15.3.3 Examples

Figure 41 illustrates the behavior of a similarity-constrained perceptron, as a function of m, for various combinations of retinal size and types of stimuli. The transformation group, in each case, consists of all horizontal and vertical translations in a square, toroidally connected retina. The stimuli considered are a pair of independent random-dot stimuli, S_a and S_b, a square stimulus S_q, and the transforms $S_{a'}$, $S_{q'}$, where the transformation employed is a shift of half the width of the retina. This guarantees that the square stimulus S_q is disjoint from its transform $S_{q'}$. All stimuli have an area R equal to one fourth of the retina. The parameters of the $A^{(1)}$ units are $x = y = 4$, $\theta = 2$.

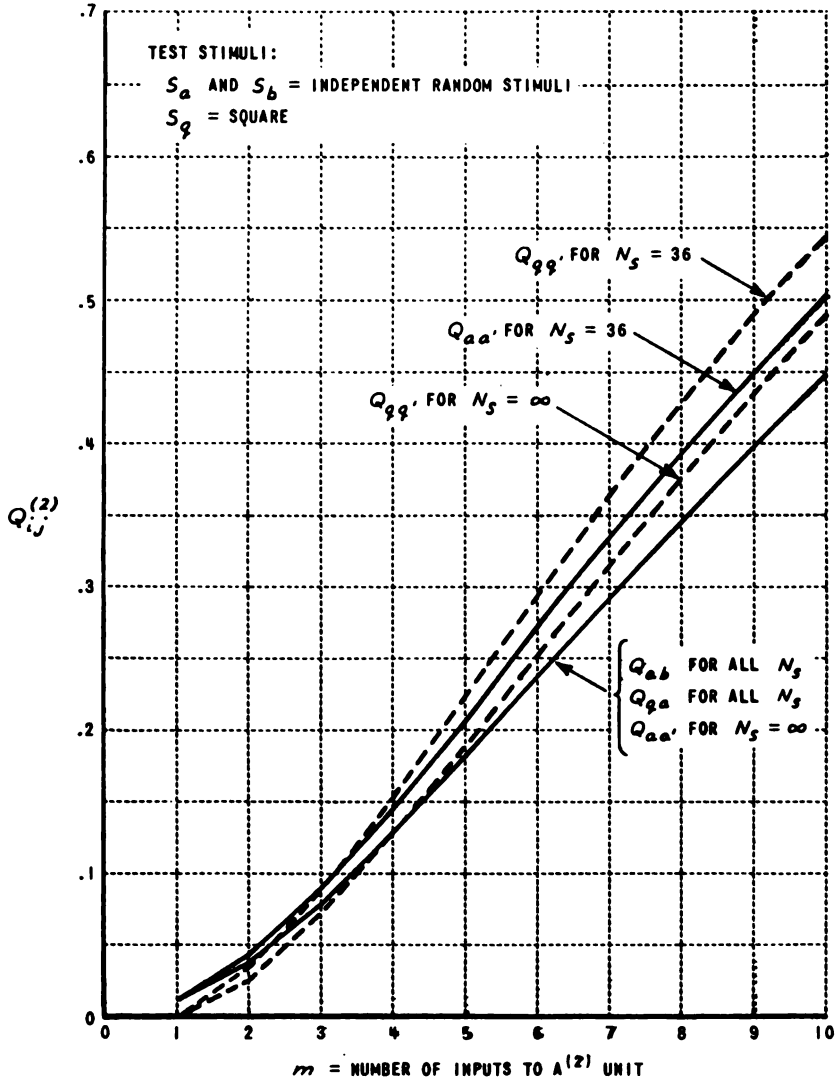


Figure 41 $Q_{ij}^{(2)}$ FOR 4-LAYER SIMILARITY-GENERALIZING PERCEPTRON. $x^{(1)} = y^{(1)} = 4$, $\theta^{(1)} = 2$, $R_i = .25$. TRANSFORMATION GROUP OF HORIZONTAL AND VERTICAL TRANSLATIONS. $S_{i'} = S_i$ DISPLACED BY HALF-WIDTH OF RETINA.

The bottom solid curve provides a baseline, with which the other conditions can be compared. This curve is identical for Q_{ab} (both stimuli random and independent), $Q_{aa'}$ (a random stimulus and its transform) where N_A is infinite, and Q_{qa} (a square stimulus vs. a random stimulus). In a small, finite retina however (specifically, with $N_A = 36$) a random stimulus will generalize more strongly to its transform than to an independent random stimulus, for any $m > 1$. This is shown by the upper of the two solid curves. The broken curves illustrate the generalization from a square to its (disjoint) transform, both for the 6 by 6 retina, and for the infinite retina. In both cases, we find that the system generalizes more strongly to a random stimulus if m is small, but that as m is increased, the perceptron begins to generalize more strongly to the disjoint transform than to a random, unrelated stimulus. For the infinite retina, the cross-over occurs between $m = 4$ and $m = 5$. This means that for a $\mathcal{T}$-system, with $m \geq 5$, g_{ij} will be positive from a square to any other square, and will be zero from a square to a random dot stimulus. Increasing the threshold of the $A^{(1)}$ units will reduce $Q_{ij}^{(2)}$ for all curves, but will increase the relative bias towards similar stimuli, and will shift the cross-over point further to the left for the $Q_{qq'}$ curves.

The difference in performance for squares as opposed to random stimuli will tend to be characteristics of any coherent stimulus patterns, provided the transformation group is one which preserves the coherence, or compactness, of the stimuli. This may be puzzling to some readers who recognize that under the connection rules employed

in these perceptrons, there is nothing unique about topologically connected, or continuous regions, which would affect the perceptron's ability to recognize them in any different way than disconnected regions. It is, after all, only the set of points to which connections happen to be made which determines the response of a perceptron, and if every S-unit were randomly interchanged with some other S-unit, a corresponding change being induced in the stimulus environment, the performance of the perceptron should not be affected at all. This will indeed be true, provided any transformation group employed in the first perceptron is replaced by a new transformation group corresponding to the rearranged retina. The essential feature of coherent stimuli with a group of coherence-preserving transformations is that the probability distribution of stimulus-intersections does not concentrate at the expected value of the intersection, as N_A and the order of the group become infinite. This permits a similarity bias to be maintained for such stimuli which cannot be maintained for random stimuli. Any group generated by a permutation operation on the points of the retina will have the same property, provided the same permutation operation is applied to the stimuli. Another way of looking at the problem is to note that with random stimuli, a sensory origin point which is close to a stimulus point, but does not coincide with it exactly, has a probability of being activated no greater than that of any other origin-point. With coherent stimuli, on the other hand, an origin-point which is close to a stimulus point has a greater probability of being activated than one which is remote from the stimulus point. Thus, for random stimuli, only a transformed origin configuration which corresponds exactly to the transformation T will help in generalizing from S

to $T(S)$. For coherent stimuli, it is sufficient that the transformed origin points should be in the neighborhood of the required transform; proximity to the required transformation is sufficient to increase the probability of being activated by $T(S)$.*

Note that as m increases, the value of $Q_{i,j}$ tends to approach unity for all curves in Fig. 41. This means that there will be a maximum similarity bias at some finite value of m , beyond which the advantage of similar over random stimuli will approach zero. By increasing the value of θ for the $A^{(1)}$ units, the location of the maximum bias can be shifted further to the right, until, with $\theta = \alpha n_4$, the maximum will occur at $m = \omega$.

15.4 Laws of Similarity-Generalization in Perceptrons

The results obtained in the previous section illustrate a number of effects which are found quite generally in perceptrons which show a capability for similarity-generalization, regardless of whether this capability is learned or intrinsic, and regardless of whether the perceptron is series-coupled or cross-coupled. Additional evidence for these general results will be found in subsequent chapters, and they appear to take on the status of empirical laws, which have now been substantiated for a rather wide variety of systems. These laws can be tentatively stated as follows:

---

* The effects noted here are directly analogous to those originally predicted for cross-coupled systems in Ref. 85.

(1) As the size of the retina increases, it becomes increasingly difficult to recognize the similarity of two random-pattern stimuli under a given transformation group, with a finite perceptron. With an infinite retina (and transformation group of infinite order) the similarity bias for random stimuli goes to zero.

(2) The similarity-bias for coherent stimuli, under a coherence-preserving transformation group, will generally be stronger than for random stimuli, and will not go to zero even for an infinite retina and transformation group of infinite order.

(3) The similarity bias of a perceptron can be increased by raising the threshold of its A-units or by increasing the number of connections to terminal A-units (i.e., generalization will be limited increasingly to the members of a similarity class, as the threshold or number of pre-terminal units is increased).

(4) Generalization to disjoint transforms of a stimulus may be less than generalization to independent random patterns, for a perceptron with weak similarity bias; generalization to disjoint transforms can be made to exceed generalization to random stimuli, however, by an increase in A-unit thresholds or by increasing the number of inputs to the terminal A-units of the network.

16. FOUR-LAYER PERCEPTRONS WITH ADAPTIVE PRETERMINAL NETWORKS

The physical universe, at a macroscopic level, is characterized by the continuity of its transformations through time. Objects do not suddenly appear out of nowhere, persist for an instant, and then vanish into nothingness. Given an appropriate time-scale, all changes appear to occur smoothly and progressively. Consequently, stimuli which are highly similar under a continuous transformation group are more likely to occur in close temporal succession than dissimilar stimuli. In this chapter, it will be shown that an initially unbiased perceptron can take advantage of this property of the physical environment to evolve a capability for similarity generalization, without any intervention by an experimenter or reinforcement control system.

The model which is presented here was developed jointly by Block, Knight, and Rosenblatt, in the hopes that its analysis would assist in the understanding of closely related problems which occur in cross-coupled systems. The similarity between the performance of this sytem and the performance of cross-coupled systems is most striking, as will be seen in later chapters. The main effects of cross-coupling will be to accelerate the adaptation process, and to make the system inherently responsive to stimulus sequences, rather than momentary stimuli. The presentation in the first parts of this chapter is essentially the same as that of Block, Knight, and Rosenblatt (Ref. 7).

16.1 Description of the Model

The perceptron to be analyzed is illustrated in Fig. 42. It is a four-layer series coupled system, with an equal number (N_a) of $A^{(1)}$ units and $A^{(2)}$ units.* Each $A^{(2)}$ unit receives a variable-valued connection from each of the $A^{(1)}$ units. In addition, each $A^{(2)}$ unit receives a fixed-value connection from one of the $A^{(1)}$ units. For convenience, the $A^{(1)}$ and $A^{(2)}$ units are placed in one-to-one correspondence, with the fixed connection to each $A^{(2)}$ unit originating from its "mate" in the $A^{(1)}$ layer. The threshold of the $A^{(1)}$ units is $\theta^{(1)}$, and the threshold of the $A^{(2)}$ units is $\theta^{(2)}$. To simplify notation, we will use the symbol θ to designate $\theta^{(2)}$, unless otherwise indicated. The fixed connections from $A^{(1)}$ to $A^{(2)}$ units all have values $\geq \theta$. For specificity, we assume that all of these fixed values are exactly equal to θ. The variable-valued connection from an $A^{(1)}$ unit a_i to an $A^{(2)}$ unit a_j has a value $c_{ij}(t$ at time t. The symbol c_{ij} will be used to designate values of $A^{(1)}$ to $A^{(2)}$ connections, and v_{ir} to designate values of $A^{(2)}$ to R-unit connections. The input connections to the $A^{(1)}$ units may be organized according to any of the models (e.g., binomial or Poisson) which were discussed in Part II. Signal transmission times, τ_{ij}, are assumed to be equal to zero, for all connections. It is assumed that stimuli occur at times t, $t + \Delta t$, $t + 2\Delta t$, etc.

---

* The numbers of units need not be equal for systems of this type to work; the constraint is introduced in order to simplify the analysis. It is equally satisfactory to organize the perceptron with m variable valued connections and 1 fixed value connection to each $A^{(2)}$ unit, with origins chosen at random.

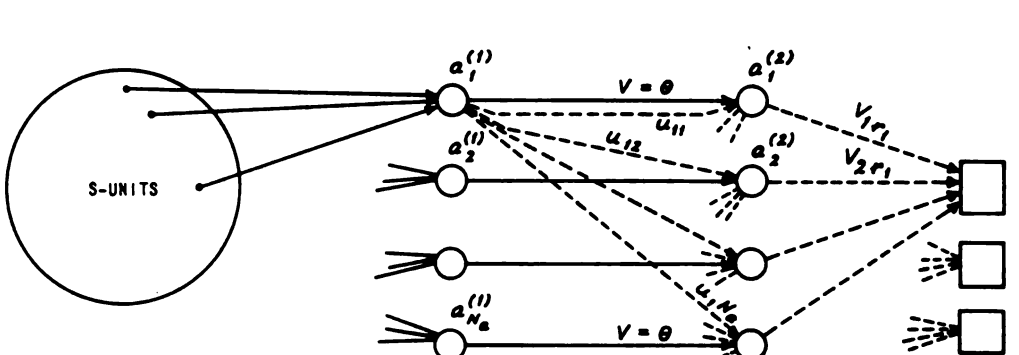


Figure 42 AN ADAPTIVE FOUR-LAYER PERCEPTRON (SOLID LINES = FIXED VALUE CONNECTIONS, BROKEN LINES = VARIABLE VALUE CONNECTIONS)

The variable values u_{ij} are assumed to be initially equal to zero, and change with time as follows: If unit $a_r^{(1)}$ is active at time t and $a_s^{(2)}$ is active at $t+\Delta t$, then u_{rs} receives an increment $(\eta \cdot \Delta t)$, and all connections u_{ij} decay by a quantity $(\delta \cdot \Delta t) u_{ij}$. The values of the $A^{(2)}$ to R-unit connections may be varied by any one of the usual reinforcement rules. Note that under these rules, the values u_{ij} will always be non-negative, so that if the "mate" of a given $A^{(2)}$ unit is active, the $A^{(2)}$ unit will always be active. In the subsequent analysis, it will be shown that with a suitable sequential organization of the environment, these dynamic rules can lead to the development of a perceptron organization closely analogous to that of the similarity-constrained perceptrons of the previous chapter.

16.2 General Analysis

16.2.1 Development of the Steady-State Equation

As in the last chapter, our main concern will be to find the values of $Q_{ij}^{(2)}$, which will permit further analysis to proceed along the lines employed for elementary perceptrons. Unlike the perceptrons of Chapter 15, however, the values of $Q_{ij}^{(2)}$, and consequently the G-matrix of the perceptron, are stochastic variables, depending upon the prior history of the system.

The set of A-units in the $A^{(1)}$ layer responding to S_i will be denoted by $A^{(1)}(S_i)$; the set responding to both S_i and S_j is $A^{(1)}(S_i) \cap A^{(1)}(S_j)$. For a perceptron with a known connection

scheme for the $A^{(1)}$ layer (or for a sufficiently large perceptron) the fraction of $A^{(1)}$ units responding to both S_i and S_j will be $Q_{ij}^{(1)}$, and is equal to the number of elements in $A^{(1)}(S_i) \cap A^{(1)}(S_j)$ divided by N_a. These quantities are fixed for all time.

Let $\alpha_\lambda^{(1)}(t)$ denote the total input signal to the unit $a_\lambda^{(2)}$ at time t, in response to stimulus S_i.* Then

$$\alpha_\lambda^{(i)}(t) = \theta a_\lambda^*(S_i) + \sum_{r=1}^{N_a} u_{r\lambda}(t)\, a_r^*(S_i) \tag{16.1}$$

where $$a_\lambda^*(S_i) = \begin{cases} 1 \text{ if } S_i \text{ activates } a_\lambda^{(1)} \\ 0 \text{ otherwise} \end{cases}$$

This represents the sum of the signal arriving at a_λ on its fixed connection, and all of the signals arriving on the variable-valued connections at time t. Let

$$\beta_\lambda^{(i)} = \theta a_\lambda^*(S_i) \tag{16.2}$$

$$\gamma_\lambda^{(i)}(t) = \sum_{r=1}^{N_a} u_{r\lambda}(t)\, a_r^*(S_i) \tag{16.3}$$

Then

$$\alpha_\lambda^{(i)}(t) = \beta_\lambda^{(i)} + \gamma_\lambda^{(i)}(t) \tag{16.4}$$

---

* The indices i, j, and k will be used throughout this chapter to designate various stimuli, and the indices r and λ will be used to designate particular A-units.

Note that $\beta_\lambda^{(i)}$ is θ or 0 depending on whether $a_\lambda^{(1)}$ is in $A^{(1)}(S_i)$ or not; it is invariant with time. On the other hand, $\gamma_\lambda^{(i)}(t)$ represents the effect of the variable $A^{(1)}$ to $A^{(2)}$ connections.

Now suppose that at time t_0 stimulus S_j occurs, and at time $t_0 + \Delta t$ stimulus S_k occurs. Then the consequent change in $u_{r\lambda}$ will be

$$u_{r\lambda}(t_0 + 2\Delta t) - u_{r\lambda}(t_0 + \Delta t) = (\eta \cdot \Delta t)\, a_r^*(S_j)\, \phi(\alpha_\lambda^{(k)}(t_0 + \Delta t)) - (\delta \cdot \Delta t)\, u_{r\lambda}(t_0 + \Delta t) \tag{16.5}$$

where $$\phi(x) = \begin{cases} 0 \text{ for } x < \theta \\ 1 \text{ for } x \geq \theta \end{cases}$$

From (16.3) and (16.5) we get

$$\gamma_\lambda^{(i)}(t_0 + 2\Delta t) - \gamma_\lambda^{(i)}(t_0 + \Delta t) = \sum_{r=1}^{N_a} \left[u_{r\lambda}(t_0 + 2\Delta t) - u_{r\lambda}(t_0 + \Delta t)\right] a_r^*(S_i)$$

$$= (\eta \cdot \Delta t)\, \phi\left[\alpha_\lambda^{(k)}(t_0 + \Delta t)\right] \sum_{r=1}^{N_a} a_r^*(S_j)\, a_r^*(S_i) - (\delta \cdot \Delta t) \sum_{r=1}^{N_a} u_{r\lambda}(t_0 + \Delta t)\, a_r^*(S_i)$$

Hence

$$\gamma^{(i)}(t_0 + 2\Delta t) - \gamma^{(i)}(t_0 + \Delta t) = (\eta \cdot \Delta t)\, \phi\left[\alpha^{(k)}(t_0 + \Delta t)\right] N_a Q_{ij}^{(i)} - (\delta \cdot \Delta t)\, \gamma^{(i)}(t_0 + \Delta t) \tag{16.6}$$

where, for brevity, the subscript λ has been suppressed. It must be remembered that γ and α, in these equations, refer to any particular $A^{(2)}$ unit, $a_\lambda^{(2)}$.

Now suppose the sequence of stimuli $\{S_{j_0}, S_{j_1}, \ldots, S_{j_M}\}$ occurs at the successive times $t, t+\Delta t, \ldots, t+M\Delta t$. In Equation (16.6) we take $t_0 = t + m\Delta t$, $[m = 0, 1, 2, \ldots, (M-1)]$, $j = j_m$, $k = j_{m+1}$, and obtain

$$\gamma^{(i)}(t+(m+2)\Delta t) - \gamma^{(i)}(t+(m+1)\Delta t) = (\eta\cdot\Delta t)\,\phi\left[\alpha^{(j_{m+1})}(t+(m+1)\Delta t)\right] N_a Q^{(1)}_{ij_m} - (\delta\cdot\Delta t)\,\gamma^{(i)}(t+(m+1)\Delta t) \tag{16.7}$$

Summing on m from 0 to $M-1$ we get the change in $\gamma^{(i)}$ due to the entire sequence of stimuli:

$$\gamma^{(i)}(t+(M+1)\Delta t) - \gamma^{(i)}(t+\Delta t) = \sum_{m=0}^{M-1}\left\{(N_a\eta\Delta t)\,\phi\left[\alpha^{(j_{m+1})}(t+(m+1)\Delta t)\right] Q^{(1)}_{ij_m} - (\delta\cdot\Delta t)\,\gamma^{(i)}(t+(m+1)\Delta t)\right\} \tag{16.8}$$

We now divide by $M\Delta t$ and let Δt approach zero* to obtain

$$\frac{d\gamma^{(i)}}{dt} = \sum_{m=0}^{M-1}\frac{(N_a\eta)}{M}\,\phi\left(\alpha^{(j_{m+1})}(t)\right) Q^{(1)}_{ij_m} - \delta\gamma^{(i)}(t) \tag{16.9}$$

---

* An alternative treatment is possible in which difference equations are carried throughout, rather than converting to a differential equation. The true solution for $\gamma^{(i)}$ obtained from such an approach is a fluctuating function, the local time-average of which corresponds to the solution of the differential equation, which is obtained here. As long as η and δ are sufficiently small, the differential equation, which is somewhat easier to manage, yields a close approximation to the true solution of the finite difference equation.

Let F_{jk} be the number of times the stimulus pair $S_j S_k$ occurs in the given sequence $S_{j_0}, S_{j_1}, \ldots, S_{j_M}$; also, let $f_{jk} = F_{jk}/M$ be the average frequency of the pair $S_j S_k$. Then from (16.9) we get

$$\frac{d\gamma^{(i)}}{dt} = \sum_{j=1}^{n} \sum_{k=1}^{n} (N_a \eta) f_{jk} \phi(\alpha^{(k)}(t)) Q_{ij}^{(i)} - \delta\gamma^{(i)}(t) \qquad (16.10)$$

where n , as usual, represents the number of distinct stimulus patterns in the environment. Defining the matrix $C = QF$, with elements

$$c_{ij} = \sum_{k=1}^{n} Q_{ij}^{(i)} f_{kj}$$

we have from (16.10)

$$\frac{d\gamma^{(i)}}{dt} = (N_a \eta) \sum_{j=1}^{n} c_{ij} \phi(\beta^{(j)} + \gamma^{(j)}(t)) - \delta\gamma^{(i)}(t) \qquad (16.11)$$

This gives us a non-linear system of differential equations for $\gamma^{(1)}(t), \ldots, \gamma^{(n)}(t)$ with initial conditions $\gamma^{(i)}(0) = 0$.

If the frequencies f_{kj} vary with t , then the coefficients c_{ij} are time-dependent, but in any case they are non-negative and bounded; ϕ is non-negative, monotone increasing in γ , bounded and continuous on the right. It will be assumed here that the c_{ij} are constants (corresponding to fixed frequencies, f_{kj}).

In preparation for discussing the solution of (16.11), consider the equilibrium equation

$$\gamma^{(i)} = \frac{N_a \eta}{\delta} \sum_{j=1}^{n} c_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)}\right) \tag{16.12}$$

This corresponds to a solution of (16.11) for the steady-state condition in which the rate of gain (represented by the first term of 16.11) is exactly counterbalanced by the rate of decay. But the system of equations (16.12) may have more than one solution. However, we shall show that there is a unique minimal solution (by which we mean a solution none of whose components $\gamma^{(i)}$ exceed the corresponding components of another solution); and this minimal solution is obtained in a finite number (at most n) of iterations of (16.12), starting with all $\gamma^{(j)} = 0$ on the right-hand side of the equation, finding the new values of $\gamma^{(j)}$ from (16.12), putting these back into the right-hand side, and so on. That is, we take $\gamma_0^{(i)} = 0$ and

$$\gamma_{n+1}^{(i)} = \frac{N_a \eta}{\delta} \sum_{j=1}^{n} c_{ij}\, \phi\left(\beta^{(j)} + \gamma_n^{(j)}\right) \tag{16.13}$$

We shall prove first that this process terminates in at most n iterations. This can be seen from the following considerations. Since the right-hand side of (16.13) is non-negative and $\gamma_0^{(i)} = 0$, it follows that $\gamma_1^{(i)} \geq \gamma_0^{(i)}$. Now since the right side of (16.13) is a non-decreasing function of the γ's, it follows that $\gamma_2^{(i)} \geq \gamma_1^{(i)}, \ldots,$ $\gamma_{n+1}^{(i)} \geq \gamma_n^{(i)}$. Therefore, also $\phi\left(\beta^{(j)} + \gamma_{n+1}^{(j)}\right) \geq \phi\left(\beta^{(j)} + \gamma_n^{(j)}\right)$; that is, successive ϕ's cannot decrease. If, at a particular step, no ϕ

increases, then we are at a solution. The ϕ 's have only the values zero or 1, so even if only a single ϕ changes at each step, the process terminates in at most n steps.

The solution thus obtained will be denoted by $\gamma^{(i)*}$ We shall now prove that this solution is minimal. Let $\tilde{\gamma}^{(i)}$ be any solution of the equilibrium equation (16.12). Then for the iteration process (16.13), we have $\gamma_0^{(i)} \leq \tilde{\gamma}^{(i)}$, for all i . Since the right-hand side of (16.13) is a monotone function of $\gamma^{(i)}$, we have

$$\gamma_1^{(i)} = \frac{N_a \eta}{\sigma} \sum_{j=1}^{n} C_{ij}\, \phi\left(\beta^{(j)} + \gamma_0^{(j)}\right) \leq \frac{N_a \eta}{\sigma} \sum_{j=1}^{n} C_{ij}\, \phi\left(\beta^{(j)} + \tilde{\gamma}^{(j)}\right) = \tilde{\gamma}^{(i)}$$

Similarly, $\gamma_n^{(i)} \leq \tilde{\gamma}^{(i)}$, hence $\gamma^{(i)*} \leq \tilde{\gamma}^{(i)}$. Hence $\gamma^{(i)*}$ is minimal.

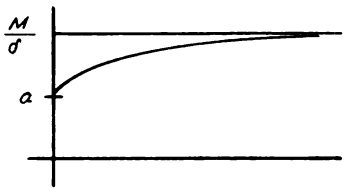
To avoid consideration of a special pathological case, we now make a mild assumption. Consider the sum $\frac{N_a \eta}{\sigma} \sum_{j \in k} C_{ij}$ taken over a subset R of the possible values of j ($1, 2, \ldots, n$). We assume that no such sum is equal to θ . This is not a serious assumption, since by a small change in $\frac{N_a \eta}{\sigma}$ this requirement can always be satisfied.

Now suppose that the $\gamma^{(i)}(t)$ satisfy the system of differential equations (16.11) and the initial conditions $\gamma^{(i)}(0) = 0$ Then we assert that the $\gamma^{(i)}(t)$ are non-decreasing and $\lim_{t \to \infty} \gamma^{(i)}(t) = \gamma^{(i)*}$. That is, the solution obtained by the iterative process (16.13) is indeed the solution of the differential equation (16.11), with initial conditions zero in each case.

First we shall show that

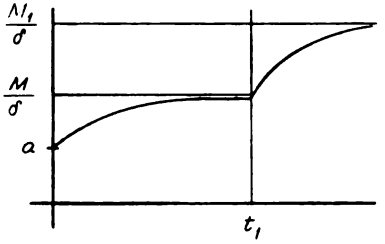
$\frac{d\gamma^{(i)}}{dt} \geq 0$. Moreover, if $\gamma^{(i)}(t) > 0$, then $\frac{d\gamma^{(i)}}{dt} > 0$. (α)

As a preliminary step, consider the nature of the solution of the equation $\frac{dx}{dt} = M - \delta x$, where M and δ are positive constants, and $x(0) = 0$, where $0 \leq a < \frac{M}{\delta}$. The solution, $x = \frac{M}{\delta} - e^{-\delta t}\left(\frac{M}{\delta} - a\right)$, has the appearance of the following curve:



(a)

The solution approaches M/δ monotonely from below, and $dx/dt > 0$ for all $t > 0$. If at time $t = t_1$ we replace M by $M_1 > M$ the solution appears as



(b)

as t goes from 0 to t_1 the solution approaches M/δ monotonely from below; as t increases beyond t_1 the solution approaches M_1/δ monotonely from below. The solution is continuous; so is its derivative, except at t_1, where the left and right hand derivatives are not equal, but both are positive.

If instead of $\frac{M}{\delta} > a \geq 0$, we take $M = a = 0$, the solution is $x(t) = 0$ for $0 \leq t \leq t_1$.

We now proceed to the proof of (α). Let $M_i^{(t)} = N_a \eta \sum_{j=1}^{n} c_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)}(t)\right)$. Then (16.11) can be written

$$\frac{d\gamma^{(i)}}{dt} = M^{(i)}(t) - \delta\gamma^{(i)}(t) \tag{16.14}$$

where here and in the following paragraph, i is a generic index of the set $(1, 2, \ldots, n)$, while j and k will refer to specific indices to be defined below.

Each equation $M^{(i)}(t)$ can take on at most 2^n possible values. Let k be a specific value of i and suppose first that $M^{(k)}(0) = 0$ The only times at which $M^{(i)}(t)$ can change its value are when one of the $\gamma^{(i)}$ (indeed one whose corresponding $\beta^{(i)} = 0$) reaches the value θ Suppose the first time at which this happens is $t_1 > 0$. Suppose then that $\gamma^{(j)}(t_1) = \theta$. Since in the interval $0 < t < t_1$ all $\frac{d\gamma^{(i)}}{dt} \geq 0$ we have $M^{(i)}(t_1) \geq M^{(i)}(t_0)$. Thus the solution $\gamma^{(k)}(t)$ appears as in Figure (b) above; in particular, for all k such that $M^{(k)}(0) > 0$ we have $\gamma^{(k)}(t_1) < \frac{M^{(k)}(0)}{\delta} \leq \frac{M^{(k)}(t_1)}{\delta}$; and for the others

$\gamma^{(i)}(t_1) \leq \frac{M^{(i)}(t_1)}{\sigma}$. Furthermore, since both the left and right derivatives of $\gamma^{(i)}(t_1)$ are positive we have, for $t > t_1$ and sufficiently close to t_1 , $\gamma^{(i)}(t) > \theta$, so that it will not be until t_2 , with $t_2 > t_1$, that there will again be a $\gamma^{(i)}(t)$ having the value θ . In the interval $t_1 < t < t_2$ we have the same pertinent conditions as we had in the interval $0 < t < t_1$; namely, $\frac{d\gamma^{(i)}}{dt} = M^{(i)}(t_1) - \sigma\gamma^{(i)}(t)$, with initial values $\gamma^{(i)}(t_1) \leq \frac{M^{(i)}(t_1)}{\sigma}$ and in particular $\gamma^{(k)}(t_1) < \frac{M^{(k)}(t_1)}{\sigma}$. Thus in the interval $t_1 < t < t_2$ we again have $\frac{d\gamma^{(i)}}{dt} \geq 0$, and $\frac{d\gamma^{(k)}}{dt} > 0$. The same argument applies to successive intervals $(t_2, t_3), (t_3, t_4)$, and so on. Since the $\gamma^{(i)}(t)$ are monotone there are at most n such intervals.

If $M^{(k)}(0) = 0$, then $\gamma^{(k)}(t) = 0$ for $0 < t < t_1$. If $M^{(k)}(t_1) > 0$, then we use the previous argument starting at $t = t_1$; otherwise $\gamma^{(k)}$ remains zero at least until t_2 , and so on. In any case, the statement (α) has been proven.

Next we shall show that

$$\lim_{t \to \infty} \gamma^{(i)}(t) = \gamma^{(i)\#}; \quad i = 1, 2, \ldots, n. \qquad (\beta)$$

Since, from the proof of (α) it is clear that each $\gamma^{(i)}(t)$ is monotone and bounded, $\lim_{t \to \infty} \gamma^{(i)}(t)$ exists; call it $\gamma^{(i)\#}$; it is a sum of the form $\frac{N_a \eta}{\sigma} \sum_{j \in R} c_{ij}$, which was assumed at the outset to be unequal to θ , and thus $\gamma^{(i)\#} \neq \theta$. Therefore, $\phi(\beta^{(j)} + \gamma^{(j)})$ is continuous when $\gamma^{(j)} = \gamma^{(j)\#}$. Letting $t \to \infty$

in equation (16.11) we see that $\gamma^{(i)\#}$ is a solution of the equilibrium equation (16.12). Hence $\gamma^{(i)\#} \geq \gamma^{(i)*}$, since $\gamma^{(i)*}$ is minimal. We next show that for all $t \geq 0$, $\gamma^{(i)}(t) \leq \gamma^{(i)\#}$.

Note that initially $\gamma^{(i)}(0) \leq \gamma^{(i)*}$. Suppose that t_1 is the first time at which some $\gamma^{(k)}(t) = \gamma^{(k)*}$. From (16.11) and the fact that ϕ is non-decreasing we see that at t_1,

$$\frac{d\gamma^{(k)}}{dt} = N_a \eta \sum_j C_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)}(t_1)\right) - \delta\gamma^{(k)}(t_1)$$

$$\leq N_a \eta \sum_j C_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)*}\right) - \delta\gamma^{(k)}(t_1)$$

$$= \delta\gamma^{(k)*} - \delta\gamma^{(k)}(t_1) = 0$$

i.e., $\frac{d\gamma^{(k)}}{dt} \leq 0$ at $t = t_1$.

If $\gamma^{(k)*} > 0$, we have from (α) that $\frac{d\gamma^{(k)}}{dt} > 0$ at t_1, which is a contradiction. Suppose that $\gamma^{(k)*} = 0$, so that also $t_1 = 0$. Then, as long as no $\gamma^{(i)}(t)$ reaches a non-zero $\gamma^{(i)*}$, we have

$$M^{(k)}(t) = N_a \eta \sum_{j=1}^{n} C_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)}(t)\right) \leq N_a \eta \sum_{j=1}^{n} C_{ij}\, \phi\left(\beta^{(j)} + \gamma^{(j)*}\right) = \delta\gamma^{(k)*} = 0.$$

Hence over this period $\gamma^{(i)}(t) = 0$. But no non-zero $\gamma^{(i)*}$ can ever be attained by $\gamma^{(i)}(t)$, since, by the above argument, we would have $\frac{d\gamma^{(i)}}{dt} \leq 0$ at the first time this occurs, in contradiction to (α) .

Hence if $\gamma^{(i)*} > 0$, then $\gamma^{(i)}(t) < \gamma^{(i)*}$; and if $\gamma^{(i)*} = 0$, then $\gamma^{(i)}(t) = \gamma^{(i)*}$. In general, $\gamma^{(i)}(t) \leq \gamma^{(i)*}$.

Hence $\gamma^{(i)*} = \lim \gamma^{(i)}(t) \leq \gamma^{(i)*}$, and (β) follows.

From this point on, we shall be concerned with the steady-state values $\gamma^{(i)*}$, and for brevity we shall drop the $*$. In the terminal condition, the A-unit $a_\lambda^{(2)}$, whose history we have been following up to this point, is activated by S_i if $\beta_\lambda^{(i)} + \gamma_\lambda^{(i)} \geq \theta$. The set of $A^{(2)}$ units which are activated by stimulus S_i are denoted by $A^{(2)}(S_i)$. In the initial state, the set $A^{(2)}(S_i)$ is denoted by $A_0^{(2)}(S_i)$, and in the terminal state by $A_\infty^{(2)}(S_i)$. The expected fraction of $A^{(2)}$ units which are activated by both S_i and S_j will be $Q_{ij}^{(2)}$ and is equal to the expected number of units in $A^{(2)}(S_i) \cap A^{(2)}(S_j)$ divided by N_a .

Once the $Q_{ij}^{(2)}$ are known, the behavior of the perceptron in its terminal (steady state) condition can be predicted. To determine these terminal values of $Q_{ij}^{(2)}$, we can proceed as follows. First, the set of $A^{(1)}$ units is broken into the smallest possible cells of the Venn diagram which represents the sets of units responding to different stimuli (c.f., Fig. 43). For the units in each of these cells, there is a characteristic $\mathcal{B}$ -vector. For each such $\mathcal{B}$ -vector, we solve equation (16.12) for the terminal values of $\gamma^{(i)}$. Here we assume $f_{\lambda j}$ to be given, and

$Q_{ik}^{(1)}$ can be obtained from previous equations (as in Chapter 6). Initially, $Q_{ij}^{(2)} = Q_{ij}^{(1)}$. Knowing $\beta^{(i)}$ and $\gamma^{(i)}$, we can determine the region of the $A^{(2)}$ Venn diagram to which each cell of $A^{(2)}$ units moves. Thus we obtain the complete terminal distribution of A-units in the Venn diagram of $A^{(2)}$, and hence in particular the $Q_{ij}^{(2)}$. It can be seen that the motion will be for A-units to tend to go into higher-order intersections, but that points which are initially outside all the $A^{(2)}(S_i)$ will stay outside all the $A^{(2)}(S_i)$.

16.2.2 A Numerical Example

To clarify the above description, an illustrative example is worked out here numerically. Suppose there are three stimuli, S_1, S_2, and S_3, which initially activate sets of $A^{(2)}$ units (or sets of $A^{(1)}$ units, which will be equivalent under starting conditions) shown in the Venn diagram of Figure 43(a). Here the $Q_{ij}^{(1)}$ matrix, and the initial value of the $Q_{ij}^{(2)}$ matrix is

$$\begin{pmatrix} .3 & .1 & .1 \\ .1 & .4 & .3 \\ .1 & .3 & .6 \end{pmatrix}$$

Suppose the sequence $S_{j_0}, S_{j_1}, \ldots, S_{j_m}$, from the above analysis, is $S_1\ S_2\ S_2\ S_1\ S_2\ S_3\ S_1\ S_3\ S_1\ S_2$. This is repeated over and over during the training, or "preconditioning" of the perceptron. Then the f_{ij} matrix is

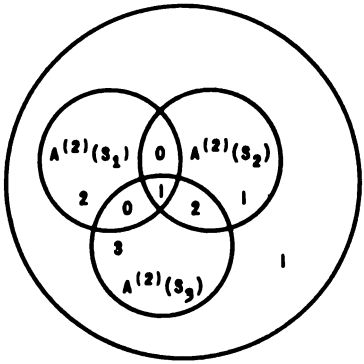


Figure 43 (a) VENN DIAGRAM OF INITIAL $A^{(2)}$ SETS, FOR ILLUSTRATIVE EXAMPLE. 10 A-UNITS, DISTRIBUTED AS SHOWN.

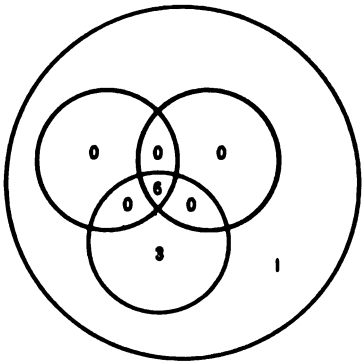


Figure 43 (b) TERMINAL VENN DIAGRAM, FOR $\eta/\delta = \theta = 1$.

$$\begin{pmatrix} 0 & .3 & .1 \\ .2 & .1 & .1 \\ .2 & 0 & 0 \end{pmatrix}$$

Consequently, we have the matrix

$$C_{ij} = \frac{1}{100} \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 3 \\ 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} 0 & 3 & 1 \\ 2 & 1 & 1 \\ 2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} .04 & .10 & .04 \\ .14 & .07 & .05 \\ .18 & .06 & .04 \end{pmatrix}$$

The equilibrium equations (16.12) then become

$$\begin{pmatrix} \gamma^1 \\ \gamma^2 \\ \gamma^3 \end{pmatrix} = \frac{\eta}{\sigma} \begin{pmatrix} .4 & 1.0 & .4 \\ 1.4 & .7 & .5 \\ 1.8 & .6 & .4 \end{pmatrix} \begin{pmatrix} \phi(\beta^1 + \gamma^1) \\ \phi(\beta^2 + \gamma^2) \\ \phi(\beta^3 + \gamma^3) \end{pmatrix} \qquad (16.15)$$

Now we begin to trace the destinations of cells of the Venn diagram of Fig. 43(a). Start with the two A-units which are activated only by S_1. Here $\vec{\beta} = \theta(1, 0, 0)$. The first iteration of (16.15) then gives

$$\begin{pmatrix} \gamma^1 \\ \gamma^2 \\ \gamma^3 \end{pmatrix} = \frac{\eta}{\sigma} \begin{pmatrix} .4 \\ 1.4 \\ 1.8 \end{pmatrix}$$

If $\eta/\sigma < \theta/1.8$, then these γ's are zero, and the points in question stay in the same resion of the Venn diagram. To be specific, let us take $\eta/\sigma = \theta = 1$. Then we get for the first approximation

$$\begin{pmatrix} \gamma^1 \\ \gamma^2 \\ \gamma^3 \end{pmatrix} = \begin{pmatrix} .4 \\ 1.4 \\ 1.8 \end{pmatrix}$$

and for the second iteration,

$$\begin{pmatrix} \gamma^1 \\ \gamma^2 \\ \gamma^3 \end{pmatrix} = \begin{pmatrix} 1.8 \\ 2.6 \\ 2.8 \end{pmatrix}$$

which is the fixed point. The two associators in question have consequently moved into the triple intersection of the Venn diagram in Fig. 43.

Continuing in this fashion with each of the eight cells of the Venn diagram, we finally arrive at the terminal distribution shown in Fig. 43(b). For this we have the terminal Q-matrix:

$$Q_{ij}^{(2)} = \begin{pmatrix} .6 & .6 & .6 \\ .6 & .6 & .6 \\ .6 & .6 & .9 \end{pmatrix}$$

The stimuli S_1 and S_2 have become indistinguishable. The G-matrix for an α-system is the same as $Q_{ij}^{(2)}$, while for a γ-system, it would be

$$\begin{pmatrix} .24 & .24 & .06 \\ .24 & .24 & .06 \\ .06 & .06 & .09 \end{pmatrix}$$

The "coagulation" of S_1 and S_2 corresponds to the fact that in the training sequence (which is reflected in the f_{ij} matrix) S_1 and S_2 follow one another quite frequently, whereas they are very rarely followed by S_3. Consequently, S_3 tends to remain distinct, in the terminal G-matrix. In the following section, it will be seen that such behavior is quite characteristic of this system.*

---

* Another numerical example will be found in Section 17.2, where the four-layer system is compared with an open-loop cross-coupled model.

16.3 Organization of Dichotomies

The general analysis of the preceding section can be applied to a large variety of particular experimental designs. To begin with, we will show that with a suitable choice of parameters for the perceptron, and a suitable sequence of stimuli, a perceptron can spontaneously dichotomize an environment into any two classes, without any control of the reinforcement process by an external agency or experimenter. The organization of the stimulus sequence will determine the particular dichotomy which is formed.*

Let the sequence of stimuli to which the perceptron is exposed be $S_{j_0}, S_{j_1}, \ldots, S_{j_M}$. In the following discussion, such a sequence will be called a "preconditioning sequence". Let P_j denote the fraction of occurrences of S_j in the given sequence, and let $P_{j,k}$ denote the number of times S_k immediately follows S_j divided by the number of times S_j occurs. Then $f_{j,k} = \frac{M+1}{M} P_{j,k} P_j$. With a sufficiently long sequence, $(M+1)/M \approx 1$, and the equilibrium equation takes the form:

$$\gamma^{(i)} = \frac{N_a \eta}{\delta} \sum_{j=1}^{n} \sum_{k=1}^{n} Q_{ij}^{(1)} P_j P_{j,k} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) \tag{16.16}$$

where P_j corresponds to the probability of S_j, and $P_{j,k}$ corresponds to the transition probability $Prob.\left\{S_j \rightarrow S_k\right\} = Prob.\left\{S_j S_k \middle| S_j\right\}$.

---

* This can be interpreted as an R-controlled reinforcement system, although it does not actually depend on the outputs of the R-units in any essential way.

EXPERIMENT 10: Take an environment, W, consisting of n stimuli, such that there is no appreciable difference in the retinal overlap of different pairs of stimuli. (With a large retina, a set a random dot stimuli will generally satisfy this condition.) Divide the stimuli arbitrarily into two classes, so that $S_1, S_2, \ldots, S_K$ are in Class X, while $S_{K+1}, \ldots, S_n$ are in Class Y. All members of a given class are equally likely to occur. Let the probability of transition to a member of the same class be p, nearly unity, and to a member of the opposite class be $1-p$, nearly zero. Let the perceptron be exposed to an extended preconditioning sequence composed according to these probabilities, without any control by the r.c.s. At the end of the preconditioning sequence, the perceptron is exposed to a short additional sequence composed in the same manner, during which R-controlled reinforcement is administered, according to the rules of the γ-system, for A-unit to R-unit connections. The values of all connections are then "frozen", and the response of the perceptron to each stimulus in W is ascertained.

It can be seen that this experiment is closely analogous to Experiment 9, in which the effects of R-controlled reinforcement were determined for an environment of horizontal and vertical bars, except for the preconditioning sequence (which would have no effect at all in a simple perceptron), and the additional condition that there is no way of determining

whether two stimuli belong in the same or opposite classes on the basis of their retinal overlap. The only thing which characterizes two members of the same class differently from stimuli of opposite classes is the difference in transition probabilities in the preconditioning sequence.

We assume $Q_{ij}^{(1)} = (q + \Delta\sigma_{ij}) / N_a$, where $\Delta > 0$, $q \geq 0$. Thus the diagonal elements of the $Q_{ij}^{(1)}$ matrix are all $(q+\Delta)/N_a$ and all other elements are q/N_a. (Note that by raising thresholds of the $A^{(1)}$ units, with a sufficient number of connections, the ratio q/Δ can be made as small as desired.) For the probabilities of stimulus-occurrence indicated in the experiment, we have

$$p_j = \begin{cases} 1/2K & \text{for } S_j \text{ in } X \\ 1/2L & \text{for } S_j \text{ in } Y \end{cases}$$

where $L = n - K$

$$p_{jk} = \begin{cases} p/K & \text{for } S_j \text{ in } X,\ S_k \text{ in } X \\ (1-p)/L & \text{for } S_j \text{ in } X,\ S_k \text{ in } Y \\ p/L & \text{for } S_j \text{ in } Y,\ S_k \text{ in } Y \\ (1-p)/K & \text{for } S_j \text{ in } Y,\ S_k \text{ in } X \end{cases}$$

Then we obtain from (16.16),

$$\gamma^{(i)} = \frac{\eta}{\delta}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} + \sum_{j=1}^{K}\sum_{k=K+1}^{n} + \sum_{j=K+1}^{n}\sum_{k=1}^{K} + \sum_{j=K+1}^{n}\sum_{k=K+1}^{n}\right] \tag{16.17}$$

$$\left[(q + \Delta\sigma_{ij})\, P_j\, P_{jk}\, \phi\left(\beta^{(k)} + \gamma^{(k)}\right)\right]$$

$$= \frac{\eta}{\delta}\left[\frac{q}{2K}\sum_{k=1}^{K}\phi\left(\beta^{(k)} + \gamma^{(k)}\right) + \frac{q}{2L}\sum_{k=K+1}^{n}\phi\left(\beta^{(k)} + \gamma^{(k)}\right) + \Delta P_i \sum_{k=1}^{n} P_{ik}\,\phi\left(\beta^{(k)} + \gamma^{(k)}\right)\right].$$

Let us now assume that S_χ is one of the stimuli of class X. Then

$$\gamma^{(\chi)} = \frac{\eta}{\delta}\left[\frac{\Delta p + qK}{2K^2}\sum_{k=1}^{K}\phi\left(\beta^{(k)} + \gamma^{(k)}\right) + \frac{\Delta(1-p) + qK}{2KL}\sum_{k=K+1}^{n}\phi\left(\beta^{(k)} + \gamma^{(k)}\right)\right] \tag{16.18}$$

We now observe the following:

i) If $\frac{\eta(\Delta p + qK)}{2\delta K^2} \geq \theta$ then $A_\infty^{(2)}(S_\chi) \supseteq \bigcup_{S_j \in X} A_0^{(2)}(S_j)$.

In words, if the stated inequality holds then, in the terminal condition, each of the stimuli of class X activates the union of all sets which were initially activated by any of the stimuli of class X. That is, each stimulus of a given class has "captured" all of the A-units that initially responded to all of the other stimuli of that class. The proof follows from

the fact that any $A^{(2)}$ unit which originally responded to any of the stimuli in class X contributes a non-zero term in $\sum_{k=1}^{K}$ in (16.18). The postulated inequality then guarantees that the A-unit will be active in the terminal state.

ii) If $\frac{\eta[\Delta(1-p)+qK]}{2K\sigma} < \theta$, then $A_{\infty}^{(2)}(S_X) \subseteq \bigcup_{S_j \in X} A_0^{(2)}(S_j)$.

In words, if the stated inequality holds then, in the terminal condition, no stimulus of class X activates any A-unit outside of the union of sets initially activated by stimuli of class X. The proof follows from the fact that, if we were to solve (16.18) by iteration, then any A-unit which is activated by none of the X-stimuli has, on the first iteration, no contribution from $\sum_{k=1}^{K}$. In virtue of the assumed inequality it will not have any contribution on any following iteration either, and α remains less than θ. Since only a finite number of iterations are involved, this unit does not become active.

iii) If the inequalities of (i) and (ii) both hold, then $A_{\infty}^{(2)}(S_X) = \bigcup_{S_j \in X} A_0^{(2)}(S_j)$.

Necessary and sufficient conditions for both (i) and (ii) to hold have been found by H. D. Block. They are

a) $\Delta > qK(K-1)$

b) $p > \left[K\Delta + qK(K-1)\right] \Big/ \Delta(K+1)$

c) $K^2/(\Delta p + qK) \leq \eta/2\theta\sigma < K/(\Delta(1-p)+qK)$

Condition a) insures that a probability $p (0 < p < 1)$ can be chosen to satisfy b). Condition b) insures that $\eta / 2\theta\delta$ can be chosen to satisfy c). The conditions can be written in the alternative form

a') $p > K/(K+1)$

b') $\Delta > qK(K-1)/\left[p(K+1) - K\right]$

c) as above.

Under the conditions indicated, if Experiment 10 is completed by exposing the perceptron to a continuation of the same stimulus sequence with R-controlled $\mathcal{T}$ -reinforcement, the first response to occur will immediately generalize to all stimuli of the same class as the one which evoked the response, since each member of the class activates the identical set of A-units, after the preconditioning sequence. Suppose a member of class X is the first stimulus to occur, and that this happens to evoke the response $r^* = +1$. Then this response will be reinforced, and will generalize immediately to all other members of class X . However, under the conditions assumed above, the intersections between the sets of A-units initially responding to stimuli of class X and stimuli of class Y were all equal to q , and it was noted that by using large thresholds, q could be made arbitrarily small relative to the measure of the responding A-sets. If each A-unit has a large number of distinct origin points (no two identical) q can, in fact, be made small relative to the product $Q_i Q_j$. Thus, with a large threshold, in a $\mathcal{T}$ -system, the generalization coefficient g_{ij} for S_i in X and S_j in Y will be negative. Consequently, any

stimuli of class Y will automatically be assigned the opposite response from stimuli of class X. Thus a completely consistent dichotomy has been created, from the time the first stimulus of the terminal training sequence occurs. Further reinforcement will only strengthen the tendencies thus established.

If the ratio η/δ is made large enough, the perceptron in Experiment 10 will ultimately arrive at a state in which every stimulus activates all A-units which ever responded to any stimulus of either class. However, in practice, the constraints on the parameters need not be as severe as those indicated in conditions a), b), and c) above, in order to obtain useful generalization effects from the system. As long as η/δ is not so large as to cause a complete merging of all A-sets for all stimuli, it remains possible to teach the "preconditioned" perceptron to discriminate all stimuli of the two classes correctly with a single corrective reinforcement for one stimulus of each class, as long as the inequality $\frac{\eta(\Delta p + qK)}{2\delta K^2} \geq \theta$ is satisfied.

16.4 Organization of Multiple Classes

Suppose we have the same kind of environment as in Experiment 10, but that the stimuli are considered to be of, say, three classes:

$$A_1, A_2, \ldots, A_K, B_1, B_2, \ldots, B_L, C_1, C_2, \ldots, C_M \qquad (K+L+M=n).$$

We assume there is not too much overlap between the different types of stimuli, an assumption which will be made more precise below. (as in the previous case, the overlap can always be reduced as far as required

by making θ sufficiently high.) The three classes will be called X, Y, and Z. We assume that the $Q_{ij}^{(1)}$ matrix is

$$Q_{ij}^{(1)} = \begin{cases} q/N_a & \text{if } S_i \text{ and } S_j \text{ are in different classes} \\ (q+r)/N_a & \text{if } S_i \text{ and } S_j \text{ are in the same class, } S_i \neq S_j \\ (q+r+s)/N_a & \text{if } S_i = S_j . \end{cases}$$

From the nature of a Q_{ij} matrix it is necessary that $q \geq 0$, $(q+r) \geq 0$, and $(r+s) \geq 0$. We assume $s \geq 0$.

Suppose that the transition probabilities are large (p) for transitions to a member of the same class, and small $(1-p)/2$ to each of the other classes. Within a class each transition is equally likely. Then

$$P_{ij} = \begin{cases} p/K & S_i \text{ in } X,\ S_j \text{ in } X \\ p/L & S_i \text{ in } Y,\ S_j \text{ in } Y \\ p/M & S_i \text{ in } Z,\ S_j \text{ in } Z \\ (1-p)/2L & S_i \text{ in } X,\ S_j \text{ in } Y;\ \text{or } S_i \text{ in } Z,\ S_j \text{ in } Y \\ (1-p)/2M & S_i \text{ in } X,\ S_j \text{ in } Z;\ \text{or } S_i \text{ in } Y,\ S_j \text{ in } Z \\ (1-p)/2K & S_i \text{ in } Y,\ S_j \text{ in } X;\ \text{or } S_i \text{ in } Z,\ S_j \text{ in } X \end{cases}$$

The probabilities of occurrence of individual stimuli are given by

$$P_j = \begin{cases} 1/3K & S_j \text{ in } X \\ 1/3L & S_j \text{ in } Y \\ 1/3M & S_j \text{ in } Z \end{cases}$$

Then Equation (16.16) becomes

$$\gamma^{(i)} = \frac{N_0 \eta}{\delta} \left[\sum_{j\epsilon X} \sum_{k\epsilon X} + \sum_{j\epsilon X} \sum_{k\epsilon Y} + \sum_{j\epsilon X} \sum_{k\epsilon Z} + \sum_{j\epsilon Y} \sum_{k\epsilon X} + \sum_{j\epsilon Y} \sum_{k\epsilon Y} \right. \tag{16.19}$$

$$\left. + \sum_{j\epsilon Y} \sum_{k\epsilon Z} + \sum_{j\epsilon Z} \sum_{k\epsilon X} + \sum_{j\epsilon Z} \sum_{k\epsilon Y} + \sum_{j\epsilon Z} \sum_{k\epsilon Z} \right] \left[Q_{ij}^{(1)} P_j P_{jk} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) \right]$$

where, for simplicity of notation, X , Y , and Z have been used for the appropriate index sets. Suppose x is in X (i.e., S_x is in X). Then (16.19) yields

$$\gamma^{(i)} = \frac{\eta}{\delta} \left\{ \frac{p(Kr+s)+qK}{3K^2} \sum_{k\epsilon X} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) \right. \tag{16.20}$$

$$\left. + \frac{(1-p)(Kr+s)+2qK}{6K} \left[\frac{1}{L} \sum_{k\epsilon Y} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) + \frac{1}{M} \sum_{k\epsilon Z} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) \right] \right\}$$

We can now assert the following:

i) If $\frac{\eta\left[p(Kr+s)+qK\right]}{3\delta K^2} \geq \theta$, then the set $A_\infty^{(2)}(S_x) \supseteq \underset{S_j \epsilon X}{U} A_0^{(2)}(S_j)$

That is, if the stated inequality holds, then every $A^{(2)}$ unit which initially responded to any stimulus in class X now responds to each stimulus in class X. This is readily proven by noting that for any $A^{(2)}$ unit which initially responds to any of the stimuli in class X there is at least one non-zero ϕ in $\sum_{A \epsilon X}$ in (16.20). The postulated inequality then guarantees that $\gamma^{(i)} \geq \theta$ for any x such that S_x is in X.

ii) If $\eta\left[\frac{(1-p)(Kr+\Delta)+2qK}{6K\theta}\right] < \theta$, then $A_\infty^{(2)}(S_x) \subseteq \bigcup_{S_j \epsilon X} A_0^{(2)}(S_j)$

That is, if the stated inequality holds, then every $A^{(2)}$ unit which did not initially respond to at least one of the stimuli of class X does not respond to any class X stimulus in the terminal state. This is proven as follows. For an A-unit which does not respond to any stimulus of class X, none of the terms in $\sum_{A \epsilon X}$ in (16.18) are present on the first iteration, which starts with $\gamma^{(j)} = 0$. The stated inequality guarantees that, even if all the other terms are present, no $\gamma^{(j)}$ for S_j will reach θ. Thus no terms in $\sum_{A \epsilon X}$ will ever be non-zero.

iii) If both of the above inequalities hold, then $A_\infty^{(2)}(S_x) = \bigcup_{S_j \epsilon X} A_0^{(2)}(S_j)$

That is, each stimulus in class X activates exactly the same set of $A^{(2)}$ units in the terminal state; and that set consists of just those A-units which originally were activated by any one of the stimuli of class X.

Necessary and sufficient conditions that the inequalities of both i) and ii) be satisfied have again been derived by Block, and are

a) $r > -\Delta/K$

b) $q < (Kr + \Delta)/K(K-1)$

c) $p > \left[2qK(K-1) + K(Kr+\Delta)\right] / (Kr + \Delta)(K+2)$

d) $3K^2/\left[p(Kr + \Delta) + qK\right] \leq \eta/\theta\sigma < 6K/\left[(1-p)(Kr + \Delta) + 2qK\right]$

Condition a) guarantees that a suitable $q \geq 0$ can be chosen in b); Condition b) guarantees that a suitable $p < 1$ can be chosen in c); Condition c) guarantees that an η/σ can be chosen to satisfy d).

If the parameters are suitably set we have seen that the response in the $A^{(2)}$ layer to any stimulus in class X is $\bigcup_{S_j \in X} A_0^{(2)}(S_j)$. Similarly for classes Y and Z . This means that a γ -system perceptron with a single R-unit will tend to assign the same response to all members of the first class of stimuli to be represented in the training sequence. All other stimuli will receive the opposite response, if the initial intersections of responding A-sets are small enough. With more than one R-unit and inhibitory connections between the R-units, so that only one can go on at one time (c.f., Chapter 20) it is thus possible for the perceptron to assign a unique response to each stimulus class. If there is too much initial overlap between the responding sets of A-units, or if condition i) is satisfied without condition ii) being satisfied, a single corrective reinforcement applied for any one stimulus of each class may still be sufficient to yield the correct response for all stimuli in the environment.

16.5 Similarity Generalization

In the experiments considered above, the nature of the stimulus classes was never explicitly stated. Clearly, they could have been similarity classes, under a suitably chosen similarity relation, and the same results would have been obtained. In order to obtain generalization over the entire class, however, it was assumed that "runs" of stimuli from each class occurred, it being much more likely that a stimulus was followed by another member of the same class than by a stimulus from a different class. After a long preconditioning sequence of this type, it might be expected that the perceptron would have seen <u>each</u> stimulus in the environment a great number of times. We now consider the generalization of a similarity relation to stimuli which have <u>not</u> occurred during the preconditioning sequence.*

EXPERIMENT 11: Consider an environment of stimuli $S_1, \ldots, S_2, \ldots, S_n$ and their transforms $T(S_1), T(S_2), \ldots, T(S_n)$ where T is any transformation in which the measure of fixed points is zero. Let the perceptron be exposed to a preconditioning sequence, consisting of stimuli followed by their transforms, i.e., a sequence of the form $\{S_{k_1}, T(S_{k_1}), S_{k_2}, T(S_{k_2}), \ldots, S_{k_M}, T(S_{k_M})\}$ where the subscripts $k_1, k_2, \ldots$ are picked at random from the set of integers 1 through n. Now consider a pair of test stimuli, S_x and S_y, and their transforms $T(S_x)$ and $T(S_y)$, none of which occured during the preconditioning sequence. Let one response be associated to S_x and the opposite response to S_y, by means of an error correction procedure. Now test the perceptron to determine its response to $T(S_x)$ and $T(S_y)$.

---

* This is directly analogous to the phenomenon of similarity generalization originally predicted for cross-coupled systems in Rosenblatt, Ref. 85.

It is predicted that if this experiment is performed with random dot stimuli in the preconditioning sequence, with a finite retina, and S_x and S_y are any other stimuli (e.g., a square and a triangle, or two letters of the alphabet) the transforms $T(S_x)$ and $T(S_y)$ will each tend to activate the appropriate response, which was associated to S_x and S_y, respectively. In other words, the perceptron will have learned that <u>any</u> two stimuli which are similar under the transformation T are to be treated as equivalent, even though the stimuli have never been seen before.

To begin with, we consider the following problem, which is essentially a special case of Experiment 11, performed with only a single test stimulus.

Consider the stimuli S_1, S_2, ..., S_K and their transforms $S_{K+1} = T(S_1)$, $S_{K+2} = T(S_2)$, ..., $S_{2K} = T(S_K)$. For example, $S_1, \ldots, S_K$ may be in the left half of the field, and T a transformation which moves them to the right half of the field. S_x $(x = 2K+1)$ is not shown during the preconditioning sequence, but is a test stimulus to be applied later. $S_{x'} = T(S_x)$, $x' = 2K + 2 = n$. Let us assume S_x intersects $S_1, \ldots, S_L$ $(L \leq K)$ to a larger extent than it does the others and hence $S_{x'}$ intersects mainly the stimuli $S_{K+1}, \ldots, S_{K+L}$. These relationships are illustrated in Figure 44.

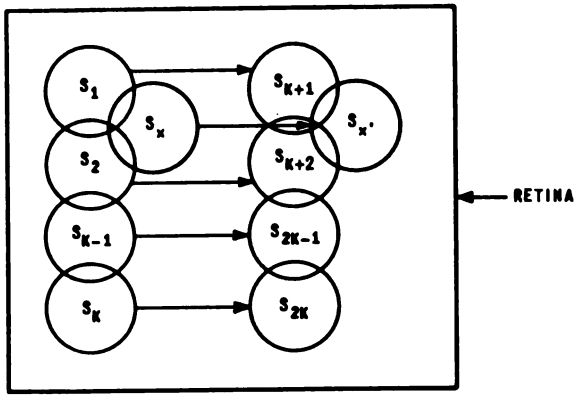


Figure 44 RELATIONSHIP OF TEST STIMULUS TO PRECONDITIONING STIMULI AND TRANSFORMS

Specifically, consider the conditions

$$Q_{xj}^{(1)} = \begin{cases} (q + s\sigma_{xj})/N_a & j > L \\ (q + r)/N_a & j \leq L \end{cases}$$

$$Q_{x'j}^{(1)} = \begin{cases} q/N_a & j \leq K \\ (q + r)/N_a & K+1 \leq j \leq K+L \\ (q + s\sigma_{x'j})/N_a & j > K+L \end{cases}$$

In the preconditioning sequence, a stimulus S_i is picked at random from $S_1, \ldots, S_K$, and this is followed by its transform, $T(S_i)$. Then another stimulus is picked at random from $S_1, \ldots, S_K$ and this is followed by its transform, and so on. Then

$$p_j = \begin{cases} 1/2K & j \leq 2K \\ 0 & j > 2K \end{cases}$$

$$p_{jk} = \begin{cases} 1 & j \leq K, \ k = K+j \\ 1/K & j > K, \ k \leq K \\ 0 & \text{otherwise} \end{cases}$$

We also specify that no A-unit is activated by more than μ ($\mu < K/L$) of the stimuli $S_1, S_2, \ldots, S_K$.

From Equation (16.16) we obtain

$$\gamma^{(x)} = \frac{N_a \eta}{2K\delta} \left[\sum_{j=1}^{K} Q_{x_j}^{(1)} \phi\left(\beta^{(K+j)} + \gamma^{(K+j)}\right) + \frac{1}{K} \sum_{j=K+1}^{2K} \sum_{k=1}^{K} Q_{x_j}^{(1)} \phi\left(\beta^{(k)} + \gamma^{(k)}\right) \right] \quad (16.21)$$

$$\gamma^{(x)} = \frac{\eta}{2K\delta} \left[(q+r) \sum_{j=1}^{L} \phi\left(\beta^{(K+j)} + \gamma^{(K+j)}\right) + q \sum_{j=L+1}^{K} \phi\left(\beta^{K+j} + \gamma^{K+j}\right) + q \sum_{j=1}^{K} \phi\left(\beta^{(j)} + \gamma^{(j)}\right) \right] \quad (16.22)$$

Hence we have the following results:

i) If $\eta(q+r)/2K\delta \geq \theta$, then $A_\infty^{(2)}(S_x) \supseteq A_0^{(2)}(S_x) + \bigcup_{j \leq L} A_0^{(2)}(T(S_j))$.

In words, if the stated inequality holds, then, in the terminal state, S_x activates all those elements originally activated either by itself or by any of the transforms $T(S_1), \ldots, T(S_L)$.

ii) If $\frac{\eta q}{2K\delta}(K+\mu-L) < \theta$, then $A_\infty^{(2)}(S_x) \subseteq A_0^{(2)}(S_x) + \bigcup_{j \leq L} A_0^{(2)}(T(S_j))$.

iii) If both inequalities hold, then $A_\infty^{(2)}(S_x) = A_0^{(2)}(S_x) + \bigcup_{j \leq L} A_0^{(2)}(T(S_j))$.

Thus far, we have considered the generalization of a response from S_χ to the transforms $T(S_1)$, $T(S_2)$, etc. Suppose a response is associated to $T(S_\chi)$; we are then interested in determining whether there is any generalization in the reverse direction, i.e., to $T^{-1}(S_{\chi'})$. We can obtain $\gamma^{(\chi')}$ from Equation (16.21), with χ replaced by χ', which yields:

$$\gamma^{(\chi')} = \frac{\eta}{2K\sigma} \sum_{j=1}^{K} \left[q\,\phi\left(\beta^{(K+j)} + \gamma^{(K+j)}\right) + \frac{Lr + Kq}{K}\,\phi\left(\beta^{(j)} + \gamma^{(j)}\right)\right]$$

Consequently,

iv) If $\frac{\eta}{2K\sigma}\left[q(K+\mu) + \frac{Lr\mu}{K}\right] < \theta$, then $A_\infty^{(2)}(S_{\chi'}) = A_0^{(2)}(S_{\chi'})$.

If inequalities i), ii), and iv) all hold, then the stimulus S_χ generalizes to $T(S_1) \ldots T(S_L)$, but the transform $T(S_\chi) = S_{\chi'}$ does not generalize to the stimulus S_χ. Necessary and sufficient conditions that all three inequalities hold are easily found: (With $r \geq 0$, then iv) implies ii)).

a) $r > \frac{qK(K+\mu-1)}{K-L\mu}$

b) $\frac{2K}{q+r} \leq \frac{\eta}{\theta\sigma} < \frac{2K^2}{K(K+\mu)q + Lr\mu}$

In particular, let $L = 1$. Then $A_\infty^{(2)}(S_\chi) = A_0^{(2)}(S_\chi) + A_0^{(2)}(T(S_1))$. Thus, due to the intersection between $A_0^{(2)}(S_{\chi'})$ and $A_0^{(2)}(T(S_1))$, the test stimulus generalizes to its transform, even though neither the test stimulus nor its transform has occurred during the preconditioning sequence. Under these conditions, the perceptron will behave in much the same manner as the specially constrained similarity-biased perceptron of Chapter 15. The actual magnitude of the bias thus induced, in a simple discrimination experiment, can be calculated as follows.

Let S_z be another test stimulus, like S_x, but its chief intersection is with S_2, say also $q + r$. Then if conditions a) and b) are satisfied, (with $L = 1$), $A_\infty^{(2)}(S_z) = A_0^{(2)}(S_z) + A_0^{(2)}(T(S_z))$ and $A_\infty^{(2)}(T(S_z)) = A_0^{(2)}(T(S_z))$. Suppose the perceptron has zero initial values on the $A^{(2)}$ to R-unit connections. Let S_x be shown, and all active A-R connections reinforced by $+1$. Then let S_z be shown, and all active A-R connections reinforced by -1. Now if the perceptron is shown $T(S_x)$ (which it has never seen before) the input to the R-unit is equal to the number of A-units in $A_\infty^{(2)}(T(S_x)) \cap \left[A_\infty^{(2)}(T(S_1)) \cup A_\infty^{(2)}(S_x)\right]$ minus the number of A-units in $A_\infty^{(2)}(T(S_x)) \cap \left[A_\infty^{(2)}(T(S_2)) \cup A_\infty^{(2)}(S_z)\right]$, which in general is positive; while if it is shown $T(S_z)$ the signal to the R-unit is negative. Thus the discrimination which was taught for S_x and S_z carries over to $T(S_x)$ and $T(S_z)$.

In the above analysis, it was postulated that the test stimuli should have larger intersections with some of the preconditioning stimuli than with others. This assumption is crucial for the predicted effect to occur. The reader will recall from the discussion of the last chapter, that in a perceptron with an infinite retina, no similarity bias could be obtained between random stimuli because the distribution of their intersections had zero variance. The same situation holds here. If the preconditioning stimuli are random dot patterns, and the retina is infinite, then every preconditioning stimulus will have exactly the same intersection with the test stimulus S_x, and the required bias cannot occur. In a finite retina, however, the intersections will be binomially distributed (as in the analysis of Chapter 15), and the predicted effect will be obtained.

We also note an advantage, as before, if compact, coherent stimuli are employed for preconditioning and as test stimuli. In this case, even in an infinite retina, the distribution of intersections will have non-zero variance, and the test stimulus will tend to be more closely related to some preconditioning stimuli than to others. As long as two test stimuli, S_x and S_z, do not intersect the same sets of preconditioning stimuli to the same degree, they can be discriminated in the terminal state of the system (provided the required parametric conditions are satisfied), but each will generalize to its transform. Thus the claim made for the performance of such a system in Experiment 11 has been verified in principle. Quantitative studies of actual cases are not yet complete, but similar experiments with cross-coupled systems (to be presented in Chapter 19) suggest that highly satisfactory results can, in fact, be obtained in practice.

The asymmetrical generalization from S to $T(S)$, but not from $T(S)$ to S can, of course, be overcome by employing a symmetrical preconditioning sequence, in which a stimulus is as likely to be followed by the inverse transformation, $T^{-1}(S)$ as by $T(S)$.

For instance, take $\{S_1, \ldots, S_n\} = \{S_1, \ldots, S_K ; S_{K+1}, \ldots, S_n\} = \{S_1, \ldots, S_K ; T(S_1), \ldots, T(S_K)\}$ where $K = n/2$. Let

$$Q_{ij}^{(1)} = (q + \Delta\sigma_{ij})/N_a$$

$$P_j = 1/2K$$

$$P_{j\ell} = \begin{cases} p & j \le K, \quad \ell = K + j \\ p & j > K, \quad \ell = j - K \\ (1-p)/(2K-1) & j \le K, \quad \ell \ne K + j \\ (1-p)/(2K-1) & j > K, \quad \ell \ne j - K \end{cases}$$

Let $\omega = p - (1-p)/(2K-1)$; then the $P_{j\ell}$ can be expressed as follows. For $1 \le j \le K$, $1 \le \ell \le K$, we have

$$P_{j\ell} = P_{j+K,\,\ell+K} = r$$

$$P_{j,\,K+\ell} = P_{j+K,\,\ell} = r + \omega\delta_{j\ell}$$

where $r = (1-\omega)/2K = (1-p)/(2K-1)$. This means that the transition probability from a stimulus to its transform, or vice versa, is $r + \omega$, while for any two unrelated stimuli, the transition probability is r.

Then from (16.16) we have

$$\gamma^{(i)} = \frac{N_a \eta}{2K\sigma}\left[\sum_{j=1}^{K}\sum_{\ell=1}^{K} + \sum_{j=1}^{K}\sum_{\ell=K+1}^{2K} + \sum_{j=K+1}^{2K}\sum_{\ell=1}^{K} + \sum_{j=K+1}^{2K}\sum_{\ell=K+1}^{2K}\right] \cdot Q_{ij}^{(1)} P_{j\ell}\, \phi\left(\beta^{(\ell)} + \gamma^{(\ell)}\right)$$

$$= \frac{N_a \eta}{2K\sigma}\left[\sum_{j=1}^{K}\sum_{\ell=1}^{K} Q_{ij}^{(1)} P_{j\ell}\, \phi(\alpha^{(\ell)}) + \sum_{j=1}^{K}\sum_{\ell=1}^{K} Q_{ij}^{(1)} P_{j,\,\ell+K}\, \phi(\alpha^{(\ell+K)})\right.$$

$$\left. + \sum_{j=1}^{K}\sum_{\ell=1}^{K} Q_{i,\,j+\ell}^{(1)} P_{j+\ell,\,\ell}\, \phi(\alpha^{(\ell)}) + \sum_{j=1}^{K}\sum_{\ell=1}^{K} Q_{i,\,j+K}^{(1)} P_{j+K,\,\ell+K}\, \phi(\alpha^{(\ell+K)})\right]$$

Assuming $S_i \epsilon \{ S_1, \ldots, S_K \}$ we have

$$\gamma^{(i)} = \frac{\eta}{2K\sigma} \sum_{j=1}^{K} \sum_{k=1}^{K} \Big[(q + s\sigma_{i,j})(r\phi(\alpha^{(k)})) + (r + w\sigma_{j,k})\,\phi(\alpha^{(k+K)}) + q\left((r + w\sigma_{j,k})\,\phi(\alpha^{(k)}) + r\,\phi(\alpha^{(k+K)})\right)\Big]$$

$$\gamma^{(i)} = \frac{\eta}{2K\sigma} \sum_{j=1}^{K} \sum_{k=1}^{K} \Big\{ 2qr\left[\phi(\alpha^{(k)}) + \phi(\alpha^{(k+K)})\right] + qw\,\sigma_{j,k}\left(\phi(\alpha^{(k+K)}) + \phi(\alpha^{(k)})\right) + s\sigma_{i,j}\left(r\,\phi(\alpha^{(k)}) + (r + w\sigma_{j,k})\,\phi(\alpha^{(k+K)})\right)\Big\}$$

$$\gamma^{(i)} = \frac{\eta}{2K\sigma}(2Kqr + qw + sr) \sum_{k=1}^{K} \left[\phi(\alpha^{(k)}) + \phi(\alpha^{(k+K)})\right] + \frac{\eta s w}{2K\sigma}\,\phi(\alpha^{(K+i)})$$

Thus if p (or w) is nearly 1 and s/q is large, S_i will generalize to its transform, and conversely $T(S_i)$ will generalize to S_i, since

$$\gamma^{(K+i)} = \frac{\eta}{2K\sigma}(2Kqr + qw + sr) \sum_{k=1}^{K} \left[\phi(\alpha^{(k)}) + \phi(\alpha^{(k+K)})\right] + \frac{\eta s w}{2K\sigma}\,\phi(\alpha^{(i)})$$

To get the specific form of the conditions for such generalization to occur, we extract the term for $k = 1$ in $\sum_{k=1}^{K}$ and put it with the second term. This gives the first required inequality,

$$(\eta/2K\sigma)(2Kqr + qw + sr + sw) \geq \Theta$$

or, replacing r and w in terms of p, and $2K$ by n, we get the condition

i) If $\eta(q+\Delta p)/\delta n \geq \theta$ then $A_\infty^{(2)}(S_i) \supseteq A_0^{(2)}(S_i) + A_0^{(2)}(T(S_i))$.

The second required inequality turns out to be

$$(\eta(K-1)/2K\delta)(2Kqr + qw + \Delta r) < \theta$$

or, replacing r and w in terms of p, we get

ii) If $\eta(n-2)\left[q(n-1)+\Delta(1-p)\right]/2n(n-1)\delta < \theta$, then $A_\infty^{(2)}(S_i) \subseteq A_0^{(2)}(S_i) + A_0^{(2)}(T(S_i))$.

iii) If both inequalities hold, then $A_\infty^{(2)}(S_i) = A_0^{(2)}(S_i) + A_0^{(2)}(T(S_i))$.

Necessary and sufficient conditions that both inequalities hold, given $n > 4$, are

a) $p > (n-2)/(3n-4)$

b) $q < \Delta\left[p(3n-4) - n + 2\right]/(n-1)(n-4)$

c) η/δ must be so chosen as to satisfy i) and ii).

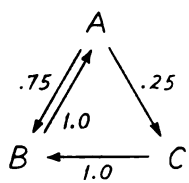
For $n = 4$, these conditions are satisfied if $p > 1/4$ and

$$4/(q+\Delta p) = \frac{\eta}{\theta\delta} < 12/\left[3q + \Delta(1-p)\right]$$

16.6 Analysis of Value-Conserving Models

In dealing with simple perceptrons, a single value-conserving model, the γ -system, has been considered. In this system, the total value of the set of input connections to an A-unit is conserved. In four-layer and cross-coupled perceptrons two types of value-conserving systems are of interest: the γ -system, defined as beofe, (where the sum of the input values is held constant) and the Γ -system, where value is conserved over the set of output connections from an A-unit, rather than the inputs. In the perceptrons to be considered in the following chapters, this second system appears to offer important advantages in performance, and will generally be preferred over the γ -system.

The most important difference between the γ-system and the Γ -system is that the latter tends to activate those A-units which would respond to the most probable successor of the present stimulus, whereas the γ -system tends to activate the set of A-units which respond to the stimulus for which the present stimulus is the most probable predecessor. The difference between these two situations can be seen from the following example. Suppose there are three stimuli, A, B, and C, with transition probabilities as shown in the following diagram:



In this case, with the γ-system, we would expect the set of A-units responding to stimulus A to become most closely associated to the set responding to stimulus C, since A is the only possible predecessor of C, whereas B can be preceded by either A or C. In a Γ-system, on the other hand, the set responding to A would be most closely coupled to the set responding to B, and might even develop inhibitory connections to the set responding to C, since B is the most common successor of A. Thus the Γ-system tends to be predictive, tending to anticipate the most likely successor of the present stimulus, whereas the γ-system tends to anticipate the stimulus which is most likely to be preceded by the present stimulus. As shown above, this latter choice is not necessarily a good prediction of the next event.

16.6.1 Analysis of γ-systems

The differential equation for the γ-system is identical with (16.11), except that the constants C_{ij} are now equal to

$$C_{ij} = \sum_{k=1}^{n} \left(Q_{ik}^{(1)} - Q_i^{(1)} Q_k^{(1)} \right) f_{kj}$$

The negative term, $-Q_i Q_k$, is familiar from previous analyses of the γ-system, and represents the quantity substracted to balance the gain in value of the active connections. It will be recalled that for a Poisson model, $Q_{ik} - Q_i Q_k$ is always equal to or greater than zero, so that the expected value of C_{ij} will remain positive, and the previous analysis (Section 16.2.1) applies without modification. More generally, however, and for a binomial model in particular, the C_{ij} may be negative, and the previous analysis must be reexamined to see how this affects the situation.

To begin with, it no longer follows that the solution will be monotone, since different combinations of positive and negative C_{ij}'s may be picked up in equation (16.11), depending on which ϕ's are currently non-zero. Since the solution is non-monotone, it also does not follow that a solution will occur in n steps, or that the solution of the iteration equation (16.13) is minimal.

While we are unable, at this time, to provide any short-cut method of finding the steady state solution (if one exists) for the γ-system, it is possible to compute a time-dependent solution by the following procedure. We note, first, that the solution is piecewise exponential, as in the case of the α-system, and that the time constants for all $\gamma^{(j)}$ are equal. This means that we can readily determine which $\alpha^{(j)}$ will be the first to cross the level of θ, by computing the initial asymptotes, $M_0^{(j)}$ for all j. The $\gamma^{(j)}$ with the highest value of $|M_0^{(j)}|$ will change most rapidly. If the initial value of $\alpha^{(j)} = \theta$, and $M_0^{(j)}$ is negative, $\phi(\alpha^{(j)})$ will immediately go to 0. If no M is negative, then the first change to occur will be for some ϕ to change from 0 to 1, and this will occur for that j for which $M_0^{(j)}$ is greatest. Having thus obtained the first discontinuity point, t_1, we can compute the values of all $\gamma^{(i)}(t_1)$, and determine the next ϕ to change.* This is done by computing the function

$$\xi_k^i = \frac{\frac{M_k^{(i)}}{\sigma} - \gamma^{(i)}(t_k)}{(\theta - \beta^{(i)}) - \gamma^{(i)}(t_k)} \qquad (16.23)$$

---

* Joseph has pointed out that singularities are possible. For example, with $\theta = 1$, $\sigma = 1$, $\beta_1 = 1$, and $\beta_2 = 0$, if $C = \begin{pmatrix} 1 & 1 \\ 3 & -3 \end{pmatrix}$ we have (at $t = \ln 3/2$) $\gamma_1 = 1/3$, $\gamma_2 = 1$. But then $\dot{\gamma}_1 = 2 - \gamma_1$ while $\dot{\gamma}_2 = -\gamma_2$. Thus γ_2 immediately falls below 1, hence back to the original equation, which brings it back to 1 again. While γ_2 thus fluctuates about 1, the future history of γ_1 is not determined.

for all i. Note that ξ_i will be greater than 1 only if the numerator and denominator agree in sign, and $(M/\sigma - \gamma) > |\theta - \beta - \gamma|$. If these conditions are met (i.e., if $\xi_k^i > 1$), $\phi(\alpha^{(i)})$ will change value sometime before $\gamma^{(i)}$ reaches its new asymptote. Thus, by finding the value (or values) of i for which ξ_k^i is maximum, at the discontinuity time t_k, we can always determine the next ϕ to change. Introducing this new ϕ gives us a new set of asymptotes, $M_{k+1}(\gamma^{(i)})$, and the process can be continued. The values of the $\gamma^{(i)}(t)$ at the discontinuity times can be readily calculated from the exponential solution:

$$\gamma^i(t_{k+1}) = \frac{M_k}{\sigma} - e^{-\sigma(t_{k+1} - t_k)}\left(\frac{M_k^{(i)}}{\sigma} - \gamma^i(t_k)\right) \qquad (16.24)$$

where the discontinuity time, t_{k+1}, is obtained by solving the equation for the next $\gamma^{(i)}$ to cross threshold, that is

$$(t_{k+1} - t_k) = -\frac{1}{\sigma}\log\left(\frac{\beta - \theta + \frac{M_k^{(i)}}{\sigma}}{\frac{M_k^{(i)}}{\sigma} - \gamma^{(i)}(t_k)}\right) . \qquad (16.25)$$

16.6.2 Analysis of Γ -systems

The Γ -system is similar to the γ -system, except that after each increment of reinforcement, the total value is restored to its former level by subtracting the net gain uniformly from the set of output

connections from an A-unit, instead of the input connections. The differential equation now takes the form

$$\frac{d\gamma^{(i)}}{dt} = N_a \eta \sum_{j=1}^{n} \sum_{k=1}^{n} \left[\phi(\alpha^{(k)}) - Q_k^{(2)}\right] Q_{ij}^{(1)} f_{jk} - \sigma \gamma^{(i)} \qquad (16.26)$$

The same uncertainties as to existence of steady state solutions and difficulties of computation occur here as in the case of the γ-system analysis. A time-dependent solution can again be computed, piecewise, by the same procedure as above. In chapter 19, we shall reconsider the Γ-system, in connection with cross-coupled perceptrons.

16.7 Functionally Equivalent Models

In Ref. 41, Joseph has presented an analysis of a perceptron with "binodal A-units", which is now seen to be functionally equivalent to a variation of the system analyzed above. In the binodal model, there is only a single layer of A-units, but each A-unit receives two logically distinct sets of input connections and has a separate threshold for each set. The first set of connections is fixed in value, and activates the A-unit according to the usual rules. The second set consists of a single connection from every sensory point in the retina, and is variable in value. The reinforcement rule for these variable connections is that if the A-unit is active at time t, and the retinal origin point of one of the variable connections is active at $t+1$, the variable connection gains an increment in value. At the same time, all variable connections tend to decay at a fixed rate, σ. This is equivalent to a four-layer model in which each $A^{(2)}$ unit receives its fixed connection from an $A^{(1)}$ unit with a normal number of input connections and threshold θ, and receives variable connections from N_a other $A^{(1)}$ units, each having a single excitatory input connection, and a threshold of 1. The main difference

from the above analysis would then be that the $A^{(2)}$ unit responds to the logical sum, rather than the algebraic sum, of the inputs from the fixed connections and the variable connections, i.e., the $A^{(2)}$ unit is active if its fixed connection (the $\mathcal{B}$-component) is active, or if the sum of the variable connections (the $\mathcal{V}$-component) $\geq \theta$. As this writer had previously predicted on heuristic grounds, Joseph has successfully demonstrated that similarity generalization will tend to occur in the binodal model, after a preconditioning sequence analogous to those discussed above. In this system, the set of fixed connections acts as a "template", and the variable connections tend to adapt themselves to an origin configuration which resembles the fixed set under the transformation T. The reader is referred to Reference 41 for a quantitative analysis.

While it was assumed that the models analyzed in the preceding sections had a complete set of connections (from every $A^{(1)}$ unit to every $A^{(2)}$ unit), a system which merely has a large number of input connections to each $A^{(2)}$ unit, originating from randomly selected $A^{(1)}$ units, can be seen to be equivalent in all of its essential properties. For such a system, the $Q_{ij}^{(2)}$ matrix, representing the expected values of the fractions of $A^{(2)}$ units responding to S_i and S_j, would have the same equations as before, except that N_a must be replaced by the number of variable connections to each $A^{(2)}$ unit.

In the following chapter, it will be shown that a form of weakly cross-coupled system, in which there are no closed loops, is also virtually equivalent to the model analyzed in this chapter, and can be represented by the same equations, with a slight reinterpretation of the $\mathcal{B}$-component of the input signals to the A-units.

17. OPEN-LOOP CROSS-COUPLED SYSTEMS

The most interesting features of cross-coupled perceptrons are those which result from the possibility of closed feed-back loops, or cycles, in the network. It is possible, however, to design a cross-coupled system with no closed loops, and such a system has a number of important features, including the ability to act as an adaptive similarity-generalizing system equivalent to the perceptrons of Chapter 16, and increased economy and versatility in general classification problems of the sort considered in Chapter 5. These properties will be considered briefly, in this chapter, before proceeding to closed-loop systems, which represent a more challenging problem in analysis.

17.1 Similarity-Generalizing Systems: An Analog of the Four-Layer System

The three-layer perceptron shown in Fig. 45 is directly comparable to the four-layer system considered in the last chapter. The A-units are divided into two subsets, called A' and A". All A-units receive fixed connections from the retina, but only the A" units have connections to the R-units, the A'units sending their output signals to the A" units. Each A'unit is connected (in a fully-coupled model) to all A" units, and each A" unit is connected to all A' units. The rule for modifying the connections from A' to A" units is identical with the rule for modifying $A^{(1)}$ to $A^{(2)}$ connections, in the four-layer system considered previously: If the origin of the connection is active at time t, and the terminus is active at t+1, the connection gains a quantity η . All inter-A-unit connections decay at a rate δ , as before.

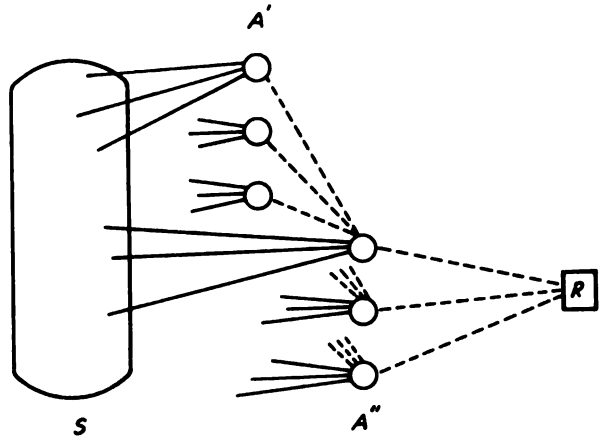


Figure 45 OPEN-LOOP CROSS-COUPLED SYSTEM (COMPARE Figure 42). BROKEN LINES INDICATE VARIABLE CONNECTIONS.

Clearly, the only differerence between this model and the previous one is that the β -component, instead of originating from one of the $A^{(1)}$ units, comes direct from the retina, and consequently can take on more than two values. The differential equation (16.11) and the equilibrium equation (16.12) thus apply without modification to this system (where the A' set is equated with the $A^{(1)}$ set, and the A'' set with the $A^{(2)}$ set). The additional freedom in choice of β -values means that the sets designated $A_\beta^{(2)}(S_i)$, representing sets of units whose β -value in response to S_i is $+1$, must now be fractionated into subsets for each possible value of β , and the history of each such subset (having a given β -vector) must be followed separately. Thus the full designation

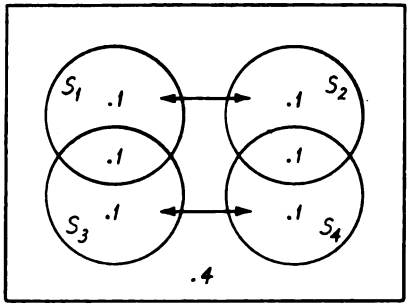
of such a subset would be $A_0^{(2)}(\beta_i, S_i)$. Apart from this further fractionation of the A-set, the same analysis holds as in the last chapter, and much the same results would be expected.

17.2 Comparison of Four-Layer and Open-Loop Cross-Coupled Models

A numerical comparison of the performance of the perceptrons considered in this and the preceding chapter will be based on the following experiment:

EXPERIMENT 12: Take an environment of four stimuli, $S_1 \ldots S_4$, each having retinal area $R = .2$. The intersections C_{13} and C_{24} are each equal to $.1$, and all other intersections are zero. The perceptron is exposed to the following sequence, which is repeated until a steady state is attained: $(S_1 S_2 S_1 S_2 S_1 S_2 S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4 S_3 S_4 S_3 S_4 S_3 S_4)$. This sequence can be considered to consist of two events, the first consisting of the alternating pair $S_1 S_2 S_1 S_2 \ldots$ with a duration of 10τ , and the second consisting of $S_3 S_4 S_3 S_4 \ldots$, also with duration of 10τ . A matrix of Q_{ij} functions is obtained at the beginning and end of the preconditioning procedure, to compare steady state with initial conditions.

The relationship among the four stimuli can be seen from the following Venn-diagram of the retinal sets, where the double-headed arrows indicate the oscillating pairs of stimuli, and the number in each cell indicates its area.



The initial and terminal Q-matrices have been computed for a four-layer and open-loop cross-coupled perceptron, as a function of the parameter $N_a \eta / \sigma$. In both models the parameters of the $A^{(1)}$ units (or of all A-units, in the cross-coupled case) were $\alpha = 3$, $\gamma = 0$, and $\theta = 2$, with a binomial model. In the four-layer model, $\theta^{(2)}$ was also taken to be 2 , so that the systems are directly comparable.

The Q-matrices obtained in this experiment are shown in Tables 5 and 6. The important Q-functions are also shown graphically in Fig. 46, as a function of the parameter $N_a \eta / \sigma$. Note that for both models, there is a considerable parametric range within which generalization is much greater for stimuli which belong to the same event than for stimuli from different events. This gain in generalization between S_1 and S_2 , and between S_3 and S_4 is more than sufficient to offset the handicap of the intersections between S_1 and S_3 , and between S_2 and S_4 , which gives the system an initial disadvantage. The cross-coupled model, while it follows a similar history, has a considerably greater "useful range" than the four-layer model. For the four-layer system, the range of

TABLE 5

Q-MATRICES FOR FOUR-LAYER α-PERCEPTRON IN EXPERIMENT 12

(PARAMETERS: $x = 3$, $y = 0$, $\theta = 2$)

INITIAL Q-MATRIX:

$$\begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .104 & .000 & .034 \\ .034 & .000 & .104 & .000 \\ .000 & .034 & .000 & .104 \end{pmatrix}$$

TERMINAL MATRICES FOR:

$77.0 < N_a \eta/\delta < 88.9$

$$\begin{pmatrix} .104 & .070 & .034 & .000 \\ .070 & .174 & .000 & .034 \\ .034 & .000 & .104 & .070 \\ .000 & .034 & .070 & .174 \end{pmatrix}$$

$88.9 < N_a \eta/\delta < 166.6$

$$\begin{pmatrix} .174 & .140 & .034 & .000 \\ .140 & .174 & .000 & .034 \\ .034 & .000 & .174 & .140 \\ .000 & .034 & .140 & .174 \end{pmatrix}$$

$N_a \eta/\delta < 166.6$

$$\begin{pmatrix} .314 & .280 & .034 & .280 \\ .280 & .314 & .280 & .034 \\ .034 & .280 & .314 & .280 \\ .280 & .034 & .280 & .314 \end{pmatrix}$$

TABLE 6

Q-MATRICES FOR OPEN-LOOP CROSS-COUPLED α-PERCEPTRON IN EXPERIMENT 12

(PARAMETERS: $x = 3$, $y = 0$, $\theta = 2$)

INITIAL Q-MATRIX:	$\begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .104 & .000 & .034 \\ .034 & .000 & .104 & .000 \\ .000 & .034 & .000 & .104 \end{pmatrix}$
TERMINAL MATRICES FOR:	
$38.5 < N_a \eta/\delta < 44.5$	$\begin{pmatrix} .122 & .018 & .034 & .000 \\ .018 & .104 & .000 & .034 \\ .034 & .000 & .122 & .018 \\ .000 & .034 & .018 & .104 \end{pmatrix}$
$44.5 < N_a \eta/\delta < 77.0$	$\begin{pmatrix} .122 & .036 & .034 & .000 \\ .036 & .122 & .000 & .034 \\ .034 & .000 & .122 & .036 \\ .000 & .034 & .036 & .122 \end{pmatrix}$
$77.0 < N_a \eta/\delta < 83.3$	$\begin{pmatrix} .174 & .082 & .034 & .000 \\ .082 & .122 & .000 & .034 \\ .034 & .000 & .174 & .082 \\ .000 & .034 & .082 & .122 \end{pmatrix}$
$83.3 < N_a \eta/\delta < 88.9$	$\begin{pmatrix} .183 & .097 & .034 & .027 \\ .097 & .140 & .027 & .034 \\ .034 & .027 & .183 & .097 \\ .027 & .034 & .097 & .140 \end{pmatrix}$
$88.9 < N_a \eta/\delta < 117.6$	$\begin{pmatrix} .192 & .131 & .034 & .036 \\ .131 & .192 & .036 & .034 \\ .034 & .036 & .192 & .131 \\ .036 & .034 & .131 & .192 \end{pmatrix}$
$117.6 < N_a \eta/\delta < 166.6$	$\begin{pmatrix} .210 & .176 & .034 & .072 \\ .176 & .210 & .072 & .034 \\ .034 & .072 & .210 & .176 \\ .072 & .034 & .176 & .210 \end{pmatrix}$
$166.6 < N_a \eta/\delta < 235.2$	$\begin{pmatrix} .262 & .228 & .034 & .176 \\ .228 & .262 & .176 & .034 \\ .034 & .176 & .262 & .228 \\ .176 & .034 & .228 & .262 \end{pmatrix}$
$N_a \eta/\delta > 235.2$	$\begin{pmatrix} .314 & .280 & .034 & .280 \\ .280 & .314 & .280 & .034 \\ .034 & .280 & .314 & .280 \\ .280 & .034 & .280 & .314 \end{pmatrix}$

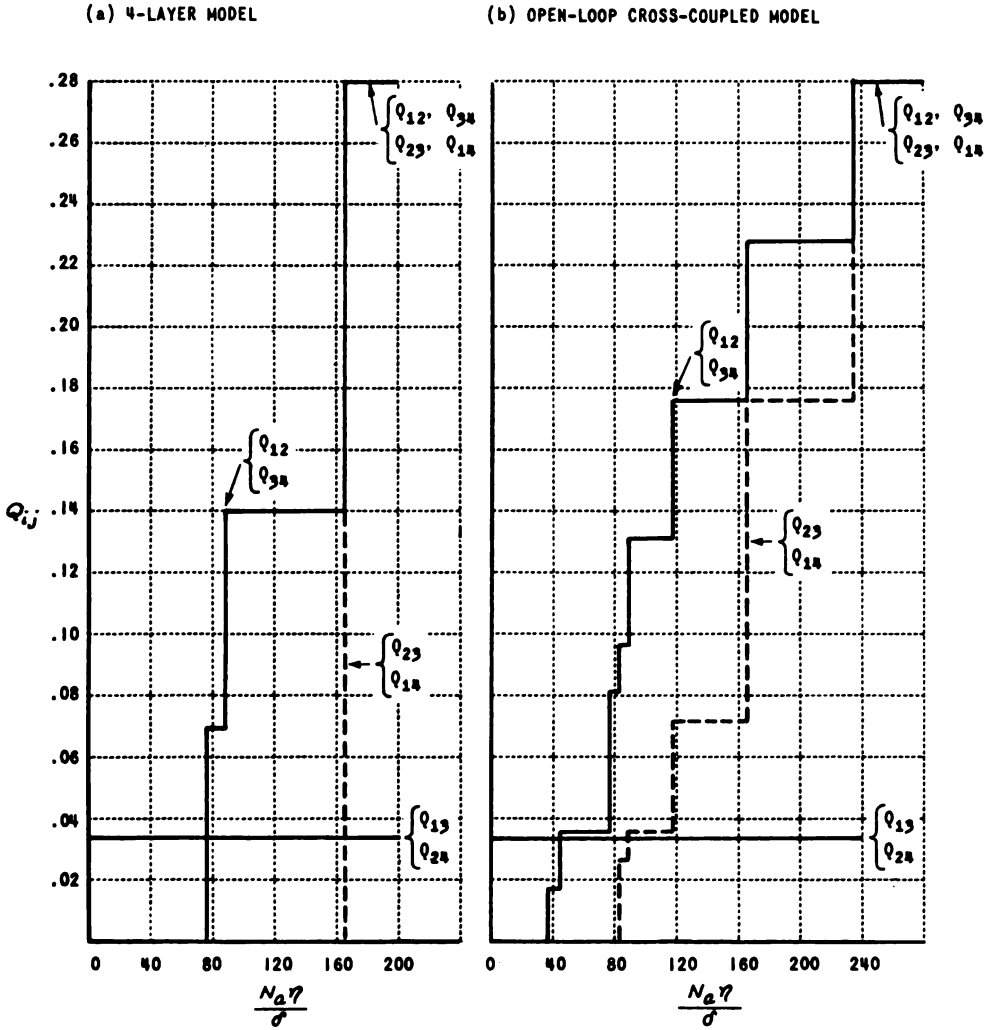


Figure 46 COMPARISON OF 4-LAYER AND OPEN-LOOP CROSS-COUPLED α-PERCEPTRONS ON EXPT. 12. (χ = 3, y = 0, θ = 2 FOR BOTH SYSTEMS)

$N_a \eta / \sigma$ for which the system tends to classify events "correctly" is 77.0 to 166.6, while for the cross-coupled model this range is extended to 38.5 to 238.2. Thus the cross-coupled model begins to show the generalization effects earlier, and saturates later than the four-layer system. Moreover, the transition occurs more gradually, in eight steps for the cross-coupled system as opposed to three for the four-layer model.

The matrices shown here assume α-system reinforcement. A γ or Γ-system, with the four-layer model, eliminates all $A^{(2)}$ activity immediately, in this experiment. In the cross-coupled model, however, activity is not completely eliminated, and the terminal Q-matrices obtained for a γ-perceptron are shown in Table 7. Note that the bias favoring Q_{13} and Q_{24} is eliminated for most values of $N_a \eta / \sigma$, and that the "dynamic range" is greater than in the α-system. The Γ-system, illustrated in Table 8, is similar to the γ-perceptron for small values of $N_a \eta / \sigma$, but it appears to "saturate" more easily.

While the performance of the cross-coupled perceptron closely resembles the system in Chapter 16, it is a somewhat more satisfying model from the standpoint of biological plausibility and parsimony, since it does not require the assumption of a special set of fixed connections from $A^{(1)}$ to $A^{(2)}$ units in addition to the variable connections - an assumption which was necessary, in the four-layer system, to provide a "template" for the organization of similar $A^{(1)}$ units to be connected to each $A^{(2)}$ unit, and in order to prevent all connections from decaying to zero value. In the present scheme, all S-A connections are fixed, and all other connections variable, yielding a conceptually simpler organization.

TABLE 7

Q-MATRICES FOR OPEN-LOOP CROSS-COUPLED $\mathcal{T}$-PERCEPTRON IN EXPERIMENT 12

(Parameters: $\alpha = 3$, $y = 0$, $\theta = 2$)

INITIAL Q-MATRIX:

$$\begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .104 & .000 & .034 \\ .034 & .000 & .104 & .000 \\ .000 & .034 & .000 & .104 \end{pmatrix}$$

TERMINAL MATRICES FOR:

$0 < \frac{N_a \eta}{\delta} < 68.7$

$$\begin{pmatrix} .008 & .000 & .001 & .000 \\ .000 & .008 & .000 & .001 \\ .001 & .000 & .008 & .000 \\ .000 & .001 & .000 & .008 \end{pmatrix}$$

$68.7 < \frac{N_a \eta}{\delta} < 85.8$

$$\begin{pmatrix} .009 & .002 & .002 & .002 \\ .002 & .009 & .002 & .002 \\ .002 & .002 & .009 & .002 \\ .002 & .002 & .002 & .009 \end{pmatrix}$$

$85.8 < \frac{N_a \eta}{\delta} < 101$

$$\begin{pmatrix} .019 & .012 & .008 & .008 \\ .012 & .012 & .008 & .008 \\ .008 & .008 & .019 & .012 \\ .008 & .008 & .012 & .012 \end{pmatrix}$$

$101 < \frac{N_a \eta}{\delta} < 152$

$$\begin{pmatrix} .022 & .022 & .014 & .014 \\ .022 & .022 & .014 & .014 \\ .014 & .014 & .022 & .022 \\ .014 & .014 & .022 & .022 \end{pmatrix}$$

$152 < \frac{N_a \eta}{\delta} < 303$

$$\begin{pmatrix} .025 & .025 & .017 & .017 \\ .025 & .025 & .017 & .017 \\ .017 & .017 & .025 & .025 \\ .017 & .017 & .025 & .025 \end{pmatrix}$$

$\frac{N_a \eta}{\delta} > 302$

$$\begin{pmatrix} .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \end{pmatrix}$$

TABLE 8

Q-MATRICES FOR OPEN-LOOP CROSS-COUPLED Γ-PERCEPTRON IN EXPERIMENT 12

(Parameters: $\alpha = 3$, $\gamma = 0$, $\theta = 2$)

INITIAL Q-MATRIX:	$\begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .104 & .000 & .034 \\ .034 & .000 & .104 & .000 \\ .000 & .034 & .000 & .104 \end{pmatrix}$
TERMINAL MATRICES FOR:	
$0 < \frac{N_a \eta}{\delta} < 58.5$	$\begin{pmatrix} .008 & .000 & .001 & .000 \\ .000 & .008 & .000 & .001 \\ .001 & .000 & .008 & .000 \\ .000 & .001 & .000 & .008 \end{pmatrix}$
$58.5 < \frac{N_a \eta}{\delta} < 77.8$	$\begin{pmatrix} .009 & .002 & .002 & .002 \\ .002 & .009 & .002 & .002 \\ .002 & .002 & .009 & .002 \\ .002 & .002 & .002 & .009 \end{pmatrix}$
$77.8 < \frac{N_a \eta}{\delta} < 88.5$	$\begin{pmatrix} .019 & .012 & .008 & .008 \\ .012 & .012 & .008 & .008 \\ .008 & .008 & .019 & .012 \\ .008 & .008 & .012 & .012 \end{pmatrix}$
$88.5 < \frac{N_a \eta}{\delta} < 92.0$	$\begin{pmatrix} .022 & .015 & .014 & .014 \\ .015 & .015 & .014 & .014 \\ .014 & .014 & .022 & .015 \\ .014 & .014 & .015 & .015 \end{pmatrix}$
$92.0 < \frac{N_a \eta}{\delta} < 131$	$\begin{pmatrix} .025 & .025 & .020 & .020 \\ .025 & .025 & .020 & .020 \\ .020 & .020 & .025 & .025 \\ .020 & .020 & .025 & .025 \end{pmatrix}$
$131 < \frac{N_a \eta}{\delta} < 181$	$\begin{pmatrix} .028 & .028 & .026 & .026 \\ .028 & .028 & .026 & .026 \\ .026 & .026 & .028 & .028 \\ .026 & .026 & .028 & .028 \end{pmatrix}$
$\frac{N_a \eta}{\delta} > 181$	$\begin{pmatrix} .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \\ .030 & .030 & .030 & .030 \end{pmatrix}$

It will be seen in Chapter 19 that this system, with the addition of a unit time-delay (all $\tau_{ij} = 1$) performs identically to a closed loop fully cross-coupled perceptron for the first two cycles of operation. By further extension of the network along the same lines, it will be shown that additional cycles of closed-loop activity can be duplicated.

17.3 Reduction of Size Requirements for Universal Perceptrons

In the case of simple perceptrons, it was demonstrated that in order to obtain a "universal perceptron", in which a solution exists for any classification of n stimuli, at least n A-units are required (Theorem 3, Corollary 2, Chapter 5). Now consider an open-loop cross coupled perceptron, constructed as follows: Let the A-units be numbered in series $a_1, a_2, \ldots, a_{N_a}$ and let $N_a = N_s$ (the number of S-points). The last of these units, a_{N_a}, has an output connection to an R-unit. Each A-unit has a variable-valued connection from every S-point, plus one connection for every A-unit prior to itself in the series; i.e., a_j receives a connection from every S-point and from $a_1, a_2, \ldots, a_{j-1}$.

It has been demonstrated by Cameron* that for small values of n ($n = 2^{N_S}$) only $\log_2(n)$ A-units are required in order to obtain a universal perceptron, in which a solution exists for all of the 2^n possible classification. This was demonstrated by explicit construction for n as large as 8. At some higher value of n, this ceases to be true, although the maximum n for which the observation holds true has not yet been determined.

---

* S. Cameron, personal communication.

A lower bound for the number of A-units required for a universal perceptron in such a system has been obtained by Joseph (although it is not a least upper bound). The analysis (given in the Appendix of Ref. 41) is based on the Hay-Joseph theorem that the maximum number of orthants achievable by linear combinations of r vectors in n-space is approximately $M(n, r) = \frac{n^{r-1}}{(r-1)!}$ where n is large, and r is small relative to n. An upper bound for the number of dichotomies achievable with N_a A-units is found to be $M(2^{N_a}, N_a+1)\, M(2^{N_a}, N_a+2) \ldots M(2^{N_a}, N_a+N_a)$. It is shown that for large N_a the number of possible dichotomies is increasing at a much greater rate than the number of achievable dichotomies, so that there must be some point at which the system ceases to act as a universal perceptron.

18. Q-FUNCTIONS FOR CROSS-COUPLED PERCEPTRONS

A general cross-coupled perceptron is illustrated in Figure 47. It consists of three layers of units, with complete freedom of interconnection among the A-units. Due to the likelihood of closed circuits of connections within the network, this is called a closed-loop system.

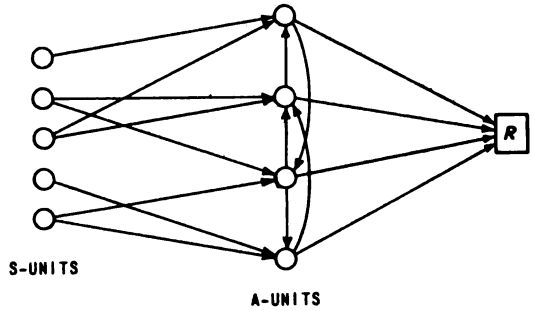


Figure 47 TYPICAL CONNECTIONS IN A CLOSED-LOOP CROSS-COUPLED PERCEPTRON

In passing from open-loop to closed-loop networks, several fundamentally new considerations enter into the analysis. In the first place, the state of the network at time t becomes a function, not only of the present sensory input and the momentary values of the connections, but of the preceding sequence of inputs and past activity states as well.

The dependence of the system's state upon time-sequences of previous states means that the transmission time, τ_{ij} , which previously played no part or only a minor part in the analysis of system performance, now becomes a parameter to be reckoned with at all times. The question of network stability is also a fundamental one; some cross-coupled networks, once triggered, will explode into total activity which prevents any further stimuli from making any impression at all, others will oscillate, and others will settle down to a stable steady-state condition. In this chapter, we begin by re-examining the concept of Q-functions, in order to provide a means of measuring the response of the network to sequences of stimuli, and comparing its response quantitatively for different stimulus sequences. These new Q-functions will be found to encompass the functions analyzed in Chapter 6 as a special case.

18.1 Stimulus Sequences: Notation

In Chapter 4, a stimulus was defined as any set of input signals to sensory units of a perceptron, excluding the null stimulus. In practice, these signals are generally taken to be 1 or zero. For present purposes, the null stimulus (all signals equal to zero) will be re-admitted as a stimulus, and will be symbolized by 0 when it occurs as part of a sequence. A stimulus sequence, $\mathcal{S}_i = (S_{i_1}, S_{i_2}, \ldots, S_{i_m})$ can be an arbitrary series of stimuli which are assumed to occur at successive discrete times $t_1, t_1 + \Delta t, t_1 + 2\Delta t, \ldots, t_1 + (m-1)\Delta t$. An arbitrary set of stimulus sequences can be taken to comprise a stimulus-sequence world, for a given perceptron, in the sense of Definition 26 of Chapter 4.

In this and the subsequent chapters, it will be assumed that the transmission time, $\tau_{i,j}$ is equal to Δt for all connections, $c_{i,j}$, and this transmission time will be symbolized in abbreviated form by τ. Consequently, if a stimulus S_i occurs at time t, the response to this stimulus in the A-system occurs at time $t+\tau$, and Q_i is interpreted to mean the probability that an A-unit is activated at time t if S_i occurs at time $t-\tau$. In a cross-coupled perceptron, however, Q_i is not a well-defined quantity, since in addition to signals from the retina, an A-unit may receive signals from other A-units at time t, so that the response at time t depends both on $S(t-\tau)$ and on the activity state of the association system at $t-\tau$. Q_i is therefore redefined to apply to sequences $\mathcal{A}_i$ of length m, which begin at time $t-m\tau$, and terminate at $t-\tau$, with the association system assumed to be totally inactive, or "silent" at time $t-m\tau$. In this case, for a sequence of length 1, Q_i is interpreted in the usual manner, and is represented by the equations of Chapter 6, without modification. For a general sequence of length m, we use the notation Q_{i_m} to designate the probability that an A-unit is active at time t, given that the sequence $\mathcal{A}_i$ began at time $t-m\tau$, so that the m^{th} member of the sequence occured at $t-\tau$. More generally, we can write Q_{i_r} to designate the probability that an A-unit is active at time t if the sequence $\mathcal{A}_i$ began at $t-r\tau$, where r may be less than, equal, or greater than m. If r is less than m, this is equivalent to the probability of response to a truncated sequence, containing only the first r stimuli of the sequence $\mathcal{A}_i = (S_{i_1}, S_{i_2}, \cdots, S_{i_r}, \cdots, S_{i_m})$. If $r > m$, we adopt the convention that the sequence $\mathcal{A}_i$ is understood to have been augmented by the addition of $r-m$ null stimuli, yielding the sequence $(S_1, S_2, \cdots, S_m, O_1, \cdots, O_{r-m})$. In other words, it is assumed that the sequence $\mathcal{A}_i$ began at $t-r\tau$, and that no other inputs occurred

through time $t-\tau$, the probability of A-unit activity then being determined for time t. In a simple perceptron, this probability would, of course, be zero for $r > m$; in a cross-coupled system, however, the presence of persistent cycles of activity, or reverberating loops in the A-system, may maintain $Q_{i_r} > 0$ for an indefinite period.

Q_{ij} is redefined in a manner analogous to Q_i. Where $\mathcal{S}_i$ and $\mathcal{S}_j$ are any two sequences, we define

$Q_{i_\mu j_\nu}$ = probability that an A-unit responds at time t if $\mathcal{S}_i$ begins at $t-\mu\tau$, and also responds at time t if $\mathcal{S}_j$ begins at $t-\nu\tau$.

It is again assumed that the A-system is "silent" at the start of each sequence for which the Q-function is defined, and that if μ or ν is greater than m, the corresponding sequence is augmented by a sufficient number of null stimuli at the right-hand end. Q-functions with arbitrary numbers of subscripts can be generated by an obvious extension of the above definition.

In contexts where no ambiguity can arise, the notation Q_{ij} will be used to denote $Q_{i_m j_{m'}}$, i.e., the probability that an A-unit responds immediately after the termination of $\mathcal{S}_i$ and also responds immediately after the termination of $\mathcal{S}_j$. Note that it is not required that the sequences $\mathcal{S}_i$ and $\mathcal{S}_j$ be commensurate, i.e., the lengths m and m' may be different for the two sequences, without requiring any redefinition of Q_{ij}.

Generalization coefficients, $g_{i_\mu j_\nu}$, can be defined analogously to Q-functions. For example, in an alpha system, we would have $E(g_{i_\mu j_\nu}) = Q_{i_\mu j_\nu}$, where $g_{i_\mu j_\nu}$ is a measure of the increment added to the output signal of the A-set responding after ν stimuli of the sequence $\mathcal{S}_j$, as a result of an α-reinforcement after the μ^{th} stimulus of the sequence $\mathcal{S}_i$. Again, if the second-order subscripts are suppressed, it will be assumed that $g_{ij} = g_{i_m j_{m'}}$ = the effect of a reinforcement immediately after the termination of $\mathcal{S}_j$ upon the signal which follows immediately after the termination of $\mathcal{S}_i$. If reinforcements are always applied and measured immediately after the end of stimulus sequences, the performance of the perceptron in learning responses to such sequences can be derived from the resulting G matrix, in precisely the same manner as was done for elementary perceptrons in Part II. Thus a knowledge of the Q-functions for a cross-coupled perceptron permits us to predict the performance of such systems in discrimination and generalization experiments.

18.2 Q_i Functions and Stability

The rigorous analysis of Q_{i_ν} for a cross-coupled perceptron with a finite number of A-units presents the identical difficulty which was encountered in the case of Q-functions for multi-layer systems (Section 15.1). The probability Q_{i_1} is, of course, identical to the function Q_i defined for the first stimulus of the sequence $\mathcal{S}_i$ in accordance with the equations of Chapter 6; but the probability Q_{i_2} already depends upon the distribution of numbers of A-units which respond to the first stimulus, S_{i_1}. In order to avoid consideration of these distributions, the Q-functions obtained here will always represent limits for large networks, where it can be assumed that the actual proportion of A-units responding after S_{i_μ} is equal to Q_{i_μ}. It should be noted that due to the assumption that the sequence $\mathcal{S}_i$ starts with a "silent" perceptron, $Q_{i_0} = 0$.

A number of alternative topological models might be considered. For convenience, the following analysis takes up the case of a perceptron in which both the connections from the retina to the A-units and the "internal" connections to each A-unit are constrained as in the binomial model of Chapter 6. In this model, we have five parameters for each A-unit:

θ = threshold of A-unit

x_A = number of excitatory connections from the S-set, or retina

y_A = number of inhibitory connections from the retina

x_a = number of excitatory connections from other A-units

y_a = number of inhibitory connections from other A-units

In the present chapter, we shall be concerned only with perceptrons in which all input connections to A-units are fixed in value, regardless of where they originate. Systems with modifiable couplings between A-units will be considered in the following chapter. It is assumed that each of the above sets of connections has its origin points assigned at random from a uniform probability distribution over the S-set or the A-set, as required. This results in the following equation for $Q_{i,\nu}$:

$$Q_{i,\nu} = \sum_{E_A - I_A + E_a - I_a \geq \theta} P_1(E_A)\, P_2(I_A)\, P_3(E_a)\, P_4(I_a) \qquad (18.1)$$

where

$$P_1(E_A) = \binom{x_A}{E_A} (R_{i\nu})^{E_A} (1-R_{i\nu})^{x_A - E_A}$$

$$P_2(I_A) = \binom{y_A}{I_A} (R_{i\nu})^{I_A} (1-R_{i\nu})^{y_A - I_A}$$

$$P_3(E_a) = \binom{x_a}{E_a} (Q_{i\nu-1})^{E_a} (1-Q_{i\nu-1})^{x_a - E_a}$$

$$P_4(I_a) = \binom{y_a}{I_a} (Q_{i\nu-1})^{I_a} (1-Q_{i\nu-1})^{y_a - I_a}$$

$R_{i\nu}$ = fraction of S-units activated by $S_{i\nu}$.

Taking $Q_{i0} = 0$, Q_{i1} can thus be developed recursively in terms of $Q_{i\nu-1}$ up to any value of ν .

For a Poisson model, in which the number of output connections from each A-unit is constrained but the number of inputs is a random variable (or in which both ends of a set of connections are picked at random) equation (18.1) still applies, but the probability functions P_1, P_2, P_3, and P_4 must be redefined, in a manner analogous to Chapter 6. It is also possible, of course, to have some kinds of connections (e.g., the internal excitatory connections) distributed binomially, while the other sets of connections are organized according to a Poisson model, so that $P_1, \ldots, P_4$ need not all be of the same type. For present purposes, however, we shall continue to concentrate on the pure binomial model defined above. All major conclusions undoubtedly apply to Poisson and mixed systems equally well.

One of the first questions to be raised about such a system concerns the stability of the activity-level, and the possible tendency of the system to burst into total activity in response to a transient stimulus (which would, of course, preclude any possibility of learning or discrimination of different stimuli). Figure 48 illustrates the response to a transient stimulus (i.e., a sequence of length 1) for a number of representative cases. Figure 49 presents the response of a number of networks to a steadily maintained stimulus, or a sequence of stimuli all of which have the identical area. (Note that it follows from Equation (18.1) that the actual sequence of stimuli does not affect $Q_{i\nu}$, so long as the stimulus area, $R_{i\mu}$, is fixed for each $S_{i\mu}$. Thus any two sequences for which the succession of $R_{i\mu}$ are equivalent will yield the same value of $Q_{i\nu}$.)

Figure 48(a) illustrates the effect of the size of the "trigger stimulus" upon the transient response of the system. Note that the final activity level is independent of R_i; it is also independent of x_A and y_A, so long as $x_A \geq \theta$. Figure 48(b) shows the effect of varying the ratio of internal excitation to internal inhibition (x_a and y_a). For a purely excitatory system, total activity of the network is likely to occur, in which all A-units become and remain active. As the inhibitory component is increased, a lower level of stable activity results, and with still further increase in y_a relative to x_a, the initial transient activity will die away entirely. Figure 48(c) shows that the effect of increasing the threshold of the A-units is similar to the effect of increasing the internal inhibitory component. It should be noted that all of these Q_i functions in response to transient phenomena in a cross-coupled system are identical to the succession of Q-functions for successive layers of a multilayer perceptron (as discussed in Chapter 15). For infinite N_a the equations for $Q_i^{(\nu)}$ and $Q_{i\nu}$ are identical, where ν in the first case denotes the layer, and in the second the cycle of activity in the A-system.

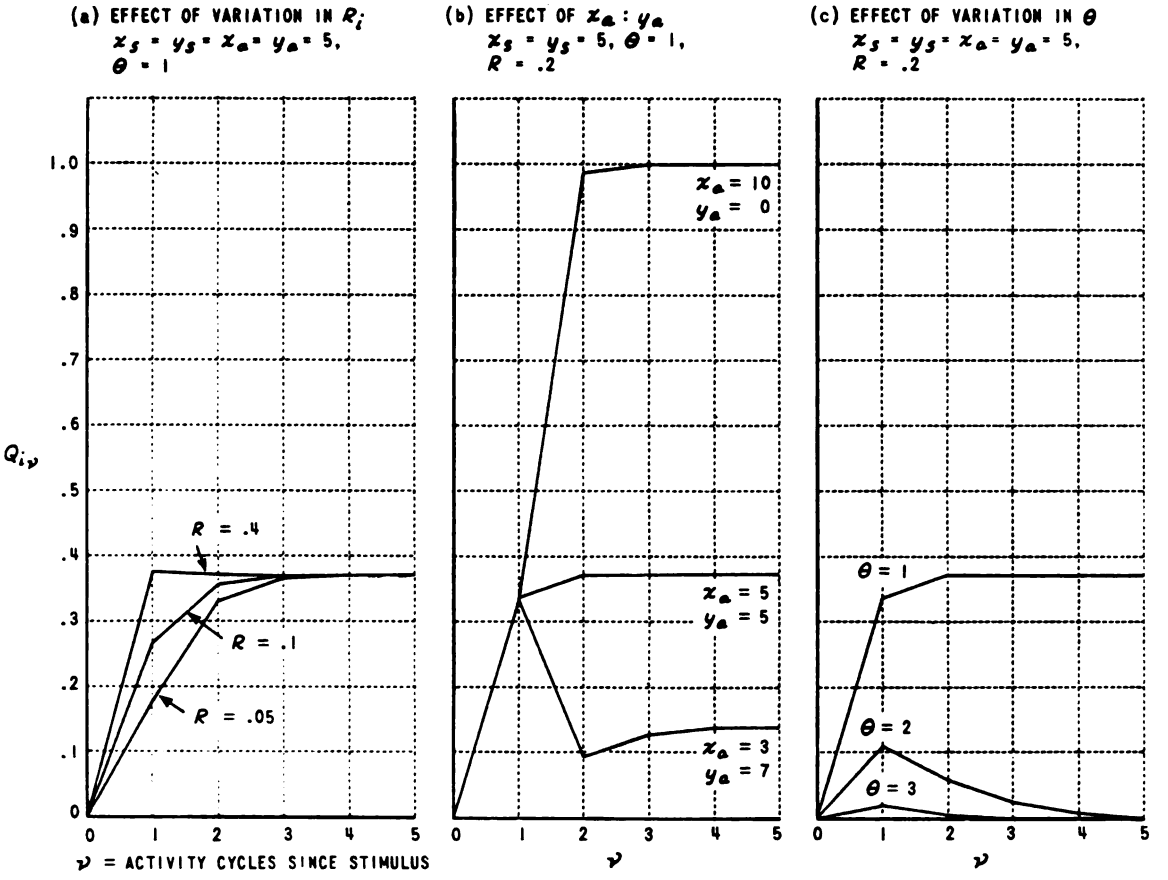


Figure 48 RESPONSE OF A CROSS-COUPLED BINOMIAL PERCEPTRON TO A TRANSIENT STIMULUS

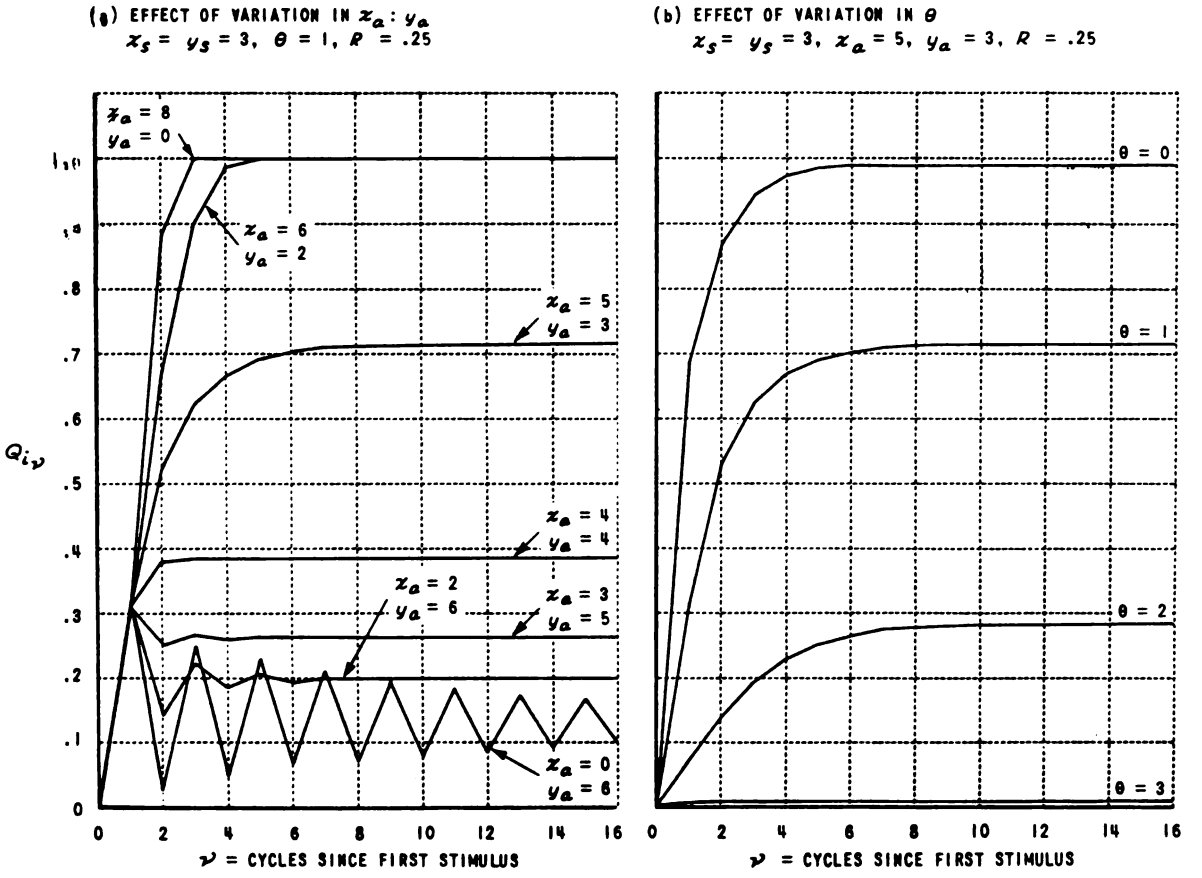


Figure 49 RESPONSE OF A CROSS-COUPLED BINOMIAL PERCEPTRON TO A STIMULUS SEQUENCE WITH CONSTANT R

Figure 49(a) shows that as the internal inhibitory component is increased to the point where the terminal steady-state level of the system is below the value of Q_i for the initial impulse from the retina, a damped series of oscillations occurs, which becomes pronounced as y_a is increased. Changing the threshold (as in Fig. 49(b)) also serves to reduce the asymptotic activity level, but does not cause the qualitative alteration from a monotonic to an oscillating sequence, as does the increase in y_a. A sequence which is either monotone or oscillating for one value of θ will remain monotone or oscillating as θ is changed.

18.3 Q_{ij} Functions

The function $Q_{i_\mu j_\nu}$ for a binomial-model cross-coupled perceptron can be calculated by an extension of the treatment employed in the preceding section. The resulting equation (again assuming large N_a) is:

$$Q_{i_\mu j_\nu} = \sum_{\substack{\alpha_i \geq \theta \\ \alpha_j \geq \theta}} P_1(E_A^i, E_A^j, E_A^c)\, P_2(I_A^i, I_A^j, I_A^c)\, P_3(E_a^i, E_a^j, E_a^c)\, P_4(I_a^i, I_a^j, I_a^c) \tag{18.2}$$

where $\alpha_i = E_A^i + E_A^c - I_A^i - I_A^c + E_a^i + E_a^c - I_a^i - I_a^c$

$\alpha_j = E_A^j + E_A^c - I_A^j - I_A^c + E_a^j + E_a^c - I_a^j - I_a^c$

The above notation for excitatory and inhibitory signal components received from the "unique" and "common" sets of sensory points and A-units active at $t - \tau$ is an obvious extension of the notation employed previously (c.f., Chapter 6). For the multinomial probabilities, we have

$$P_1(E_A^i, E_A^j, E_A^c) = \frac{x_A!}{E_A^i!\, E_A^j!\, E_A^c!\, (x_A - E_A^i - E_A^j - E_A^c)!} (U_A^i)^{E_A^i} (U_A^j)^{E_A^j} (C_A)^{E_A^c} (1 - U_A^i - U_A^j - C_A)^{x_A - E_A^i - E_A^j - E_A^c}$$

$$P_2(I_A^i, I_A^j, I_A^c) = \frac{y_A!}{I_A^i!\, I_A^j!\, I_A^c!\, (y_A - I_A^i - I_A^j - I_A^c)!} (U_A^i)^{I_A^i} (U_A^j)^{I_A^j} (C_A)^{I_A^c} (1 - U_A^i - U_A^j - C_A)^{y_A - I_A^i - I_A^j - I_A^c}$$

$$P_3(E_a^i, E_a^j, E_a^c) = \frac{x_a!}{E_a^i!\, E_a^j!\, E_a^c!\, (x_a - E_a^i - E_a^j - E_a^c)!} (U_a^i)^{E_a^i} (U_a^i)^{E_a^j} (C_a)^{E_a^c} (1 - U_a^i - U_a^j - C_a)^{x_a - E_a^i - E_a^j - E_a^c}$$

$$P_4(I_a^i, I_a^j, I_a^c) = \frac{y_a!}{I_a^i!\, I_a^j!\, I_a^c!\, (y_a - I_a^i - I_a^j - I_a^c)!} (U_a^i)^{I_a^i} (U_a^j)^{I_a^j} (C_a)^{I_a^c} (1 - U_a^i - U_a^j - C_a)^{y_a - I_a^i - I_a^j - I_a^c}$$

where C_A = proportion of S-points activated both by S_{i_μ} and S_{j_ν} .

$U_A^i = R_{i_\mu} - C_A$ where R_{i_μ} is the proportion of S-points activated by S_{i_μ} .

$U_A^j = R_{j_\nu} - C_A$ where R_{j_ν} is the proportion of S-points activated by S_{j_ν} .

$C_a = Q_{i_{\mu-1}\, j_{\nu-1}}$

$U_a^i = Q_{i_{\mu-1}} - Q_{i_{\mu-1}\, j_{\nu-1}}$

$U_a^j = Q_{j_{\nu-1}} - Q_{i_{\mu-1}\, j_{\nu-1}}$

For arbitrary values of μ and ν, $Q_{i_\mu j_\nu}$ can again be calculated by a recursive operation, assuming that the perceptron is "silent" prior to the start of each sequence. If the two sequences $\mathcal{S}_i$ and $\mathcal{S}_j$ are incommensurate (or if $\mu \neq \nu$) the values of C_a are thus taken to be zero up to the time that both sequences have begun. (This is equivalent to extending the shorter sequence by adding a sufficien t number of null stimuli at the beginning to make it equal in length to the longer sequence.)

Two questions are of particular importance concerning these functions. The first is the question of the sensitivity of the system to pertubations in a sequence of stimuli; this determines how well a cross-coupled perceptron can discriminate one stimulus sequence from another. The second question is the dependence of the present state of the system upon stimuli from the remote past; this is of importance in order to guarantee a sufficiently consistent response to a present stimulus so that it can be correctly identified, and also in justifying an approximation to the perceptron's performance by means of an analysis of finite sequences (as will be done in the following chapter). Figures 50 and 51 present the results of an investigation of these questions.*

In Figure 50 the effect of a perturbation in the stimulus sequence is illustrated. In each case the sequence $\mathcal{S}_1$ is assumed to consist of 17 stimuli ($A_1, A_2, \ldots, A_{17}$). In the other sequences, one or more "perturbation stimuli" are introduced in place of some of the "A" stimuli;

---

* The data for these illustrations were computed by W. Eisner, on the Burroughs 220 computer at Cornell University.

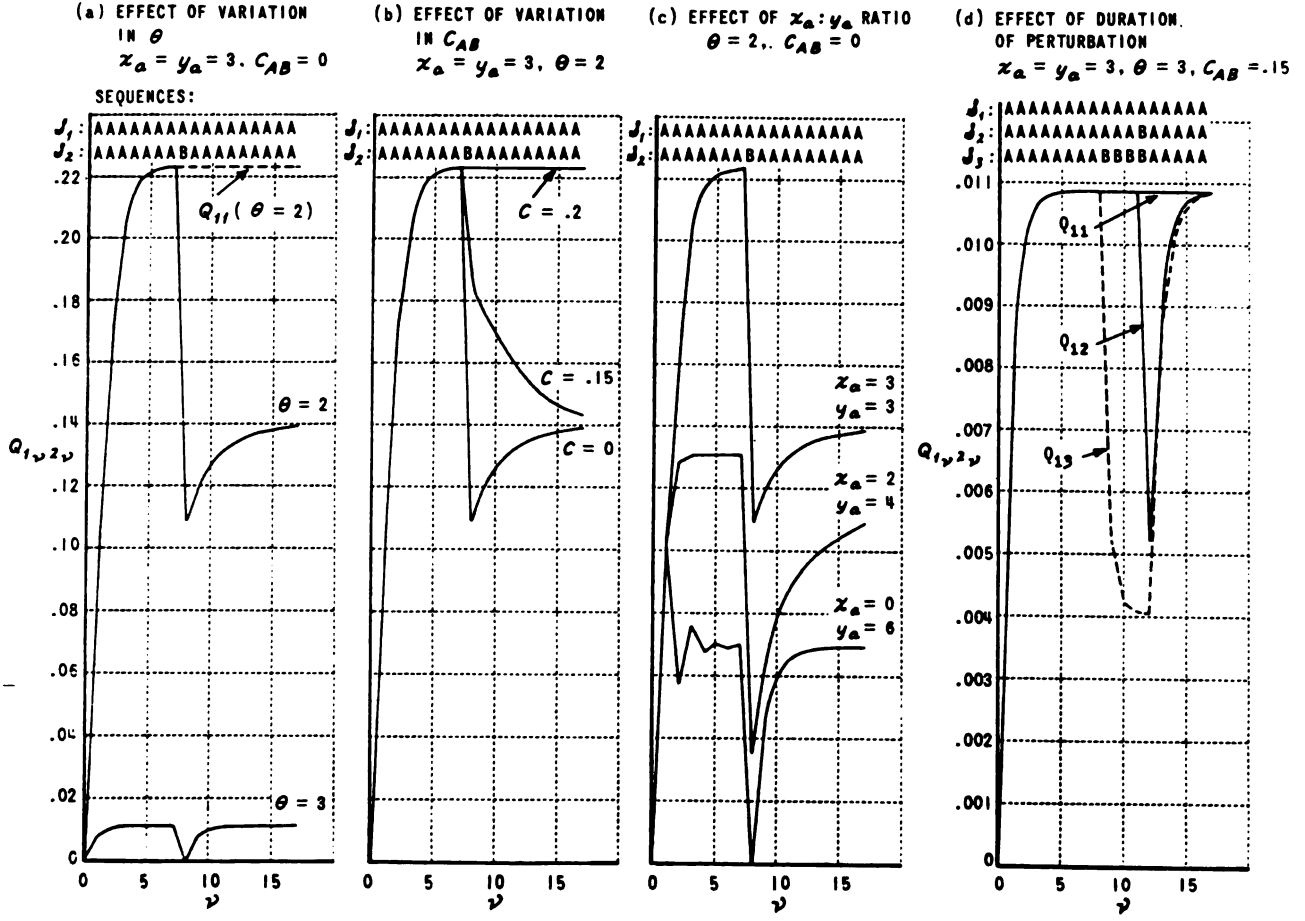


Figure 50 $Q_{i\nu j\nu}$ FOR CROSS-COUPLED PERCEPTRONS: EFFECT OF PERTURBATION IN STIMULUS SEQUENCE
$x_s = 3$, $y_s = 0$, $R = .2$ FOR ALL STIMULI

these are denoted by the letter "B" in the figure. In figure 50(a), a single "B" stimulus is introduced, in place of the eighth "A" stimulus, with C_{AB} (the intersection between the "B" stimulus and the corresponding "A" stimulus, A_B) being zero. We find that with $\theta = 2$, Q_{12} is abruptly reduced as soon as the "B" stimulus occurs, and then approaches a new asymptotic level, considerably below the Q_{11} level. With a threshold of 3, however, the curve following the perturbation returns to the Q_{11} level, so that three or four stimuli after the perturbation it is impossible to tell from the active A-set that the perturbation occurred. If the location of the "B" stimulus in the sequence is changed, the same type of Q_{12} curve is found, with the deflection merely being displaced in time, but not changed in magnitude. Figure 49(b) shows that the same asymptotic level is approached regardless of the value of C_{AB} , as long as the "A" and "B" stimuli are not identical ($C < .2$). In general, it appears that the asymptotic value of Q_{12} depends on the parameters of the network, but is independent of the magnitude of the perturbation.

Figure 50(c) shows that as the internal inhibitory component is increased, the asymptotic value of Q_{12} approaches the asymptotic value of Q_{11} , in much the same manner as when the threshold is increased. Finally, Figure 50(d) illustrates the effect of increasing the duration of the perturbation up to four "B" stimuli. Note that the return curve following the perturbation is practically identical in all cases.

Figure 51 demonstrates the effects of introducing null stimuli at the beginning of each stimulus sequence, in place of the initial "A" stimuli. The curves obtained are very similar to those obtained with a

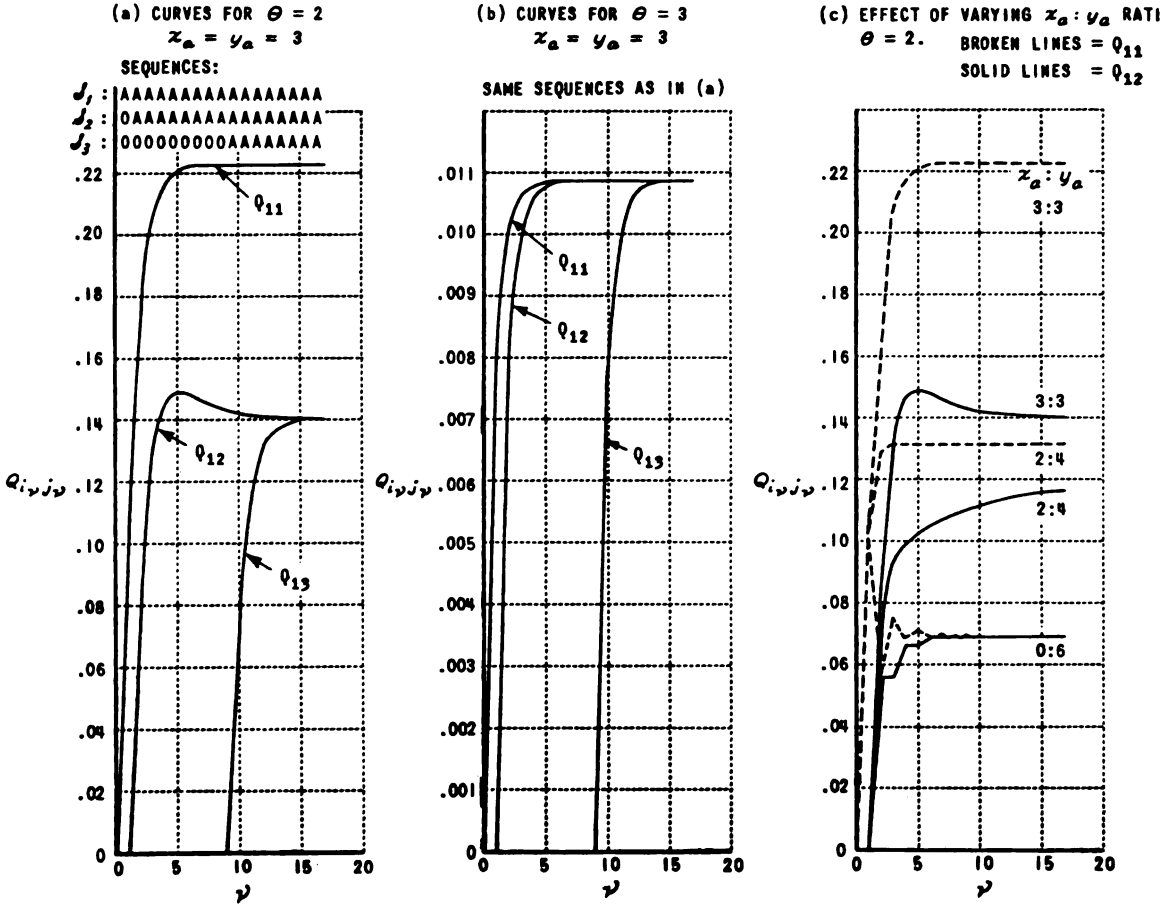


Figure 51 $Q_{i\nu j\nu}$ FOR CROSS-COUPLED PERCEPTRONS: EFFECT OF ABBREVIATION OF STIMULUS SEQUENCES
$x_s = 3$, $y_s = 0$, $P_A = .2$

perturbation of the α_1 sequence, and it is again found that by increasing the threshold or the value of y_a the A-set responding to the altered sequence can be made to approach the set responding to the original, unaltered sequence.

These results demonstrate that there are two distinct conditions which may be found in a cross-coupled perceptron, depending on the choice of parameters. With small θ, or small values of y_a, any perturbation or variation in the stimulus sequence will cause the system to follow a unique course for all subsequent time, and the A-set which is active at time t depends on the entire sequence at all times prior to t, rather than on the most recent stimuli. By increasing θ or y_a, however, such a perceptron can be converted into the second type, in which only the most recent stimuli appreciably affect the current state of the A-system, and stimuli which are sufficiently remote in time have a negligible effect. By lowering θ or y_a slightly, the duration of the noticeable aftereffects of a sequence perturbation can be increased, while still permitting an ultimate return to the A-states associated with the unperturbed sequence. This means, in effect, that the perceptron has a "short term memory" for sequences of a length commensurate with the time for the $Q_{i,j}$ curve to return to its "normal" level, and such sequences can be discriminated by the system. In discriminating such sequences, the most recent stimuli will tend to dominate, and differences which occur in the remote past will be harder to recognize. With the first type of perceptron, however, which is obtained abruptly when the threshold becomes low enough (or y_a becomes low enough) even the most remote stimuli have about the same effect as the most recent stimuli, and the current A-state gives relatively little information about what the present stimuli actually are. Thus, in order to guarantee an adequate degree of correlation between the activity state and the current stimuli, it is necessary to maintain thresholds or inhibitory components at a sufficiently high level; a perceptron of the first type is unlikely to be of much practical value.

19. ADAPTIVE PROCESSES IN CLOSED-LOOP CROSS-COUPLED PERCEPTRONS

In Chapter 18, cross-coupled perceptrons with fixed connection networks were analyzed to determine their stability and characteristic responses to sequences of stimuli. In earlier chapters, four-layer and open-loop cross-coupled perceptrons were analyzed to show that an adaptive preterminal network could vastly improve the capabilities of such systems for similarity generalization. We now turn to the consideration of cross-coupled perceptrons with adaptive interconnections between the A-units, and will attempt to show that the same phenomena can be found here, in a more general and more efficient form. The cross-coupled system not only recognizes sequences of stimuli of arbitrary length, but tends to accellerate its adaptation process due to positive feedback effects within the system. It will be shown later that the closed-loop cross-coupled system is equivalent to an infinitely extended open-loop system, analogous to the one described in Chapter 17.

The first attempt to demonstrate similarity generalization in cross-coupled systems was that of Rosenblatt, in Ref. 85. This was a partially analytic and partially heuristic argument, based upon a study of the similarities of origin-point configurations of the A-units under an arbitrary transformation T. While the general predictions in this paper were correct, and have subsequently been demonstrated in simulation experiments, the method of analysis failed to yield quantitative predictions of the terminal state of the system, after a prolonged period of pre-conditioning. The method employed here is basically different, and yields a more general, as well as more accurate, result. In the following sections, the time-dependent evolution equations for the cross-coupled system will first be developed in their most general form, and specific applications will then be made to

systems in which the assumptions and initial conditions are simplified, to permit a more complete analysis. In the final sections, several similarity generalization experiments will be presented, and performance will be compared with that of multi-layer perceptrons.

19.1 Postulated Organization and Dynamics

The perceptrons to be analyzed in this chapter will be assumed, for convenience, to be fully cross-coupled, that is, there is a connection from every A-unit to every other A-unit and to itself as well. It can be shown that the conclusions which we shall reach for such a system can be extended to any perceptron for which the number of cross-coupling connections per A-unit is large, and the termini of the connections are assigned at random.

Connections from S to A-units are assumed to be fixed in value, and connections from A to R-units are modifiable according to any of the usual reinforcement rules. (We shall not be concerned here with the reinforcement of A-R connections, but shall concentrate upon the evolution of the association network itself.) The A-units are assumed to be simple, with threshold θ, and output signals $a^* = 1$ or 0. The transmission time for all connections is a constant τ. Stimuli are assumed to occur at intervals of the transmissiom time, τ.

Interconnections among A-units are assumed to be variable, according to the same rule employed for the four-layer system of Chapter 16; namely, if a_i is active at time t, and a_j is active at time $t + \tau$, the value of the connection c_{ij} is increased by a quantity $\eta \cdot \Delta t$, and at the same time, all values v_{ij} decay by the quantity $\sigma \Delta t (v_{ij})$. The time unit, Δt, will generally be considered large relative to τ. In symbols, we have

$$\Delta v_{ij}(t) = \begin{cases} (\eta - \delta v_{ij})\,\Delta t & \text{if } a_i^*(t-\tau)\,a_j^*(t) = 1 \\ -\delta \Delta t\,(v_{ij}) & \text{otherwise} \end{cases} \tag{19.1}$$

thus the total signal, $\alpha_i(t)$, received by the A-unit a_i at time t consists of a fixed-connection component, $\beta_i(t)$, originating from the retina, and a variable component, $\gamma_i(t)$, coming from those A-units which were active at $t-\tau$.

19.2 The Phase Space of the A-units

Let us suppose that the environment of a cross-coupled perceptron consists of exactly n admissible stimulus sequences. In order to obtain a G-matrix for this perceptron, and predict its performance, it is necessary to know how its A-units will respond to each of the admissible sequences, including the response to the 1st, 2nd, ... , m^{th} member of the sequence. We will use the notation $a_i^*(S_{j_\nu})$ to denote the output signal of the unit a_i following the ν^{th} stimulus of the sequence $\mathcal{S}_j$. If the sequence $\mathcal{S}_j$ begins at $t-\nu\tau$, the stimulus S_{j_ν} will occur at $t-\tau$, and the input to the unit a_i at time t is given by

$$\alpha_i^{(j\nu)} = \beta_i^{(j\nu)} + \gamma_i^{(j\nu)}(t) \tag{19.2}$$

where $\beta_i^{(j\nu)}$ is the sum of the signals received from the retina following the occurrence of S_{j_ν} and $\gamma_i^{(j\nu)}(t)$ is the sum of the signals received from other A-units at time t, given that $\mathcal{S}_j$ began at $t-\nu\tau$. Knowing $\alpha_i^{(j\nu)}$, we can readily determine $a_i^*(S_{j_\nu})$, since

$$a_i^*(S_{j_\nu}) = \begin{cases} 1 \text{ if } \alpha_i^{(j\nu)} \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

In the perceptrons to be considered here, $\beta_i^{(j\nu)}$ is a constant, while $\gamma_i^{(j\nu)}(t)$ is a time-dependent variable (as in the four-layer perceptrons of Chapter 16).

It will be convenient to represent each abbreviated sequence consisting of the first ν members of any of the original n sequences by a full sequence of length ν. If m is the maximum sequence length, this results in a set of at most mn sequences. Let N be the number of such sequences, and let them be numbered from $\mathcal{A}_1$ through $\mathcal{A}_N$. Then in terms of these new sequences, we can obtain all of the $a_i^*(S_{j\nu}) = a_i^*(\mathcal{A}_k)$ where $\mathcal{A}_k$ is the sequence corresponding to the first ν members of the original sequence $\mathcal{A}_j$. The notation $a_i^*(\mathcal{A}_k)$ means the signal from a_i following the last member of sequence $\mathcal{A}_k$. Similarly, we have $\alpha_i^{(k)} = \beta_i^{(k)} + \gamma_i^{(k)}(t)$.

All of the information necessary to predict the response of an A-unit a_i at time t can now be obtained from the $2N$ numbers $(\beta_i^{(1)}, \beta_i^{(2)}, \ldots, \beta_i^{(N)}, \gamma_i^{(1)}(t), \ldots, \gamma_i^{(N)}(t)) = (\beta_i, \gamma_i(t))$. Thus the set of all possible signals (divided into retinal and internal components) which might affect the activity of a_i at time t, can be represented by a vector of $2N$ components, which depends on t. The space of all such vectors can be mapped into a Euclidean $2N$-space, where each point represents a possible A-unit, or set of A-units, of the perceptron. This will be called the phase space of the A-units. For a large, or infinite perceptron, there is likely to be some concentration of A-units at each point in this phase space at time t. Thus, at time t, there is a probability density associated with each point in the phase space. The state of the entire association system at a given time, t, can then be represented by a probability density distribution over the phase space of the A-units.

For convenience of notation, parentheses for superscripts of α , β , and γ components will hereafter be omitted, with the understanding that the symbol β_i^r means the β-component for unit a_i from stimulus sequence S_r . If exponents are required, they will be expressed by the notation $(\beta_i^r)^k$, which would be β_i^r to the k^{th} power. It should be remembered that with the symbols α , β , and γ , subscripts always denote A-units, whereas superscripts indicate stimulus sequences.

19.3 The Assumption of Finite Sequences

In analyzing the performance of a perceptron, it will generally be our objective to predict the condition of the association system in the limit, as the length of the preconditioning sequence becomes infinite. This means that there are generally an infinite number of possible sequences in the environment, and the phase space of the A-units is properly represented by an infinite dimensional Euclidean space. To justify later assumptions, however, it is necessary to assume that the preconditioning sequence is actually composed of a mixture of a finite number of subsequences of finite length. While this assumption will be carried through the analysis of the following section, it will be shown later that it is possible to drop the assumption in the case of periodic preconditioning sequences.

Justification for an assumption of finite sequences can be found in one of two ways. First, we may assume that only the m stimuli prior to time t can have any appreciable effect on the activity state of the A-system at time t . In this case, we need consider only sequences of length m as possible determinants of $a_i^*(t)$. Note that this assumption applies only to

the activity state of the system, and not to the values of the connections or memory state of the network, which clearly depends on all prior time. Such an assumption appears to be supported by the analyses of the last chapter, which show that for suitable parameters, only the most recent stimuli affect the activity state of the system at time t, progressively more remote stimuli making a progressively smaller contribution, which soon becomes negligible. Specifically, it has been shown that with suitable parameters, it makes no significant difference to assume that the sequence began at time $t - m\tau$, rather than at some earlier time, which is equivalent to the assumption of a finite universe of sequences of length m, in place of the universe of infinite sequences.

An alternative approach, for which a rigorous analysis rather than a mere approximation is possible, is the following: Assume that the activity of the A-units is "quenched" after every m stimuli; i.e., the perceptron is shown only sequences of length m, and at the end of each such sequence, its activity is interrupted by setting all $a_i^* = 0$, so that the next sequence begins with the perceptron in a "silent" state, as required. Let us analyze the performance of such a perceptron (for which the dimension of the phase space is finite) and then let m approach infinity. The limiting behavior of such a system should correspond to a perceptron in which the sequences are uninterrupted. For specificity, and to permit a rigorous analysis, this type of interrupted-activity system will be assumed in the following analysis, although it will be shown later that the results can be extended to a more general case.

In keeping with the above assumption, it will be assumed that there are a total of N possible subsequences which comprise the preconditioning sequence of the perceptron, symbolized $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N$. The phase space therefore has dimension $2N$, and it is assumed that no stimulus

sequence (i.e., no subsequence) has more than m members (where m is finite). By selecting both η and δ sufficiently small, it can be guaranteed that the change in the memory state of the perceptron during a single sequence of length m is negligible, or infinitesimal, so that the output signal $a_i^*(\mathcal{J}_k)$ depends only on $\mathcal{J}_k$ and the memory state of the system at the start of the sequence, and does not depend on changes in the memory state which occurred during the sequence $\mathcal{J}_k$ itself.

19.4 General Analysis: The Time-Dependent Equation

Given the probability density over the phase space of the A-units at time t, it is possible to obtain the Q-functions $Q_{i_\mu j_\nu} = Q_{ij}$ for any pair of sequences (of length μ and ν, respectively) by integrating the probability density over the region of phase space for which $a^*(\mathcal{J}_i)\, a^*(\mathcal{J}_j) = 1$ That is, we integrate over the region for which $\alpha^i \geq \theta$ and $\alpha^j \geq \theta$ The subscript denoting particular A-units is suppressed here, since we are concerned only with the density of such A-units, and not with their individual identity.

The object of a general analysis of the evolution of the association system in such a perceptron is to describe the "flow" of A-units in thi s phase space, so as to obtain the density function at time t as a function of the initial distribution and the stimulus sequences to which the perceptron has been exposed. The system can be represented by a sort of hydrodynamic model; the probability density in the phase space is treated as a sort of compressible fluid, in which convection phenomena occur, but in which there is no diffusion, since it will be seen that the A-units which initially occupy a given point in phase space will always move together, in unison,

rather than following unique paths. Throughout this analysis, it will be assumed that we are dealing with finite stimulus sequences (as described in Section 19.3), and that the rate of flow (the length of the velocity vector) for all points in the phase space is infinitesimal over the duration of the longest sequence. The history of the perceptron, then, consists of an endless sequence of such finite sub-sequences, so that at a given point in time, the perceptron can be assumed to be exposed to a mixture of all possible sequences, each weighted according to its probability. The velocity vector for a given point in phase-space at time t then depends on the combination of velocity components contributed by each of the stimulus sequences to which the perceptron is exposed.

We have seen that each A-unit, a_i, is characterized by a set of coordinates in phase space at time t, namely $(\beta_i^1, \beta_i^2, \ldots, \beta_i^N, \gamma_i^1, \gamma_i^2, \ldots, \gamma_i^N)$. For the given A-unit, the β-components are fixed for all time, while the γ-components depend on t. Thus, to follow the history of this A-unit (or point in phase space) we must determine the velocity vector $\dot{\gamma}_i = (\dot{\gamma}_i^1, \dot{\gamma}_i^2, \ldots, \dot{\gamma}_i^N)$ as a function of time for the point (β_i, γ_i).

We consider first the effect of the reinforcement which occurs for the last stimulus in a sequence S_q upon the component γ_i^r. To be specific, suppose sequence S_q occurs at time t, and S_r occurs at $t + \Delta t$, and assume the transmission time $\tau << \Delta t$. Then the (infinitesimal) change in γ_i^r due to having reinforced the last stimulus in sequence S_q at time t will be denoted by $\Delta_q^r(\beta_i, \gamma_i(t))$. It is a function of the location of the point in phase space whose motion is being

traced, at time t . Note that although only the effect due to the last stimulus of the sequence $\mathcal{S}_q$ is considered, all abbreviated sequences are present among the N possible sequences, so that if we know the effect of reinforcing the terminal stimulus in each case, the effect of all possible reinforcements can be calculated.

A notation for the sequence corresponding to $\mathcal{S}_q$ with its terminal member omitted (i.e., the sequence $\mathcal{S}_q$ abbreviated by one stimulus) will be required. We shall use the symbol $\mathcal{S}_{q'}$ to denote such an abbreviated sequence. The change in the memory state due to the last stimulus of sequence $\mathcal{S}_q$ is then attributable to the modification of the values of those connections which originate in the set of A-units which respond to $\mathcal{S}_{q'}$ and which terminate in the set of A-units responding to $\mathcal{S}_q$. From equation (19.1) we see that each such connection gains a quantity of value $(\eta - \sigma\nu)\,\Delta t$, while all other connections lose a quantity $-\sigma\nu\,\Delta t$

Figure 52 illustrates the relationship of the A-unit sets which are involved in this transaction, and shows the increments to γ^r which result from the occurrence of $\mathcal{S}_q$ at time t . The sets responding at time t and $t-\tau$ are designated $A_q(t)$ and $A_{q'}(t)$, respectively. The set $A_{r'}(t+\Delta t)$ is the set responding to the preterminal stimulus of sequence $\mathcal{S}_{r'}$. The measures of these sets are $Q_q(t)$, $Q_{q'}(t)$ and $Q_{r'}(t+\Delta t)$. Since it was assumed that all A-units are interconnected, the measure of the set of connections for which $\Delta\nu = (\eta - \sigma\nu)\Delta t$ is $Q_{q'}(t)$ for $a_i \in A_q(t)$, and the measure of the set of connections for which $\Delta\nu = -\sigma\nu\Delta t$ is $1 - Q_{q'}$. If $a_i \notin A_q(t)$, all of its input connections lose $-\sigma\nu\Delta t$. But we are particularly interested in the change in γ_i^r,

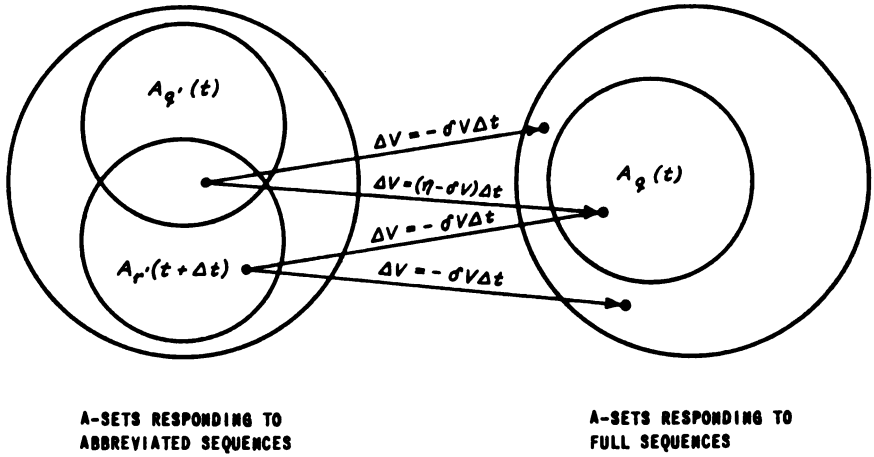


Figure 52 EFFECT OF REINFORCING SEQUENCE J_q UPON γ^r

which is the sum of the changes of value for all connections originating in the set $A_{r'}(t+\Delta t)$, and terminating on the arbitrary unit a_i, whose coordinates are (β_i, γ_i). These connections can be divided into three subsets:

(1) Connections which originate from the intersection $A_{q'}(t) \cap A_{r'}(t+\Delta t)$ and terminate in $A_q(t)$ change by $(\eta - \sigma v)\Delta t$

(2) Connections which originate from the set $A_{r'}(t+\Delta t) - \{A_{q'}(t) \cap A_{r'}(t+\Delta t)\}$ and terminate in $A_q(t)$ change by $-\sigma v \Delta t$.

(3) All connections which originate from the set $A_{r'}(t+\Delta t)$ and terminate outside of $A_q(t)$ change by $-\sigma v \Delta t$.

Now let us consider the difference equation

$$\Delta_q^r(\beta_i, \gamma_i(t)) = \gamma_i^r(t+\Delta t) - \gamma_i^r(t) \tag{19.3}$$

for the A-unit a_i whose location at time t is $(\beta_i, \gamma_i(t))$. Since $\gamma_i^r = \sum_{a_j \epsilon A_{r'}} v_{ji}$, we can make the substitutions:

$$\gamma_i^r(t) = \sum_{a_j \epsilon A_{r'}(t)} v_{ji}(t)$$

$$\gamma_i^r(t+\Delta t) = \sum_{a_j \epsilon A_{r'}(t+\Delta t)} v_{ji}(t+\Delta t) = \sum_{a_j \epsilon A_{r'}(t)} v_{ji}(t) + \sum_{a_j \epsilon A_{r'}(t)} \Delta v_{ji}$$

$$+ \sum_{a_j \epsilon \{A_{r'}(t+\Delta t) - A_{r'}(t)\}} v_{ji}(t+\Delta t) - \sum_{a_j \epsilon \{A_{r'}(t) - A_{r'}(t+\Delta t)\}} v_{ji}(t+\Delta t)$$

Making these substitutions yields:

$$\Delta_q^r(\beta_i, \gamma_i(t)) = \sum_{a_j \epsilon A_{r'}(t)} \Delta v_{ji}(t) + \sum_{a_j \epsilon \Delta A_{r'}} \pm v_{ji}(t+\Delta t) \tag{19.4}$$

where $\Delta A_{r'} = \{A_{r'}(t+\Delta t) - A_{r'}(t)\} + \{A_{r'}(t) - A_{r'}(t+\Delta t)\}$, that is, the set of A-units added or subtracted from the set $A_{r'}(t)$ during the period Δt . The first sum represents the change of value of the set of connections

which originate in $A_{r'}(t)$ and are reinforced at time t due to sequence s_q. This change in value is readily obtained from the components listed above, and is given by

$$\sum_{a_j \in A_{r'}(t)} \Delta v_{ji}(t) = \begin{cases} \left[N_a \eta \, Q_{q'r'}(t) - \sigma \sum_{a_j \in A_{r'}(t)} v_{ji}(t) \right] \Delta t & \text{for } a_i \in A_q(t) \\ -\sigma \Delta t \sum_{a_j \in A_{r'}(t)} v_{ji}(t) & \text{for } a_i \notin A_q(t) \end{cases}$$

which may be combined in the form

$$\sum_{a_j \in A_{r'}(t)} \Delta v_{ji}(t) = \left[N_a \eta \, Q_{q'r'}(t) \, \phi\left(\beta_i^q + \gamma_i^q(t)\right) - \sigma \gamma_i^r(t) \right] \Delta t \qquad (19.5)$$

where, as before, $\phi(\alpha) = 1$ for $\alpha \geq \theta$, and 0 otherwise, and $\gamma_i^r(t)$ has been substituted for $\sum_{a_j \in A_{r'}(t)} v_{ji}(t)$.

The second sum in (19.4) represents the value of the set of connections which originate from the incremental set, $\Delta A_{r'}$. For this sum, it will be convenient to substitute the symbol $\Delta_q^* \gamma_i^r(t)$. Thus, (19.4) becomes

$$\Delta_q^r\left(\beta_i, \gamma_i(t)\right) = \left[N_a \eta \, Q_{q'r'}(t) \, \phi\left(\beta_i^q + \gamma_i^q(t)\right) - \sigma \gamma_i^r(t) \right] \Delta t + \Delta_q^* \gamma_i^r(t) \qquad (19.6)$$

where the subscript i indicates that the subscripted variable is a component of the vector (β, γ) for the unit a_i.

Now suppose each possible "conditioning sequence", $\mathcal{S}_q$, occurs with a probability P_q , and that a statistically uniform mixture of all such sequences occurs at time t . This supposition is justified by our assumption that the length of each sequence is infinitesimal, relative to the rate of change in the memory-state of the perceptron. In that case, we obtain from (19.6)

$$\Delta^r(\beta_i, \gamma_i(t)) = \sum_q P_q \Delta_q^r(\beta_i, \gamma_i(t)) \qquad (19.7)$$

$$= N_a \eta \Delta t \left[\sum_q P_q Q_{q',r'}(t) \phi\left(\beta_i^q + \gamma_i^q(t)\right) \right] - \delta \Delta t \, \gamma_i^r(t) + \Delta^* \gamma_i^r(t)$$

where $\Delta^* \gamma_i^r(t)$ = value added or subtracted due to connections originating from the combined incremental set due to all $\mathcal{S}_q$. If we now divide both sides by Δt and allow Δt to approach zero, we obtain the differential equation for the velocity component $\dot{\gamma}^r(t)$ for the unit a_i ;

$$\frac{d\gamma_i^r(t)}{dt} = N_a \eta \sum_q P_q Q_{q',r'}(t) \phi\left(\beta_i^q + \gamma_i^q(t)\right) - \delta \gamma_i^r(t) + \frac{d^*\gamma_i^r(t)}{dt} \qquad (19.8)$$

where $\frac{d^*\gamma_i^r}{dt} = \lim_{\Delta t \to 0} \frac{\Delta^* \gamma_i^r(t)}{\Delta t}$. *

Note that the quantity $\Delta^* \gamma_i^r(t)$ is zero except at those times that new A-units are added to the set $A_{r'}(t)$, since it represents the sum of the values in the incremental set $\Delta A_{r'}$. Again, we note that for sequences of length 2 or less, the set $A_{r'}(t)$ never changes, since new units can be added to the set only if $\phi(\alpha^{r'})$ changes from 0 to 1, and for

---

* Strictly speaking, this is either zero, or fails to exist. However, this expression will be restated below in terms of delta-functions (see Equation 19.9).

sequences of length 2, $\phi(\alpha^{r'}) = \phi(\beta^{r'})$, which is constant. Similarly, for sequences of length 2 or less, $Q_{q',r'}(t)$ is constant. Consequently, for these conditions, the equation (19.8) is equivalent to (16.11), except that $Q_{q',r'}$ takes the place of $Q_{q',r}$. In the general case, however, $d^* \gamma_i^r(t)/dt$ is not always zero; at those times that new A-units are added to the set $A_{r'}$, an unknown increment to the value of γ_i^r occurs, which depends upon the values of the connections from those units whose $\alpha^{r'}$ has just become equal to θ. This quantity is exceedingly difficult to calculate, as it depends upon detailed correlation of the β-vectors for the new transmitting units and the β-vector for the receiving unit, a_i. Fortunately, it can be shown that the steady-state solution to (19.8) does not depend upon the actual value of the last term, even though it affects the rate of convergence to the steady-state condition.

In the general case, the solution of (19.8) is discontinuous, unlike the solution of (16.11), which was always continuous despite its discontinuous derivative. From the above discussion as to the nature of $\Delta^* \gamma_i^r(t)$, it becomes clear that (19.8) can be rewritten in terms of Dirac delta-functions:

$$\frac{d\gamma_i^r}{dt}(t) = N_a \eta \sum_q P_q Q_{q',r'}(t) \phi(\beta_i^q + \gamma_i^q(t)) - \sigma \gamma_i^r(t) + \sum_k \sigma(t - t_k^*) \Delta^* \gamma_i^r(t_k^*) \quad (19.9)$$

where t_k^* is any time at which one or more of the $\phi(\alpha_j^{r'})$ changes from 0 to 1 or vice versa.

19.5 Steady State Solutions

Consider the equilibrium equation corresponding to (19.9). If an equilibrium exists at time t, then no $\phi(\alpha)$ can change its value at time t, and thus the last term of (19.9) is zero at this time. Thus, a steady state solution must correspond to a solution to the equation

$$\frac{d\gamma_i^r(\infty)}{dt} = N_a \eta \sum_q P_q Q_{q'r'}(t)\, \phi\left(\beta_i^q + \gamma_i^q(\infty)\right) - \sigma \gamma_i^r(\infty) = 0 \tag{19.10}$$

which gives

$$\gamma_i^r(\infty) = \frac{N_a \eta}{\sigma} \sum_q P_q Q_{q'r'}(\infty)\, \phi\left(\beta_i^q + \gamma_i^q(\infty)\right) \tag{19.11}$$

or, substituting for $Q_{q'r'}$,

$$\gamma_i^r(\infty) = \frac{N_a \eta}{\sigma} \sum_q \left[P_q \phi\left(\beta_i^q + \gamma_i^q(\infty)\right) \sum_j \frac{1}{N_a} \phi\left(\beta_j^{q'} + \gamma_j^{q'}(\infty)\right) \phi\left(\beta_j^{r'} + \gamma_j^{r'}(\infty)\right) \right] \tag{19.12}$$

Note that the terminal vector (β, γ_∞) of an A-unit (in a given system) depends only on the starting vector (β, γ_o) so that we can also write in place of (19.12),

$$\gamma_i^r(\infty) = \frac{N_a \eta}{\sigma} \sum_q \left[P_q \phi\left(\beta_i^q + \gamma_i^q(\infty)\right) \sum_{(\beta, \gamma_o)} P(\beta, \gamma_o) \phi\left(\beta^{q'} + \gamma^{q'}(\infty)\right) \phi\left(\beta^{r'} + \gamma^{r'}(\infty)\right) \right] \tag{19.13}$$

where $P(\beta, \gamma_o)$ is the probability that an A-unit is initially situated at the point (β, γ_o) in the phase space. Thus, in this form, the steady-state solution requires no knowledge of the individual A-units and their connections,

but depends only on the initial point-mass distribution over the phase space. The corresponding time-dependent differential equation represents the velocity vector for an element of probability-mass in this phase space.

Now a possible solution of (19.13) can be found by the following iterative procedure: Assume that initially, the values of all A-A connections are zero, so that $\gamma_0 = 0$ for all units, and (19.13) depends only on the β -vectors. Begin by inserting $\gamma_0 = 0$ for all γ's on the right-hand side of (19.13), and compute the resulting approximation for $\gamma_i^r(\beta, \gamma_\infty)$, for all possible β -vectors (or for all units, a_i). The first approximation for γ_∞ is then inserted on the right-hand side, to obtain the next approximation, etc. If we let $\gamma_{(z)}$ represent the result of the z^{th} iteration, we have

$$\gamma_{i(z+1)}^r = \frac{N_a \eta}{\sigma} \sum_q \left[P_q \phi\left(\beta_i^q + \gamma_{i(z)}^q\right) \sum_{\beta_j} P(\beta_j)\, \phi\left(\beta_j^{q'} + \gamma_{j(z)}^{q'}\right) \phi\left(\beta_j^r + \gamma_{j(z)}^{r'}\right) \right] \qquad (19.14)$$

We will now attempt to show that this iteration must converge in a finite number of steps to the solution of the differential equation (19.9), for equivalent initial conditions.

We first show that the iteration process itself converges in less than $N_\beta N$ steps (where N = the number of stimulus sequences, and N_β = the number of β -vectors for which $P(\beta) > 0$) . On the first iteration, it is clear that the γ's can only increase, since they start out from zero, and are set equal to a non-negative quantity. But introducing this quantity for the next iteration can only increase the ϕ's from zero to 1; it cannot cause any ϕ to decrease. Consequently, on the next iteration, the γ's can again only

increase, and similarly for each subsequent iteration. Since γ^r is non-decreasing, $\phi(\beta^r + \gamma^r)$ is non-decreasing, for all r. But γ^r can change only when some ϕ changes, and each ϕ can change at most once (from 0 to 1). But there are at most $N_\beta N$ ϕ-functions, $\phi(\alpha_i^r)$. If all of these are initially zero, the system is already at a solution, and no further changes will occur. Therefore, at most $n < N_\beta N$ ϕ-functions can change, and the process must converge in less than $N_\beta N$ iterations.

Let the end result of this process be $\gamma_i^{r^*}$ for any unit a_i. We now wish to prove that $\gamma_i^{r^*}$ is a solution of the differential equation (19.9).

To begin with, we prove that γ^{r^*} is a minimal solution of the equilibrium equation (19.13).

Let $\tilde{\gamma}_i^r$ be any solution of the equilibrium equation. Then for the iteration process, we have $\gamma_i^r(0) \leq \tilde{\gamma}_i^r$ for all r and all β_i. Since the right-hand side of (19.13) is a monotone non-decreasing function of γ_i^r, we have

$$\gamma_{i(1)}^r = \frac{N_a \eta}{\delta} \sum_q \left[P_q \phi\left(\beta_i^q + \gamma_{i(0)}^q\right) \sum_{\beta_j} P(\beta_j) \phi\left(\beta_j^{q'} + \gamma_{j(0)}^{q'}\right) \phi\left(\beta_j^{r'} + \gamma_{j(0)}^{r'}\right) \right]$$

$$\leq \frac{N_a \eta}{\delta} \sum_q \left[P_q \phi\left(\beta_i^q + \tilde{\gamma}_i^q\right) \sum_{\beta_j} P(\beta_j) \phi\left(\beta_j^{q'} + \tilde{\gamma}_j^{q'}\right) \phi\left(\beta_j^{r'} + \tilde{\gamma}_j^{r'}\right) \right] = \tilde{\gamma}^r(\beta_i)$$

Similarly, $\gamma_{i(n)}^{r} \leq \tilde{\gamma}_i^{r}$, and hence $\gamma_i^{r*} \leq \gamma_i^{r}$. Hence γ_i^{r*} is minimal.

Now consider the differential equation, (19.9). As long as no ϕ changes value, all δ functions are zero, and (19.9) simplifies to

$$\frac{d\gamma_i^r}{dt} = N_a \eta \sum_q P_q Q_{q'r'}(t)\, \phi\left(\beta_i^q + \gamma_i^q(t)\right) - \delta\, \gamma_i^r(t)$$

$$= N_a \eta \sum_q P_q\, \phi\left(\alpha_i^q(t)\right) \sum_{\beta_j} P(\beta_j)\, \phi\left(\alpha_j^{q'}(t)\right) \phi\left(\alpha_j^{r'}(t)\right) - \delta\, \gamma_i^r(t)$$

where $\alpha_j^r(t) = \beta_j^r + \gamma_j^r(t)$. Thus, while the ϕ's are constant, the differential equation is of the form $\frac{d\gamma}{dt} = M - \delta\gamma$, where

$$M = N_a \eta \sum_q \left[P_q\, \phi(\alpha_i^q) \sum_{\beta_j} P(\beta_j)\, \phi(\alpha_j^{q'})\, \phi(\alpha_j^{r'}) \right]$$

Thus, during this time, there is an exponential approach to the limit M/δ , analogous to the solution discussed in Chapter 16 (pg. 355). Now suppose at time t_l^* one of the ϕ's changes. At this point, the last term in (19.9) is infinite, and the solution is discontinuous, since the value of the connections from the incremental set $\Delta A_{r'}$ has just been added to γ_i^r Consequently, the solution takes the form shown in Figure 53.

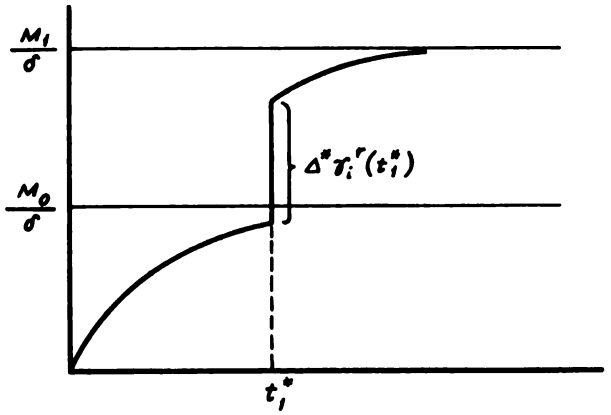


Figure 53 FORM OF SOLUTION FOR CROSS-COUPLED α-SYSTEMS

where

$$M_i = N_a \eta \sum \left[P_\gamma \phi\left(\alpha^\delta(t_1^*)\right) \sum P(\beta_j)\, \phi\left(\alpha_j^{\delta'}(t_1^*)\right) \phi\left(\alpha_j^{r'}(t_1^*)\right) \right],$$

$$\alpha_j^\Delta(t_1^*) = \beta_j^\Delta + \left[\frac{M_0(\gamma_j^\Delta)}{\delta} - e^{-\delta t}\, \frac{M_0(\gamma_j^\Delta)}{\delta} \right] + \Delta^*\, \gamma_j^\Delta(t_1^*)$$

The middle term of this expression represents the value of γ_j^Δ at time $t_1^* - dt$, just prior to the discontinuity. The magnitude of $\Delta^* \gamma_i^r$ remains unknown, but we know that it must be non-negative, since it consists of values of A-unit interconnections which began at zero and can only have changed in a positive direction. As in the case of the iterative

process, there are at most $N_\beta N$ times, t_k^*, at which these discontinuities can occur, and each new limit $\frac{M_k}{\delta} \geq \frac{M_{k-1}}{\delta}$. Moreover, the solution remains monotone increasing, despite its discontinuities. This last conclusion can be seen from the fact that the increment $\Delta^* \gamma_i^r$ comes from the values of a set of connections whose origins are now active for one more stimulus sequence than previously. Since no previously active A-units have become inactive (all ϕ's being monotone increasing) the values of these connections will not diminish, and will, in fact, tend to increase. Thus the new limit for γ_i^r can be no lower than its present value.

Now consider the first step of the iterative process. This yields for $\gamma_{i(1)}^r$ the value of the first asumptotic level, M_0/δ, for all γ_i^r in the differential equation. This means that if any ϕ changes in the differential equation prior to reaching the level M_0/δ, this ϕ must also change in the first step of the iterative process. (If no ϕ changes prior to the level M_0/δ then no ϕ will ever change, and we are at a solution for both equations*). But the new level, M_1/δ, is a positive monotonic function of the ϕ's, and the next step of the iteration process, $\gamma_{i(2)}^r$, corresponds to the level M_1/δ which would have resulted had every γ actually attained its asymptotic level M_0/δ. Thus $\gamma_{i(2)}^r \geq \gamma_i^r(t_1^*)$ for every r. But from the same argument, it follows that $\gamma_{i(3)}^r \geq \gamma_i^r(t_2^*)$, and in general, $\gamma_{i(k)}^r \geq \gamma_i^r(t_{k-1}^*)$ Consequently, $\gamma_i^{r*} \geq \gamma_{i(\infty)}^r$, and the solutions of the two equations are indeed identical.

---

* It is assumed that M is not identically equal to θ, in which case the solutions might coincide only for $t = \infty$

19.6 Analysis of Finite-Sequence Environments

The term "finite-sequence environment" will be used for any system in which the stream of activity is periodically interrupted, either by actively setting all a_i^* to zero, or by introducing sequences of null stimuli of sufficient duration to allow all A-unit activity to die out of its own accord. The latter possibility exists only for systems in which the internal connection values are sufficiently small, or contain a sufficient inhibitory component, to guarantee that activy will, in fact, die away. Some idea of the conditions for this to occur may be gained from Section 18.2, and Figure 47. For convenience (and because it can always be realized, regardless of choice of parameters) the interrupted activity model will be considered here. In either case, finite-sequence environments are directly analyzeable by the method of Section 19.5. Several examples are given here, based on the same stimulus environment as in Experiment 12. It will be recalled that this consisted of four stimuli, with areas $R = .2$, and intersections C_{13} and $C_{24} = .1$, all other intersections being zero. As in the example in Chapter 17, we will consider a binomial perceptron with parameters $\alpha = 3$, $\gamma = 0$, and $\theta = 2$, for all A-units.

EXAMPLE 1: Suppose the preconditioning sequence consists of an endless repetition of the subsequence: $S_1 S_2 S_3 S_4 \;/\; S_1 S_2 S_3 S_4 \;/ S_1 S_2 S_3 S_4 /\ldots$, where the symbol / is used to indicate points at which activity is interrupted. Then for this environment there are actually four possible sequences to be considered in the analysis, namely

$$\mathcal{S}_1 = (S_1)$$
$$\mathcal{S}_2 = (S_1 S_2)$$
$$\mathcal{S}_3 = (S_1 S_2 S_3)$$
$$\mathcal{S}_4 = (S_1 S_2 S_3 S_4)$$

each occurring with probability $P_i = .25$. The β -vectors for these four sequences correspond to the signals received from the terminal stimulus in each sequence, and are listed in Table 9, together with their probabilities. β -vectors consisting only of 1's and 0's represent A-units which will always remain inactive.

The initial Q-matrix for this experiment is precisely the same as that found for the corresponding terminal stimuli in Chapter 17, namely,

$$Q_{(0)} = \begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .104 & .000 & .034 \\ .034 & .000 & .104 & .000 \\ .000 & .034 & .000 & .104 \end{pmatrix}$$

It is found that no change occurs in this matrix for $N_e \eta / \delta < 117.6$. Let us therefore consider the case in which $N_e \eta / \delta = 160$. In the open-loop system of Chapter 17, the sequence of Experiment 12 yielded the terminal Q-matrix:

$$Q_{(\infty)} = \begin{pmatrix} .210 & .176 & .034 & .072 \\ .176 & .210 & .072 & .034 \\ .034 & .072 & .210 & .176 \\ .072 & .034 & .176 & .210 \end{pmatrix}$$

If we now compute the terminal matrix for a fully cross-coupled system, from Equation (19.14), we obtain:

$$Q_{(\infty)} = \begin{pmatrix} .104 & .000 & .034 & .000 \\ .000 & .152 & .000 & .130 \\ .034 & .000 & .104 & .000 \\ .000 & .130 & .000 & .152 \end{pmatrix}$$

TABLE 9

β-VECTORS FOR STIMULI OF EXPERIMENT 12

(Parameters of A-units: $\alpha = 2$, $\gamma = 0$)

β	$P(\beta)$
0000	.064
0001	.048
0010	.048
0100	.048
1000	.048
0011	.024
0110	.024
1001	.024
1100	.024
0101	.072
1010	.072
0111	.030
1011	.030
1101	.030
1110	.030
1111	.036
0003	.001
0030	.001
0300	.001
3000	.001
0303	.001
3030	.001
0002	.012
0020	.012
0200	.012
2000	.012
0202	.018
2020	.018
0201	.027
0102	.027
2010	.027
1020	.027
0203	.003
0302	.003
2030	.003
3020	.003
0012	.003
0021	.003
0120	.003
0210	.003
1200	.003
2100	.003
1002	.003
2001	.003
0103	.003
0301	.003
1030	.003
3010	.003
0212	.003
2120	.003
0121	.003
1210	.003
2021	.003
1202	.003
1012	.003
2101	.003
1212	.003
2121	.003
1112	.006
1121	.006
1211	.006
2111	.006
0112	.006
0211	.006
1021	.006
2011	.006
1102	.006
1201	.006
1120	.006
2110	.006

The only change which occurs in this case is that the set A_2 gains a larger intersection with the set A_4 . There is no tendency here for the A-sets responding to adjacent pairs of stimuli to merge, as would be the case in a four-layer model, or an open-loop cross-coupled network with zero transmission times. This is shown even more strikingly in the following example.

EXAMPLE 2: For the same parameters as Example 1, let us extend the basic subsequence to 8 stimuli, using as the preconditioning sequence:

$$S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4 / S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4 / \ldots .$$

The sequences for this environment are now

$$\mathcal{S}_1 = (S_1) \qquad \mathcal{S}_5 = (S_1 S_2 S_1 S_2 S_3)$$
$$\mathcal{S}_2 = (S_1 S_2) \qquad \mathcal{S}_6 = (S_1 S_2 S_1 S_2 S_3 S_4)$$
$$\mathcal{S}_3 = (S_1 S_2 S_1) \qquad \mathcal{S}_7 = (S_1 S_2 S_1 S_2 S_3 S_4 S_3)$$
$$\mathcal{S}_4 = (S_1 S_2 S_1 S_2) \qquad \mathcal{S}_8 = (S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4)$$

Each sequence occurs with probability $P_j = .125$. The initial Q-matrix again depends only on the terminal stimuli, and takes the form:

$$Q_{(0)} = \begin{pmatrix} .104 & .000 & .104 & .000 & .034 & .000 & .034 & .000 \\ .000 & .104 & .000 & .104 & .000 & .034 & .000 & .034 \\ .104 & .000 & .104 & .000 & .034 & .000 & .034 & .000 \\ .000 & .104 & .000 & .104 & .000 & .034 & .000 & .034 \\ .034 & .000 & .034 & .000 & .104 & .000 & .104 & .000 \\ .000 & .034 & .000 & .034 & .000 & .104 & .000 & .104 \\ .034 & .000 & .034 & .000 & .104 & .000 & .104 & .000 \\ .000 & .034 & .000 & .034 & .000 & .104 & .000 & .104 \end{pmatrix}$$

For the terminal matrix (again with $N_a \eta/\theta = 160$) we now have

$$Q_{(\infty)} = \begin{pmatrix} .104 & .000 & .104 & .000 & .104 & .000 & .104 & .000 \\ .000 & .174 & .000 & .174 & .000 & .174 & .000 & .174 \\ .104 & .000 & .174 & .000 & .174 & .000 & .174 & .000 \\ .000 & .174 & .000 & .174 & .000 & .174 & .000 & .174 \\ .104 & .000 & .174 & .000 & .174 & .000 & .174 & .000 \\ .000 & .174 & .000 & .174 & .000 & .174 & .000 & .174 \\ .104 & .000 & .174 & .000 & .174 & .000 & .174 & .000 \\ .000 & .174 & .000 & .174 & .000 & .174 & .000 & .174 \end{pmatrix}$$

This corresponds to an oscillating condition, in which each A-unit (after giving its original unaltered response to the first stimulus of the sequence) responds either 1, 0, 1, 0, 1, 0, 1 or 0, 1, 0, 1, 0, 1, 0 to the remaining seven stimuli of the sequence.

In contrast to previous models, there appears to be a failure to associate successive stimuli, and an association of every alternate stimulus instead. Actually, appearances are misleading here; a strong association of successive stimuli is masked by the appearance of these stimuli in the test sequence (which is identical, in this experiment, with the preconditioning sequence). In other words, the perceptron "predicts" the A-set for the next stimulus at precisely the time that this stimulus actually appears, and consequently the effect of the prediction is not detected. The following experiment reveals these "hidden associations" in a striking fashion.

EXPERIMENT 13: Using the same four stimuli as in Experiment 12, the perceptron is shown the preconditioning sequence S_1, S_2, S_3, S_4 / S_1, S_2, S_3, S_4/.... . It is then tested with the sequence $S_1, 0, 0, 0 \ldots$, and the Q-matrix for all subsequences (from both preconditioning and test sequences) is obtained.

If this experiment is performed with $N_a \eta/\delta = 100$, and all other parameters as before, it is found that on presenting the test sequence $(S_1, 0, 0, \ldots)$ the perceptron recapitulates the identical sequence of active sets A_1, A_2, A_3, A_4 which would have been activated had the preconditioning sequence occurred in full. After A_4, the system lapses into inactivity, since the preconditioning sequence is interrupted at this point.*

19.7 Analysis of Continuous Periodic Environments

Up to this point, it has been assumed that the activity of the perceptron is interrupted at least once every m stimuli. We now turn to the case of a continuous, unbroken sequence of stimuli, where the activity of the association system is allowed to run on without interruption. To begin with, the case of a periodic stimulus sequence will be considered, where the preconditioning sequence takes the form:

$$S_1 S_2 S_3 \ldots S_m S_1 S_2 S_3 \ldots S_m \ldots$$

the period of the sequence being m. Such an environment can be considered as being composed of a set of m subsequences, each of length $m+1$. Specifically, we have the subsequences:

$$\begin{aligned} \mathcal{S}_1 &= (S_1 S_2 S_3 \ldots S_m S_1) \\ \mathcal{S}_2 &= (S_2 S_3 \ldots S_m S_1 S_2) \\ &\vdots \\ \mathcal{S}_m &= (S_m S_1 S_2 S_3 \ldots S_m) \end{aligned}$$

---

* This "hallucinatory recall" effect, in which the perceptron, cued by the initial stimulus of the sequence, reproduces the identical sequence of internal states which would have been activated had the stimuli continued in their usual order, is suggestive of some of Penfield's observations on hallucinatory recall of stereotyped sequences induced by electrical stimulation of brain foci in epileptics (Ref. 68).

Each sequence occurs with probability $1/m$, and each sequence begins and ends with the same stimulus.

Now since the preconditioning sequence is assumed to extend indefinitely into the past, at any arbitrary time t, the antecedent sequence for the first and last stimulus of any $(m+1)$-subsequence is the same; consequently $\gamma_i^1 = \gamma_i^m$ for all t. But this means that there are, in fact, only a finite number (m) of γ's for any A-unit, a_i, so that the steady-state value of γ_i^r can be computed exactly by equation (19.14), where the sequence s_r, is interpreted to mean the sequence s_{r-1} in the set of m subsequences specified above.

Several special cases are of particular interest. Consider first the case of a steadily maintained stimulus, $(S, S, S, \ldots)$. Substituting in (19.14), we have

$$\gamma'_{i(\gamma+1)}^{(\infty)} = \frac{N_a \eta}{\delta} \phi(\beta_i' + \gamma_{i(\gamma)}) \sum_{\beta_j} P(\beta_j) \phi(\beta_j' + \gamma'_{j(\gamma)})$$

and it is readily seen that the set of active units can never change from the initial set, since this equation yields zero unless $\phi(\beta_i') = 0$ for the first iteration. Thus for a steady stimulus, we have

$$Q_{ii}^{(\infty)} = Q_{ii}(0)$$

Next, consider the alternating sequence $S_1 S_2 S_1 S_2 \ldots$. In this case, (19.14) takes the form

$$\gamma_i^r(z+1) = \frac{N_a \eta}{2\sigma}\left[\phi\left(\beta_i^1 + \gamma_i^1(z)\right)\sum_{\beta_j} P(\beta_j)\,\phi\left(\beta_j^2 + \gamma_j^2(z)\right)\phi\left(\beta_j^{r'} + \gamma_j^{r'}(z)\right)\right.$$

$$\left. + \phi\left(\beta_i^2 + \gamma_i^2(z)\right)\sum_{\beta_j} P(\beta_j)\,\phi\left(\beta_j^1 + \gamma_j^1(z)\right)\phi\left(\beta_j^{r'} + \gamma_j^{r'}(z)\right)\right]$$

In this case, if either $\phi(\beta_i^1)$ or $\phi(\beta_i^2) = 1$, γ_i^r will generally be non-zero, and the system will tend to form a union of the sets initially responding to S_1 and S_2 (provided $Q_{12}(0) \neq 0$).

Finally, consider the stimulus sequence of Experiment 12, consisting of a period of alternation of S_1 and S_2 followed by an alternation of S_3 and S_4 , as described in Chapter 17. Rather than compute the entire 20 by 20 Q-matrix for Experiment 12, we present here a "miniaturized version" of this experiment, based upon the eight-stimulus sequence employed in Example 2 of the preceding section. For the continuous environment, the eight sequences will be:

$$\mathcal{J}_1 = (S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4 S_1) \qquad \mathcal{J}_5 = (S_3 S_4 S_3 S_4 S_1 S_2 S_1 S_2 S_3)$$

$$\mathcal{J}_2 = (S_2 S_1 S_2 S_3 S_4 S_3 S_4 S_1 S_2) \qquad \mathcal{J}_6 = (S_4 S_3 S_4 S_1 S_2 S_1 S_2 S_3 S_4)$$

$$\mathcal{J}_3 = (S_1 S_2 S_3 S_4 S_3 S_4 S_1 S_2 S_1) \qquad \mathcal{J}_7 = (S_3 S_4 S_1 S_2 S_1 S_2 S_3 S_4 S_3)$$

$$\mathcal{J}_4 = (S_2 S_3 S_4 S_3 S_4 S_1 S_2 S_1 S_2) \qquad \mathcal{J}_8 = (S_4 S_1 S_2 S_1 S_2 S_3 S_4 S_3 S_4)$$

It is found that in this experiment, there is no choice of parameters which will yield an increase in Q_{12}, Q_{34}, Q_{56}, and Q_{78} without producing a corresponding increase in the set of A-units responding jointly to all stimulus sequences. It can also be shown that no matter how far the period of the preconditioning sequence is extended (by increasing the duration of $S_1 S_2$ alternation and also increasing the duration of $S_3 S_4$ alternation) the system will never be able to selectively combine the sets (A_1, A_2) and (A_3, A_4) as in previous models. There is, nonetheless, a "predictive" effect which would be revealed if the stimuli were suddenly cut off, as in Experiment 13.

From this example (and those of the preceding section) it is clear that the condition for selective merging of A-sets for temporally adjacent stimuli is not as easily satisfied as in the four-layer system, or open-loop systems with zero transmission time. Experiment 14, however, illustrates a simple modification of the preconditioning sequence by which such a merger can be obtained.

EXPERIMENT 14: The same four stimuli are employed as in Experiments 12 and 13. The preconditioning sequence, however, takes the form: $S_1 S_1 S_2 S_2 S_1 S_1 S_2 S_2 S_1 S_1 S_3 S_3 S_4 S_4 S_3 S_3 S_4 S_4 S_3 S_3$, repeated ad infinitum. The terminal Q-matrix is obtained as before, for the twenty possible sequences of duration 21.

In this case, it is found that there will be a tendency for the sets A_1 and A_2 to merge, and for the sets A_3 and A_4 to merge in a separate "cell assembly".* What happens here is that the A-units responding to S_1 tend to be associated to the two most common successors of S_1 in the preconditioning sequence: namely, S_1 itself, and S_2. Similarly, S_2 is associated both to S_2 and S_1. Thus, when S_1 occurs at the start of the sequence it tends to be followed (coincident with its second appearance) by the combined set (A_1, A_2). When the first S_2 stimulus appears, A_2 combines with the "predicted" A_1 set, and the combined (A_1, A_2) set tends to persist until the first occurrence of S_3, at which point it may combine with the new A_3 set, or may become inactive, depending upon the magnitude of $N_a \eta/\delta$. In order to prevent the original set from persisting indefinitely (since each A-set tends to predict itself, on the following cycle) $N_a \eta/\delta$ must be kept small enough so that the γ-components alone are insufficient to activate A-units whose β-components are zero. In this case, only part of the original A-sets will be activated in the absence of the actual stimulus, but a bias will still remain in the direction of the desired combination of A-sets.

---

* The term "cell assembly" seems appropriate here, as the sets which are formed in the terminal state of a cross-coupled perceptron bear a close resemblence in organization and functional properties to the cell assembly concept proposed by Hebb, in Ref. 33.

In general, if each stimulus which forms part of an "event" can occur with equal probability after any other stimulus in the same event, then all of the A-sets responding to these stimuli will tend to merge, at least in part, and will be evoked by any stimulus of the event-class. This is essentially the same effect which was found for four-layer perceptrons in Chapter 16.

Actually, with the β-vectors corresponding to those in Table 9, (for A-units with only three retinal connections) the system is not well behaved in Experiment 14 regardless of the choice of threshold and $N_a \eta/\delta$ With larger numbers of connections and the possibility of higher thresholds, however, it seems likely that the desired effect could be obtained with the preconditioning sequence given in the experiment. A γ-perceptron (or a Γ-perceptron) would probably be somewhat better behaved in this experiment, as it would tend to inhibit the sets of A-units characteristic of the first "event" once the second event began. In the α-system, there is a strong tendency for all A-sets to merge whenever $N_a \eta/\delta$ is sufficient to permit the merger of the desired sets,

19.8 Analysis of Continuous Aperiodic Environments

If the preconditioning sequence is not periodic, some sort of approximation procedure must be used, if Equation (19.14) is to be applied. Two possibilities suggest themselves: First, the aperiodic sequence (if it is statistically uniform throughout) can be approximated by a periodic sequence if the period is sufficiently long to encompass all likely juxtapositions and short subsequences of stimuli. Second, we can consider all

subsequences of length m, assigning a probability to each, and analyze the system as though we were dealing with a finite-sequence environment, consisting of the various m-sequences in an appropriate frequency mixture. In this case, the analysis should converge to a correct solution as m becomes large, provided the original sequence is statistically uniform. If the statistical composition of the original preconditioning sequence changes over time, neither of these methods are applicable, and it seems likely that accurate solutions can then be obtained only by actually simulating the system and observing its behavior empirically.

In the experiments which are of primary concern at this time, it is always possible to assume a statistically uniform preconditioning sequence, so that one of the two methods described above can be applied. In practice, this problem is likely to be soluble only for relatively small numbers of stimuli in the environment, as the Q-matrices rapidly become too large to handle in currently available digital computers. For long stimulus sequences and large numbers of stimuli, digital simulation remains the preferred technique, and this offers the additional advantage of being applicable to small perceptrons or systems where the assumption of infinitesimal transmission time is inadmissible. In the preceding examples, where theoretical values (rather than empirical values) of Q_{ij} were used, N_a was implicitly taken to be very large.

19.9 Cross-Coupled Perceptrons with Value-Conservation

The two types of value-conserving systems, γ-systems and Γ-systems, which were considered in section 16.6, are also of interest in cross-coupled systems. The Γ-system, which tends to strengthen connections to the A-set responding to the most likely successor of the present stimulus, while developing inhibitory connections to the A-units responding to unlikely successors, appears to be the more promising of the two. In most environments, however, both systems will probably show similar phenomena, provided transitions between stimuli can occur symmetrically in either direction. The analysis of the γ-system, which is somewhat more familiar from previous work, will be considered first.

19.9.1 Analysis of γ-systems

In the γ-perceptron, the total value of the set of input connections to each A-unit is conserved. Specifically, (assuming the system to be fully coupled) the change in the value of connection c_{ij} is given by

$$\Delta V_{ij} = a_j^*(t)\left[a_i^*(t-\tau) - \frac{1}{N_a}\sum_k a_k^*(t-\tau)\right]\cdot\eta\,\Delta t \qquad (19.15)$$

Instead of (19.19), this leads to the differential equation:

$$\frac{d\,r_i^r}{dt} = N_a\eta\sum_q P_q\left(Q_{q,r}(t) - Q_q(t)\,Q_r(t)\right) \qquad (19.16)$$

$$-\delta\,r_i^r(t) + \sum_k \delta(t-t_k^*)\,\Delta^*\,r_i^r(t_k^*)$$

Since $Q_{g'r'} - Q_{g'} Q_{r'}$ may be negative, the former proof of convergence again breaks down, since γ^r need not be monotonic. As in the case of the four-layer system, the approach will be to try to obtain a time-dependent solution for the γ's. The task is complicated in this case, however, by the presence of the unknown quantities $\Delta^* \gamma_i^r(t)$ in the equation, which we have not hitherto had to evaluate.

For the γ-system, any equilibrium equation must be of the form:

$$\begin{aligned} \gamma_i^r(\infty) &= \frac{N_a \eta}{\delta} \sum_g P_g \phi(\alpha_i^g) \left[Q_{g'r'} - Q_{g'} Q_{r'} \right] \\ &= \frac{N_a \eta}{\delta} \sum_g P_g \phi(\alpha_i^g) \left(\sum_{\beta_j} P(\beta_j) \left[\phi(\alpha_j^{g'}) \phi(\alpha_j^{r'}) - \phi(\alpha_j^{r'}) \sum_k P(\beta_k) \phi(\alpha_k^{g'}) \right] \right) \\ &\equiv \frac{M[\gamma_i^r(A^*)]}{\delta} \end{aligned} \tag{19.17}$$

Where A^* = set of active A-unit sets, A_j, for which the value of $\gamma_i^r(\infty)$ is computed. As long as all ϕ's remain fixed, the γ's will tend exponentially towards such an equilibrium condition, as in previous models. Now consider the set of units whose ϕ's change value at time t_k^*. We wish to find the asymptotic value of the change in γ_i^r due to adding or subtracting this set of active units to the set $A_{r'}$ at time t_k^*. This is equal to the difference between the asymptotic value of γ_i^r based on the new set of active units $A_{r'}(t_k^*)$ and the asymptotic value based on the old set of active units $A_{r'}(t_{k-1}^*)$. Specifically, from (19.17), and with an obvious extension of previous notation,

$$\frac{1}{\delta} M\left[\Delta^* \gamma_i^r(A^*_{k-1})\right] = \frac{1}{\delta} M\left[\gamma_i^r\left(A^*(t^*_k)\right)\right] - \frac{1}{\delta} M\left[\gamma_i^r\left(A^*(t^*_{k-1})\right)\right]$$

$$= \frac{N_a \eta}{\delta} \sum_j P_j \phi\left(\alpha_i^r(t^*_{k-1})\right)\left(\sum_{\beta_j} P(\beta_j)\left[\phi\left(\alpha_j^{r'}(t^*_{k-1})\right)\left\{\phi\left(\alpha_j^{r'}(t^*_k)\right) - \phi\left(\alpha_j^{r'}(t^*_{k-1})\right)\right\}\right.\right.$$

$$\left.\left. - \left\{\phi\left(\alpha_j^{r'}(t^*_k)\right) - \phi\left(\alpha_j^{r'}(t^*_{k-1})\right)\right\} \sum_{\alpha} P(\alpha_\alpha)\, \phi\left(\alpha_\alpha^{r'}(t^*_{k-1})\right)\right]\right)$$

$$= \frac{N_a \eta}{\delta} \sum_j P_j \phi\left(\alpha_i^r(t^*_{k-1})\right)\left(\sum_{\beta_j} P(\beta_j)\left[\phi\left(\alpha_j^{r'}(t^*_k)\right) - \phi\left(\alpha_j^{r'}(t^*_{k-1})\right)\right]\left[\phi\left(\alpha_j^{r'}(t^*_{k-1})\right) - Q_{j'}(t^*_{k-1})\right]\right) \qquad (19.18)$$

With this equation for the asymptotic value of the "incremental set" of A-units which become active (or inactive) at time t^*_k, it becomes possible to compute the time-dependent solution in much the same manner as for the four-layer perceptron. To begin with, we obtain the functions ξ^i_α (defined in equation 16.23) for all α, and thus determine the next α_k for which $\phi(\alpha_k)$ will change. This gives us the values of $\phi\left(\alpha_j^{r'}(t^*_k)\right)$ which are required in equation (19.18). We then compute the <u>actual</u> value of $\Delta^* \gamma_i^r(t^*_k)$ as follows. The contribution, $\Delta^* \gamma_i^r$, being composed of a number of individual values, ν_{jl}, will approach its asymptotic value exponentially, with the same time-constant as the γ's. Thus, if we can determine the value of the set of contributing connections at the start of the interval (time t^*_{k-1}) we can determine its value at time t^*_k. Now the value at t^*_{k-1} is simply the sum of the $\nu_{jl}(t^*_{k-1})$ for all j such that $\phi(\alpha_j^{r'})$ changes at t^*_k. We will use the notation $\Delta_0 \gamma_i^r(t^*_k)$ for this starting value. Specifically,

$$\Delta_0 \gamma_i^r(t^*_k) = N_a \sum_{\beta_j} P(\beta_j)\left[\phi\left(\alpha_j^{r'}(t^*_k)\right) - \phi\left(\alpha_j^{r'}(t^*_{k-1})\right)\right] \nu_{jl}\,(t^*_{k-1}) \qquad (19.19)^*$$

* To avoid computing $\nu_{ji}(t)$, an approximation is required, e.g., $\nu_{ji}(t) \approx \sum_k \nu_{ki}(t) = 0$.

Then, by analogy to (16.24), we have

$$\Delta^* \gamma_i^{r\prime}(t_k^*) = \frac{M\left[\Delta^* \gamma_i^r(A_{k-1}^*)\right]}{\delta} - e^{-\delta(t_k^* - t_{k-1}^*)}\left(\frac{M\left[\Delta^* \gamma_i^r(A_{k-1}^*)\right]}{\delta} - \Delta_0 \gamma_i^r(t_k^*)\right) \tag{19.20}$$

Thus, the complete solution for γ_i^r at time t_k^* (including the discontinuity at the terminal end of the interval) is given by:

$$\gamma_i^r(t_k^*) = \frac{M\left[\gamma_i^r(A_{k-1}^*)\right] + M\left[\Delta^* \gamma_i^r(A_{k-1}^*)\right]}{\delta} - e^{-\delta(t_k^* - t_{k-1}^*)}\left(\frac{M\left[\gamma_i^r(A_{k-1}^*)\right] + M\left[\Delta^* \gamma_i^r(A_{k-1}^*)\right]}{\delta} - \gamma_i^r(t_{k-1}^*) - \Delta_0 \gamma_i^r(t_k^*)\right) \tag{19.21}$$

The value of the dicontinuity time, t_k^*, is obtained as before, from equation (16.25).

This completes the analysis of the cross-coupled γ-system. While no cases have actually been computed at the present time, it seems likely that this system will generally be better behaved than the α-system, particularly in such problems as Experiment 14, where there is a tendency for all A-sets to merge under α-system dynamics.

19.9.2 Analysis of Γ-systems

In the Γ-system, where the value is conserved over the set of output connections from each A-unit, the change in the value of the connection α_{ij} is now

$$\Delta \alpha_{ij} = a_i^*(t-\tau)\left[a_j^*(t) - \frac{1}{N_a}\sum_k a_k^*(t)\right]\eta\,\Delta t \tag{19.22}$$

This leads to the differential equation and equilibrium equation, respectively,

$$\frac{d\gamma_i^r}{dt} = N_a \eta \sum_g P_g \left[\phi\, \alpha_i^g(t) - Q_g(t)\right] Q_{g'r'}(t) - \delta\, \gamma_i^r(t) + \sum_k \delta(t - t_k^*)\Delta^* \gamma_i^r(t_k^*) \tag{19.23}$$

$$\gamma_i^r(\infty) = \frac{N_a \eta}{\delta} \sum_g P_g \left[\phi(\alpha_i^g) - Q_g\right] Q_{g'r'} \tag{19.24}$$

From these equations, a solution for $\gamma_i^r(t_k^*)$ can clearly be computed along the same lines as in the previous section, for the γ-system. Specifically, the asymptotic value for the connections from the difference set takes the form:

$$\frac{M\left[\Delta^* \gamma_i^r(A_{k-1}^*)\right]}{\delta} = \frac{N_a \eta}{\delta} \sum_g P_g \left[\phi\left(\alpha_i^g(t_{k-1}^*)\right) - Q_g\left(t_{k-1}^*\right)\right] \cdot \sum_{g'} P(\beta_j)\, \phi\left(\alpha_j^{g'}(t_{k-1}^*)\right) \left[\phi\left(\alpha_j^{r'}(t_k^*)\right) - \phi\left(\alpha_j^{r'}(t_{k-1}^*)\right)\right] \tag{19.25}$$

$\Delta_o \gamma_i^r(t_k^*)$ and $\Delta^* \gamma_i^r(t_k^*)$ are computed by equations (19.19) and (19.20) without any modification, so that the final solution can be obtained as before from Equation (19.21).

Due to its apparent superiority as a predictive system, and since it appears to have the same advantages in stability of the A-set organization as the γ-system, this model seems likely to be the most versatile system analyzed thus far.

19.10 Similarity Generalization Experiments

The consideration which first drew attention to the importance of cross-coupled perceptrons was the prediction by Rosenblatt (Ref. 85) that such networks would be capable of improving their performance in similarity

generalization, as a result of prolonged exposure to an environment in which stimuli are more likely to be succeeded by their transforms than by unrelated stimuli. In Chapter 16, it was shown that a suitably organized four-layer perceptron has such a capability, and the above analysis shows that for sequences in which the activity of a cross-coupled perceptron is interrupted after every other stimulus, its performance should be equivalent to the four-layer model. Thus the original prediction appears to be upheld.

The mathematical analysis of cross-coupled networks has been completed too recently to permit detailed examples of similarity generalization to be worked out at this time. A series of simulation experiments have been completed, however, employing a program written by Trevor Barker for the IBM 704. In this program a fully coupled network of 102 association units is represented, with γ-system dynamics. The model differs from those analyzed above, in that the values do not decay. This leads to "instability" of the system (a tendency to go into terminal oscillatory modes with massive A-unit activity, unrelated to the stimuli which are presented), unless some additional measures are taken to limit the growth of the connection values. The program was therefore modified for bounded values. In order to prevent the tendency of the γ-system to turn off most of the initially responding A-units after the first few preconditioning stimuli, a further modification was included to permit half-integer values for θ. Thus the values of the cross-coupling connections have no effect until the magnitude of γ is at least equal to $1/2$.

Even in this modified program, performance is considerably poorer than might be expected of the decaying value models, since the system ultimately goes to a saturation condition, with all values either at the upper or lower bound. Prior to this saturation state, however, (and to a lesser degree even in its saturated condition) similarity generalization can be successfully demonstrated, as in the following experiments.

Figure 54 shows the results of two experiments, with five excitatory and five inhibitory retinal connections to each A-unit, $\theta = 1.5$, $\eta = .005$, and an upper bound of .2 for all values. In each case, the preconditioning sequence consisted of random stimuli, alternating with their transforms. The transform, $T(S)$, consisted of a displacement of S by half the width of the retina. The retina itself was a 4 by 36 mosaic (144 points), and all stimuli covered one fourth of these points. In the first experiment, the preconditioning stimuli consisted of random "salt and pepper patterns", in which any combination of points is equally likely. In the second experiment, the stimuli were constructed by a "blob generating program" which produces coherent, but randomly shaped patterns such as those illustrated in the figure. The test stimuli, in each case, consisted of the same set of ten coherent patterns (rectangular designs). After being exposed to the preconditioning sequence $S_1, T(S_1), S_2, T(S_2), S_3, T(S_3), \ldots,$ activity of the A-system is interrupted, and a G-matrix is computed for the twenty sequences:

$$\mathcal{S}_1 = S_1 S_1 \qquad T(\mathcal{S}_1) = T(S_1) T(S_1)$$

$$\mathcal{S}_2 = S_2 S_2 \qquad T(\mathcal{S}_2) = T(S_2) T(S_2)$$

$$\vdots \qquad\qquad \vdots$$

$$\mathcal{S}_{10} = S_{10} S_{10} \qquad T(\mathcal{S}_{10}) = T(S_{10}) T(S_{10})$$

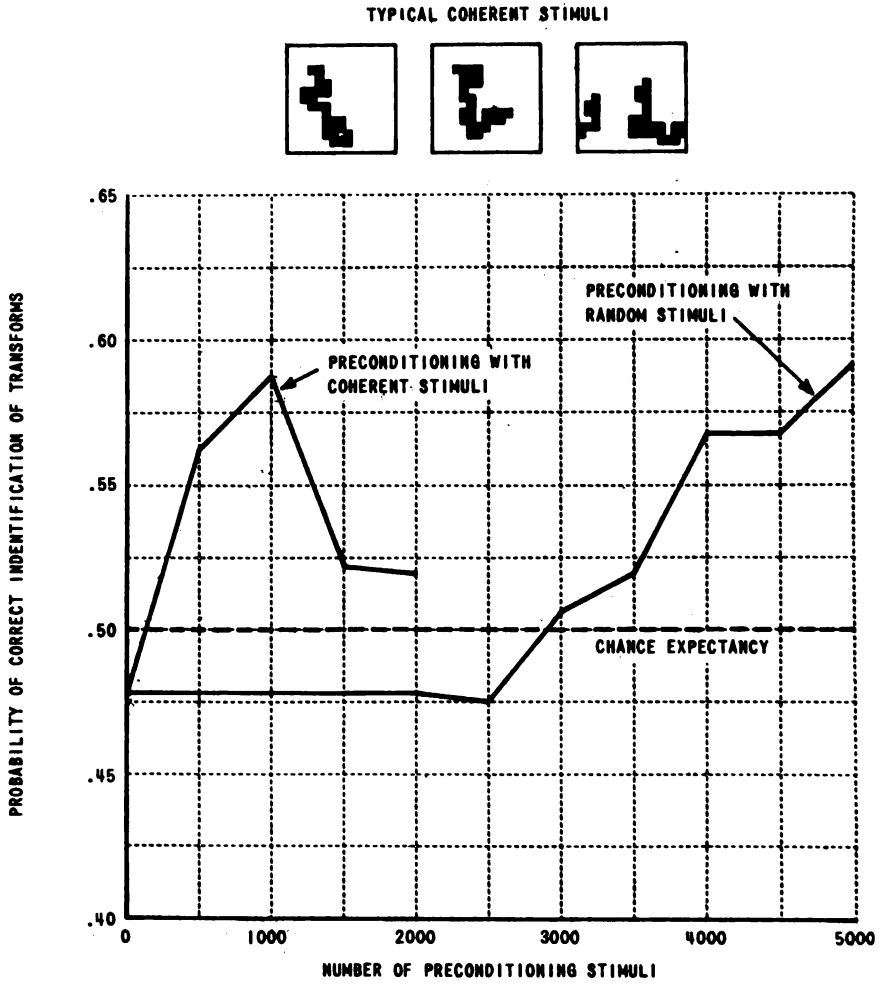


Figure 54 CROSS-COUPLED PERCEPTRON SIMULATION EXPERIMENTS

This G-matrix indicates which of the ten transforms would be identified correctly if the perceptron were trained to recognize their images, by means of a single reinforcement. Sequences of duration 2 are used, to provide time for impulses to propagate over the cross-connections before testing the response.

The curves show the mean performance of ten perceptrons over the set of ten test transforms, as a function of the number of preconditioning stimuli. In the case of the coherent stimuli, note that learning is both more rapid, and saturation is reached more quickly than with the random stimuli (where the saturation condition has not been reached even after 5000 preconditioning stimuli). While the peak performance level is less than .60, a statistical evaluation of the data reveals that the trend is definitely significant. All ten perceptrons, individually, showed a trend in the expected direction, so that the chance of obtaining these results accidentally would be less than .001. It should be noted that since the expected generalization coefficient, g_{if}, from a stimulus to its disjoint transform is negative (in a γ-system) these perceptrons had to overcome an initial negative bias before achieving even the "chance" level of 50% correct identifications.

These experiments confirm the predicted tendency of cross-coupled perceptrons to generalize on the basis of similarity, in a suitably organized environment. They also indicate the advantage of coherent over random stimuli, which is more pronounced in larger retinas than that illustrated. Doubling the number of retinal points would virtually eliminate the trend which is found for random stimuli, while the coherent stimulus curve would be relatively unaffected. All of these results are consistent with the laws of similarity generalization which were tentatively proposed in Section 15.4.

Until further empirical studies are completed, the theoretical results obtained for cross-coupled systems should still be interpreted with caution. There is at present no knowledge of the variance in performance over perceptrons, and how this relates to the size of the system; nor can we estimate the effects of finite stimulus sequences, in which the assumption of an infinitesimal rate of reinforcement per stimulus is not fully justified. The equations of the preceding sections represent limiting behavior for large values of N_a, very gradual memory modification, and very long training sequences. The assumption of large N_a can be obviated by writing the equations with empirical β-vectors measured for a particular perceptron, but in this case the results can be generalized only by means of an empirical sampling procedure, with many such perceptrons. The given equations will probably be found to yield correct qualitative results, but considerable work is still required to test their quantitative accuracy.

19.11 Comparison of Cross-Coupled and Multi-Layer Systems

In similarity generalization experiments, it has already been observed that there is a marked similarity between the performance of the four-layer perceptron of Chapter 16, the open-loop cross-coupled system of Chapter 17, and the closed-loop cross-coupled systems considered above. All of these systems are capable of learning to associate patterns which occur frequently in temporal succession, and abstracting the principle of similarity from a transformation sequence (in which stimuli alternate with their transforms). All of these systems will tend to work better with coherent patterns than with random point patterns. In all cases, the constant $N_a \eta / \delta$ determines the nature of the terminal G-matrix which is obtained, for a

given environment. Actually, an exact equivalence is found between the performance of the fully cross-coupled system in finite-sequence environments, with sequences of two stimuli, and the performance of the open-loop system of Chapter 17 with $\gamma = 1$. Suppose the system of Chapter 17 is extended to include an infinite number of A-sets, each with identical connections from the retina, and with variable connections to each unit in the k^{th} A-set from each member of the $k-1^{th}$ A-set (and allowing unit time delay in transmission). It can then be shown that the states of the k^{th} A-set for the first k stimuli in the sequence will correspond exactly to the states of the equivalent fully cross-coupled model (having all S-A connections equivalent to those in the open-loop model). Thus, the fully cross-coupled model, considered through all time, is equivalent to the output of an infinitely extended open-loop model, of the type discussed in Chapter 17.

While these similarities would lead us to expect basically similar behavior in most problems for these different types of systems, some noteworthy differences do exist between the cross-coupled system and multilayer systems with finite numbers of layers. First of all, there is an inherent sequence-dependence in the cross-coupled model, which makes its present state a function of the recent succession of events, (i.e., stimuli) rather than just the last event to occur. This means that all cross-coupled systems have some capability for temporal pattern recognition, even without variation in the transmission times of the input connections. Secondly, the cross-coupled systems are likely to reach their terminal condition more rapidly, and with initially accelerating rates of adaptation, since the differential equation depends on changes both in the transmitting and receiving sets of A-units, while in the four-layer model, the differential equation

depends only on changes in the receiving set, the transmitting set being fixed for all time. The dependence on both receiving and transmitting sets makes the cross-coupled system more subject to "instability" phenomena, and probably tends to reduce the "dynamic range" of the system (as a function of $N_a \eta / \delta$) in most cases. These phenomena have not yet been studied sufficiently to present conclusive quantitative results at this time.

A more important difference than any of the above may be potentially present, although this remains in the realm of speculation at present. In a value-conserving cross-coupled perceptron, where there is the possibility of developing pronounced inhibitory interaction between A-sets, there is a tendency to develop "cell assemblies" (in Hebb's sense), and these cell-assemblies tend to rival one another for dominance at all times. It seems possible that such a phenomenon may provide a basis for figure-ground separation in complex sensory fields, where it is desired that the system attend to one object, or component of the input situation, and ignore the remainder. This will be discussed further in Part IV. If such an effect can be demonstrated, many of the remaining problems in the design of a perceiving system would be solved.

20. PERCEPTRONS WITH CROSS-COUPLED S AND R-SYSTEMS

A number of interesting effects may be obtained by cross-coupling the S-units or R-units of a perceptron. Several such systems are considered briefly in this chapter. The first section deals with cross-coupled sensory systems; the second section deals with cross-coupled R-systems. Detailed analyses are not presented here, although several analytic studies are available in the referenced literature.

20.1 Cross-coupled S-units

If the sensory units are arranged in a two dimensional array, or retina, then it has been proposed that inhibitory interconnections between each S-unit and its nearest neighbors will tend to inhibit activity most strongly in the center of a field of illumination, and less around the edges. Such a system should lead to accentuated edges or boundaries for a visual pattern, reducing the relatively redundant information coming from interior regions.* Systems utilizing this principle have been proposed by Taylor (Ref. 99), by Inselberg, Löfgren, and von Foerster (Ref. 4), and by a number of others. The Inselberg-Löfgren-von Foerster treatment includes a more detailed quantitative analysis than was hitherto available, including cases in which the probability of interconnection of two units is a Gaussian function or an exponential function of the distance between them.

While it appears that contour detectors can indeed be constructed by this means, it should be noted that some information is lost in the process: namely, the indication of the direction of the illumination gradient

* See also Chapter 23, on visual analyzing mechanisms.

across the contour. Thus if a square patch of illumination is operated upon by the network to yield a square outline, there is no way to tell whether the inside of the square was light and the outside dark, or vice versa. The contour-detectors proposed by Rosenblatt in Ref. 79, which consist of A-units with circular or elliptical distributions of origin points, with slightly different centers for excitatory and inhibitory origin clusters, still preserve this gradient information.*

A somewhat more interesting possibility has been demonstrated by Inselberg, et all, if three layers of units with anisotropic connections are superimposed on one another, with a rotation of the axes of symmetry by 60° in the successive layers. With such a system, it appears to be possible to construct a network from which there is zero output from a straight-line stimulus (regardless of its orientation) but a non-zero output from a curved line. Such systems clearly deserve more study as possible stimulus analyzing mechanisms for reducing the input data to a perceptron.

Systems with excitatory interconnections between S-units are of relatively little interest, as such a network would generally lead only to a spread of activity from the stimulus region. The only useful function which such connections might have would be in smoothing irregular or broken images, by filling in holes and gaps; such an application, however, seems to be of questionable utility at the present time.

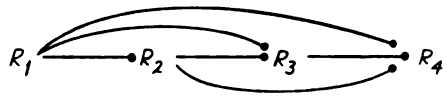
20.2 Cross-coupled R-units

Inhibitory interconnections between R-units may be useful in several ways. One application is to guarantee that no more than one R-unit can be "on" at any time. For this purpose, all R-units are given inhibitory

* See also Hubel, Ref. 113, for relevant biological evidence.

interconnections to all the others; whichever unit first goes on, inhibits all the others, holding them off. Such a system will tend to "hang up" in this state, until the positive signal to the first R-unit is reversed, permitting some other unit to come on. If the speed of response of an R-unit is proportional to the magnitude of its input signal, such a scheme can be used to select the R-unit with maximum input from a given stimulus.

In R-controlled reinforcement systems, inhibitory connections between R-units may sometimes be employed to guarantee that a unique response is associated to each new stimulus in succession. Suppose there are four stimuli, which activate disjoint or nearly disjoint sets of A-units. Let there be four R-units, with inhibitory connections as follows:



In this scheme, unit R_i inhibits (absolutely) all successive R-units (R_{i+1}, R_{i+2}, . . .) . Now if stimulus S_1 occurs, and transmits an initially positive signal to all R-units, only R_1 can go on. With an R-controlled value-conserving system (in which the sum of values over all connections is held constant) S_1 will then develop an excitatory signal to R_1 , and negative signals to all other R-units. At the same time (since we have assumed essentially disjoint A-sets) the value-conserving system will guarantee that the R_1 response generalizes negatively to all other stimuli. Thus, when S_2 occurs, it will tend to turn off R_1 , but will try to turn on R_2 , R_3 and R_4 . Of these, only R_2 can remain on, due to the inhibitory coupling, so that S_2 (or whichever stimulus occurs second in the sequence) will become associated to R_2 . Similarly, S_3 is associated to R_3 , and S_4 to R_4 .

This scheme becomes somewhat less trivial if it is applied to the four-layer perceptrons of Chapter 16, subsequent to a preconditioning sequence in which the perceptron has learned to associate a unique A-set to each similarity class of stimuli in a given environment. The above method can then be employed to assign a unique response to each class of stimuli (provided the terminal A-sets have sufficiently small intersections).

While the interconnection schemes proposed here for S and R-units are occasionally useful for control purposes, they do not introduce any fundamentally new properties of importance. The most striking phenomena to be found in cross-coupled systems are the similarity generalizing capabilities of the cross-coupled association systems -- with the tantalizing possibility of a figure-ground mechanism still to be investigated in future work.

PART IV

BACK-COUPLED PERCEPTRONS AND PROBLEMS FOR FUTURE STUDY

21. BACK COUPLED PERCEPTRONS AND SELECTIVE ATTENTION

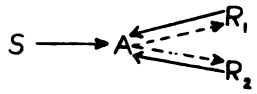
In Parts II and III of this volume, we have tried to establish the fundamental properties of two topological classes of perceptrons: series-coupled and cross-coupled systems. While the possible configurations of these two types of perceptrons have by no means been exhausted, the most general forms of series-coupled and cross-coupled networks appear to be sufficiently well understood so that their principles can now be applied to the analysis of more elaborate systems. The most general network is achieved with the addition of back-coupling (Definition 26, Chapter 4), so that layers of units which are relatively remote from the sensory end of the perceptron can modify the activity of layers which are relatively close to the sensory end. Given this additional mode of coupling, then virtually all perceptrons of interest, however elaborate their structure, can be regarded as compounds or modifications of the types previously considered.

The modulating effect of back-coupling upon the behavior of a perceptron will be considered qualitatively in this chapter. It will be seen that while the analysis of such systems can frequently be carried out in terms of already established principles, their behavior possesses a new order of sophistication. In particular, the psychological phenomena of selective attention and "cognitive set" now begin to emerge. A related exposition of these ideas can be found in Rosenblatt, Ref. 79, Chapter X.

21.1 Three-Layer Systems With Fixed R-A Connections

21.1.1 Single Modality Input Systems

The first case to be considered is the class of three-layer perceptrons having fixed-value connections from the R-units back to the A-units. For simplicity, it is assumed that there is no cross-coupling within any of the three layers. Such a perceptron with two R-units can be represented by the symbolic diagram:



where solid arrows represent fixed-value connections, and broken lines represent variable-valued connections. In particular, assume that there is a connection from every R-unit back to every A-unit, half of these connections, chosen at random, having the value +1, and the other half having the value -1. In the following section it will be assumed that the R-units are of an "on-off" variety (having the outputs 1 or 0, rather than +1 and -1) although analogous effects can be found for simple R-units. It is also assumed, for the sake of avoiding impossible closed-loop situations, that all connections have a short time delay, τ ; a stimulus, however, is generally assumed to be held on the retina for a time $T >> \tau$.

The signal c^*_{rl} which is fed back to an A-unit a_l from the response unit r is given by the linear function

$$c^*_{rl} = r^* \nu_{rl}$$

Thus c^*_{rl} is equal either to ν_{ri} or 0, depending on whether $r^* = 1$ or 0. The effect of these feedback signals on the set of A-units responding to a given stimulus is shown in Figure 55. The symbol β is used to represent the component of the input signal, α, which comes to the A-unit from the retina. It is assumed that there are two R-units, so that there are four disjoint sets of A-units with roughly $N_a/4$ units in each set, corresponding to the four possible combinations of $\nu_{r_1 l}$ and $\nu_{r_2 l}$. These sets of A-units are represented by the four quadrants of the diagram. The circles indicate the values of β_i received from the given stimulus, in relation to the threshold, θ_i. The A-units in the innermost circle, for which $\beta \geq \theta + 2$, will always be on when the given stimulus occurs, regardless of the condition of the R-units. Those units for which $\theta \leq \beta < \theta + 2$ will be on except when they receive an inhibitory signal from both R-units simultaneously. The units for which $\beta = \theta - 1$ must receive a net excitatory signal from one or both of the R-units in order to go on, and those units for which $\beta = \theta - 2$ will only go on (in the presence of the given stimulus) if they receive an excitatory feedback signal from both R-units at once. Units for which $\beta < \theta - 2$ will never respond to this stimulus. The magnitudes of these sets can be calculated from tables of Q-functions (c.f., Chapter 6 and Reference 87). The shaded area in Figure 55 shows the sets which respond to the given stimulus when $(r^*_1, r^*_2) = (1, 1)$.

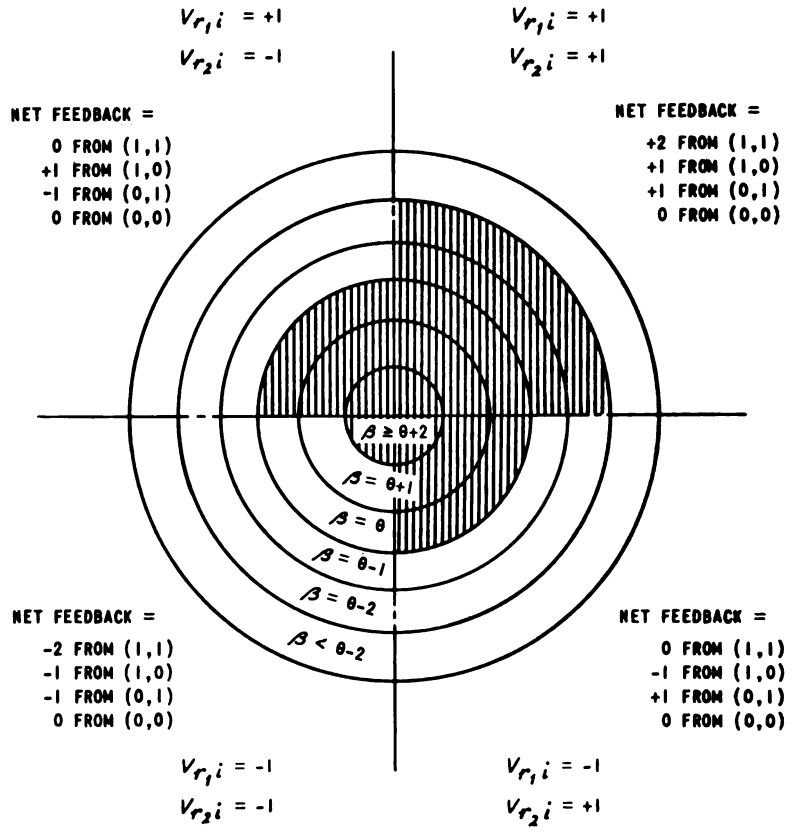


Figure 55 EFFECT OF FEEDBACK ON ACTIVITY OF A-SET, IN RESPONSE TO A GIVEN STIMULUS, FOR PERCEPTRON WITH 2 R-UNITS. SHADING SHOWS ACTIVE A-SETS FOR THE RESPONSE STATE r_1^*, r_2^* = (1,1).

Now suppose there are two stimuli, S_1 and S_2. S_1 is trained to give the response combination $(r_1^*, r_2^*) = (1, 0)$, while S_2 is associated to the response code $(0, 1)$. We assume that the retinal sets representing the two stimuli are completely disjoint. Having trained the perceptron, let us now present both stimuli simultaneously (i.e., a composite image, $S_1 \cup S_2$, is projected on the retina). Under these conditions, a series-coupled perceptron might equally well give the response combinations $(0,0), (0,1)$, $(1,0)$ or $(1,1)$. The present system, however, will tend to respond either with $(1,0)$ or with $(0,1)$. In other words, it will tend to correlate those R-states which go with one of the two stimuli, rather than giving a partial response to each. This can be understood by reference to Figure 56, where the A-sets responding to each of the two stimuli are shown. For convenience, the sets responding to S_1 are assumed to be disjoint from the sets responding to S_2, and the diagram is simplified by assuming that the set which is active for the composite $S_1 S_2$ stimulus (in the presence of a given R-state) is equal to the union of the sets responding to S_1 and S_2 alone. This last assumption is not generally warranted; but the qualitative conclusions reached will still be correct. The shading shows the reinforced sets for S_1 and S_2.

At the moment that $S_1 S_2$ appears on the retina, both R-units will be off, so that there is zero feedback to the A-system, and the total signal coming to each R-unit from the A-system will be approximately zero (consisting of a positive signal from one stimulus, and an approximately equal negative signal from the other stimulus). Suppose initially, both R-units go on. In this case, the sets of A-units responding when $R^* = (1,1)$ will become active, and the total signal to each R-unit will still be approxi-

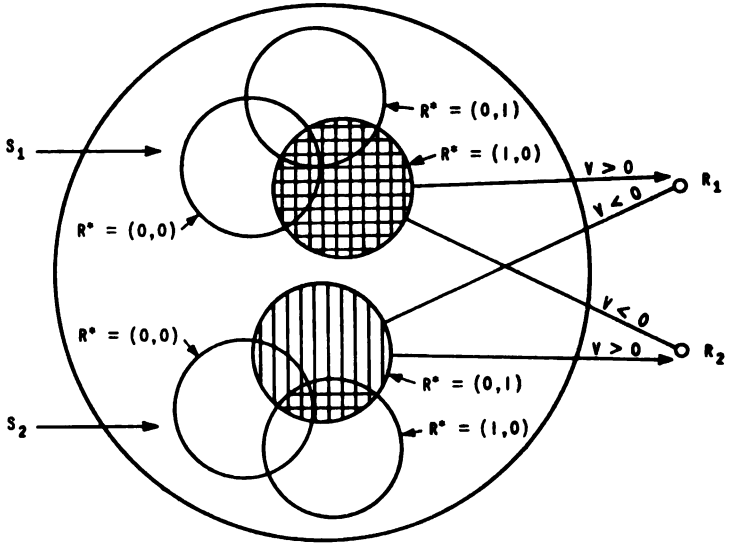


Figure 56 A-SETS RESPONDING TO THE STIMULI S_1 AND S_2, FOR THREE RESPONSE CONDITIONS. SHADED AREAS SHOW REINFORCED SETS, AND DOUBLE HATCHING SHOWS REINFORCEMENT WHICH GENERALIZES TO THE CONDITION $R^* = (1,0)$.

mately zero, so that the response state is unstable. Alternatively, suppose the R-state goes to (1,0) . In this case, the signal to the R-units comes from the double-hatched regions of the Venn-diagram in Figure 56, and the S_1 set becomes "dominant". If this occurs, the response (1,0) will tend to remain stable, and may even persist after the stimuli are removed (provided some of the A-units have thresholds ≤ 1). Similarly, if the R-state goes to (0,1) , then the S_2 set becomes dominant, and its response will tend to persist.

If either stimulus has been trained to give the response (0,0) in the above experiment, the R-units will tend to "hang up" in their initial condition, and no other response can ever occur to the joint stimulus $S_1 S_2$ On the other hand, it is possible to produce an oscillating or cyclical response by training a given stimulus to give the response (1,1) when the present response is (0,0) , then conditioning the (1,1) set to give the response (1,0), conditioning this set to give (0,1) , and finally associating the response (0,0) to the A-set responding for (0,1) . In this case, as long as the stimulus is held on the retina, the R-units will cycle through the four responses in succession.

The important tendency which has been demonstrated for this system is a tendency to correlate the output of the R-units so that they all apply to a single stimulus, when a composite stimulus occurs at the retina. This now provides the basis for the following experiment:

EXPERIMENT 15: Using a four-R-unit perceptron, and a universe of squares and triangles of equal area in all positions on the retina, train the system to give the responses $(r_1^*, r_2^*) = (1, 0)$ for a triangle, and $(0, 1)$ for a square; $(r_3^*, r_4^*) = (1, 0)$ for a stimulus in the top half of the retina, and $(0, 1)$ for a stimulus in the bottom half. After training with an error-correction procedure, test the response of the perceptron to the stimuli S_x = triangle in the top half of the field and square in the bottom half, and S_y = square in the top half with triangle in the bottom half.

In this experiment, the first pair of responses are used for square/triangle discrimination, and the second pair for top/bottom discrimination. For the time being, assume that the error correction procedure is modified by forcing the correct R^* condition whenever a correction is applied. (This assumption will be dropped in Section 21.2.) It is predicted that a back-coupled system, organized as above, will tend to give one of the two responses (1, 0, 1, 0) or (0, 1, 0, 1) for stimulus S_x (signifying "triangle, top" or "square, bottom", respectively), but will give one of the two responses (1, 0, 0, 1) or (0, 1, 1, 0) for stimulus S_y (signifying "square, top" or "triangle, bottom"). In other words, the system should give a consistent description of one of the two stimuli, in terms of shape and location, and ignore the other stimulus; it will not name the shape of one and the position of the other, even though both shapes and both positions are simultaneously present.

That the predicted effects will tend to occur can be seen by referring to Figure 57, where it is assumed that the S_x combination (top triangle and bottom square) occurs. Reinforcement is shown by cross-hatching. The relative sizes of the intersections in the Venn diagram are drawn to suggest the relative intersections of the A-sets for the response states of interest. Note that the set responding when $R^* = (1, 0, 0, 0)$ tends to have a relatively large intersection with the $(1, 0, 1, 0)$ set, due to the fact that three of the four R-units are in identical states. The combined intersection of the $(1, 0, 0, 0)$ set with the sets which are reinforced to yield the "top" response $(1, 0)$ on r_3 and r_4 is greater than the combined intersection with the sets which were reinforced for the "bottom" response. If the triangle first becomes dominant with respect to the r_1, r_2 pair of responses (yielding the condition $1, 0, 0, 0$) the activated set which has been most heavily reinforced, shown by cross-hatching, will now tend to evoke the "top" response from r_3 and r_4, since the "top triangle" set now carries considerably greater weight than the "bottom square" set. Thus a consistent configuration on all four R-units is induced. If $(0, 1, 0, 0)$ should occur, however, the system will have an opposite bias for r_3 and r_4, tending to evoke the condition $(0, 1, 0, 1)$. If S_y should occur instead of S_x, the biases will be found to favor the $(1, 0, 0, 1)$ or $(0, 1, 1, 0)$ conditions, as predicted.

Experiment 15 illustrates the simplest conditions under which "selective attention" might be said to occur in a perceptron. In a complex field, with more than one trained stimulus present, rather than giving a conflicting mixture of responses, the perceptron tends to pick a single familiar "object" and respond to this object to the exclusion of everything

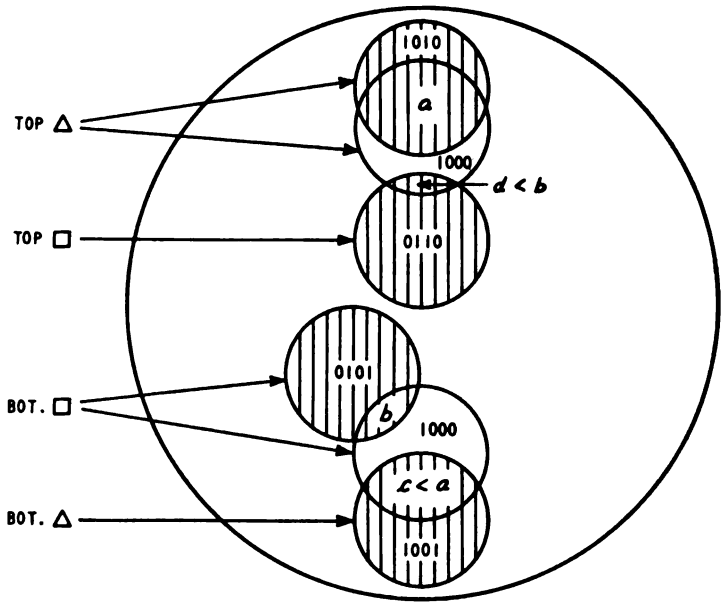


Figure 57 SETS AFFECTING THE TRANSITION FROM THE RESPONSE STATE (1,0,0,0) WHEN THE COMBINED STIMULUS "TOP TRIANGLE" AND "BOTTOM SQUARE" OCCURS. SHADING SHOWS REINFORCED SETS, AND THE MEASURES OF THE INTERSECTIONS WITH THE (1,0,0,0) SETS ARE DENOTED BY THE LETTERS a, b, c, AND d. THE VENN DIAGRAM IS DRAWN SO AS TO EMPHASIZE THE PROBABLE MAGNITUDES INVOLVED.

else. By adding additional responses, a complete description might be obtained of the shape, size, position, etc., of a single object in the field. The particular object which is selected, however, depends on chance factors, such as the relative amounts of reinforcement which have been applied to different A-sets, or momentary noise within the network. In the following section, it will be shown how a stimulus in a different modality, such as a spoken word, can be made to direct the attention of the perceptron towards a selected object or region in the visual field.

21.1.2 Dual Modality Input Systems

The perceptron which is illustrated in Figure 58 is similar to the one which was described in the preceding section, except that it possesses two sensory input systems, one visual (a retina) and the other auditory (e.g., a filter system). There is a set of A-units for each of these input sets, designated A_v for the visual association system, and A_a for the auditory association system. Again, there are four R-units, each one receiving variable-valued connections from all A-units in both sets, and sending a set of fixed value connections back to all the A-units. As before, half of the feedback connections from each R-unit are assumed to be excitatory, and the remainder inhibitory, with values ± 1 .

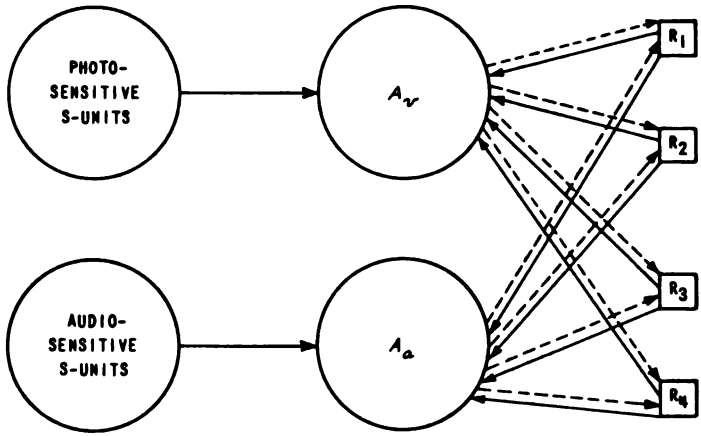


Figure 58 ORGANIZATION OF A DUAL MODALITY PERCEPTRON, WITH 4 R-UNITS (BROKEN LINES INDICATE VARIABLE-VALUED CONNECTIONS)

With this system, the following experiment can be performed:

EXPERIMENT 16: Using a dual-modality input system (visual and auditory), with four R-units, train the perceptron to distinguish square/triangle and top/bottom, using the same code and stimuli as in Experiment 15. Then, selecting four discriminable audio-patterns, SQ, TR, T, and B, train the perceptron by means of the audio-input to associate the responses for "square",

"triangle", "top" and "bottom" to these four stimuli. In testing the perceptron, a composite visual stimulus, consisting of a triangle in the top half of the field and a square in the bottom half, is used. Simultaneously with the visual input, the audio-pattern SQ, TR, T, or B is presented, and the response of the perceptron is observed for each of these four conditions.

From the discussion of Experiment 15, it is clear that the visual section of the perceptron will tend to give a consistent response of (1,0,1,0) or (0,1,0,1), representing "top triangle" or "bottom square", respectively. The effect of adding the audio-stimuli is to add an additional bias to the R-units, favoring one of the four "concepts", square, triangle, top, or bottom. For example, if the TR stimulus is applied (which has been independently associated to the composite response $r_1^*, r_2^* = 1,0$) there will be an auxiliary positive signal to r_1, and an inhibitory signal to r_2, coming from the A_a set. There will be no bias introduced on r_3 and r_4. Consequently, the system will be biased to give the initial response (1,0,0,0), which we have seen tends to transform itself into the stable condition (1, 0, 1, 0) for the given stimulus.

Thus the results which are predicted for Experiment 16 are that when the audio-pattern TR is given, the perceptron will give the composite response indicating the shape and position of the triangle; when SQ is presented, the perceptron will indicate the shape and position of the square; for the audio-input T, it will indicate the shape and location of the top visual pattern; and for B, it will indicate the shape and location of the bottom pattern. An audio-command can therefore be used to direct the

attention of the visual system to a specified location or a specified shape, and the output of the perceptron will be a consistent description of the indicated object.

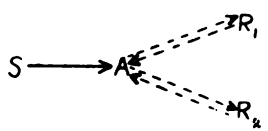
While it is possible by means of the above procedure to assign "names" to visual objects or events, and direct the attention of the perceptron by means of these names, it should be noted that the association is actually much too complete for this to serve as a model for linguistic "naming behavior". For the perceptron, there is no difference (at the response level) between the name for an object and the object itself. Thus the audio-symbol TR and the visual image of a triangle both turn on the same response combination (1,0,..) in the experiment considered above. If it is desired to retrain the system to associate some other visual pattern (say, "trapezoid") with the TR symbol, it is necessary to completely eliminate the previous association of triangles to (1,0,..) and train trapezoids to give this response instead. Words and visual patterns are part of the same conceptual class, for this perceptron, and cannot be re-associated as distinct entities, but can only be used as raw material for building up new conceptual classes. The distinction between the name and the visual object becomes important in practice if we wish to tell the perceptron to "look for the square" when there is no visual square present. The audio-symbol "look" might be used to start an automatic scan or hunting process, but to stop the process when a square is found, the perceptron must be capable of distinguishing between the audio-symbol for "square" (which it must remember for the duration of the search process to tell it what it is looking for) and the visual pattern of a "square", which must stop the search when it appears. A perceptron which is capable of distinguishing between symbols and objects, and is not subject to these criticisms, will be considered in Section 21.3.

21.2 Three-Layer Systems With Variable R-A Connections

In the previous examples, the existence of a bias towards one of the two consistent response configurations when part of the R^* state is achieved, is due to the fact that reinforcement is applied only in the presence of the correct response. This means that whenever a corrective reinforcement is applied, the reinforcement control system must first "force" the desired response configuration. But in a simple error-correction procedure, as this concept has been used previously, the corrective reinforcement would normally be applied only when the response is wrong, and this would tend to reduce the indicated bias quite drastically. For example, in Figure 56, it can be seen that if S_2 had been negatively reinforced in the presence of the $R^* = (1, 0)$ state, this negative reinforcement would tend to cancel the effect of the S_1 signal. One method of eliminating this problem, which leads to a system which appears to be generally better-behaved (on the basis of a qualitative examination of its properties) is to make use of adaptive back-connections, rather than fixed-value connections, from the R to A-units.

21.2.1 Fixed Threshold Systems

The first model to be considered corresponds topologically to the model treated in Section 21.1.1, but differs in having variable connections, so that its symbolic diagram is of the form:



The forward connections, from A to R-units, are assumed to follow the usual α-system dynamics, subject to error-correction procedures. The back-connections, however, are subject to the Γ-system rule which was introduced for cross-coupled perceptrons. This means that the total value of the set of feedback connections from each R-unit remains constant, but that if both termini (the R-unit and the A-unit) are active in succession, the connection value is incremented by a positive quantity, η. At the same time, a proportional decay occurs in all active R-A connections, so that in the absence of reinforcements, they tend to approach zero exponentially. The net change in value of connection c_{ri} at time t is therefore*

$$\Delta c_{ri}(t) = r^*(t-\tau)\left[a_i^*(t)\eta - \frac{\eta}{N_a}\sum_j a_j^*(t) - \delta c_{ri}(t)\right] \tag{21.1}$$

Assuming, as before, that each stimulus persists for a time $T \gg \tau$, the result of this rule is to raise the value of the feedback signal to all S-units which respond to the current stimulus, from the active R-units, and at the

---

* Note that in this equation decay occurs only when $r_i^* = 1$. This means that the feedback signals from different R-units will have approximately equal weight, regardless of the relative frequency with which the R-units are used. The transmission delay, τ, is included only for conformity to previous models, and plays no essential role here.

same time to develop inhibitory connections to the A-units which are not currently active. The decay guarantees that the entire system will tend towards a dynamic equilibrium, at which the expected rate of gain just balances the rate of decay.

The effect of this system is illustrated in Figure 59, which shows the condition after associating stimulus S_1 to the response (1,0) and S_2 to the response (0,1), by an error correction procedure. This corresponds to the same conditions as Figure 56. The sets which respond when $R^* = (0,0)$ are shown by the large circles. If these sets are initially reinforced to yield the appropriate response for each stimulus, then when the composite stimulus appears, they will try to turn on opposite responses, with about equal strength. Such a condition, however, will be an unstable one. If one of the sets, say S_1, carries slightly greater weight than the other, the condition illustrated in the figure will arise. With r_1 on, excitatory

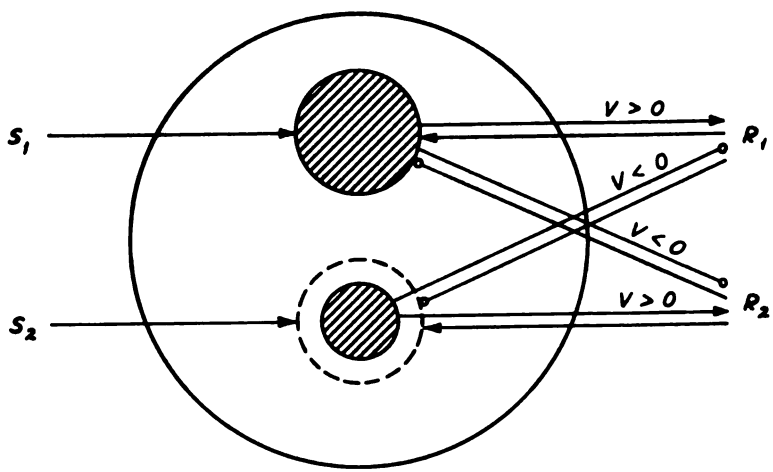


Figure 59 A-SETS RESPONDING TO THE COMPOSITE STIMULUS $S_1 S_2$. SHADING SHOWS ACTIVE A-SETS FOR THE RESPONSE STATE (1,0). (COMPARE Figure 56).

signals will be transmitted back to the S_1 set, and inhibitory signals to all other A-units, including the S_2 set. Thus the S_1 set remains unchanged, but the S_2 set is diminished. Alternatively, if S_2 should gain an advantage, the S_2 set will tend to remain unchanged, and the S_1 set will be reduced.

If we assume that the universe consists of a large number of stimuli in each class, as in Experiments 15 and 16, the set of A-units responding to S_1 would generally not be perfectly preserved, but would be shifted to include more units which respond to many stimuli in the S_1 class, and to eliminate those units which respond only to S_1. Thus there is an additional tendency, in this system, to convert the sets of A-units for different stimuli which have been associated to the same response, to sets which are nearly identical. It is clear that if the procedures of Experiments 15 and 16 are carried out with this system (but with the usual error-correction practice of reinforcing in the presence of the wrong responses only, rather than forcing the correct response) the results predicted in Section 21.1 will be obtained, but with less chance of confusion or erroneous bias due to conflicting active sets. The special property of the variable feedback system can be characterized as a tendency to activate the A-units responding to one of the previously trained parts of a complex stimulus, while suppressing those A-units which respond to the remaining parts.

21.2.2 Servo-Controlled Threshold Systems

In all perceptrons considered thus far, the thresholds of the A-units have been assumed to be invariant over time. It is possible to vary the effective threshold of an A-unit by adding an excitatory or inhibitory component to its input signal. If this is done for all A-units in the system, the result will be to increase or decrease the proportion of units which respond to a given stimulus. If all signals and thresholds are quantized, then the change in the active set will occur by sudden jumps; for example, the addition of $\Delta\theta = +1$ will suddenly activate all A-units whose α-signal was equal to $\theta_i - 1$. Such a condition would be hard to utilize effectively for the control of activity. On the other hand, if each A-unit has a threshold θ_i selected at random from some continuous distribution, say a Gaussian distribution, then there will always be some A-units whose thresholds θ_i are just below the present value of α_i, and others whose thresholds are just above the present value of α_i. In this case, a slight change in θ will always yield a corresponding change in the size of the active A-set, and the size of the active set will vary in an approximately continuous fashion as θ is changed continuously.

Figure 60 shows a back-coupled perceptron in which the amount of activity is continuously monitored by a servomechanism, which controls the magnitude of the thresholds so as to keep the total activity constant. If the fraction of active units falls below the desired level, the servo-system transmits an excitatory signal to all A-units (equivalent to $\Delta\theta < 0$) while if the activity rises above the desired level, an inhibitory signal (equivalent to $\Delta\theta > 0$) is transmitted to all A-units.

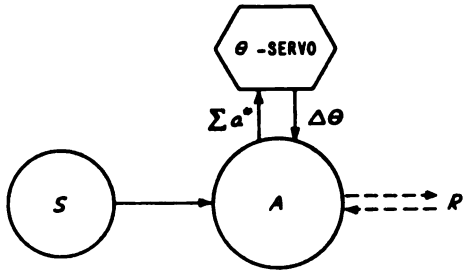


Figure 60 BACK-COUPLED PERCEPTRON WITH SERVO-CONTROLLED THRESHOLDS.

Such a system is likely to have advantages in many types of perceptrons. Attached to a series-coupled perceptron, for example, the θ -servo can guarantee that regardless of stimulus size or intensity, the level of A-unit activity will be optimum. In a cross-coupled system, it can be used to prevent "blow-ups" of activity, by providing an active mechanism for counterbalancing the growth of excitatory weights. It is worth noting that the θ -servo can substitute for inhibitory connections from the retina to A-units, since it generally yields the condition that if stimulus S_x is a subset of stimulus S_y (on the retina), the corresponding active association set $A(S_x)$ will not be a subset of $A(S_y)$. In the back-coupled system, the Θ -servo yields particularly interesting results.

Figure 61(a) shows the condition of the A-set for the same stimuli as in Figure 59, with the R-units in the (0,0) state, so that there is no feedback. The large circles show the sets which respond to S_1 and S_2 alone,

normalized by the action of the servomechanism. When the composite stimulus appears, it is no longer possible for the union of the sets $A(S_1)$ and $A(S_2)$ to remain active, however; consequently the active sets are reduced to those units (shown by the shaded areas of the diagram) for which $\beta_i \geq \theta_i + \Delta\theta$. Under these conditions there is still no bias favoring the S_1 response or the S_2 response; both sets are still in balance, and either response might occur. As before, however, this condition tends to be unstable, and (assuming that S_1 and S_2 have been associated to the same response codes as previously) either (1,0) or (0,1) will tend to occur.

Figure 61(b) shows the stable state of the system in which the response (1,0) has become dominant. The servo-system is now obliged to adjust to the effect of the excitatory signal fed back to the $A(S_1)$ set, and the inhibitory signal to the $A(S_2)$ set. The result is that the active set is nearly identical to the set which would be active for S_1 alone, the $A(S_2)$ set being virtually obliterated by the combined effect of the negative feedback and the increased threshold. It seems likely that by strengthening the excitatory feedback component (ν_1 in the diagram) sufficiently, the active set can be made to coincide perfectly with the set responding to S_1 alone. Thus the effect of selecting the (1,0) response configuration is to enable the perceptron to respond exclusively to the S_1 stimulus, completely

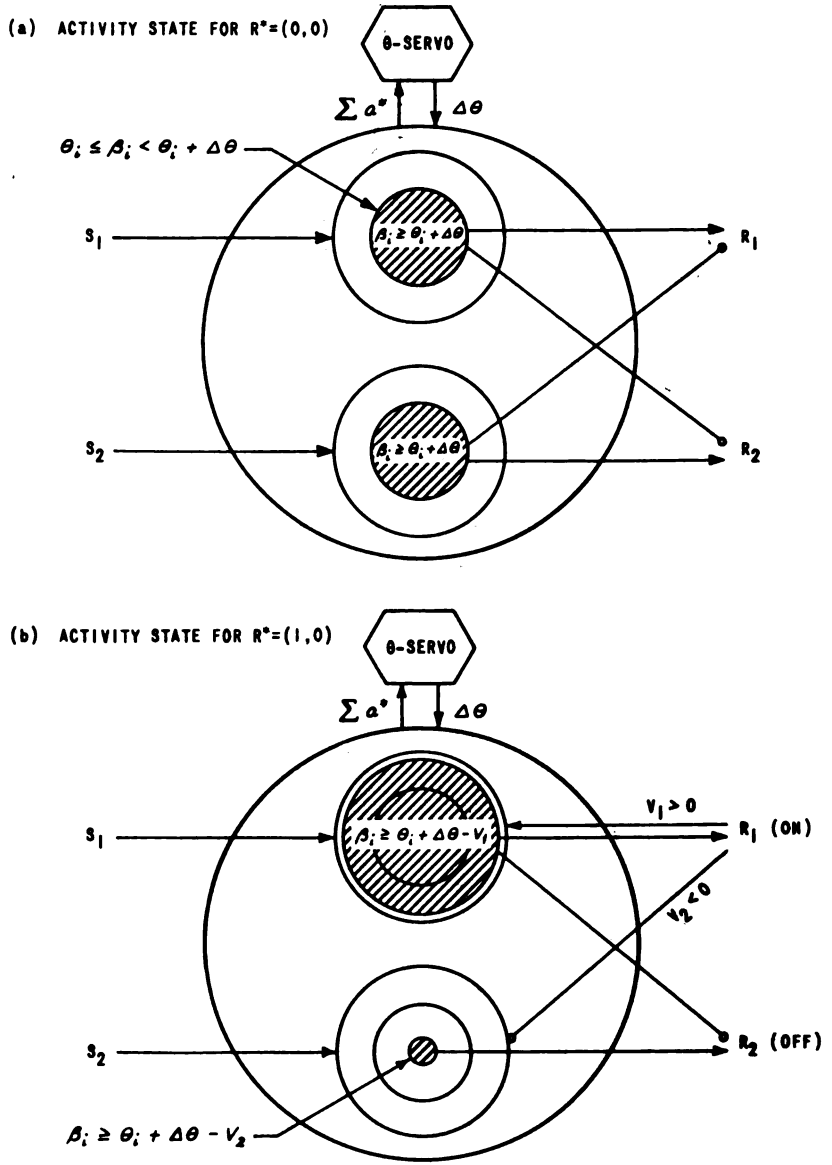


Figure 61 ACTIVE A-SETS FOR COMPOSITE $S_1 S_2$ STIMULUS, IN SERVO-CONTROLLED BACK-COUPLED SYSTEM. ACTIVE SETS SHOWN BY SHADED AREAS.

free from interference by the presence of S_2. Reversal of the R^* state would, of course, lead to a reversal of the A-state. These phenomena are highly suggestive of reversible perspective and figure-ground reversal in psychological experiments, where one of two ways of perceiving a complex figure dominates to the exclusion of the other.

In a dual-modality perceptron, the above system will work in a similar fashion, assuming that separate θ-servos are employed for the visual and auditory channels. Thus by giving the audio symbol for square or triangle, top or bottom, in Experiment 16, the perceptron can be directed to attend to one of the two objects present, and will develop an A-unit state which corresponds closely to the state which would be expected if only the indicated object was present in the field.

21.3 Linguistic Concept Association in a Four-Layer Perceptron

In Section 21.1.2, it was noted that although names can be associated to objects or visual events in a three-layer back-coupled model, so as to permit the experimenter to direct the attention of the perceptron selectively to a named object in a compound field of stimuli, the associations formed tend to be associations of particular stimuli, rather than universals. It is not possible to change the name of an object (or a class of objects) without actually undoing the previous perceptual organization of the stimulus world for the given perceptron, and then reconstructing it in a new form. Words and visual patterns are not distinguished, at the response level, but are amalgamated into a common concept.

A perceptron which is capable of first forming auditory and visual concepts, or universals, and then associating these with one another, and which can change its "linguistic associations" without disrupting its perceptual organization, is illustrated in Figure 62. The system has a visual input and an audio-input, as in Figure 58. It is also equipped with a θ -servo, and the back connections to the A_v set are variable, as in Section 21.2. For present purposes no back-connections to the A_a set are required. There are two distinct sets of R-units: one set, R^v , receives its primary inputs from the A_v system, and can be associated to visual stimuli. The second set, R^a , receives its primary inputs from the audio-system, and can be trained to respond to sound patterns, or words. (By using a spectrum of γ_{ij} for the S_a to A_a connections, or by means of a cross-coupled A_a -set, the system can be taught to recognize sound sequences, so that it need not be restricted to momentary sound patterns.)

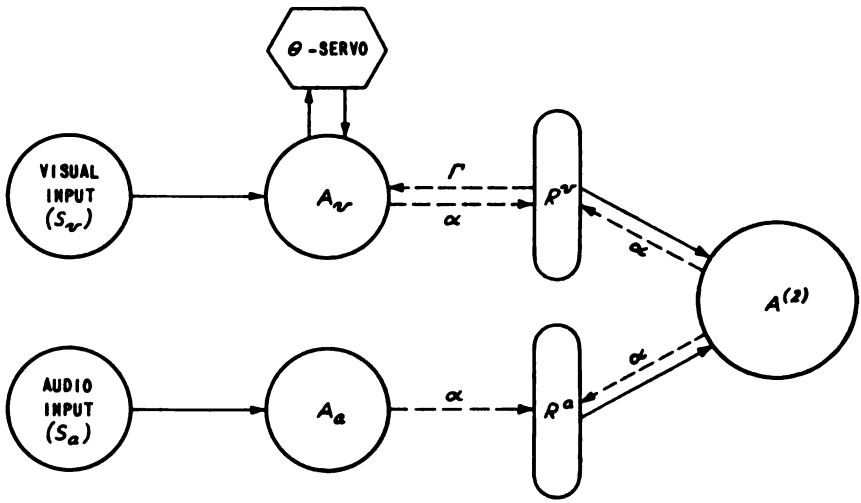


Figure 62 A DUAL-MODALITY PERCEPTRON FOR LINGUISTIC CONCEPT ASSOCIATION.

Thus far, we have what amounts to two mutually independent perceptrons, one for visual stimuli, and the other for auditory stimuli. Each of these perceptrons can form classes and generalizations by means of an error-correction procedure applied to the appropriate response sets. The added feature, however, is the extra association layer, which, in this system, comes after the R-units. The A-units in this set receive fixed connections from the R-units (which form a sort of retina for a second-order perceptron) and send back variable-valued α-system connections to the R-units. It is assumed that each R-unit (in both sets) receives connections from all of the $A^{(2)}$ units, and that the values of these connections can be corrected by an error-correction procedure, just as with the connections from the $A^{(1)}$ layer.

Suppose the perceptron has already been trained to recognize several kinds of visual objects (say squares and triangles) and has also been trained to recognize several spoken words ("square" and "triangle") for a variety of intonations, voice qualities, etc. During this training, the $A^{(2)}$ to R-unit back-connections have not been reinforced. Now let the perceptron hear the word "triangle", without any visual stimulus being present. The result will be an appropriate code-configuration in the R^{*} units, which will induce a characteristic state of the $A^{(2)}$ system, identifying the spoken word. By means of an error correction procedure, the perceptron can now be biased to give the R^{*} code for a triangle, and will hereafter tend to prefer this response to any others when the word "triangle" occurs. Consequently, when a composite stimulus is presented, as in Experiment 16, together with the spoken word "triangle", the system will tend to give the

R^{ν} response to the triangle, and due to the feedback connections to the A_{ν} set, and the action of the θ-servo, it will selectively augment the inputs to those A-units which respond to the triangle, while tending to suppress activity of A-units responding to other stimuli. Since all idiosynchratic forms of the spoken word, and all forms of the triangle-pattern, have been associated to identical response codes, the association will generalize immediately over both the audio class and the visual class of stimuli, without having to train the system with multiple examples of each.

Thus the four-layer perceptron can be made to direct its attention in response to spoken commands in much the same way as the previous models, but without requiring a modification of the A-R connections, or "perceptual organization" of the network, in forming the linguistic association. By a similar procedure, the $A^{(2)}$ to R^{e} connections can be reinforced in the presence of a visual pattern to create a bias, or "expentancy", favoring the perception of the word corresponding to the perceived object. By replacing the α-system back-connections from $A^{(2)}$ to the R-units with Γ-system connections (as in Equation 21.1)the association can be made to occur in a relatively spontaneous fashion, by presenting the visual image together with its spoken name. The result will be a reinforcement of the connections from the $A^{(2)}$ set which responds jointly to the visual and auditory codes; since this set will have many units in common with the separate audio and visual $A^{(2)}$ sets, the reinforcement will tend to generalize, to yield the desired result.

This system can be used for the problem of searching for a named object which is not currently present in the visual field. For this task, one must assume that the R^{∞} units are of a "flip-flop" variety, which tend to go on and stay on when they receive a sufficient input signal, until they are specifically cut off by a strong inhibitory signal. The system is taught to initiate an automatically controlled search or scan procedure in response to the spoken word "search". It is also trained (at the $A^{(2)}$ level) to turn off the search response whenever a coincidence occurs between a spoken name-code, and the visual object-code, but to leave the search-state alone when either the name or object, but not both, are present. Thus, given the command "Search for square", the word "search" initiates the search activity, and the word "square" sets the system to anticipate a square pattern. When a square appears in the field, the $A^{(2)}$ set corresponding to the combined object-code and word-code is activated, and transmits a strong inhibitory signal to the search response, turning it off. It would be possible to go a step farther, by training the perceptron (which has now isolated the set of A_{ν} units responding to the square) to continuously center the image of the square in the retina, using two continuous R-units to measure x and y displacements of the image from the center of the field (as in Section 10.2). Such a system, having found a moving stimulus, will track it and tend to keep it centered without being confused by the presence of extraneous objects in the field.

22. PROGRAM-LEARNING PERCEPTRONS

In the last chapter, we have seen that a back-coupled perceptron can be made to attend selectively to parts of a complex field, suppressing A-unit activity corresponding to objects other than the one attended to. In the last few paragraphs, it was also shown that such a perceptron can be made to anticipate decisions which are to be made at a future time, and execute them when the appropriate perceptual conditions are met. This lays the basis for the learning of sequential programs of responses in perceptrons.

Programmed activity is, of course, of supreme importance in carrying out logical sequences or algorithms, as in a digital computer. It also appears to provide a possible basis for the recognition of highly complex stimulus configurations, which depend on relations of simpler parts, rather than a fixed overall shape. The recognition of a human form, or an animal, is of this variety. It is also possible that the recognition of abstract topological relations -- a problem which has hitherto defied all perceptrons analyzed -- can be performed by means of a suitable programmed sequence of observations. This writer has become increasingly convinced that a passive filter-type system (such as a simple perceptron) cannot be designed which will economically recognize topological abstractions and relations such as "A and B are disjoint" or "A is inside B" or "A is a closed curve". On the other hand, a perceptron which can attend selectively to part of the stimulus pattern at a time, and carry out a sequence of observations under program-control, seems to offer a potential solution to this problem.

22.1 Learning Fixed Response Sequences

A perceptron of the back-coupled or cross-coupled variety can be taught to execute a fixed, stereotyped sequence of responses without introducing any new features in the system. If the sequence R_1^*, R_2^*, R_3^* is required, for example, when stimulus S_1 occurs, but the inverse sequence (R_3^*, R_2^*, R_1^*) when S_2 occurs, it is only necessary to associate the required responses to the succession of A-states which follow the stimulus in the cross-coupled system, or to the A-states which result from the interaction of the retinal input and the R-A feedback, in the back-coupled system. Of these two approaches, the cross-coupled system is more versatile, since it can be triggered by a momentary stimulus, and will not return to an identical state if the same response condition should occur at different points in the sequence. The cross-coupled system, however, requires that the response sequence occur with exact timing of each element. If the triggering or execution of each response takes an indeterminate amount of time, then a closed-loop system of the type shown in Figure 63 would be more appropriate. This system (which is also applicable to the recognition of strings of sensory events, such as words or speech sounds, where each element of the sequence is of indeterminate duration) employs an $A^{(2)}$ system with units which tend to lock on once they are activated, unless specifically triggered. These units are of the same variety as the "flip-flop R-units" employed in the R^* set in Section 21.3. The $A^{(2)}$ set is cross-coupled, with fixed connections, and feeds back (with fixed connections) to the $A^{(1)}$ set.

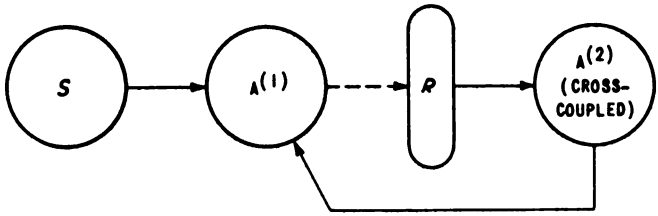


Figure 63 FOUR-LAYER PERCEPTRON FOR RECOGNITION AND CONTROL OF R-SEQUENCES WITH ELEMENTS OF INDETERMINATE DURATION.

When a response occurs in the R-set, it immediately triggers the $A^{(2)}$ system to assume some characteristic state. The parameters of the cross-coupling at the $A^{(2)}$ level can be so picked (e.g., by making all interconnections inhibitory) that the system will immediately assume a steady state, which will be held until some subsequent response occurs. When the second response of the sequence occurs, it finds the effective thresholds of the $A^{(2)}$ units modified by the cross-coupling signals from the units which are already on. Consequently, the $A^{(2)}$ state which occurs will depend not only on the new response, but also on the previous $A^{(2)}$ state. Unlike the previous cross-coupled systems, however, it does not depend on the time-lapse since the previous input, since the $A^{(2)}$ state has held steady over the interval.

By means of the feedback to the $A^{(1)}$ set, the $A^{(2)}$ state (and consequently the response sequence) can be made to modify the response of the $A^{(1)}$ system to the present stimulus. Thus a distinct succession of responses can be associated to the stimulus, each new $A^{(1)}$ state signifying the joint information that the stimulus is present, and that a particular succession of responses has occurred in the past. To terminate such a sequence, it is possible to assume that one of the R-units has inhibitory connections to all $A^{(2)}$ units, so that when the end of the sequence is recognized, the $A^{(2)}$ system can be reset to its inactive state, by turning on this response.

22.2 Conditional Response Sequences

In the last section, the response sequences learned by the perceptron were assumed to be of a fixed, stereotyped variety, such as the utterance of a given word or phrase, or the execution of a particular sequence of movements. Of more general interest, is the possibility of conditional response sequences, where the execution of the next step depends upon the realization of a set of conditions at the present time.

In a limited sense, we have already demonstrated the possibility of conditional responses in the perceptron of Figure 63, where the next response depends not only upon the preceding R-sequence, but also upon the continuation of the initiating stimulus. A more interesting case, however, would be one in which the next response depends upon the recognition of some condition which results from the preceding activity of the perceptron itself. For example, if the perceptron is equipped with a move-

able appendage by means of which it can apply pressure to external objects, we might ask it to push aside any object placed in front of it. Such objects might have their movement blocked, either to the right of to the left, in which case the perceptron might first bring its "pushing arm" into contact with the left side of an object and try pushing to the right, but if it finds that the object remains stationary, it must reverse the position of its arm, and push to the left.

Such a decision program still seems to be within the capability of a perceptron of the type just described. It must recognize (through its visual inputs) the conditions "no object present","object present to right of arm location", "object present to left of arm location", arm in contact with left side but object stationary", "arm in contact with left side and object moving", etc. The recognition of the contact conditions might be facilitated by the inclusion of pressure transducers on the arm, providing an auxiliary sensory input to the association system. An appropriate response sequence must then be associated to each of these conditions. For example, if the condition "arm in contact with left side but object stationary" is recognized, the response sequence might be

1. Retract arm
2. Shift arm position to right
3. Extend arm

This would then yield the condition "object present to left of arm location", for which the response would be

1. Shift arm to left until it contacts object
2. Apply pressure

The conditions of "moving" and "stationary" objects can, of course, be recognized by a perceptron with time delays from the retina to the $A^{(1)}$ units, so that there is nothing in the above description which cannot be done, in principle, by perceptrons which have already been analyzed.

22.3 Programs Requiring Data Storage

In all of the sequential programs considered above, the next step has been determined entirely by the conditions at the previous step, and a knowledge of how many steps have already occurred in the current sequence. More elaborate programs require a conditional response based on information which was available several steps previously, but is no longer present in the sensory input. The perceptrons considered so far can solve such problems only by anticipating all possible sequences of conditions, and learning a unique response sequence for each special case. This rapidly becomes impractical, as the sequences become more involved. An example of such a problem is counting. In counting from zero upwards, we first produce a sequence of single digits, from one through nine; we then add a second digit (a one) and reset the low order digit to zero. The one in second place is held fixed, while the low order digits are recycled, and is then changed to two, and so forth. At an advanced stage in this procedure, we may be holding three or four high-order digits "in memory" while modifying the low-order digits. To perform such a program expeditiously, an internal storage mechanism is required, which can be set to hold a given item of information and read out or altered whenever required. Such a memory mechanism is much more like a conventional digital computer memory than anything yet encountered in perceptron theory.

While it is fairly easy to contrive systems which employ rigidly determined gating mechanisms and more-or-less conventional computer memory logic to provide a temporary storage device for a perceptron, no realy satisfying solution has been found to date. A biological system undoubtedly employs something more subtle than a coded address system which transmits its stored information on command, but the similarity in logical requirements nonetheless suggests that there might be a similarity in structure at this particular point. It should be remembered, however, that human ability to perform complex algorithms without extensive practice and learning time does not begin to approach that of a digital computer. The human computer also tends to rely heavily on such external aids as pencil and paper to augment his memory for relevant data, and with the aid of an external transcription of its outputs, a perceptron can also be made to perform rather elaborate logic (in the manner of section 22.2).

Some possible cues as to the nature of temporary data storage in the human brain come from introspective observations of recall of strings of digits, words, or melodies, and such exercises as attempting to count in binary up to the point where one loses track of the number on which one is operating. In all of these cases, recall is helped by rhythmic grouping of elements, and by visualization or auditory imagery of the elements in a continuously recurrent sequence. It seems likely that an active memory, such as a reverberating loop system, which continuously rewrites itself on every rehearsal of the stored information, is involved.

22.4 Attention-Scanning and Perception of Complex Objects

The preceding sections have dealt with the phenomena of program learning with respect to response sequences. A capability for program learning is also useful for the direction of attention over a sensory field, and the perception of a complex pattern or object by noting its parts and the relations between them. The possibility of directing attention selectively to part of the visual field was already observed in the last chapter. A program-controlled perceptron could, therefore, be taught to direct its attention successively to different parts of the field in some systematic order, e.g., to scan from left to right, or top to bottom. It is also plausible (although it remains to be demonstrated) that a back-coupled perceptron can be taught to shift its field of attention along a contour, or edge of a figure, so that the association set, at any one time, responds only to part of the contour. Such a system, by starting at one point on a curve and following it in one direction, could determine whether the curve is closed or open by indicating whether the scan process returns to its starting point without having lost the contour at any time.

In the recognition of a complex structured object, such as a man (regardless of posture, angle of view, etc.) a program of observations might note significant parts and the transitions between them. There should, for example, be a head joined to the shoulders, and by following a path from one of the hands, the system should successively come to a forearm, shoulder, and torso. The reader may recognize a similarity between this suggestion and Hebb's concept of a "phase sequence" (Ref. 33). The phase sequence consists of a progression of cell-assemblies, each of which represents some

elementary perceptual fragment, the entire sequence representing a perception of a complex stimulus or experience. In the perceptron, however, the progression of states is assumed to be under the control of a learned program, which directs the attention of the system in such a way as to make first one set of A-units, then another set achieve dominance, by the mechanisms described in Chapter 21. A sequence-recognizing system, such as the five-layer perceptron shown in Figure 64, would be required for the direction of the scanning process and for the recognition of the total configuration from its parts. This system employs an $A^{(2)}$ layer of the same type as in Figure 63 (cross-coupled, with fixed interconnections, and A-units which hold their state until triggered by a sufficiently strong signal to change). The $A^{(2)}$ set in this model, however, has variable-valued connections both to a new $R^{(2)}$ set, which can learn to recognize complex patterns from sequences of parts, and also back to the $R^{(1)}$ units, so that the system can be taught to direct its attention in a systematic manner to look for anticipated parts of the complex.

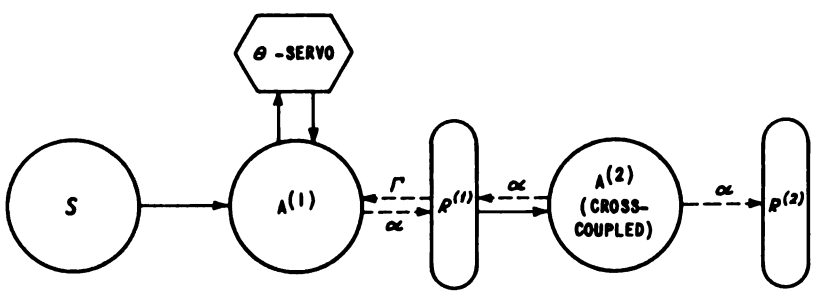


Figure 64 FIVE-LAYER PERCEPTRON FOR RECOGNITION OF COMPLEX PATTERNS BY ATTENTION SCANNING PROGRAMS. (BROKEN ARROWS INDICATE VARIABLE CONNECTIONS).

22.5 Recognition of Abstract Relations

It is apparent that the perceptrons proposed above are already stretching the limits of what has been firmly established analytically and experimentally. While there is good reason to think that the proposed systems would work in principle, they are highly speculative, and we are far from being able to describe their performance in quantitative terms. Nonetheless, one further venture in extrapolation seems to be of interest: As was previously noted, the recognition of abstract topological relations (or metric relations, for that matter) cannot be performed economically by a perceptron which is required to grasp the relation instantaneously from a complex pattern. The relation "A is inside of B", for example, would require that the system be trained with all possible cases of "A inside B" and "A outside B", even after it has been taught to identify patterns "A" and "B" correctly. It seems more likely that a program-controlled perceptron, having been taught to recognize patterns A and B, can determine whether A is inside of B by means of a directed scanning process.

Suppose we show the perceptron a complex field, containing a circle and a square, both of which it has previously been taught to identify, and we ask the system to indicate whether the circle is inside or outside the square. This question could be answered by means of two attention sweeps, beginning at the circle and first sweeping to the right, then returning to the circle and sweeping to the left. If an edge of the square is encountered on one of the two sweeps but not on the other, then the circle is "outside" the square; if an edge is encountered both to the right of the circle and to the

left, the circle is "inside" the square. A somewhat more elaborate program would determine whether a known figure (e.g., a square or triangle) is inside or outside of an arbitrary closed curve.

In the recognition of topological relations or metric relations (A is larger than B, or A is above B), and in programs which call for attention scanning, it would probably help considerably to introduce geometric constraints into the S-A and A-A connections of the perceptron. In the models which have been of primary interest up to this point, there is no way of telling, apart from learned association, that activity of a particular A-unit refers to a particular region of the sensory field. The A-unit space is non-topological in character; it has no well-defined geometry or dimensionality. This means that, apart from learning, there is no way of telling from observations on the state of the A-units, what are the topological or geometrical properties of the stimulus which is present on the retina. While it seems likely that a geometrically constrained organization of A-unit connections (e.g., increasing the probability of interconnection between A-units whose retinal fields lie in close proximity to one another) would be helpful, there is still no indication of what are the best constraints, or what gain in performance can actually be realized by such means.

23. SENSORY ANALYZING MECHANISMS

The term "sensory analyzing mechanism" will be used for any signal transmission unit or network which detects and transmits information about selected parts or features of a total stimulus pattern. Such mechanisms can frequently be used to reduce the amount of information which the perceptron must be prepared to evaluate. They are particularly useful in highly organized environments (such as the familiar visual environment, or an environment of printed words or spoken language) where purely random stimuli are unlikely to occur or are of little interest. Thus a mechanism which detects boundaries of a solid image or describes gradients and contrasts in the visual field, or performs a Fourier analysis of an audio input, or which encodes speech into a sequence of phonemes, would be considered a sensory analyzing mechanism. A simple sensory unit which detects the level of illumination at a given point, or an A-unit which samples the illumination over a selected set of points are also sensory analyzing mechanisms.

In most models considered thus far, little attempt has been made to optimize the sensory analyzing mechanisms employed. The random origin configurations which have generally been employed can be shown to be far from optimum. In this chapter, various methods of improving this primitive organization will be considered, particularly with respect to visual and auditory systems. For the most part, these mechanisms are assumed to take the form of built-in constraints, such as were considered briefly in the d.i.d. models of Section 7.2.2, and the similarity-constrained perceptrons of Section 15.3. The existence of such mechanisms in biological organisms is supported by an increasing amount of evidence, such as Lettvin's studies of

frog vision (Ref. 51), Sutherland's studies of octopus vision (Ref. 98), Gibson and Walk on depth perception (Ref. 24), Sauer's work on bird navigation (Ref. 90), and Hubel's work on cat vision (Ref. 113). Since most of these mechanisms appear to be hereditary rather than learned, it seems likely that they may be realized either by simple spatial constraints in the distributions of connections in the sensory network, or else by simple "typological constraints" governing the kinds of cells which may be interconnected.

23.1 Visual Analyzing Mechanisms

A number of basic strategies for processing visual information have been proposed. Some of these are so closely tied to digital computer processes that they are of little interest for a biological model, while others require such a degree of logical precision and so large a system as to be biologically implausible (e.g., Refs. 16, 17, 71). The techniques to be considered here are grouped under four main headings: (1) Local property detectors; (2) Hierarchical retinal field organizations; (3) Sequential programs (centering and scanning methods); and (4) Sampling of sensory parameters. The possible advantages of each of these methods will be considered (largely in a qualitative fashion), and the problem of an optimum mixture of analyzing mechanisms (somewhat analogous to the "mixed strategy" problem in game theory) will be discussed.

23.1.1 Local Property Detectors

The term "local property detector" will be used for any mechanism or neuron which responds to some particular feature of the stimulus pattern at a particular location (for example, brightness, color,

contour direction, etc.). Contour detectors and other types of property detectors have been described by Culbertson (Ref. 17), Taylor (Ref. 99), Inselberg, Löfgren, and von Foerster (Ref. 4), and others. Lettvin and associates (Ref. 51) have described four mechanisms (for detection of contrast, convexity, or small spot detection, moving edge detection, and dimming detection) which appear to map into four distinct layers of the frog's tectum. Of particular interest for present purposes is the series of experiments described by Hubel (Ref. 113), in which the cells of a cat's visual cortex are shown to respond to lines and bars in particular positions and orientations, or to stimuli moving in particular directions.

The visual property detectors which appear on an a priori basis to be of maximum value for pattern recognition in an ordinary terrestrial environment (where the main purpose of the system is to detect and recognize coherent physical objects) include the following:

1) Brightness and color detection and measurement
2) Contour and gradient detection
3) Curvilinearity detection and measurement
4) Detection of angles, intersections, and discontinuities of lines and boundaries
5) Spot detection
6) Sensing of textures, and measurement of texture gradients
7) Velocity and accelleration detection and measurement

In order to recognize stimulus patterns or objects, information of the types listed above must somehow be combined for different parts of the retina, to provide an indication of the total configuration. This has been the main job of the association units, in the perceptrons considered thus far. In all cases considered in previous chapters, the A-units have formed combinatorial functions of information coming from "local intensity detectors" (the S-units); thus the only property detectors employed have been of the first type. The perceptron illustrated in Figure 65 introduces an additional layer of A-units immediately following the S-units, which can detect additional properties of the types indicated above. The $A^{(2)}$ layer, having its origin points in the $A^{(1)}$ layer, now responds to combinations of local properties such as lines and gradients, rather than merely to points of light.

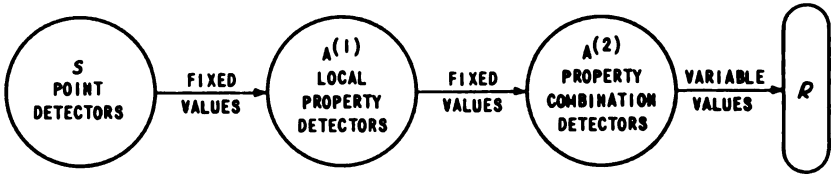
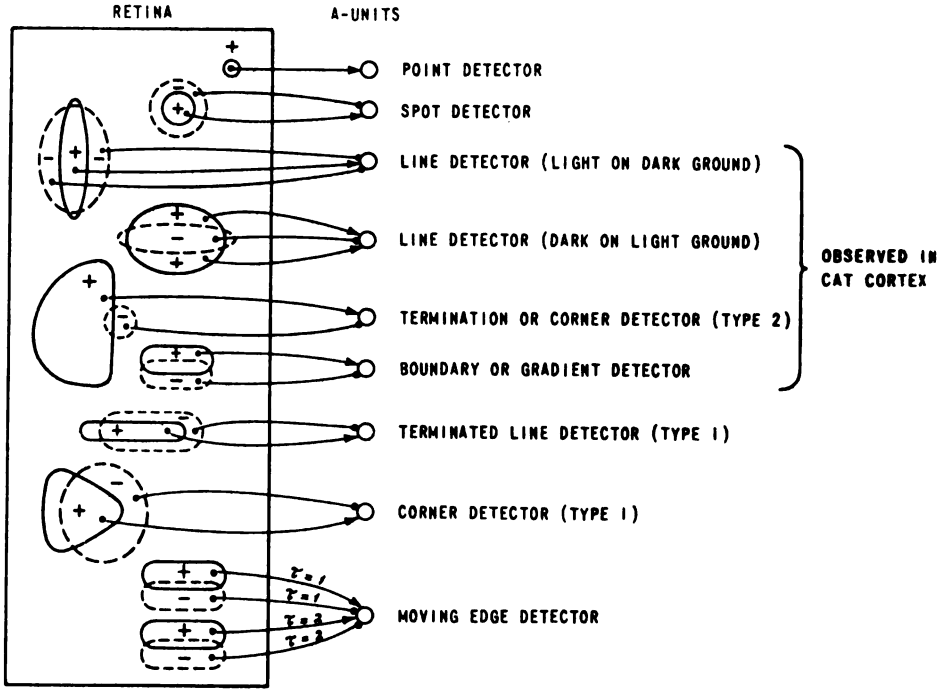


Figure 65 ORGANIZATION OF A PERCEPTRON EMPLOYING LOCAL PROPERTY DETECTORS.

The organization of origin fields for A-units serving as property detectors of various types is illustrated in Figure 66. The single-connection "point detector" serves merely as a logical relay for information which could be obtained equally well directly from the retina. The concentric field organization of the spot detector appears to be found (in the case of the cat) more characteristically in the retinal ganglion cells than in the visual cortex (Ref. 113). The various forms of line detectors and the "Type 2" termination detector have all been observed in the cat's cortex by Hubel. Hubel has also reported units which respond only to moving stimuli, although the organization appears to be different from that suggested in Fig. 66(a), for the "moving edge detector". There is some evidence that the movement detectors in the cat rely more upon the simultaneous summation of "off" signals from uncovered retinal points and "on" signals from retinal points which have just been covered by the displaced stimulus.

The use of the Type 2 termination detectors is illustrated in Fig. 66(b). An $A^{(2)}$ unit which receives connections both from a termination detector and a line detector crossing the same field can recognize that the line approaches the inhibitory spot of the termination detector, but does not cross it. The same termination detector, taken in conjunction with lines at different angles, can serve to indicate termination of any one of the lines, so that there is considerable saving by this method. In fact, if there are k discriminable angles for straight lines, and r discriminable translates of each line, (so that there are about r^2 distinguishable termination-points scattered over the retina) then a system which employs Type 1 termination detectors would require a total of $r^2 k$ $A^{(1)}$ units to guarantee a detector

(a) ORGANIZATION OF SENSORY FIELDS OF A $^{(1)}$ UNITS. BROKEN LINES INDICATE FIELDS OF INHIBITORY ORIGIN POINTS; SOLID LINES INDICATE EXCITATORY FIELDS.



(b) TYPICAL A $^{(2)}$ COMBINATIONS. POSITION OF RETINAL FIELDS OF A $^{(1)}$ UNITS IS SHOWN RELATIVE TO FIXED AXES, FOR EACH UNIT.

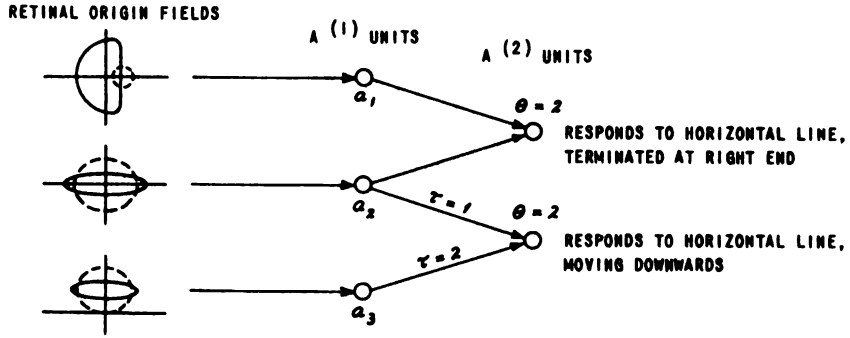


Figure 66 ORIGIN FIELD ORGANIZATIONS FOR LOCAL PROPERTY DETECTORS

for each combination of angle and termination point. The use of Type 2 detectors in conjunction with line detectors (as in Fig. 66(b)) would require only $r^2 + r\ell$ $A^{(1)}$ units, to convey the same information. If r and ℓ are both equal to 100, this means that 10^6 $A^{(1)}$ units are required with Type 1 units, and 2×10^4 with Type 2 units. This may indicate why the Type 2 configuration appears to be found in the cat, rather than the Type 1 configurations.

Figure 66(b) also demonstrates the multiple use of the same elementary property detectors ($A^{(1)}$ units) for a number of more complex functions at the $A^{(2)}$ level. Thus, the unit a_2 is employed both in a terminated line detector and also as part of a moving line detector. Since movement detection can thus be obtained quite economically at the $A^{(2)}$ level, the type of moving edge detector illustrated in Figure 66(a) would tend to be obviated. Hubel's observations on the cat suggest that (although more complex organizations may remain to be discovered) the most prominent types of property detectors in the visual cortex are of very simple types, such as the line and boundary detectors and Type 2 termination detectors illustrated in Figure 66(a). In all of these cases, a single excitatory and inhibitory field, with simple constraints on the density of connections of each type, is sufficient to yield the mechanism indicated.

The actual advantages which might be realised by means of various types of property detectors have been investigated for several simple discrimination problems, with the results shown in Table 10. Two types of environments were considered: the first consists of the letter "T"

in right-side-up and upside-down orientations, and the second consists of the letter "L", also right-side-up and upside-down. Each letter can appear in all translational positions. The problem of discriminating the right-side-up "T" from the upside-down "T" is considered for a variety of retinal sizes ranging from 20 x 20 to 1000 x 1000. The retina is assumed to be torroidally connected in all cases. With both the T and the L , the horizontal line is taken to be nine units long, while the height of the letter is ten units. The thickness of the lines is one unit, throughout. The perceptrons considered are of the type shown in Figure 65, with the assumption that all inputs to A-units are excitatory. Rather than attempting to find optimum parameters for the various types of property detectors, the number of $A^{(2)}$ inputs is always the minimum number which will permit the discrimination to be achieved. Other parameters (and the introduction of inhibitory connections) would undoubtedly permit more economical solutions, but this serves to illustrate basic principles.

The table gives the probabilities of finding $A^{(2)}$ units which will discriminate between a given stimulus of the "positive" class (say the upright position) and all members of the opposite class. The origin points of the $A^{(2)}$ units are assumed to be chosen at random from among the $A^{(1)}$ units. The first line of the table, in which the $A^{(1)}$ units are simple point detectors, corresponds to the case of a simple perceptron, where each A-unit receives its input connections directly from the retina. For such a system, it can easily be seen that at least two excitatory origins and a threshold of 2 are required in order to distinguish between the upright and upside-down "L", while three excitatory origins and a threshold

TABLE 10. COMPARISON OF STIMULUS ANALYZING MECHANISMS FOR 5 DISCRIMINATION PROBLEMS. $A^{(2)}$ ORIGINS SELECTED AT RANDOM

(UNDERLINING INDICATES BEST SYSTEM FOR EACH PROBLEM. ALL STIMULI ARE 9 X 10.)

$A^{(1)}$ ORIGIN CONFIGURATION	$\theta^{(1)}$	NO. OF CONNECTIONS TO $A^{(2)}$ UNIT ($= \theta^{(2)}$)	PROBABILITY OF USEFUL $A^{(2)}$ UNIT				
			L vs. Γ IN 20 X 20 RETINA	T vs. $\perp$ IN 20 X 20 RETINA	T vs. $\perp$ IN 40 X 40 RETINA	T vs. $\perp$ IN 100 X 100 RETINA	T vs. $\perp$ IN 1000 X 1000 RETINA
SINGLE POINT, IN RANDOM LOCATION	1	2	9.00×10^{-4}	0	0	0	0
	1	3	NOT COMPUTED	1.15×10^{-4}	1.79×10^{-6}	7.34×10^{-9}	7.34×10^{-15}
4 POINTS IN 2 X 2 SQUARE	1	2	3.19×10^{-3}	0	0	0	0
	1	3	NOT COMPUTED	3.97×10^{-4}	6.20×10^{-6}	2.54×10^{-8}	2.54×10^{-14}
16 POINTS IN 4 X 4 SQUARE	1	2	7.15×10^{-3}	0	0	0	0
	1	3	NOT COMPUTED	1.40×10^{-3}	3.49×10^{-5}	1.43×10^{-7}	1.43×10^{-13}
5 POINTS IN LINE, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	2	2	2.75×10^{-4}	2.75×10^{-4}	1.93×10^{-5}	4.95×10^{-7}	4.95×10^{-13}
10 POINTS IN LINE, HORIZ. OR VERT. WITH EQUAL PROB.	2	2	1.50×10^{-4}	1.50×10^{-4}	2.81×10^{-5}	7.20×10^{-7}	7.20×10^{-13}
10 POINTS IN 2 X 5 BAR, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	3	2	9.00×10^{-4}	9.00×10^{-4}	5.63×10^{-5}	1.44×10^{-6}	1.44×10^{-12}
20 POINTS IN 2 X 10 BAR, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	3	2	7.00×10^{-4}	7.00×10^{-4}	8.75×10^{-5}	2.24×10^{-6}	2.24×10^{-12}
20 POINTS IN 4 X 5 BAR, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	5	2	1.50×10^{-3}	1.50×10^{-3}	9.38×10^{-5}	2.40×10^{-6}	2.40×10^{-12}

of 3 are required to distinguish the upright from the upside-down "T".* The figures in the first two columns of the table are influenced by small-retina effects, which disappear for the 40 x 40 and larger retinas.

Several general conclusions can be drawn from this table. First of all, it is clear that the value of different types of property detectors depends upon the stimuli to be discriminated as well as the size of the retina. For the discrimination of the L-shaped stimuli, which require only two points or blobs for discrimination, the best results are obtained with large (4 x 4) square origin point configurations for the $A^{(1)}$ units, while for the T's a slightly elongated (4 x 5) configurations with a high threshold is preferable, since it permits the use of only two $A^{(1)}$ units instead of three per $A^{(2)}$ unit. Note that the advantage of the rectangular origin configuration over the 4 x 4 square is pronounced only for large retinal sizes, however; for a smaller retina than 20 x 20, the square configuration might actually be preferable. For the conditions considered in this analysis, the following equation for the probability of a useful $A^{(2)}$ unit shows the effect of increasing retinal size:

$$P = \frac{m}{(rN_A)^k} \tag{23.1}$$

---

* The reader may find it instructive to examine the Q-matrices for a binomial perceptron in these problems, and satisfy himself that they are consistent with the geometrical requirement that three inputs and a threshold of 3 are required to discriminate between the upright and upside down "T", in all translational positions.

where N_s = number of S-points in retina

m = number of useful combinations of $A^{(1)}$ origin configurations for an $A^{(2)}$ unit

r = number of admissible rotational positions for each $A^{(1)}$ configuration

k = number of input connections to each $A^{(2)}$ unit

For a large retina, P clearly becomes small very rapidly, and the situation is aggravated by the requirement of many inputs for each $A^{(2)}$ unit. Thus for the discrimination of the upright and upside-down T , which requires three point inputs, P goes from 10^{-4} for a 20 x 20 retina to about 10^{-16} for a 1000 x 1000 retina. The use of 4 x 5 bars as line detectors instead of point configurations, while it improves the probability by more than three orders of magnitude, still leaves a requirement for over 10^{12} $A^{(2)}$ units if the T is to be discriminated reliably in the large retina. Even with optimum parameters, the required number of $A^{(2)}$ units is inadmissibly large. Nonetheless, the recognition of the position of a 9 x 10 "T" in a 1000 x 1000 field is certainly well within the limits of human vision. Some additional means must therefore be found, to provide an economical solution for this problem without introducing a brainful of special "T-detectors". The principles discussed in the following section, combined with the use of property detectors, will be seen to yield a radical improvement in the recognition of small stimuli.

23.1.2 Heirarchical Retinal Field Organizations

The "retinal field" of an A-unit is the region of the retina in which its origin points may be found. In a multi-layer system, the retinal field of an $A^{(2)}$ unit is the union of the retinal fields of the $A^{(1)}$ units which are connected to the $A^{(2)}$ unit; in general, the retinal field of an $A^{(k)}$ unit is the union of the retinal fields of the connected $A^{(k-1)}$ units. In a perceptron with a heirarchical retinal field organization, the retinal fields of the A-units tend to increase in area, the greater the logical distance of the A-unit from the retina. For example, the $A^{(1)}$ units may have highly localized origin configurations for the detection of local properties (as in Table 10); the $A^{(2)}$ units could then detect combinations of properties over a somewhat larger field (responding to small, compact figures or parts of larger patterns); and a layer of $A^{(3)}$ units might then be added to respond to combinations of sub-figures over the entire retina. While the general principle of organization is from small to large retinal fields as the A-units increase in depth, it is not required that all A-units at a given level have retinal fields of the same size; there may be $A^{(1)}$ units, for example, whose fields are larger than the smallest $A^{(1)}$ fields, provided the expected size of the retinal fields increases with increasing depth.

Such a system is clearly much closer to the organization of the mammalian visual system than the uniform origin distributions which were considered in previous models. A brief consideration was given to constrained origin fields in Section 7.2.2, where it was found that no appreciable gain in performance was obtained with large stimuli, such as the squares and triangles of Experiment 7. The effects of employing cons-

trained retinal fields for the $A^{(2)}$ units in the perceptron of Figure 65 will now be considered, for the range of retinal sizes shown in Table 10. It was found in the preceding section that as the retina becomes large relative to the size of the stimuli, the probability of finding a useful $A^{(2)}$ unit becomes inadmissibly small in the unconstrained system. Table 11 shows the effect of limiting $A^{(2)}$ retinal fields to a 20 x 20 region of the retina (located at random in a larger retina). Again, it should be remembered that the parameters have not been optimized, and that appreciably better results might be obtained with larger numbers of inputs to the $A^{(2)}$ units, and the inclusion of inhibitory connections. Nonetheless, a comparison with Table 10 illustrates the marked improvement in the size of the system necessary to achieve recognition in a large retina. The first column of probabilities (for the 20 x 20 retina) is, of course, identical to the corresponding column of Table 10, and the first line corresponds to a three-layer model with constrained origin fields for the A-units. In the case of the 1000 x 1000 retina, using the best of the $A^{(1)}$ origin configurations, a gain of more than five orders of magnitude is obtained, bringing the discrimination problem for the first time within the capacity of a human-sized brain model. Note, however, that the best $A^{(1)}$ origin configuration has shifted from the 4 x 5 bar with $\theta = 5$ to the 4 x 4 square with $\theta = 1$.

The probability P' of finding a useful $A^{(2)}$ unit in this system is given by the following equation, which is analogous to (23.1):

$$P' = \frac{m}{(rN_s')^k} \cdot \frac{N_s'}{N_s} = \frac{m}{r^k N_s'^{k-1} N_s} \qquad (23.2)$$

TABLE II. COMPARISON OF STIMULUS ANALYZING MECHANISMS FOR PERCEPTRONS WITH 20 X 20 RETINAL FIELDS FOR A $^{(2)}$ UNITS. DISCRIMINATION OF $\top$ VS. $\perp$.

(UNDERLINING INDICATES BEST SYSTEM FOR EACH PROBLEM. STIMULI ARE 9 X 10.)

A $^{(1)}$ ORIGIN CONFIGURATION	$\theta^{(1)}$	NO. OF CONNECTIONS TO A $^{(2)}$ UNIT (= $\theta^{(2)}$)	PROBABILITY OF USEFUL A $^{(2)}$ UNIT			
			20 X 20 RETINA	40 X 40 RETINA	100 X 100 RETINA	1000 X 1000 RETINA
SINGLE POINT, IN RANDOM LOCATION IN RETINAL FIELD	1	3	1.15×10^{-4}	2.86×10^{-5}	4.59×10^{-6}	4.59×10^{-10}
4 POINTS IN 2 X 2 SQUARE	1	3	3.97×10^{-4}	9.92×10^{-5}	1.59×10^{-5}	1.59×10^{-9}
16 POINTS IN 4 X 4 SQUARE	1	3	1.40×10^{-3}	$\underline{5.58 \times 10^{-4}}$	$\underline{8.94 \times 10^{-5}}$	$\underline{8.94 \times 10^{-9}}$
5 POINTS IN LINE, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	2	2	2.75×10^{-4}	7.72×10^{-5}	1.24×10^{-5}	1.24×10^{-9}
10 POINTS IN LINE, HORIZ. OR VERT. WITH EQUAL PROB.	2	2	1.50×10^{-4}	1.12×10^{-4}	1.80×10^{-5}	1.80×10^{-9}
10 POINTS IN 2 X 5 BAR, HORIZ. OR VERT. WITH EQUAL PROBABILITY	3	2	9.00×10^{-4}	2.25×10^{-4}	3.60×10^{-5}	3.60×10^{-9}
20 POINTS IN 2 X 10 BAR, HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	3	2	7.00×10^{-4}	3.50×10^{-4}	5.60×10^{-5}	5.60×10^{-9}
20 POINTS IN 4 X 5 BAR HORIZ. OR VERTICAL WITH EQUAL PROBABILITY	5	2	$\underline{1.50 \times 10^{-3}}$	3.75×10^{-4}	6.00×10^{-5}	6.00×10^{-9}

where m, r, and k are defined as for equation 23.1, N_s = number of S-points in the retina, and N_s' = number of S-points in the retinal field of an $A^{(2)}$ unit. Taking the ratio of equations (23.2) and (23.1), we obtain the relative advantage of the constrained retinal field system over the unconstrained system:

$$\frac{P'}{P} = \left(\frac{N_s}{N_s'}\right)^{k-1} \tag{23.3}$$

Thus the advantage increases exponentially with the number of connections required to each $A^{(2)}$ unit, and with the ratio N_s/N_s'. Both of these effects can be seen in Table 11.

Clearly, if the system is required to recognize a stimulus of diameter D, the size of the retinal field cannot be taken smaller than D, without loss of performance; the above equations assume that the retinal field is large enough so that boundary effects can be neglected. The optimum size, then appears to be on the order of D, the expected stimulus diameter. We now have the problem of how to deal with universes of stimuli which vary in diameter from very small to very large patterns. The best choice of a distribution of retinal field sizes for the $A^{(2)}$ units will generally be one which guarantees the same likelihood of finding a useful $A^{(2)}$ unit for all stimuli. For the particular case in which the stimulus diameter distribution is uniform between the limits D_{min} and D_{max}, this can be approximately realized by taking

$$\text{Prob}\,(A = D^2) = 1/D^2\,\Sigma \tag{23.4}$$

where A = fraction of retinal area in an $A^{(2)}$ retinal field

$$\Sigma = \sum_{D_{min}}^{D_{max}} \frac{1}{D^2} \quad \text{(where } D \text{ is measured in retinal diameters)}$$

Table 11 suggests that stimuli of the complexity of alphabetic characters ranging in size from .01 to 1 retinal diameter can be recognized by a system the size of the human brain (10^{10} units) by employing a four-layer model, with a suitable combination of property detector configurations and a suitable distribution of $A^{(2)}$ field diameters. The recognition problem can be made considerably more difficult, however, by adding additional degrees of freedom to the stimulus organizations. Consider, for example, the following environment: Let W consist of two classes of composite stimuli. Each stimulus consists of two 9 x 10 T's , which may be located at any position in the retinal field, provided they are at least 10 retinal units apart. If both T's are right-side-up or if both are upside-down, the stimulus is a member of the positive class; if one is right-side-up and the other is upside-down, the stimulus is in the negative class. Let us consider the probability of finding a useful $A^{(2)}$ unit for this dichotomy.

If these stimuli are to be differentiated by A-units with random-point origin configurations (all excitatory, as in the previous examples) then six connections and a θ of 6 is required for each $A^{(2)}$ unit. By employing one of the line-detector mechanisms of Table 10, 4 inputs and a θ of 4 are required. The constrained-field system of Table 11 (with 20 x 20 retinal fields for the $A^{(2)}$ units) cannot be employed here, as the combined stimulus

pattern may cover the entire retinal field. The best that can be done is to employ the $A^{(1)}$ configuration of 4 x 5 bars, which yields a probability of 6×10^{-25} of finding a useful $A^{(2)}$ unit, with a 1000 x 1000 retina. (For the single random point configuration -- the worst case -- the probability is 7.34×10^{-33}.)

By employing a five-layer topology, it is possible to take advantage of the fact that each stimulus actually consists of two organized sub-patterns, each having quite small dimensions relative to the retina. Assume the $A^{(2)}$ units to have 20 x 20 retinal fields, as in Table 11, while the $A^{(3)}$ units have two excitatory input connections, chosen at random from among the $A^{(2)}$ units. Thus the $A^{(1)}$ units serve as local property detectors, the $A^{(2)}$ units serve as sub-pattern detectors, and the $A^{(3)}$ units integrate this information over the whole retinal field. (In this particular problem, the performance could be improved further by taking a larger number of input connections for each $A^{(3)}$ unit, but as before, we are trying to demonstrate basic principles rather than find optimum organizations.) This five-layer system is compared with the four-layer system in Table 12. For moderate numbers of connections to the $A^{(3)}$ units in this system, the probability of a useful $A^{(3)}$ unit (with $\theta = 2$) can be closely approximated by the binomial probability:

$$P'' = \binom{n^{(3)}}{2} P'^{2} (1-P')^{n^{(3)}-2} \tag{23.5}$$

where P' = probability of a useful $A^{(2)}$ unit for "sub-figure" discrimination, and

$n^{(3)}$ = number of (excitatory) input connections to an $A^{(3)}$ unit

TABLE 12. 4-LAYER SYSTEM WITH UNCONSTRAINED A $^{(2)}$ FIELDS VS. 5-LAYER SYSTEM WITH 20 X 20 A $^{(2)}$ FIELDS AND 2 INPUTS TO EACH A $^{(3)}$ UNIT. ($\top\top$ AND $\perp\perp$ VS. $\perp\top$).

A $^{(1)}$ ORIGIN CONFIGURATION	$\theta^{(1)}$	$\theta^{(2)}$	PROBABILITY OF USEFUL A $^{(2)}$ OR A $^{(3)}$ UNIT			
			100 X 100 RETINA		1000 X 1000 RETINA	
			4 LAYERS	5 LAYERS	4 LAYERS	5 LAYERS
SINGLE POINT, IN RANDOM LOCATION	1	3	7.34×10^{-21}	2.11×10^{-11}	7.34×10^{-33}	2.11×10^{-19}
4 POINTS IN 2 X 2 SQUARE	1	3	2.54×10^{-20}	2.53×10^{-10}	2.54×10^{-32}	2.53×10^{-18}
16 POINTS IN 4 X 4 SQUARE	1	3	1.43×10^{-19}	7.99×10^{-9}	1.43×10^{-31}	7.99×10^{-17}
5 POINTS IN LINE, HORIZ. OR VERT. WITH EQUAL PROB.	2	2	1.24×10^{-15}	1.54×10^{-10}	1.24×10^{-25}	1.54×10^{-18}
10 POINTS IN LINE, HORIZ. OR VERT. WITH EQUAL PROB.	2	2	1.80×10^{-15}	3.24×10^{-10}	1.80×10^{-25}	3.24×10^{-18}
10 POINTS IN 2 X 5 BAR, HORIZ. OR VERT. WITH EQUAL PROB.	3	2	3.60×10^{-15}	1.30×10^{-9}	3.60×10^{-25}	1.30×10^{-17}
20 POINTS IN 2 X 10 BAR, HORIZ. OR VERT. WITH EQUAL PROB.	3	2	5.60×10^{-15}	3.14×10^{-9}	5.60×10^{-25}	3.14×10^{-17}
20 POINTS IN 4 X 5 BAR, HORIZ. OR VERT. WITH EQUAL PROB.	5	2	6.00×10^{-15}	3.60×10^{-9}	6.00×10^{-25}	3.60×10^{-17}

Thus with 25 inputs to each $A^{(3)}$ unit the probabilities for the five-layer systems could be increased by a factor of about 300. Note that even under these conditions, however, while the problem becomes soluble for a brain-sized system in the case of a 100 by 100 retina, it is still unmanageable in the 1000 x 1000 retina.

The difficulty of this problem for the large retina should not surprise us; it is unlikely that a human subject, asked to perform the indicated discrimination with tachistoscopically presented stimuli, could do appreciably better than chance, where the two T's each subtend only 1/100 of the central visual field, and are located at random relative to one another. Even the case of the 100 x 100 retina (where the T's subtend 1/10 the diameter of the field) would probably yield marginal results, if the subject were not permitted time to scan the field or shift his attention during the exposure. On the other hand, if the T's were constrained to lie relatively close to one another (say within a 40 x 40 subfield) the problem would probably not be difficult. This problem, however, could readily be handled by a five-layer perceptron in which the $A^{(3)}$ retinal fields were constrained to a 40 x 40 region, while limiting the $A^{(2)}$ fields to 20 x 20, as before. Thus it appears that a heirarchical organization with three association layers is competitive with human visual performance, with respect to resolution of detailed figures, and recognition of complexes of sub-figures, under conditions in which no scanning or shifting of attention is allowed.

If we were to complicate the problem by adding a third "T", again placing the stimuli in the positive class if all T's face the same way, and the negative class if some face up and some down, the probabilities of finding suitable $A^{(2)}$ and $A^{(3)}$ units would again fall by many orders of magnitude. For this problem, it is unlikely that any purely spatial and parametric constraints on the network would permit a solution with only 10^{10} units, with a retina appreciably greater than the size of the stimuli. It is also unlikely that a human subject, under tachistoscopic conditions, could do much better. Thus for complex organizations of organized sub-figures, each of which has several degrees of freedom independently of the others, some additional strategy must be sought to improve recognition capability. The use of sequential observations seems to be indicated at this point.

23.1.3 Sequential Observation Programs

The perceptrons considered in the last two sections, while facilitating the discrimination of small patterns in which fine details provide the essential information, are still far from optimum. For one thing, the number of A-units required remains very large; for another thing, the learning time would be correspondingly great, if the discrimination must be learned for all combinations of figural elements. These difficulties can be drastically reduced by the employment of a program-learning perceptron, such as the models considered in the last chapter. In particular, a system of the type described in Section 22.4, with a selective attention mechanism which permits it to attend to one detail or sub-figure at a time, is likely to prove useful in dealing with complex stimuli. Such a system can be employed in at least two basic ways:

1) It can be taught to recognize the presence of a sub-pattern (a spot or region of in which the fine structure is particularly dense) without having to classify it or differentiate it precisely. It can then direct the visual centering mechanisms to bring this pattern to the center of the retina, where high-resolution is possible, and where the system is taught to differentiate the type of pattern more precisely.

2) The perceptron may be taught to examine each of a number of retinal regions in turn (either by a systematic scanning procedure, by following boundaries, or by directing attention to those sub-fields in which the fine structure is particularly dense). This will result in the recognition of a definite sequence of details, which, in its entirety, serves to identify the complex stimulus organization.

The recognition of small objects in a large field may best be achieved by the first of these methods, while the discrimination of complex organizations (e.g., individual faces) requires the second method. In employing the second method, it would be particularly helpful if the perceptron could shift its field of attention systematically in a given direction, with the direction of attention shift provided as an additional piece of information to the association system at all times. In this case, the general configuration of the letter "A" followed by the letter "B" followed by "C" could be recognized by starting from the left of the field, shifting attention right to the first "detail", then right again to the second detail, and then right again to the third. The recognition of this complete

sequence would indicate the ABC configuration regardless of the actual positions of the letters in the field or their relative distances. It seems likely that the general problem of relation-recognition will ultimately yield only to sequential programs of this type.*

23.1.4 Sampling of Sensory Parameters

A fourth basic strategy for simplifying the sensory data which the perceptron must deal with is that of independent sampling of sensory parameters. In a general visual input system, five parameters are of interest: the intensity of illumination at a point, the frequency or color of the illumination, the time at which it occurs, and the x and y coordinates of the location of the point on the retinal surface. Each of these variables may be varied independently of the others. If we required a retina of 1000 lines resolution (i.e., 10^6 points), with sensitivity to 10 frequency bands, 10 levels of illumination, and 10 time delays for the outputs of each S-point, a total of 10^9 retinal points would be required to provide a sensory unit for each combination of values.

If it is actually required to discriminate between any two patterns, no matter how minute the difference between them, then there is no way of escaping this requirement. In general, however, we are satisfied with approximate information, and it is only under special conditions of "good observation" that we expect to obtain the highest resolution from the system. We can take advantage of this by means of the following organization.

---

*One sequential mechanism which may greatly improve performance is to take a sequence of "looks" at a given stimulus, with different fixation points selected at random, accepting a majority decision for the final response. The gains which might be expected, assuming independence between "looks", have been discussed in Reference 79, pp. 156-157.

Suppose we limit the number of retinal points to 10^6. To each of these S-points, x and y coordinates are assigned at random (from a uniform distribution over the whole field, rather than just points on a 1000 by 1000 lattice). In addition, a frequency drawn at random from the sensitivity range of the system is assigned to each S-point, and a threshold and time delay are similarly assigned at random. Now, if the perceptron sees a moving figure, with a variety of shading and color variation, it will be less precise in its judgement as to the exact position of the figure at time t, or the color of a given point in the retinal field at time t, than would be the case with the "complete" system with 10^9 S-points. If, however, we "fix" the position of the figure on the retina, and provide maximum contrast between illuminated and non-illuminated points (i.e, sharpen the figure to a black and white silhouette), and observe it for long enough to permit all time delays to propagate, then we have just as good shape-definition as in the system with 10^9 S-points, since all 10^6 retinal points will contribute one bit of information. Alternatively, if the entire field is illuminated at maximum intensity with a given frequency of light, this frequency can be discriminated to one part in 10^6, or five orders of magnitude better than the previous model. The same will be true with respect to intensity discrimination if the field is illuminated with white light, all frequency components being present with the same intensity. Similarly, the velocity, acceleration, and higher derivatives of the velocity of a moving object can be discriminated much better with the 10^6 element randomized-parameter system, provided the moving image consists of a sharp black and white pattern. Finally, we note that if we wish to specify the exact retinal coordinates of a square, the edges of which are alligned with the lattice points in the first

model, we can expect a maximum accuracy of one part in 1000, whereas with the random configuration (where some of the points will fall virtually on the boundary of the square regardless of its location) we could expect to improve the performance by several orders of magnitude.

What is sacrificed in this system is the ability to provide full information about individual retinal points, and the ability to provide maximum precision of discrimination in the case of shaded, moving figures. It would be difficult, for example, to precisely locate the boundary of a moving cloud, or to state the exact colors of specified points in a continuously varying mixture of colored lights; these are precisely the conditions, however, under which a human observer would also encounter difficulty, whereas if we optimize the conditions of observation by providing stationary figures and sharp contrast, resolution far in excess of the "fixed lattice system" can be obtained. Note that there is a trade-off between the resolution obtainable in one parameter and the resolution in other parameters; we cannot simultaneously optimize conditions for observing position and velocity, or color and intensity, for example. An interesting analogy can be drawn to the limitations on simultaneous observation of related variables in quantum mechanics, although there is no reason to suppose that the analogy is anything other than coincidental.

23.1.5 "Mixed Strategies" and the Design of General Purpose Systems

In the preceding sections, it has been demonstrated that the kind of network organization which is best suited for one stimulus environment or discrimination problem may be far from optimum for a different problem. The upright and upside-down T's , for example, might best be discriminated by a specially designed T -detector; but in this case every other letter, or combination of lines which might be encountered would have to have its own special detector mechanism, and the system would be useless in a general environment. Thus the question arises, if we know only the general character of an environment, but cannot anticipate all discriminations that the perceptron may be required to learn, what is the best combination of stimulus analyzing mechanisms to provide a good "general purpose" system?

This problem (on which no real analysis has been done to date) seems to be related, at least superficially, to the mixed strategy problem in game theory. The object of the game is to minimize the probability that any discrimination problem likely to arise in nature will be insoluble, subject to constraints on the size of the system, admissible learning times, etc. In Equation (23.4) a proposed solution for the distribution of $A^{(2)}$ fields was presented, for the special case in which the stimulus diameters are uniformly distributed. A more general solution should also consider the best mixture of line-detectors, spot-detectors, point-combination detectors, etc., among the $A^{(1)}$ units, the number of layers to be employed and the distribution of retinal fields among them, etc.

A few general rules seem to have emerged from studies thus far. For one thing, it seems to be inadvisable to seek highly specialized property detectors in the early stages of the network. A few basic types, such as line and boundary detectors, spot detectors, termination detectors, and movement detectors are certainly helpful, and yield appreciable improvements over random-point combinations. But higher-level organizations seem to be achieved better either by a mixture of simple properties at a greater logical depth in the network (as in the five-layer system considered in Section 23.2.2) or else by learning, at the R-unit level. For another thing, the extension in depth of a heirarchical retinal field system is useful for a limited number of levels, but extension much beyond three association layers seems unlikely to improve capabilities appreciably in systems the size of the human brain. Recognition problems which cannot be dealt with by a five-layer heirarchical structure, due to the large number of small details which must be considered in solving the problem, are best handled by a sequential system, rather than by continuing to increase the depth of the network.

It is questionable whether analytic procedures will be able to make much headway in dealing with this problem, although a combined attack with simulation techniques and analysis wherever applicable should yield considerably better information concerning the optimum organization for a given visual universe.

23.2 Audio-Analyzing Mechanisms

The sensory analyzing mechanisms which are best suited to an auditory input system are in some respects similar to those which have been considered for visual inputs. The difference in character of typical auditory patterns (speech in particular), where temporal organization largely takes the place of spatial organization, leads to a number of distinctive requirements. The following sections consider several of these special problems.

23.2.1. Fourier Analysis and Parameter Sampling

In principle, a number of possible sensory representations could be used for auditory material, including the continuous measurement of the amplitude of a waveform; spectral analysis, with the amplitudes given for all frequency components as a function of time; and various "reduced information" systems, such as the indication of zero-crossings, or the outputs of special filter systems. In the human auditory system, phase information appears to be disregarded, and a Fourier analysis into spectral components is employed. In a system designed to simulate human performance in speech recognition, musical recognition, and related problems, a presentation of the actual waveform would burden the system with a great deal of excessive information. The same word spoken with slightly different phase relations between the frequency components, for example, would present completely different wave-shapes, which the perceptron would have to learn to identify. Thus the spectral analysis of the audio input seems preferable.

With a Fourier analyzed input, the important sensory parameters to be represented by an S-point are the frequency, amplitude (or threshold), and time relative to the present (generally represented by connection delays). With these three variables, the principle of independent sampling of sensory parameters, discussed in Section 23.1.4, is again applicable. If the system is required to discriminate 100 frequencies, 100 time delays, and 100 amplitudes, for example, then a total of 10^6 frequency-threshold-delay combinations would be required with a "complete lattice" system. Using independently sampled parameters, on the other hand, a system with only 1000 S-units could discriminate 1000 frequencies in an intense sustained tone or mixture; it could discriminate 1000 amplitude levels in a "white noise" mixture sustained for the duration of the maximum time delays; or it could place the occurrence of an intense "pip" of white noise to a precision of one part in 1000 in time. Under less optimum conditions, the accuracy of discrimination in separate dimensions would be reduced, but the composite organization could still be discriminated readily from an appreciably different organization.

23.2.2 A Phoneme-Analyzing Perceptron

An introductory discussion of the phenomena of speech perception can be found in the chapters by Licklider and Miller in Ref. 112. Perceptrons for speech recognition and the association of names with objects or events have been discussed in Section 21.3. In these systems, it is assumed that a complete word must be learned as a primitive pattern, without preliminary analysis into significant sounds, or phonemes. In this section, a more sophisticated perceptron, capable of phonemic analysis, will be described.

The possible improvement in efficiency which can be obtained by analyzing a word into a sequence of phonemes can be highly significant. If we consider a hypothetical (and rather unnatural) language in which there are 100 allophones (or functionally equivalent sounds) for each phoneme, and a word of five phonemes consists of an independent choice of one of the allophones for each phoneme, then the word may appear in any one of $100^5 = 10^{10}$ possible forms. For a perceptron with a high degree of sensitivity to differences in sound patterns, this would mean that the discrimination of two words would require an enormous number of utterances (perhaps many millions) in order to generalize to all equivalent pronounciations (allomorphs) which might occur. (In actuality, the correlation between choices of allophones for different phonemes, in ordinary speech, would greatly reduce the sample size required, but the example will serve for illustrative purposes.) On the other hand, if each phoneme were first recognized by a distinct R-unit, and the outputs of the R-units taken as the input for a word recognizing perceptron, this second perceptron would receive an invariant sequence for each word, and in principle a single utterance of each word (morphene) would be sufficient for complete generalization. The phoneme-recognizing units would each have to distinguish a set of 100 allophones from a universe of 500 (assuming that only five phonemes are involved, so that the learning at this level might be achieved quite readily.

In actuality, the recognition of a phoneme is not as simple as the above discussion suggests, since a single speech sound cannot, in general, be recognized independently of its context. The preceding and subsequent sounds may completely alter the sound of a vowel, for example. Thus a phoneme-recognizing perceptron must itself be a sequence-recognition device, rather than a momentary-pattern recognizer.

A perceptron which appears to be capable of analyzing a sequence of words, so as to spontaneously develop an internal code for the phonemes employed is illustrated in Figure 67. It is a five layer perceptron, with variable connections between the $A^{(1)}$ and $A^{(2)}$ layers, and between the $A^{(2)}$ layer and the R-units. The $A^{(2)}$ layer can be thought of as playing the role of "R-units" for the first three layers of the perceptron, and will eventually learn the phoneme code to be employed. At the same time, it serves as the "sensory system" for the last three layers, which act as a three-layer perceptron for word-recognition. The $A^{(3)}$ system may either be organized as a cross-coupled system, or its input connections may be given a spectrum of delays; in either case, it is capable of recognizing sequences of inputs, rather than just momentary patterns. If the $A^{(3)}$ units are cross-coupled (particularly with inhibitory connections) and are of the "flip-flop" variety, tending to remain in their present "on" or "off" state until receiving a super-threshold signal, then the $A^{(3)}$ system will tend to go to a state characteristic of the sequence of input patterns regardless of the duration of the individual patterns in the sequence. This is particularly true if the $A^{(2)}$ system goes through a sequence of states (A, B, C, . . .) where each state is "held" without variation for a time greater than the "settling-down time" of the $A^{(3)}$ system (which should normally be no greater than two or three transmission delays, for the conditions given). Thus a "word" encoded into a sequence of phonemes by the $A^{(2)}$ units would lead to a fixed state of the $A^{(3)}$ system upon its termination, regardless of the actual duration of the phonemes.*

* This effect, as well as some of the others discussed in this section, might be employed to advantage in a visual system which is required to recognize sequences of stimuli, such as successively presented letters or signals.

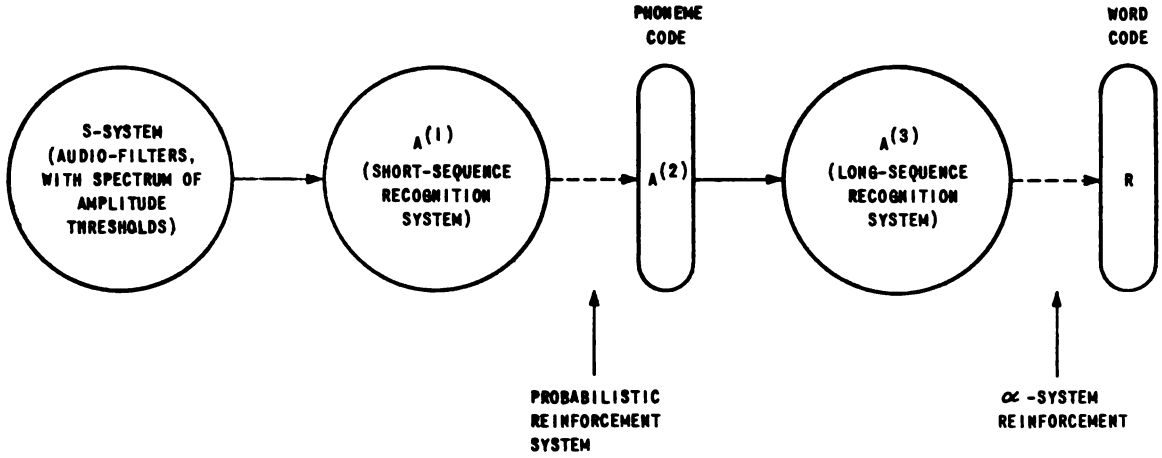


Figure 67 ORGANIZATION OF A PHONEME-ANALYZING PERCEPTRON. SOLID LINES INDICATE FIXED CONNECTIONS, BROKEN LINES INDICATE VARIABLE-VALUED CONNECTIONS. NUMBERS OF $A^{(1)}$ AND $A^{(3)}$ UNITS >> NUMBER OF $A^{(2)}$ UNITS.

The reinforcement rule for the $A^{(3)}$ to R-unit connections is a conventional α-system rule, so that an error correction procedure may be employed to teach the system to recognize words. The reinforcement rule for the $A^{(1)}$ to $A^{(2)}$ connections, however, is a probabilistic one, defined as follows:

1. With each connection, c_{ij}, from an $A^{(1)}$ to an $A^{(2)}$ unit is associated a time-dependent probability, $P_{ij}(t)$, called the instability coefficient of the connection.

2. Reinforcement at the preterminal level ($A^{(1)}$ to $A^{(2)}$ network) is applied only upon the decision of the reinforcement control system, or experimenter. Otherwise, the values of these connections remain unchanged.

3. If preterminal reinforcement is applied at time t, all instability coefficients are changed by the amount $\Delta P_{ij} = a_i^* \epsilon - \delta P_{ij}(t), [0 < \epsilon < 1]$. If no reinforcement is applied at time t, $\Delta P_{ij} = -\delta P_{ij}(t)$.

4. If reinforcement is applied, assume that the current activity states of all $A^{(2)}$ units are "wrong", and apply the correction $\Delta v_{ij} = 1^{-a_j^*} \cdot (a_i^* \eta)$ with probability $P_{ij}(t)$. (This is equivalent to an α-system error correction applied probabilistically.)

The actual training procedure can best be described in terms of the following experiment:

Assume a language, L, possessing three phonemes, A, B, and C, with k allomorphs of each phoneme. Time is quantized in units Δt . Each phoneme persists for a duration Δt , unless otherwise indicated. Let L consists of the six words, AB, BA, AC, CA, BC, CB. Assume some output code, $R^*(w)$ is assigned to each word, w . Then the procedure for training the perceptron is as follows:

Present a randomly chosen allomorph of the first word (AB), and observe the response of the perceptron. If this is correct, go on to the next word (BA); if it is incorrect, present AB again, this time applying a (quantized) error-correction reinforcement to the terminal connections ($A^{(2)}$ to R-units). Again test the response to the word AB. If the response is now correct, go on to the next word; otherwise, present the word again, this time reinforcing the preterminal network ($A^{(1)}$ to $A^{(2)}$ connections) and leaving the terminal network unaltered. Then apply a second correction to the terminal network, and retest the response to AB. Continue alternating between reinforcements applied to the terminal network and reinforcements applied to the preterminal network, until AB elicits the correct response. Then go on to the next word (BA) and repeat the same procedure. Continue cycling through the complete vocabulary until all words have been learned correctly.

A very limited amount of experimental work has been done with this system, using coin-tossing experiments and pecil-and-paper simulation techniques to investigate performance for the three-phoneme language considered above. Note that in this experiment, the perceptron is never given a single phoneme in isolation, but always as part of a two-phoneme word. Moreover, the perceptron is never corrected for "mistakes" in a single phoneme; reinforcements applied to the preterminal network are

maintained for the duration of an entire word, regardless of whether one or both phonemes are causing the difficulty. Nonetheless, it is found that as long as the number of $A^{(2)}$ units is greater than the number of phonemes ($N_A^{(2)} = 5$ has been found to work well), the system tends to form a phoneme-code at the $A^{(2)}$ level; i.e., after a period of training, each phoneme (A, B, and C) activates a different set of $A^{(2)}$ units, and all allophones of a given phoneme tend to activate the identical set of $A^{(2)}$ units.

These results can be obtained in a very short training sequence (generally less than one complete run through the 6-word vocabulary) with a suitable choice of the parameters ϵ and δ (which determine the rate of growth and decay of the instability coefficients, p_{ij}). On the other hand, no deterministic system has been found which will yield comparable results, although something like a dozen alternatives have been tried. A rough heuristic explanation for the observed effect can be given as follows: When the system arrives at some state in which the activities of the $A^{(2)}$ units constitute a phoneme-code for the language, new words can generally be learned with at most one or two reinforcements of the terminal network, so that there is little occasion to reinforce the preterminal connections. Consequently, the instability coefficients, p_{ij}, all decay towards zero, and the probability of disrupting the learned code, even if a reinforcement of the terminal network does fail to correct an error, is negligible. On the other hand, if any two phonemes are assigned the same code, there will be repeated confusions of words which can only be distinguished by means of the undiscriminated phonemes. Consequently, the preterminal network will frequently be reinforced for words containing these phonemes, but not for other words. Therefore, the connections originating from $A^{(1)}$ units which are activated by one of the conflicting phonemes will tend to acquire large instability coefficients, leading eventually to the modification of the $A^{(2)}$ responses to these phonemes. But since the corrections are applied probabilistically, the system will tend to try out arbitrary $A^{(2)}$ codes, and is

thus immune to "trapping" cycles, which tend to occur in deterministic models. In brief, the effect of the instability coefficients is to make those connections most suspectible to change which are most troublesome to the system.

It remains to be seen why the system tends to assign the same $A^{(2)}$ code to all allophones of a given phoneme, rather than merely making up totally unique codes for every input pattern. In part, this is helped by keeping the number of $A^{(2)}$ units small, so that conflicts are likely to arise if the code is not an economical one. The main effect, however, is due to the fact that different allophones of a given speech sound are not arbitrary, independent patterns, but tend to be highly correlated in the frequency-time-amplitude picture which comes from the sensory system. Thus the conditions are ideally suited for generalization from one allophone to nearly identical sounds, from there to next-nearest neighbors, etc. In fact, the tendency would be to classify all sounds identically (due to positive g_{ij} coefficients in an α-system) were it not for the intervention of the experimenter or r.c.s., which forces the separation of significantly different sound patterns. The spontaneous clustering of "similar" sounds can be compared to the spontaneous clustering of "similar" visual stimuli discussed in Section 7.3, and demonstrated for a γ-system in Experiment 9 (Page 214).

By adding fixed back-connections from the $A^{(3)}$ to the $A^{(1)}$ units in the perceptron of Figure 67, the recognition of individual phonemes may be more readily influenced by the preceding sequence. Alternatively, variable-valued back-connections from $A^{(3)}$ to $A^{(2)}$ units might be conditioned, by a suitable training procedure, to provide a bias to the $A^{(2)}$ units, tending to favor the recognition of the most probable next phoneme, as determined by the prior sequence.

While the above discussion has concentrated on demonstrating the possibility of a self-organizing mechanism for phoneme analysis, it is also possible to employ a somewhat simpler version of the five-layer system in which the $A^{(2)}$ units are actually trained by the experimenter to emit a chosen code for each phoneme. In this case, the $A^{(2)}$ units are actually R-units, and the probabilistic reinforcement rule for the pre-terminal network is no longer necessary, an ordinary α-system error correction procedure being perfectly suitable. One might also consider the possibility of extending the five-layer system in depth, by adding another A-unit layer and terminal R-layer after the last layer of the present model. By reinforcing first the terminal connections, then the $A^{(3)}$ outputs, and finally the $A^{(1)}$ outputs (in case of failure to correct the mistake at the terminal level), the system might be expected to develop a phoneme code in the initial part of the network, a syllable code in the middle, and a code for complete words or phrases at the level of the final R-units.

23.2.3 Melodic Bias in a Cross-Coupled Audio-Perceptron

The final stimulus analyzing mechanism to be considered is one which seems likely to occur spontaneously in cross-coupled perceptrons (of the type analyzed in Chapter 19) with audio-inputs. Suppose such a perceptron is exposed to a random sequence of notes, covering a range of several octaves, and played by a variety of string and wind instruments. Each note is held long enough for the cross-connections of the association system to be reinforced, before the next note is sounded. Then, assuming that the input comes from a Fourier analyzing system, the fundamental will be associated most strongly to the major overtones of the sequences characterizing the instruments employed. Thus the main association will generally be to the octave above or below, next to intervals of a major fourth and fifth, etc. This means that the main harmonic intervals of a twelve-tone scale will tend to predominate, rather than purely random frequency associations.

Such a system will tend to respond most unambiguously to chords and combinations of notes bearing a simple harmonic relation to one another (e.g., major fifths, fourths, and octaves) while strongly discordant combinations will tend to create a conflict (particularly in a γ'-system) such that the system tends to oscillate between several alternative and mutually competitive activity states.

By adding variable-valued back-connections from R-units to A-units (as in Figure 60), and associating a different response to each fundamental tone, the perceptron can be made to emit responses corresponding to a melodic sequence, if each response in turn is suppressed shortly after it is turned on. Such a perceptron, preconditioned as above, will tend to pick a harmonically consistent sequence, probably avoiding major shifts in tonality except by means of gradual progressions.

These observations, although suggestive, should not be over-interpreted. It seems plausible that melodic and harmonic biases in music have a fundamental basis in the overtone series (as Hindemith has suggested); however, the ease of vocal transition from one note to the next, and other considerations which play no part in the above model, are undoubtedly of equal importance in the determination of musical traditions and the conditioning of musical perception.

24. PERCEPTION OF FIGURAL UNITY

In almost all tests of perceptron performance considered in previous chapters, the environment, or stimulus world, was assumed to consist of discrete objects, or events, occurring one at a time in an ordered sequence. The actual physical environment which we experience on a day-to-day basis is not of this form; the visual field, in particular, is likely to contain a large number of different objects, patterns, or constellations of objects simultaneously. In human perception, it is easy to detect and name familiar objects in an unfamiliar scene, such as a landscape or a strange room. For a perceptron, each such combination of objects represents a new "composite" stimulus. If the composite organization consists of familiar patterns which have previously been learned in isolation, then it has been demonstrated that the perceptron may attend selectively to one object or pattern, and respond consistently to this object (see Chapter 21). For the human observer, however, it is not necessary for the individual objects or component patterns in the field to have been previously learned individually; totally new and unfamiliar organizations may nonetheless be perceived as "objects". Other organizations, no matter how familiar, will always be perceived as sets of objects, rather than as single entities.

The organization of a complex field into "objects" or distinct entities is frequently ambiguous, in that the field permits many alternative constructions or organizations of "meaningful parts". Problems of reversible perspective, the interpretation of Rorschach ink blots, or the detection of alphabetic characters in collections of random lines, all serve to indicate this ambiguity. The recognition of an "object" in the human perceptual process is generally experienced as a figure-ground organization, in which the object emerges as "figure" while the rest of the field serves as "ground". Hebb, who holds the segregation of figural patterns to be an innate process, has proposed the term "primitive unity" for such figural entities (Ref. 33). The perception of such unity is clearly

essential for an organism which must move about and interact with the objects of its environment. It applies not only to spatial organizations but to temporal sequences as well; a sequence of human movements is broken up, perceptually, into acts, steps, or gestures, while speech or music is divided into words or phrases, even if the sequence of sounds is an unfamiliar one.

The Gestalt psychologists have considered the problem of figural unity from the standpoint of what constitutes a "good figure" (c.f., Ref. 44). It is assumed that certain organizational properties of the stimulus field lead to a preference for one figural organization rather than another, and considerable experimental data have been gathered on the influence of such factors as contrast boundedness, connectedness, and the like. There is no doubt that all of these factors are important determinants of figure-organization in human perception. For present purposes, however, we will attempt to work with the hypothesis that what is most readily seen as a figural entity in a given environment tends to be an organization which is likely to undergo a continuous transformation in that environment (e.g., a detachable rigid object, or surface bounded by discontinuities). Whether the patterns which are most likely to be operated on by a continuous transformation are learned or innately recognized is left open, for the time being; it seems likely that both innate and acquired biases are at work in human vision.

Posing the problem in this form suggests that the system must be sensitive to cues indicating rigid, moveable objects, or surfaces (such as the faces of a cube) whose two-dimensional projections may undergo transformations which are discontinuous at their boundaries (i.e., the object moves, but adjoining regions of the field do not, or undergo a different kind of motion). The attempt to define figural objects as connected blobs of uniform illumination (as has been advocated in several computer programs) seems quite inapplicable, except under highly contrived and artificial conditions. It seems likely that in actuality, a

combination of many different cues of "good figure" are at work simultaneously, the final organization being arrived at by an active process, typically involving a good deal of trial and error before a good "fit" is obtained.

The cues which are suggested by psychological experiments as being influential in the determination of figural organization, or the perception of separate entities, include the following:

1) Differential motion of textured or bounded regions, or sets of points in the retinal field.

2) Cues indicating differential distance or "depth" of surfaces, or sets of points.

3) Differential surface properties in a bounded region (e.g., color, texture, or type of fine-structure).

4) Contours, boundaries, or discontinuities in surface gradients.

5) Object familiarity.

These five types of information are listed in approximate order of their strength, or dominance. If two constellations of points in a visual field are seen in relative motion, then even if they are intermixed spatially, they will tend to be seen as distinct objects, and the observer will have difficulty attending to both simultaneously. This is illustrated by the view of a moving scene outside a dust-streaked train window: either the window or the outside scene can be viewed as an object, but not both in combination. An experiment by Gibson employs motion pictures of talcum powder scattered on two glass plates, one behind the other. As long as both plates are stationary, or both moved jointly, the two planes cannot be separated; as soon as differential motion is introduced, however, the picture breaks up unmistakeably into two planes, each with its own distribution of spots.

The relationship of depth to figure organization is well known, and suggests that an attack on problems of depth perception in perceptrons will also contribute a great deal to the figural unity problem. The remaining cues (contrasting surface areas, boundaries, and familiar object recognition) are the ones most generally incorporated in attempts at devising computer programs or nerve-net models for figure segregation. It should be noted that the last of these (object familiarity) is the only one demonstrated as workable in perceptrons up to this point, in the selective attention mechanisms of Chapter 21; nonetheless, this mechanism is only useable under relatively ideal conditions, in which objects are present without overlap, confusing lines, spots, or "camouflage", and where it can be assumed that a pattern which contains the sensory components of a familiar object actually represents the object, rather than a random concatination of lines or points of illumination.

In order to evaluate the performance of a perceptron in the realm of figural organization, or the "perception of unity", a suitable set of criterion experiments must be defined. This proves much more difficult than in the testing of discrimination capabilities, or the study of generalization over a given group of transformations. In the simplest case, we may require a decision as to presence or absence of a figure in a noisy field. In this case, the detection experiments discussed in Sections 7.4 and 8.4 may be employed, with little ambiguity. In the case of organized environments, however (c.f., illustration in Section 8.4) it is frequently difficult to decide on an a priori basis that a particular decision is "right" or "wrong". If the field is sufficiently ambiguous, or the context is not indicated, a particular pattern of lines might represent the letter "E" or a random pattern of cracks on concrete. To evaluate performance on such material, it may be helpful to run the same experiment with human subjects, to provide an arbitrary standard for comparison. The results, however, are always subject to interpretation, based on the intentions, experience, and additional information available to the human observers in contrast to the perceptron.

Three types of criterion experiments seem possible:

1) Description of the figure by a multi-response perceptron (e.g., "small right triangle in upper left, with cross-hatched surface").

2) Detection of familiar objects; perceptron is trained to tell whether object is present or absent.

3) "Test-point experiments" where the perceptron can attend selectively to a test-point, or the end of a pointer placed in the field, and tell whether or not the point is in contact with the figure. In this way, a description of the figure can be obtained by tracing its boundaries, or obtaining an inventory of its parts.

Little work has been done, to date, to determine the capabilities of cross-coupled and back-coupled perceptrons in experiments of these types. The detection experiment is the one most readily performed with the systems analyzed to date, and it is hoped that some data can be obtained in the near future. Series-coupled perceptrons appear to offer little hope of good performance in these problems.

Cross-coupled perceptrons have been observed to form mutually exclusive "cell assemblies" in their association systems, under the spontaneous organization rules considered in Chapter 19. It is possible that with a suitable choice of preconditioning sequence and network parameters, such cell assemblies may be related to figural organizations, so that when two or more rival figure-ground organizations are present, the A-units activated will correspond to one of these organizations in preference to the others. At present, however, this conjecture must be regarded as pure speculation, with no real evidence to support it.

The introduction of back-coupling, however, does permit the perceptron to take advantage of the first and most powerful cue as to figural organization,

namely, differential motion. A suitable organization is illustrated in Figure 68. The perceptron is a three-layer system with multiple R-units, of the "on-off" variety. Each R-unit is trained to respond to a different motion, or transform-

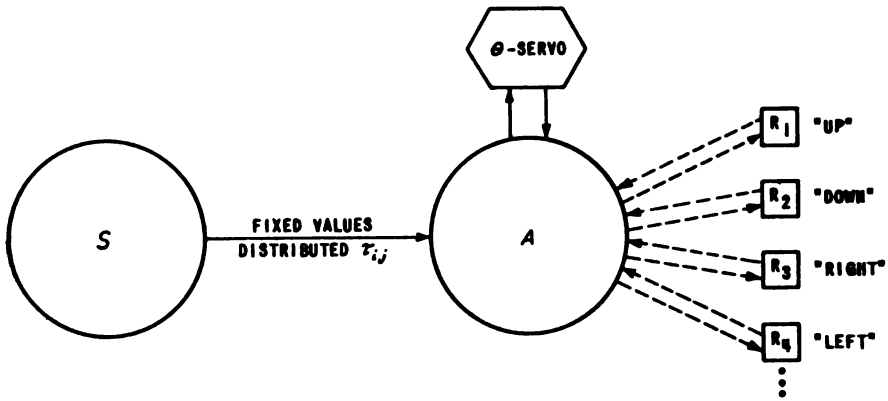


Figure 68 A PERCEPTRON FOR FIGURAL SEPARATION OF MOVING PATTERNS.

ation. The variable connections from A to R-units and from R to A-units are reinforced as in Chapter 21, for selective attention systems with variable back-coupling. Due to the spectrum of time delays, the A-units respond directly to the movement pattern as well as the shape of the stimulus. The system may be further improved by adding inhibitory interconnections between the R-units, so that only one can go on at a time. If there should be two stimulus patterns simultaneously present on the retina, moving in opposite directions (or one moving and the other stationary), the dominant response will tend to support those A-units responding to the stimulus whose motion corresponds to the R-unit, and will suppress the A-units responding to the second stimulus. The threshold servo plays the same role as in the systems of Chapter 21. If the A-system is cross-coupled, with a [illegible]-system rule, the effect will be supported by the formation of "cell assemblies" characterizing different directions or velocities of motion.

As the stimulus field becomes increasingly ambiguous in its organization (as in ink-blot patterns, for example) the field organization which results in a human observer depends less on a passive response to automatic mechanisms, and more on an active "construction" of a meaningful figure. In this process, a number of alternatives may be reviewed in quick succession, before one of them "settles in", and the field loses its ambiguity. This sort of active structuring of the field may also be possible for a perceptron with feedback loops from the R-units, if the perceptron can evaluate the strength, or decisiveness of its response, and actively perturb its response state (and hence the feedback signals to the A-units) until a strong, persistent response is obtained. This may be done by adding random Gaussian noise signals to the inputs of the R-units, resulting in frequent changes in the response state as long as the signals from the A-system are weak and indecisive.

While the above discussion indicates several possibilities which are open to experimental treatment, it is clear that much fundamental groundwork remains to be completed before the problem of figural unity can be attacked in a systematic manner. At the present time, this problem remains one of the most severe challenges to all theories of brain mechanisms.

25. VARIABLE-STRUCTURE PERCEPTRONS

All of the memory mechanisms employed in previous chapters employ a fixed network structure, in which the weights of connections are variable. It is occasionally proposed that a system in which the structure of the network itself is modifiable, with new connections being formed and old ones discarded on the basis of demonstrated utility, might lead to the evolution of a better model, with a smaller number of logical elements than would be possible for a fixed-structure perceptron with random connections. This might, for example, be a way of evolving special-purpose stimulus analyzing mechanisms of a high degree of utility for a particular environment. A model in which structural modification is possible -- i.e., in which the origins or termini of connections are changed as a result of activity -- has previously been referred to as an "evolutionary model". Apart from the possibility that such a system might provide a useful memory mechanism, or adaptive mechanism, it has been suggested that by observing the terminal states to which such a model goes, after long exposure to an environment, we might learn something about the kinds of physical constraints which could be usefully built into future systems at the outset.

25.1 Structural Modification of S-A Networks

To date, very little work has been done with evolutionary systems. Several examples have been programmed for the IBM 704, which indicate a slight improvement in some cases, but these programs have proven too costly in computer time to permit extensive experimentation. The cases illustrated here come from this group of pilot experiments.* A three-layer perceptron with a single R-unit was employed, and an α-system error correction method was used for reinforcing the terminal network.

*The programs were written by Kesler, and carried out at the AEC/NYU Computing Center.

The rules for changing the structure of the network are closely analogous to those employed for perceptrons with variable S-A connections, in Chapter 13. Each A-unit, a_i, is continuously evaluated by means of a utility measure, E_i. If the current response r^* is wrong, E_i may be increased by 1 with probability p_1, p_2, or p_3, defined as follows:

p_1 = probability of incrementing E_i if the sign of $v_{i,r}$ disagrees with the desired classification of the current stimulus, and a_i is active.

p_2 = probability of incrementing E_i if the sign of $v_{i,r}$ agrees with the desired classification of the current stimulus, and a_i is inactive.

p_3 = probability of incrementing E_i if the sign of $v_{i,r}$ disagrees with the desired classification, and a_i is inactive.

The quantities E_i are assumed to decay by an amount δE_i at each stimulus presentation time. If E_i reaches or exceeds a threshold level, θ_E, the origins of all connections to unit a_i are reassigned, and E_i is reset to zero. In most experiments, $p_1 > p_2 > p_3$, so that an A-unit is most likely to have its connections changed if the value of its output signal frequently disagrees in sign with the intended classification of the stimulus which activated the unit.

The results of several experiments (on horizontal/vertical bar discrimination) are shown in Figures 69 and 70, with the performance curves for the corresponding fixed-structure models shown for comparison. While there seems to be a slight advantage for the variable-structure systems (particularly in Figure 69, where only 20 A-units were used), the improvement over the fixed-structure system is not impressive. Nonetheless, it is possible that a more sophisticated procedure for determining which A-units are to be changed would produce better results. It also seems likely that the horizontal/vertical bar problem, which is not very demanding in the geometry of origin

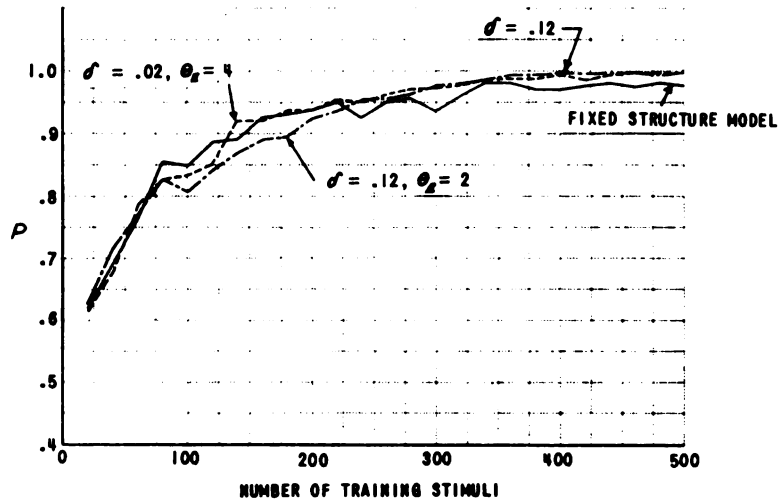


Figure 69 EVOLUTIONARY MODEL, IN HORIZONTAL/VERTICAL BAR DISCRIMINATION. MEANS OF 10 PERCEPTRONS. 50 A-UNITS, $\chi = 8$, $y = 2$, $\theta = 3$, $P_1 = .9$, $P_2 = .9$, $P_3 = .01$.

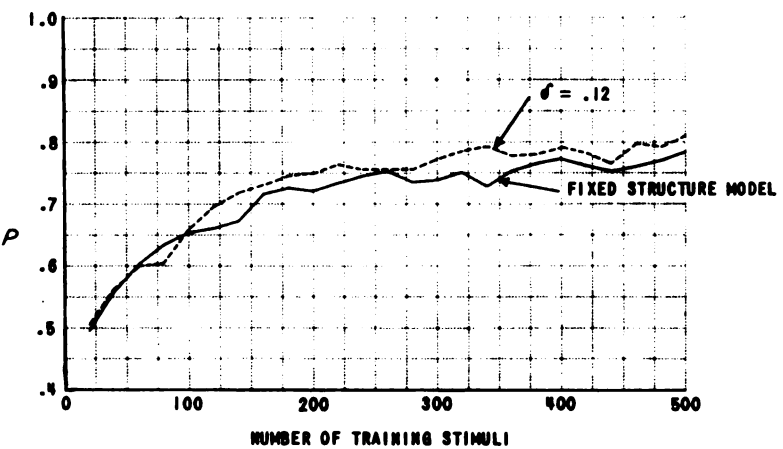


Figure 70 EVOLUTIONARY MODEL, IN HORIZONTAL/VERTICAL BAR DISCRIMINATION. MEANS OF 10 PERCEPTRONS, ZERO RESPONSES COUNTED WRONG. 20 A-UNITS, $\chi = 8$, $y = 2$, $\theta = 3$, $\theta_a = 3$, $P_1 = .9$, $P_2 = .8$, $P_3 = .01$.

configurations required for discrimination, may be a poor choice of a calibration experiment for evaluating the evolutionary model. Unfortunately, the procedure is so time-consuming for a digital computer that only a small number of experiments have proved feasible.

As a memory process, the above system seems excessively complicated. Not only are three distinct probabilities required, under three sets of logical conditions, but the E_i must be stored as an auxiliary variable for each A-unit. This is clearly implausible for a biological mechanism. The difficulties encountered seem to be common with those met in all attempts at providing a useful memory process which operates on the preterminal connections of the network (as in the variable S-A systems of Chapter 13). It is hard to see what simple criterion might be employed to identify those connections which should be changed in order to improve the final output of the R-units. It seems likely that a local information rule (Page 289) is incompatible with an efficient system of reinforcement at the preterminal levels of the network.*

25.2 Systems with Make-Break Mechanisms for Synaptic Junctions

A somewhat different kind of structural modification from the model described above is that in which there is a fixed set of "potential connections" to each unit, but these connections may be either "made" or "broken" on an all-or-nothing basis, in the manner of switches or mechanical relays. A possible application of such a mechanism to the terminal network of a three-layer perceptron is illustrated in Figure 71. The A-units are divided into a set of excitatory units (E-units) whose output is always positive, and a set of inhibitory units (I-units) whose output is always negative. All signals are of unit amplitude, and the connections from I-units to the R-unit are fixed, only the E-unit connections being modifiable. The connections from E-units to the R-unit are of the make-break variety, the reinforcement rule being as follows:

*There is some hope, however, that the "elastic perturbation" system suggested in Section 26.4 will prove applicable to this problem.

The reinforcement control system can call for a $\Delta v > 0$ or for $\Delta v < 0$. If a positive increment is required, excitatory connections with active origins are made with probability P (applied independently for each unconnected E-unit), while if a negative Δv is required, excitatory connections with active origins are broken with probability P. If the system begins with initial conditions such that the number of connected E-units just balances the number of connected I-units, and if the number of units is very large, the effect of a single reinforcement will be identical to the application of a quantized α-system reinforcement to a system with fixed A-R connections. Thus, under the error correction procedure, this system can be expected to duplicate the performance of an α-system perceptron quite closely, provided the number of A-units is large.

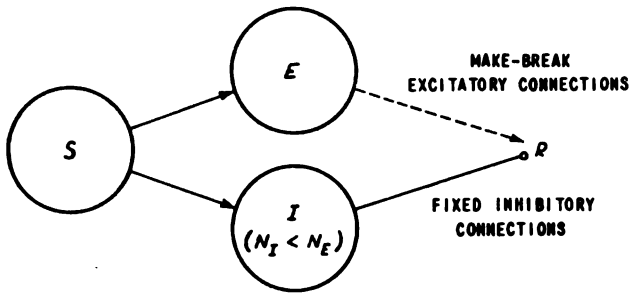


Figure 7I SIMPLE PERCEPTRON WITH MAKE-BREAK CONNECTION SYSTEM.

An alternative system is one with equal numbers of E and I-units, in which the I-connections are also variable. In this case, new connections can be made, but once established are assumed to be permanent. For $\Delta v > 0$, new E-connections are formed with probability P, as above. For $\Delta v < 0$, however, new I-connections are formed with probability P, instead of breaking E-connections. At the outset, assuming that all A-units are initially disconnected, this system again behaves in much the same way as an α-system perceptron. As the system "saturates", due to the exhaustion of available connections, the

increments to the R-unit input signal from each new reinforcement become progressively smaller. If the number of A-units is infinite, then the system never saturates entirely, new reinforcements always having some effect, although this is apt to become negligible as saturation is approached.

These models are of more interest as possible analogs for biological systems than as significantly new types of perceptrons. Their properties, short of the saturation condition, closely resemble the systems previously considered, but they do not require values which change sign, and are suggestive of a possible synaptic growth mechanism in biological memory. As engineering devices, their reliance on probabilistic mechanisms is apt to make their construction more difficult than the α-system models.

26. BIOLOGICAL APPLICATIONS OF PERCEPTRON THEORY

When the perceptron was first proposed, it was considered primarily as a model of biological memory mechanisms. As the models became more sophisticated, a number of psychological properties not directly related to memory were investigated, but the main emphasis, as a biological model, is still on the adaptive mechanisms employed, and the recording of past experience. In this chapter, the application of perceptron theory to biological problems will be considered primarily from this point of view.

26.1 Biological Methods for the Achievement of Complex Structures

The biological evidence which has been cited repeatedly throughout this volume indicates that highly organized structural constraints exist in many parts of the nervous system. Apart from the gross anatomical complexity of the brain, the mechanisms of optic nerve growth and regeneration, the stimulus analyzing mechanisms found by Lettvin in the frog and by Hubel in the cat, and the better known mechanisms of motor coordination and control indicate that organization of a rather involved type may occur even in the fine structure of the network. In perceptron theory, as it has developed to date, most emphasis has been placed on learning and memory as a means of achieving such organization. In actuality, a number of alternative procedures are possible for the creation of complex networks, satisfying a given set of logical constraints. These include:

1. Logical specification (e.g., let the i^{th} cell of the k^{th} row be connected to the i+1st cell of the k+3rd row, for all $i > k$). This is equivalent to an exact blueprint of the network.
2. Natural selection, whereby the useful sub-networks of an originally random population survive, while the others decay.
3. Simple spatial constraints (gradients, directional bias, or distributions of connections specified by a small number of parameters).

4. Typological constraints (e.g., cells of Type A can only connect to cells of Type B or C, where cell types might be distinguished by chemical properties).

Of these four mechanisms, only the last three seem to be well suited for the development of biological nerve nets. The first mechanism, logical specification of the structure, is primarily a contrivance of engineering, which is well suited to the construction of computers, but which seems to have no clear counterpart in known mechanisms of growth and maturation. It is this first method of control, however, which has been most investigated in studies of brain mechanisms during the last few decades (e.g., References 17, 57, 71).

In specifying the initial physical form of the networks in perceptron theory, most attention has been given to the third alternative; spatial constraints of a simple sort have been employed throughout. In the last chapter, limited use was made of the second and fourth methods. The use of typological constraints has thus far been used mainly to distinguish excitatory from inhibitory neurons (Section 25.2), but it seems likely that its use is relatively widespread in biological systems. In particular, Sperry's work on neural maturation and fiber regeneration, and Lettvin and Maturana on the regeneration of scrambled connections in the frog's brain, suggest a chemical control or "homing mechanism" of remarkable sensitivity.

The limited experiments performed thus far on "natural selection" as a structural control mechanism do not appear particularly promising (Section 25.1). The evolution of the network occurs too slowly, and is too subject to disruption and instability of partially achieved organizations, to be useful in any of the forms examined up to this point. It remains possible, however, that a more rapidly converging mechanism may be found, and the field remains open for future investigation. Typological constraints, on the

other hand, are likely to come into their own with the investigation of perceptrons having complex mixtures of property detectors, and other specialized A-units, all deriving their connections from a common sensory field.

26.2 Basic Types of Memory Processes

Perceptron memory mechanisms have all taken the form of modifications of the signals transmitted across synaptic junctions. There appear to be at least two basic types of memory dynamics which are useful in perceptrons. The first is a system in which values remain stable unless action is taken by a reinforcement control system, based upon an evaluation of the current response of the perceptron. The most effective method actually investigated for this purpose has been the α-system, with an error correction procedure for modifying the values of A to R-unit connections. The second type of memory is one which achieves stability only in the form of a dynamic equilibrium with a continuously active reinforcement process. This second system does not depend upon evaluation of the perceptron's output, but maintains a continuous state of adaptation in the network, based only on local activity. In practice, it seems likely that a decaying Γ-system will prove to be the best of the systems of this type which have been analyzed. The first type of mechanism permits the system to learn from an external "teacher", or by reward and punishment experienced as a result of trial and error activity. The second type permits the perceptron to acquire an internal model of the "similarity structure" of its environment, as defined by the temporal relationships of moving stimuli. It may be that more complex forms of organization (such as the recognition of connected patterns, or Gestalten) can also be achieved by means of dynamic processes of the second type, but this remains conjectural at this time.

While it is certainly conceivable that additional basic mechanisms may be required to perform the memory tasks of a complex organism, there seems to be some reason to believe that the two types of dynamics characterized above may prove sufficient for the phenomena of "adaptive behavior". The first variety permits the system to be "set" passively to any desired state, which will then be retained indefinitely. Thus any form of permanent learning can be handled, in principle, by such a system. The error correction theorems of Chapters 5 and 10 seem sufficient to demonstrate this assertion. On the other hand, any spontaneous modification process which is not to be self-defeating must ultimately achieve some sort of dynamic equilibrium with the conditions which induce the change in state; without such a mechanism (provided in the case of our four-layer and cross-coupled perceptrons by the decay term in the equations) the dynamic range of the memory variables must ultimately be exhausted, and the system will saturate. In any case, a mechanism which is to serve as a basis for generating a model of the external environment must be one which ultimately approaches a stable condition, as the model approaches a true representation of the external world. Such considerations make the second mechanism appear to be a natural complement to the first.

Two memory functions which might call for processes of a different logical character are the serial recording of experience (in the manner of a tape recorder or motion picture camera) and a temporary memory for data which are to be used in the immediate future and then forgotten (as in the "memory" of a digital computer). For the second of these phenomena, it is likely that a dynamic storage mechanism, such as pools of activity or reverberating loops which can be triggered and extinguished by a suitable control system, will prove to be the most effective storage mechanism. The problem of serial memory is a more serious one, but can only be dealt with together with the problem of selective recall and the mechanisms for its control.

It is certain that in a simple perceptron, memories are not tagged in any way which would permit their serial order to be re-established later. But the "memories" stored in a simple perceptron are in any case merely associative, rather than substantive. The nature of substantive memory in humans must be investigated more carefully in the future. While it seems unlikely that a complete image or state of the association system is stored, it is nonetheless clear that a great deal more information is retained than is represented by a simple classification of an experience as belonging to one of n categories. One alternative is that of storing a description of a large number of characteristics or dimensions, which jointly permit the reconstruction of the original experience by the active creation of a model, or image, which approximates the original state of the association system. Among the characteristics stored would be such time-tagging information as the location in which the event occurred, the time of day, the activity that the subject was engaged in, etc. An accumulation of such cues would enable a suitable search process to locate the experience in time, and to associate it with preceding or successive events in appropriate order (c.f., Reference 79, Chapter VIII). In any case, it seems likely that substantive recall is an active, creative (or recreative) process, rather than merely a passive reading-out of a memory bank.

26.3 Physical Requirements for Biological Memory Mechanisms

From the considerations just stated, it should be clear that not one but several memory mechanisms are likely to be encountered in a complex system. Limiting our attention, for present purposes, to the two basic mechanisms which have been studied in perceptrons, what can we say as to the probable physical characteristics of the memory traces?

First, as to location: it appears that the most suitable location is in the connections, or synapses, which mediate the interaction of particular pairs of neurons. Perceptrons in which the memory trace affects an entire neuron and

all of its interactions with other neurons have been investigated (Reference 79) but this has invariably involved the introduction of artificial constraints on the topology or logic of the network, in order to limit the effects of reinforcement to the desired transmission channels. In any case, systems in which the reinforcement is specific to the connections appear to be far more economical than those in which reinforcement is applied to an entire neuron, or A-unit.

A second condition is that the memory change should be reversible. Both the externally controlled error-correction procedure and the fully automatic memory processes of the cross-coupled perceptrons require reversible modifications. In the case of the error-correction procedure, two antagonistic control mechanisms seem to be called for, one of which strengthens the excitatory outputs of active A-units, and the other of which weakens excitatory outputs or strengthens inhibitory outputs. While most of our analyses have assumed that the actual sign of the value of a connection may change from positive to negative, this is clearly a non-biological artifact, introduced for convenience in analysis. The same effects could be achieved by a system in which half of the connections are always positive, and half are always negative. If the negative connections are fixed in magnitude, then only the excitatory connections need be modified, yielding a net positive signal if they exceed the strength of the fixed inhibitory component, and a net negative signal if they fall below the inhibitory strength. Alternatively, the excitatory connections might be fixed, and the inhibitory connections variable, or each type might be variable within its own dynamic range.

The requirement that the "strength" or value of a connection be modified as a consequence of the correlated activity of <u>both</u> terminal units, rather than just the transmitting unit, appears to place a unique condition on the memory process. Most metabolic processes such as growth, changes in cell chemistry, etc., which might be involved here are of a type which generally depend only upon the cell in which the change occurs, and its over-all environment,

whereas we seem to require a two-factor phenomenon, which depends upon the activity of two specific cells. This writer has previously stated the conjecture (Reference 83) that the required effect might be obtained if the production of transmitter substances depended upon an enzyme or catalyst produced in the nucleoplasm of the trans-synaptic cell, and released to the medium when that cell is stimulated to activity. The presynaptic fibers which were most recently active, being in a heightened metabolic state, would then be in the most favorable position to compete for the limited supply of this catalyst, which would then enable them to produce their transmitter substance at an increased rate in the future. The competition for metabolites in limited supply in the neighborhood of a particular cell body would tend to create a $\mathcal{T}$-system, in which the most active cells would gain at the expense of the inactive ones. Whether this is a correct description of the mechanism or not, some type of symbiotic relationship seems to be demanded between the presynaptic fibers and the post-synaptic cell, in order to provide a memory mechanism of the type analyzed in Part III of this volume.

The memory mechanism employed for error-correction learning places rather different demands on the biological system. Here the reinforcement depends not so much on the correlation of activity of the two terminal units, as on the correlation of the activity of the transmitting unit with the decisions of the reinforcement control system. It is conceivable that this might again involve the release of a catalyst in the neighborhood of the active connections, but in this case the release must be remotely controlled -- perhaps through glandular action. In one respect this is a simpler requirement, conceptually, than the former case, where the activity of two specific cells had to be considered for each connection which might be reinforced. In the present case, the general release of an excitatory or inhibitory reinforcing agent from a central source would appear to be sufficient; the recently active connections, being most metabolically active, would tend to be most strongly affected. In a second respect, however, this

mechanism presents a new problem which is more serious: the problem of limiting the effect of reinforcement to the specific response which is to be corrected.

It was demonstrated in Chapter 12 that the error correction procedure can be guaranteed to work only if the correction is limited to the erroneous responses, in a multiple response system. To achieve this condition in a biological system, it seems that a mechanism is called for which can select one response, or response component, at a time as a candidate for reinforcement, and limit the corrective action to the selected locality. In dealing with motor responses, the topographical mapping of the motor control areas of the cortex is likely to prove helpful here, particularly if we adhere to the hypothesis that the memory trace involves the release of a chemical agent which affects everything in its neighborhood. *

The proportional decay mechanism which is required for the "spontaneous" memory process is probably the easiest of the requirements to rationalize in a biological model; a chemical mechanism, in particular, would tend to exhibit decay at a rate which increases with the concentration.

At present, any treatment of the compatibility of perceptron theory with biological memory mechanisms must remain entirely speculative. It is to be hoped that as additional evidence on synaptic transmission and neurochemistry comes to light, it can be fitted into the picture. Thus far, there seem to be no serious conflicts, although there are a number of missing links. The considerations stated above do suggest several plausible hypotheses for experimental investigation.

---

*A procedure is now being investigated by which an error correction is applied to a randomly chosen set of R-units, the value increments being transient rather than permanent, unless the correction actually proves effective. It is hoped that this technique will yield an efficient reinforcement mechanism which does not depend on specification of the erroneous R-units. (see Section 26.4)

26.4 Mechanisms of Motivation

The problem of motivation for perceptrons, considered as models for biological nervous systems, has hardly been treated adequately up to this time. The reinforcement control system, which forms part of the experimental system, plays the role of a sort of deus ex machina, which not only has knowledge of right and wrong responses, but can control the distribution of reinforcement to individual R-units in the perceptron, as required. A more "natural" system with only a slight reduction of efficiency does seem to be possible, however, although at present the model proposed is a heuristic one, on which no quantitative analysis has been completed.

The proposed model for biological reinforcement mechanisms is illustrated in Figure 72. In this system, the r.c.s. is no longer external to the system, but is essentially part of the perceptron. It is assumed that the perceptron system includes a sensing device for a physiological condition which has been arbitrarily called the "discomfort level", measured by the variable D. This might be compared to Ashby's concept of "essential variables". In addition to continuously measuring the variable D, which is assumed for simplicity to be some function of the current stimulus pattern, a second mechanism (readily represented by a neuron with inhibitory input connections with a short time delay and excitatory connections with a longer time delay, both originating from the "D-detector") responds to a negative dD/dt. The corrections to this system are random perturbations applied either to active connections, or to all connections of the perceptron; the increments, however, take the form of "elastic perturbations", so that the connections tend to decay back to their previous values unless a "positive reinforcement" occurs to "fix" the new values. Thus negative reinforcement applies a slight random perturbation, which tends to disappear unless it actually proves helpful, in which case it is stabilized by a positive reinforcement.

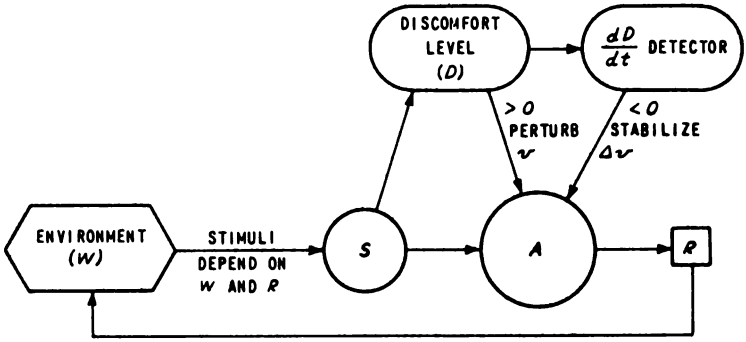


Figure 72 EXPERIMENTAL SYSTEM EMPLOYING ELASTIC PERTURBATIONS, STABILIZED BY IMPROVEMENTS IN SENSORY SITUATION (COMPARE Figure 4).

For this system to function efficiently, it is again necessary to assume some degree of temporal continuity in the environment, so that the change in D indicates a true improvement in the response of the system, rather than an irrelevant change due to a sudden alternation of the environment. Preliminary simulation experiments to evaluate this scheme are now in progress, employing the Burroughs 220 computer, and indicate that the system should work with a reasonable degree of efficiency, as compared to a system employing a more deterministic error correction procedure. The results of these experiments will be reported as soon as the data are complete. The system has the advantage that it works well with an arbitrarily large number of R-units, without requiring an individual decision as to the error of each one, as long as D is some monotone increasing function of the joint error, such as the norm of the difference vector, $\|r^* - r'\|$. Such a representation will work best when all of the R-units are continuous transducer units, so that any random value-perturbation will have a 0.5 probability of yielding an improvement.

27. CONCLUSIONS AND FUTURE DIRECTIONS

Man's intelligence is a unique phenomenon on our planet, occurring at such a level of complexity in a single species only. The lack of other similarly intelligent species is unfortunate from the standpoint of science, for it makes it difficult to tell from comparative evidence which features of human psychology are accidental products of man's peculiar biological constitution, and which are fundamental to the nature of intelligence itself. Despite this lack of comparative material, some of us believe that it may ultimately be possible to answer such questions through an understanding of the physical basis of psychological phenomena, independently of the biology of any one species. The perceptron program represents a small part of such an undertaking; it is an attempt to study the psychological properties of certain highly simplified mathematical or physical models of the central nervous system, in the hope that such a study may throw light on basic principles which can then be applied to more sophisticated models.

The use of "models" to represent complicated natural phenomena has been an essential technique in the physical sciences for many centuries. The model is a simplified theoretical system, which purports to represent the laws and relationships which hold in the real physical universe. The solar systems of Ptolemy, Copernicus, and Einstein, and the Atomic models of Democritus, Bohr, and Heisenberg represent two successions of such models, each in turn coming somewhat closer to an adequate representation of its subject matter. In some cases (the concept of an "ideal gas" for example) the model deliberately neglects certain complicating features of the natural phenomena under consideration, in order to obtain a more readily analyzed system, which will suggest basic principles that might be missed among the complexities of a more accurate representation. Such simplified models may then be refined through a series of "perturbations", which introduce the known complications one at a time, in a manner which permits the mathematician to incorporate them

into his analysis. It is this approach which has been most characteristic of the perceptron program.

Stated in simplest terms, our objective has been to discover a physical system, or abstract model, which will be capable of "perceiving" its environment, and learning to recognize those objects or events which it has perceived in the past. However, since it is our purpose to understand the actual mechanisms employed by the brain, rather than simply to construct a new type of computing device, the perceptron models are constrained in their organization and dynamic properties by what is known of the biological nervous system. Rather than attempting to "invent" or "construct" a machine which will calculate such things as similarities or geometrical properties of stimuli, the approach has been to begin with a hypothetical network of idealized neurons, or nerve cells, resembling the brain in its general organization, and then analyze the system mathematically to determine whether or not it possesses "psychological" properties of interest. Where the model is found to deviate markedly from the behavior of biological systems, modifications are suggested, and the new model that results is subjected to the same sort of analysis. In this fashion, it is hoped that the necessary conditions for a system to "perceive" in the same manner as the brain can be abstracted.

In this chapter, we will attempt to summarize the principle results which have thus far emerged from this approach, the problems which have now come to the foreground, and the means by which these problems might be attacked. The possible applications of perceptron theory to engineering devices and the construction of physical brain models will also be considered. Finally, an attempt will be made to anticipate the future relationship of the neurodynamic approach to the various alternative strategies by which the problems of understanding and simulating intelligence are being investigated.

27.1 Psychological Properties in Neurodynamic Systems

Our main conclusions deal with the properties of closed experimental systems, such as those illustrated in Figures 3, 4, and 72. It has been shown that as the topological organization of the perceptron increases in complexity, new psychological properties emerge. The principle results can be summarized as follows:

(1) A network consisting of less than three layers of signal transmission units, or a network consisting exclusively of linear elements connected in series, is incapable of learning to discriminate classes of patterns in an isotropic environment (where any pattern can occur in all possible retinal locations, without boundary effects).

(2) A three-layer series-coupled perceptron is a minimal system capable of learning to discriminate arbitrary classes of stimulus patterns or stimulus sequences. Any discrimination problem can, in principle, be solved by such a system, and any arbitrary response function can be assigned to the stimuli of a given universe.

(3) By means of an α-system error-correction procedure, a three-layer series-coupled perceptron with simple A-units and a fixed preterminal network can always be taught the solution to any classification problem or response function for which a solution exists.

(4) Equations for the learning curves of simple perceptrons under various reinforcement rules have been presented. The results indicate that for simple tasks, such as the recognition of large alphabetic characters against a plain background, the three-layer series-coupled system performs with reasonable efficiency, although it may require a lengthy training procedure with large samples of each stimulus class to guarantee recognition of all variations, or "allomorphs" of a pattern.

(5) In perceptrons with variable-valued preterminal networks, a non-deterministic reinforcement rule may be required to guarantee that the solution to a classification problem will be achieved, given that the solution exists.

(6) Generalization capabilities of three-layer series-coupled systems are poor, and in "pure generalization" experiments (where the test stimuli have no sensory points in common with the training stimuli) there is essentially no generalization capability.

(7) Series-coupled perceptrons with randomly organized origin-point configurations for the A-units tend to be highly resistant to stimulus noise and network damage; in a complex field containing mixtures of familiar stimuli, however, they are easily confused, and are incapable of responding selectively to one stimulus or object at a time.

(8) The addition of a fourth layer of signal transmission units, or cross-coupling the A-units of a three-layer perceptron, permits the solution of generalization problems, over arbitrary transformation groups.

(9) Four-layer and cross-coupled systems with suitable rules for modifying their connection values (Chapters 16, 17, and 19) are capable of learning a group of transformations which have occurred commonly in sequences of stimuli, and later recognizing the similarity of stimuli which are equivalent under the observed transformation group. This phenomenon occurs "spontaneously", without any external influence on the perceptron apart from the occurrence of stimuli.

(10) In back-coupled perceptrons, selective attention to familiar objects in a complex field can occur. It is also possible for such a perceptron to attend selectively to objects which move differentially relative to their background.

(11) By a suitable combination of geometric constraints (Chapter 23) a multi-layer perceptron can be enabled to recognize detailed patterns in high-resolution fields with markedly increased efficiency, compared to a randomly organized three-layer system. For a given universe of stimuli, there will be an optimum organization of such a system, which will rarely exceed three layers of A-units for tasks commensurate with human capabilities under tachistoscopic conditions.

(12) A number of speculative models which are likely to be capable of learning sequential programs, analysis of speech into phonemes, and learning substantive "meanings" for nouns and verbs with simple sensory referents have been presented in the preceding chapters. Such systems represent the upper limits of abstract behavior in perceptrons considered to date. They are handicapped by a lack of a satisfactory "temporary memory", by an inability to perceive abstract topological relations in a simple fashion, and by an inability to isolate meaningful figural entities, or objects, except under special conditions.

The capabilities which are outlined above, and the variety of networks and dynamic principles considered, map out a substantial territory, much of which still remains to be explored in detail. While rudimentary perceptual behavior appears to be present in these systems, it seems likely that to deal adequately with the problems of complex perceptual fields and the recognition of abstract relations between objects or events, additional principles must still be found.

27.2 Strategy and Methodology for Future Study

A number of perceptrons analyzed in the preceding chapters have been analyzed in a purely formal way, yielding equations which are not readily

translated into numbers. This is particularly true in the case of the four-layer and cross-coupled systems, where the generality of the equations is reflected in the obscurity of their implications, except for the few cases where explicit examples have been worked out. For other models, only qualitative results are available, although the way is clear for quantitative work to be initiated. Those problems which appear to be foremost at this time include the following:

(1) Theoretical learning curves for the error correction procedure. (At present, only empirical results are available, and no attempts at theoretical analysis have proven successful.)

(2) Determination of the probability that a solution exists to a given problem, for a perceptron drawn from a specified class.

(3) The development of optimum codes for the representation of complex environments, in perceptrons with multiple R-units (see Section 12.2).

(4) Development of an efficient reinforcement scheme for pre-terminal connections (c.f., Chapter 13).

(5) Optimum organization of stimulus analyzing mechanisms and networks with geometrically constrained connections (c.f., Chapter 23).

(6) Terminal performance of cross-coupled and four-layer perceptrons in generalization experiments, as a function of network parameters, reinforcement dynamics, and environment characteristics.

(7) Theoretical analysis of convergence-time and learning curves for adaptive four-layer and cross-coupled perceptrons.

(8) Quantitative studies of effects of threshold servos on system performance (c.f., Chapter 21).

(9) Quantitative studies of speech recognition and phoneme analyzing systems.

(10) Performance of back-coupled systems in selective attention and detection experiments.

(11) Quantitative studies of sequential program learning in back-coupled systems.

(12) Effect of spatial constraints in cross-coupled systems (e.g., limiting interconnections to pairs of A-units with adjacent retinal fields).

(13) Studies of possible figure-segregation (figure-ground) mechanisms.

(14) Studies of abstract concept formation, and the recognition of topological or metrical relations.

(15) Biological memory mechanisms, and studies of neurophysiology in relation to perceptron theory.

Four basic techniques are available for the study of these problems: theoretical analysis, digital simulation, the construction of physical models, and physiological experimentation. The first two problems of the above list are specifically mathematical in character. The third, while posed as a theoretical question, might best be investigated at the outset by means of simulation studies. In the case of problems (4) and (5), simulation studies seem to be indicated for preliminary exploration, although it is hoped that some theoretical formulations may ultimately be achieved. The sixth problem -- the determination of terminal performance of adaptive four-layer and cross-coupled systems -- calls in effect for a variety of explicit solutions to the steady-state equations presented in Part III. Such a program is currently being carried out both by direct computation of the equations and by simulation techniques. For the cross-coupled systems, simulation is likely to prove more economical in most cases than the numerical solution of the equations. The seventh question

again is a theoretical one, although preliminary results obtained from simulation programs should prove enlightening. The problem of threshold servomechanisms can be investigated both by theoretical means and by simulation.

It has recently been proposed that an audio-perceptron should be constructed at Cornell University to study the problem of speech recognition. Since this is a problem in which the chief interest is in performance under typical environmental conditions, rather than in theoretical problems of pattern recognition (which have all been solved on paper, insofar as spoken inputs resemble any other form of sensory sequences), it seems best to provide for convenient input to a real-time system, rather than working with simulated perceptrons and samples of digitalized speech. The problem of phoneme analysis, however, still presents enough theoretical problems and uncertainty as to the best solution, so that a digital simulation program is indicated. The system proposed in Chapter 23 is now being investigated by this means. The problems of back-coupled systems referred to in (10) are probably also best referred to an actual physical model, although a certain amount of useful simulation can be performed in checking out the general theory before such a model is built. Problem (11) is also of this character. Problem (12) is again of the type which will yield most readily to simulation at this time. It is of interest in connection with possible figure-ground mechanisms, which are included in a more general way in Problem (13).

Problems (13) and (14) are primarily speculative in character, and must await new insight into possible mechanisms, the exact nature of which is not yet clear. It is hoped that studies of the other problems, which are all well enough formulated to be investigated directly, will suggest possible approaches to these two problems, which represent the most baffling impediments to the advance of perceptron theory in the direction of abstract thinking and concept formation. The previous questions are all in the nature of "mopping-up" oper-

ations in areas where some degree of performance is known to be possible, and where suitable mechanisms can be described, at least in qualitative terms; the problems of figure-ground separation (or the recognition of unity) and topological relation recognition represent new territory, against which few inroads have been made.

The last problem -- the correlation of perceptron theory with biological evidence -- represents at once an area of investigation in its own right, and a potential source of insights into solutions to the prior problems. To date, little has been done to obtain relevant physiological data directly. Nonetheless, several hypotheses have been suggested (c.f., Chapter 26), and a great deal of useful work along the line of Hubel's studies of the cat cortex can be carried out using known laboratory techniques.

27.3 Construction of Physical Models and Engineering Applications

From a purely scientific standpoint, physical models of particular perceptron organizations seem to be indicated only for relatively advanced systems (such as the speech recognition, selective attention, and program learning perceptrons referred to above) where the theory is reasonably well known, but the actual quantitative behavior under realistic environmental conditions remains in doubt. In some cases, it may ultimately prove more economical to build a physical model than to simulate a highly parallel signal network on a sequential computer. Digital simulation, however, always has the advantage of greater versatility and adaptability to radical changes in design and dynamics of the simulated network. Its main difficulties are insufficient speed, insufficient high-speed memory, and difficulty of programming the simulation of complicated "naturalistic" environments required for some exeriments. This last disadvantage can be overcome by the design of special sensory input devices (such as audio analyzers and flying-spot scanners) for digital computers, and it is hoped that such equipment will be available in the near

future. While most problems can be investigated successfully in scaled-down versions using a computer comparable to the IBM 704 or 7090, a problem occasionally occurs which places a severe strain on the capability of even the best digital equipment now available. The study of evolutionary models, and adaptation processes in cross-coupled systems appear to be of this variety. A special purpose digital computer (such as the Mark II design proposed by C. A. L.) may ultimately prove to be the most expedient solution to these problems, although the limits of useful simulation with conventional computers have not yet been reached.

The construction of physical perceptron models of significant size and complexity is currently limited by two technological problems: the design of a cheap, mass-produceable integrator, and the development of an inexpensive means of wiring large networks of components. The Mark I (Frontispiece) employs motor-driven potentiometers for integrators, and a large patch-panel for connections - both intolerable solutions for very large systems. The integrator problem is currently being attacked by groups at Aeronutronic and Stanford Research Institute, who have developed magnetic integrators which are suitable for alpha-system perceptrons, and at Cornell University, where an electrochemical system is under investigation. While these approaches seem to offer some hope of an "intermediate" solution to the problem, an ultimate solution is more likely to come from some of the solid state work and studies of microelectronics, such as the work of Shoulders at SRI (Reference 114). This last technique offers a potential solution to the interconnection problem, as well as a possible means of fabricating large numbers of digital integrators at low cost.

Since the main emphasis in this volume has been on neurodynamic theory, rather than applications, little has been said about the engineering aspects of the field. It is clear that if the objective of a coherent theory of brain mechanisms

is achieved, it is likely to prove applicable to pattern recognition and control devices, as well as the development of advanced computing systems of many varieties. Preliminary studies have been carried out dealing with possible applications of perceptrons to photo-interpretation (Reference 116) and the recognition of events in bubble chambers (Reference 115). More abstract applications of the pattern recognition ability, such as the diagnosis of clinical syndromes or meteorological prediction, have occasionally been proposed, although little evidence has been accumulated regarding the relative suitability of perceptrons as opposed to more conventional techniques for dealing with such problems. The applications most likely to be realizeable with the kinds of perceptrons described in this volume include character recognition and "reading machines", speech recognition (for distinct, clearly separated words), and extremely limited capabilities for pictorial recognition, or the recognition of objects against simple backgrounds. "Perception" in a broader sense may be potentially within the grasp of the descendants of our present models, but a great deal of fundamental knowledge must be obtained before a sufficiently sophisticated design can be prescribed to permit a perceptron to compete with a man under normal environmental conditions.

The most important technological development which may be inherent in the future development of brain models, would be the provision of "eyes and ears" for conventional computers and automata, giving them a common universe of discourse with their operators. Current attempts at heuristic problem-solving programs (such as Newell and Simon's programs) and at automatic language translation, are hampered by a lack of common referents for symbols, which can be no more than code-numbers for the computer, but which have a wealth of associated meanings for the operator. The development of a system which, by virtue of shared sensory experience, can "comprehend" the nature of the physical referents in a descriptive statement, is probably a necessary first step to the creation of a truly useful problem-solving computer. Linguistic capability, related

to perceptual experience, is of the essence for an "intelligent" system, artificial or otherwise.

27.4 Concluding Remarks

The last four years have seen the development of perceptron theory from the study of a few primitive models to the mapping of a comprehensive field of investigation. In its present form, this theory is definitive only in its treatment of relatively simple systems, although a considerable number of more advanced systems are now understood at least in a qualitative fashion, and the way is now open to quantitative studies of well-defined problems.

As advanced perceptron models become more sophisticated in their psychological properties, it becomes more appropriate to consider them as devices capable of performing arbitrary programs of observation, response, and manipulation of data. As this condition is reached, the methodology of perceptron studies is likely to merge with that of the "heuristic program" approach to psychological functioning, advocated by Newell and Simon (Reference 62). In such programs, goal-motivated behavior becomes the main object of study, whereas in perceptrons studied to date, the behavior is motivated primarily by the present environment and state of the system. A merger of these approaches will not only open up new territory, but will be a sign of the "psychological maturity" of perceptron theory, inasmuch as it will permit the study of non-trivial problems in the psychology of thinking and problem-solving, in terms of neurodynamic systems of known physical structure.

On the other hand, the "biological maturity" of neurodynamic theory must await the solution, or at least a more promising approach, to the biological memory problem. Once this is achieved, a fruitful interaction between perceptron theory and neurophysiology can be expected; but the memory problem remains paramount in importance.

The theoretical approach presented in this volume is clearly a long way from an adequate "explanation" of the foundations of human experience. The work will have fulfilled an important purpose, however, if it has succeeded in conveying a recognition of the potential power of a mathematical study of neurodynamic systems, not only for understanding the physical mechanisms of the brain itself, but for comprehending the relationship of the cognitive process in man to the nature of the environment in which it occurs.

APPENDICES

APPENDIX A

NOTATION AND STANDARD SYMBOLS

1. Notational Conventions

While the mathematical notation employed in this volume may still be capable of further improvement, several conventions have been established which appear to work reasonably well. They include the following:

(1) Individual signal-units in the perceptron are referred to by a lower case letter to indicate the type, and a subscript to designate the particular unit in question (a_i = i^{th} A-unit). Individual stimuli are referred to by a subscripted capital (S_j), while stimulus sequences are designated by script capitals ($\mathcal{S}_j$).

(2) Numbers of signal units are designated by a capital N, with a subscript to indicate the type of unit in question (N_a = number of A-units). The number of stimuli is indicated by a small n.

(3) An asterisk is used to denote activity: a_i^* = activity state (or output signal) of the unit a_i; N_a^* = number of active A-units; $c_{i,j}^*$ = signal transmitted by connection $c_{i,j}$.

(4) Sets of units may be designated either by a subscripted capital or by a functional notation. For example, the set of A-units responding to stimulus S_i may be designated either by A_i or by $A(S_i)$.

(5) Where it is necessary to refer both to the unit receiving a signal and to the stimulus for which the signal occurs, a tensor notation is employed, with the signal unit indicated by a subscript and the stimulus by a superscript. For example, $a_i^j(t)$ = input signal to the i^{th} unit from the j^{th} stimulus at time t. An obvious extension would permit this notation to be applied to origins as well as termini of signals; thus $c_{i,j}^{*A}(t)$ would designate the signal trans-

mitted to unit j from unit i in response to stimulus S_k at time t.

(6) Whenever pairs of subscripts are used to designate a signal or connection (as in $c_{i,j}$) the first subscript indicates the origin, and the second the terminus. In generalization coefficients ($g_{i,j}$), the first subscript indicates the "recipient" and the second subscript indicates the "source" stimulus.

(7) In multi-layer systems, the layers are counted separately for each type of unit, and the number of the layer may be denoted by a superscript in parentheses (e.g., $N_a^{(2)}$ = number of units in the second association layer; $r_k^{(3)}$ = k^{th} R-unit of the third R-unit layer).

Matrix and vector notations, where employed, follow usual conventions, the particular symbols being defined in the text where they appear. The symbol σ, when it appears without subscripts, indicates a decay rate, and should not be confused with Kroneker's delta, which appears only with subscripts ($\sigma_{i,j}$), or with Dirac delta-functions, $\sigma(x)$, for which the functional notation is always used.

2. Standard Symbols

The following list includes those symbols which are used consistently throughout the text. A number of additional symbols are occasionally employed for convenience in particular expositions, and are defined where they occur.

u_i = generic symbol for the i^{th} signal-unit of a perceptron, or, in simple perceptrons, signal to the R-unit from the i^{th} stimulus.

A_i = i^{th} sensory unit

a_i = i^{th} association unit

r_i = i^{th} response unit

$c_{i,j}$ = connection from unit i to unit j.

A_i^* = output signal from A_i.

a_i^* = output signal from a_i.

r_i^θ = output signal from r_i .

$\mathcal{R}_i$ = sequence of response states occurring as outputs of a perceptron.

$c_{i,j}^\theta$ = signal transmitted to unit j from unit i , on connection $c_{i,j}$ (measured at point of arrival at the terminal unit).

$\tau_{i,j}$ = transmission time of connection $c_{i,j}$

$v_{i,j}$ = value of connection $c_{i,j}$ (occasionally abbreviated to v_i in simple perceptrons, indicating the value of the connection from a_i to the R-unit).

N_A = number of S-units

N_a = number of A-units

N_r = number of R-units

α_i = total input signal to the i^{th} unit. The signal due to stimulus S_j is designated either by $\alpha_i(S_j)$ or by α_i^j . If the tensor notation is employed, then α_i designates the vector of signals $(\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^n)$. Similarly, α^j may be used to designate the vector $(\alpha_1^j, \alpha_2^j, \ldots, \alpha_{N_a}^j)$.

β_i^j = component of α_i^j consisting of the sum of all signals originating from the S-units.

γ_i^j = component of α_i^j consisting of the sum of all signals originating from the A-units.

(The vectors β_i , β^j , γ_i , and γ^j are defined analogously to the corresponding α vectors.)

$\phi(\alpha)$ = functional notation for activity state of a simple A-unit. $\phi = 1$ if $\alpha \geq \theta$, 0 otherwise.

x = number of excitatory input-connections to an A-unit

y = number of inhibitory input-connections to an A-unit

θ = threshold (specifically, θ_i = threshold of i^{th} unit)

S_i = i^{th} stimulus

$\mathcal{S}_i$ = i^{th} sequence of stimuli

$\mathcal{S}_i'$ = i^{th} sequence of stimuli up to, but not including, the terminal stimulus

P_i = normalized retinal area (or fraction of sensory points) covered by S_i

$C_{i,j}$ = common area (retinal intersection) of stimuli S_i and S_j

W = stimulus world, or universe

n = number of stimuli in W

N = number of admissible stimulus sequences, consisting of stimuli in W

$C(W)$ = classification of stimuli in W, into two or more equivalence classes.

$R(W)$ = response function, assigning possible R-unit states to each stimulus in W

ρ_j = sign of classification of stimulus S_j (+1 or -1) in a binary classification, $C(W)$

η = increment of reinforcement per connection (typically ± 1 or 0, in quantized systems)

δ = decay rate, generally applied to decaying values, but occasionally used in connection with other quantities which are subject to exponential decay.

g_{ij} = generalization coefficient; the change in the signal to an R-unit for stimulus S_i as a result of applying a unit of positive reinforcement ($\eta = +1$) for stimulus S_j

G = matrix of generalization coefficients, g_{ij}

Q_i = probability that an A-unit, in a given class of perceptrons, responds to stimulus S_i

$Q_i^{(k)}$ = probability that a k^{th} layer A-unit responds to S_i

Q_{i_ν} = probability that an A-unit responds to the ν^{th} stimulus in sequence $\mathcal{S}_i$

Q_{ij} = probability that an A-unit responds both to S_i and to S_j

$Q_{i_\mu} Q_{j_\nu}$ = probability that an A-unit responds both to the μ^{th} stimulus of $\mathcal{S}_i$ and to the ν^{th} stimulus of $\mathcal{S}_j$

(The probability of joint response for an arbitrary number of stimuli, $Q_{ij\ldots m}$, is similarly defined. When it is understood that the environment consists of stimulus sequences, as in discussions of cross-coupled perceptrons, the subscripts of the Q-functions are always understood to refer to stimulus sequences, rather than individual stimuli.)

$\mu(x)$ = mean of the random variable x

$E(x)$ = expected value of x

$\sigma(x)$ = standard deviation of x

P = probability, particularly probability of correct performance in a given experiment.

$P_x(c)$ = notation commonly used for the probability that the random variable x has the value c; equivalent to $P(x = c)$

$T(S)$ = the transform obtained by applying transformation T to stimulus S

t = time

τ = number of stimuli (or duration, in units Δt) in a training sequence

α, γ, and Γ as prefixes indicate types of reinforcement systems.

r.c.s. = reinforcement control system.

APPENDIX B

LIST OF THEOREMS AND COROLLARIES

This appendix contains those results which have been explicitly stated in the form of theorems, for convenient reference. Theorems are numbered by chapter and theorem number, in the order in which they originally appear.

THEOREM 5.1: Given a retina with two-state (on or off) input signals, the class of elementary perceptrons for which a solution exists to every classification, $C(W)$, of possible environments, W, is non-empty.

THEOREM 5.2: Given an elementary perceptron and a classification $C(W)$, the following conditions are necessary but not sufficient for a solution to $C(W)$ to exist:

i) every stimulus must activate at least one A-unit;

ii) there should be no subset of stimuli containing at least one member of each class, such that in the union of the responding A-unit sets, every A-unit has the same bias ratio (with respect to the stimuli of the subset).

THEOREM 5.3: Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$; then in order for a solution to $C(W)$ to exist, it is necessary and sufficient that there exist some vector u in the same orthant as $C(W)$, and some vector α such that $G\alpha = u$.

COROLLARY 1: Given an elementary perceptron and a stimulus world W, then if G is singular, some $C(W)$ exists for which there is no solution.

COROLLARY 2: Given an elementary perceptron, if the number of stimuli in W is $n > N_a$, there is some $C(W)$ for which no solution exists.

COROLLARY 3: For any elementary perceptron, as the number n of stimuli in W increases, the probability that a randomly selected classification, $C(W)$, has a solution approaches zero (where $C(W)$ is chosen from a uniform distribution over the possible classifications of W).

THEOREM 5.4: Given an elementary α-perceptron, a stimulus world W, and any classification $C(W)$ for which a solution exists; let all stimuli in W occur in any sequence, provided that each stimulus must reoccur in finite time; then beginning from an arbitrary initial state, an error correction procedure (quantized or non-quantized) will always yield a solution to $C(W)$ in finite time, with all signals to the R-unit having magnitudes at least equal to an arbitrary quantity $\delta \geq 0$.

COROLLARY: Given an elementary perceptron, a stimulus world W, and any classification $C(W)$; then if a solution to $C(W)$ exists, the set of possible solutions to $C(W)$ has positive measure over the phase space of the perceptron.

THEOREM 5.5: Given an elementary α-perceptron with a finite number of memory states, a random-sequence stimulus world W, and any classification $C(W)$ for which a solution can be reached from the starting point by some reinforcement sequence, then a solution will be obtained in finite time with probability 1 by means of a random-sign correction procedure.

THEOREM 5.6: Given an elementary α-perceptron, a stimulus world W, and some classification $C(W)$ for which a solution exists, a solution can sometimes be achieved by an S-controlled reinforcement procedure. However, such a solution cannot be guaranteed for an arbitrary stimulus sequence, and may be unstable if it occurs.

THEOREM 5.7: Given an elementary perceptron with a finite number of memory states, a stimulus world W, and a classification $C(W)$ for which a solution can be reached from the starting point by some reinforcement sequence, then a solution can always be obtained in finite time by means of a random perturbation correction procedure.

THEOREM 5.8: Given an elementary γ-perceptron, a stimulus world W, and a classification $C(W)$, it is possible that a solution to $C(W)$ exists which cannot be achieved by the perceptron.

THEOREM 5.9: Given an α-perceptron, and a classification $C(W)$, a necessary and sufficient condition that the error correction procedure reach a solution (in finite time, with arbitrary starting point) is that there exists no non-zero vector $X^{\#}$ (whose components do not disagree in sign with $C(W)$) such that $b_i X^{\#} = 0$ for all i (where b_i is the bias number, defined as in Chapter 5).

COROLLARY: For an α-system, the condition that there exist no non-zero vector $X^{\#}$ such that $b_i X^{\#} = 0$ for all i is equivalent to the condition that there exist Z and U such that $GZ = U$ (where U is in the same orthant as $C(W)$).

THEOREM 5.10: Given a γ-perceptron, and a classification $C(W)$, a necessary and sufficient condition that the error correction procedure reach a solution (in finite time) is that there exists no non-zero $X^{\#}$ such that $b_i X^{\#} = c$ for all i.

COROLLARY: For a γ-system, the condition that there exist no non-zero vector $X^{\#}$ such that $b_i X^{\#} = c$ for all i is equivalent to the condition that there exist Z and U such that $GZ = U$ (where U is in the same orthant as $C(W)$).

THEOREM 7.1: Given a class of elementary α-perceptrons, a finite stimulus world W, a classification $C(W)$, and a training sequence;

then for every $\epsilon > 0$, there exists an $N_0(\epsilon)$ such that if $N_a > N_0(\epsilon)$, the probability of selecting a perceptron which will correctly identify the class of every positive stimulus will be greater than $1 - \epsilon$.

(see Page 157 for definition of positive stimulus.)

THEOREM 9.1: In a bounded α-perceptron, with S-controlled reinforcement, the probability distribution $\pi(v)$ (for the value of a particular connection) approaches a stable terminal distribution of the form $\pi(v) = c\left(\frac{p}{q}\right)^{v-L}$ where c is a normalization constant equal to $\frac{1-(p/q)}{1-(p/q)^{L-L+1}}$.

THEOREM 10.1: Given a completely linear perceptron, a stimulus world W, and a classification $C(W)$ such that the bias ratio of every S-unit is equal (and non-zero) no solution to $C(W)$ can exist.

THEOREM 10.2: Given a simple α-perceptron with simple A-units, an R-unit with a continuous monotonic sign-preserving signal generating function, a stimulus world W (in which each stimulus ultimately reoccurs) and any response function $R(W)$ for which a solution exists, then by means of the error-corrective reinforcement procedure, the given response function can always be approximated in finite time by an output vector $R(W) + \epsilon$, where ϵ is a vector $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, $|\epsilon_i| < \epsilon'$, where ϵ' may be an arbitrarily small quantity greater than zero.

LEMMA 1: Given a symmetric positive definite or positive semi-definite matrix, H, and any vector ζ, then $(\zeta, H\zeta) = 0$ only if $H\zeta = 0$.

LEMMA 2: For the same conditions as Theorem 10.2, given that a solution exists, the set of all solutions forms a hyperplane of dimension equal to the nullity of G.

COROLLARY 1: For the conditions of Theorem 10.2, and a phase space which is unbounded in all dimensions, the probability of convergence to an arbitrarily close approximation to $R(W)$ by means of a random-sign correction procedure or a random-perturbation correction procedure may be less than 1.

COROLLARY 2: Given the conditions of Theorem 10.2, and a phase space bounded in all dimensions, then (given that a solution to $R(W)$ exists in this bounded space) the response function can always be approximated by means of the random-sign correction procedure, the system converging in finite time to an approximation $R(W)+\epsilon$, ϵ a vector, where $|\epsilon_i| < \epsilon'$ for arbitrarily small $\epsilon' > 0$.

COROLLARY 3: Given the same conditions as Corollary 2, the response function can always be approximated by the random-perturbation correction procedure, the system converging in finite time to an approximation $R(W)+\epsilon$, ϵ having components of magnitude $|\epsilon_i| \leq |\eta|$ if the reinforcement is quantized, or $|\epsilon_i| \leq \epsilon' > 0$, if η is chosen from a continuous distribution around zero.

THEOREM 10.3: Given a simple perceptron with a simple R-unit, and with transmission functions for all A-R connections of the form $f(\alpha_i)v_{i_r}$, where f is any function, and given the existence of a solution to a classification function $C(W)$ for this perceptron, then if $\rho(v)$ is any polynomial of odd degree in v, there also exists a solution if the transmission function is changed to $f(\alpha_i)\rho(v_{i_r})$.

THEOREM 10.4: Given the perceptron of Theorem 10.3, if a solution exists for some transmission function $f(\alpha_i)v_{i_r}$, a solution does not necessarily exist for the transmission function $g(\alpha_i)v_{i_r}$, $g \neq f$.

THEOREM 10.5: Given a simple perceptron with A-R connections which differ in their transmission functions, or with uniform transmission functions but non-simple A-units, a response function $R(W)$ may

have a solution which is unattainable by either the error correction procedure or the random-sign correction procedure.

THEOREM 10.6: Given a simple perceptron with any mixture of transmission functions $f_j(\alpha_j, v_{jr})$ for the connections c_{jr}, and a response function $R(W)$ for which a solution exists; then there exists some transmission function $g(\alpha, v)$ which is uniform for all connections, such that a solution to $R(W)$ exists.

THEOREM 10.7: Given a simple perceptron with an R-unit which is either simple or has a continuous signal generating function, and with any combination of transmission functions from its A-units (all continuous functions of v_{ir}, equal to zero if $\alpha_i = 0$), and given a bounded phase space within which a solution exists for $R(W)$; then, if each stimulus in W ultimately reoccurs, an approximate solution $R(W) + \epsilon$ is always obtainable in finite time by the random-perturbation correction procedure.

THEOREM 12.1: Given a perceptron with more than one R-unit, and a response function $R(W)$ or a classification $C(W)$ for which a solution exists, it may be impossible to achieve this solution by an error correction procedure which applies negative reinforcement jointly to all R-units based on errors in their joint response.

THEOREM 13.1: Given a three-layer series-coupled perceptron with simple A and R-units and variable S-A connections, and a classification $C(W)$ for which a solution exists, it may be impossible to achieve a solution by any deterministic correction procedure which obeys the local information rule.

THEOREM 13.2: Given a three-layer series-coupled perceptron, with simple A and R-units, variable-valued S-A connections, bounded A-R values, and a classification $C(W)$ for which a solution exists,

then a solution to $C(W)$ can be obtained in finite time with probability 1 by means of a back-propagating error-correction procedure, given that each stimulus in W always reoccurs in finite time, and that probabilities p_1, p_2, and p_3 are all greater than 0 and less than 1.

(See Section 13.3 for definition of the back-propagating correction procedure.)

APPENDIX C
BASIC EQUATIONS

The following equations are those most likely to be referred to repeatedly, and are listed here in a somewhat different order from their appearance in the text.

(1) Generalization Coefficients

For an α-system,

$$g_{i,j} = n_{i,j}$$
$$E\,g_{i,j} = Q_{i,j} \quad \text{(normalized form)}$$

For a γ-system,

$$g_{i,j} = n_{i,j} - (1/N_a)\, n_i\, n_j$$
$$E\,g_{i,j} = Q_{i,j} - Q_i\, Q_j \quad \text{(normalized form)}$$

(2) R-unit Input Signals

For an α or γ-system,

$$u = Gx$$

where u is the vector of R-unit input signals, and $x_i = \rho_i f_i$ (f_i being the number of times S_i has been reinforced).

(3) Q-Functions

For individual stimuli, in a simple perceptron,

$$Q_i = \sum_{E=\theta}^{E_{max}} \sum_{I=\theta}^{E-\theta} P_x(E)\; P_y(I)$$

where $E_{max} = \begin{cases} z \text{ for binomial model} \\ \infty \text{ for Poisson model} \end{cases}$

$P_x(E)$ = probability that E excitatory connections to an A-unit originate from active S-points (see Equations 6.2 and 6.3)

$P_y(I)$ = probability that I inhibitory connections to an A-unit originate from active S-points (see Equations 6.2 and 6.3)

$$Q_{ij} = \sum_{\substack{E_i + E_c - I_i - I_c \geq \theta \\ E_j + E_c - I_j - I_c \geq \theta}} P_x(E_i, E_j, E_c)\, P_y(I_i, I_j, I_c)$$

where P_x and P_y are defined by Equations 6.6 and 6.7, for binomial and Poisson models.

For series-coupled perceptrons with distributed transmission times, see Sections 11.1 and 11.2 for prototype equations.

For multi-layer series-coupled systems, Q-functions for the k^{th} layer can be computed by the approximation described in Section 15.1.

For similarity-constrained four-layer perceptrons, Q_{ij} for two random or unrelated stimuli is given by:

$$Q_{ij}^{(2)} = \left[1 - \left(1 - Q_i^{(1)}\right)^m\right]^2$$

where m is the number of $A^{(1)}$ units connected to each $A^{(2)}$ unit.

For a stimulus S_i, and its transform S_i', in a similarity-constrained model,

$$Q_{ii'}^{(2)} = Q_i^{(2)} Q_{i'|i}^{(2)}$$

where $Q_i^{(2)} \approx 1-(1-Q_i^{(1)})^m$ and $Q_{i'|i}^{(2)}$ can be approximated by Equations 15.5 and 15.8 for the case of random stimulus patterns in a finite retina. In an infinite retina, with random stimuli, $Q_{ii'}^{(2)} = Q_{ij}^{(2)}$. For coherent stimuli and assuming T to be a topological transformation,

$$Q_{i'|i}^{(2)} \approx \frac{m-1}{\omega-1} + \left(1 - \frac{m-1}{\omega-1}\right)\left[1 - \left(1 - Q_{i'|i*}^{(1)}\right)^{m-1}\left(1 - \frac{Q_{ij}^{(1)}(0)}{Q_i^{(1)}}\right)\right]$$

where ω is the order of the transformation group, and $Q_{i'|i*}^{(1)}$ is given by Equation 15.6. A particular solution for the case of square stimuli can be found in Equation 15.15.

For cross-coupled perceptrons with fixed connections, $Q_{i\nu}$ and $Q_{i\mu,i\nu}$ are given by Equations 18.1 and 18.2, respectively.

For adaptive four-layer and cross-coupled systems, the terminal values of the Q-functions are obtained as a product of the iterative procedures described in Chapters 16, 17, and 19, and take the form:

$$Q_{ij} = \sum_{\beta_k} P(\beta_k)\ \phi\left(\beta_k^i + \gamma_k^i(\infty)\right)\left(\beta_k^j + \gamma_k^j(\infty)\right)$$

(4) Equations for Learning-Performance

For an error correction procedure, an upper bound on the number of corrections that will be required to achieve a solution from zero initial conditions is given by

$$N \leq nM/\alpha$$

where n is the number of stimuli in W, M is the maximum diagonal element g_{ii}, and α is the minimum of the function $z'Hz/\|z\|^2$ as defined for Theorem 4, Chapter 5. For a more general bound, see Equation 7.12.

For an S-controlled learning procedure, in an elementary perceptron, a bound on the error probability for a "positive stimulus" S_x is given by

$$P_e \leq \frac{\sigma^2(u_x)}{E^2(u_x)}$$

An improved estimate of the probability of correct response, employing a normal distribution assumption, is given by Equation 7.7.

For fixed training sequences,

$$E(u_x) = \begin{cases} TN_a \sum_j \rho_j P_j Q_{jx} & \text{for an } \alpha\text{-system} \\ TN_2 \sum_j \rho_j P_j (Q_{jx} - Q_j Q_x) & \text{for a } \gamma \text{ or } \gamma'\text{-system} \end{cases}$$

$$\sigma^2(u_x) = T^2 N_a \sum_j \sum_k \rho_j \rho_k P_j P_k (Q_{jkx} - Q_{jx} Q_{kx})$$

for an α-system, and

$$\sigma^2(u_x) = T^2 N_a \sum_j \sum_k \rho_j \rho_k P_j P_k \Big[(Q_{jkx} - Q_j Q_k Q_x) - 2Q_k(Q_{jx} - Q_j Q_x) - (Q_{jx} - Q_j Q_x)(Q_{kx} - Q_k Q_x)\Big]$$

for a γ'-system. The equation for a true γ-system is given in Equation 8.7.

For random training sequences, $E(u_x)$ is as above, and the variances are given by Equation 7.11 for an α-system, and Equation 8.14 for a γ'-system.

(5) Steady-State Equations for Four-Layer and Cross-Coupled Systems

For an adaptive four-layer α-perceptron (Chapter 16), the terminal values of the signals transmitted by the variable-valued connections are given by iterating the equation:

$$\gamma^{i}_{(\nu+1)} = \frac{N_a \eta}{\sigma} \sum_{j=1}^{n} C_{ij}\, \phi\left(\beta^{j} + \gamma^{j}_{(\nu)}\right)$$

where $\gamma^{i}_{(0)} = 0$ and $C_{ij} = \sum_{k=1}^{n} Q^{(1)}_{ik} f_{kj}$ (f_{kj} being the frequency of the sequence $S_k S_j$). This equation will converge in at most n steps to the terminal value of γ^{i}. Equations for γ and Γ-systems are presented in Chapter 16.

For an open-loop cross-coupled system, the above iteration equation applies without modification.

For a closed-loop cross-coupled α-perceptron, the iteration equation becomes

$$\gamma^{r}_{i(\nu+1)} = \frac{N_a \eta}{\sigma} \sum_{q} \left[P_q\, \phi\left(\beta^{q}_{i} + \gamma^{q}_{i(\nu)}\right) \sum_{\beta_j} P(\beta_j)\, \phi\left(\beta^{q'}_{j} + \gamma^{q'}_{j(\nu)}\right) \phi\left(\beta^{r'}_{j} + \gamma^{r'}_{j(\nu)}\right) \right]$$

which is specific to the i^{th} A-unit, or to the set of A-units having the β-vector β_i. The solutions for γ and Γ-systems are discussed in Chapter 19.

APPENDIX D

STANDARD DIAGNOSTIC EXPERIMENTS

A number of experiments have been described in the course of the text which are employed for comparison and evaluation of different perceptron models. Those experiments which are referred to by number are listed here for convenience in cross-referencing figures and discussions in the text.

EXPERIMENT 1: Horizontal/vertical bar discrimination, in 20 by 20 toroidally connected retina, with 4 by 20 bars. Stimuli occur in fixed sequence. S-controlled reinforcement is employed. (see Page 162)

EXPERIMENT 2: Same environment and procedure as Experiment 1, but with alternating positions in opposite classes. (see Page 164)

EXPERIMENT 3: Same as Experiment 1, but with stimuli occurring in random sequence. (see Page 170)

EXPERIMENT 4: Same as Experiment 3, but horizontal bars occur four times as frequently as vertical bars. (see Page 170)

EXPERIMENT 5: Same as Experiment 1, but with error-correction reinforcement. (see Page 173)

EXPERIMENT 6: Same as Experiment 5, but with stimuli occurring in random sequence. (see Page 173)

EXPERIMENT 7: Triangle/Square discrimination experiment, with error-correction procedure, in 20 by 20 retina. Random sequence, with stimuli occurring in all translational positions with equal probability. (see Page 173)

EXPERIMENT 8: Horizontal/vertical bar discrimination, with random sequences, and random-sign correction procedure. (see Page 176)

EXPERIMENT 9: Horizontal and vertical bars in random sequence, with R-controlled reinforcement. (see Page 214)

EXPERIMENT 10: "Spontaneous organization" experiment, with an environment of n stimuli, such that all pairs have equal intersections. The stimuli are divided into two classes, and the perceptron is exposed to a preconditioning sequence in which the transition probability between members of the same class is large, and the transition probability between classes is small. At the end of the preconditioning sequence, R-controlled reinforcement is applied for a brief period. (see Page 365)

EXPERIMENT 11: "Transformation learning" experiment, in which perceptron is exposed to alternating preconditioning sequence of stimuli and their transforms. After the preconditioning period, the perceptron is taught to discriminate two test stimuli, which were not previously seen, and is then tested on their transforms. (see Page 375)

EXPERIMENT 12: The preconditioning sequence consists of a repetitive sequence of four stimuli, with spatial relationships favoring the dichotomy (S_1, S_3) vs (S_2, S_4), while temporal association favors (S_1, S_2) vs (S_3, S_4). The Q-matrix is evaluated at the end of the preconditioning period. (see Page 393)

EXPERIMENT 13: "Sequence prediction" experiment. The preconditioning procedure uses a finite sequence environment with the same stimuli as in Experiment 12, but the perceptron is tested (in addition) with the stimulus S_1 followed by a sequence of null stimuli, and the Q-matrix for all subsequences is obtained. (see Page 445)

EXPERIMENT 14: Preconditioning procedure with same stimuli as in Experiment 12, but with each stimulus repeated two times whenever it occurs. The terminal Q-matrix for all subsequences is determined. (see Page 450)

EXPERIMENT 15: Selective attention experiment, for a four R-unit perceptron trained to discriminate shapes and retinal positions of stimuli, and then tested with complex stimuli combining two shapes and two positions simultaneously. (see Page 478)

EXPERIMENT 16: Selective attention in an audio-visual perceptron, trained to discriminate shapes and positions as in Experiment 15, but biased by the addition of an auditory name for the shape or position of part of the stimulus pattern. (see Page 482)

REFERENCES

1. Adrian, Bremer, Jasper, et al Brain Mechanisms and Consciousness (Symposium) Blackwell Scientific Publications, Oxford, 1954

2. Allanson, J.T. Some Properties of a Randomly Connected Neural Network Proceedings of Third London Symposium on Information Theory Butterworths, London, 1956

3. Ashby, W.R. Design for a Brain Wiley, New York, 1952

4. Babcock, Inselberg, Löfgren, vonFoerster, Weston, and Zopf Some Principles of Preorganization in Self-Organizing Systems (Technical Report No. 2, Contract Nonr 1834(21)) University of Illinois, Urbana, 24 June 1960

5. Bartlett, F. Remembering Cambridge University Press, 1954

6. Beurle, R. L. Properties of a Mass of Cells Capable of Regenerating Pulses Phil. Trans. Roy. Soc. London, B 240, No. 669, 55

7. Block, H.D., Knight, B.W., and Rosenblatt, F. Analysis of a Four-Layer Series Coupled Perceptron Paper No. 1, Cornell University Cognitive Systems Research Program Ithaca, July 1960

8. Bok, S.T. Histonomy of the Cerebral Cortex Elsevier, Amsterdam, 1959

9. Brink, F. Excitation and Conduction in the Neuron, in Stevens, S.S., Handbook of Experimental Psychology Wiley, New York, 1951

10. Bruner, J.S., Postman, L., and Rodrigues, J. Expectation and the Perception of Color Amer. J. Psychol., 1951, 64, 216-227

11. Bullock, T.H. Neuron Doctrine and Electrophysiology Science, 1959, 129, 997-1002

12. Burks, A.W., Goldstine, H.H., and vonNeumann, J. Preliminary Discussion of the Logical Design of an Electronic Computing Instrument, Part I Institute for Advanced Study, Princeton, 1947

13. Burns, B.D. The Mammalian Cerebral Cortex Edward Arnold, London, 1958

14. Cajal, S.R. Neuron Theory or Reticular Theory? Consejo Superior de Investigaciones Cientificas, Madrid, 1954

15. Clark, W.A. and Farley, B.G. Generalization of Pattern Recognition in a Self-Organizing System Proc. Western Joint Computer Conf., 1955, 86-91

16. Culbertson, J. T. The Mechanism for Optic Nerve Conduction and Form Perception: I Bull. Math. Biophysics, 1948, 10, 31-40

17. Culbertson, J. T. Consciousness and Behavior Wm. C. Brown Co., Dubuque, Iowa, 1950

18. Eccles, J.C. The Neurophysiological Basis of Mind Clarendon Press, Oxford, 1953

19. Eccles, J.C. The Physiology of Nerve Cells Johns Hopkins Press, Baltimore, 1957

20. Farley, B.G. and Clark, W.A. Simulation of Self-Organizing Systems by Digital Computer Trans. IRE Professional Group on Information Theory September 1954

21. Feller, W. An Introduction to Probability Theory and Its Applications Wiley, New York, 1950

22. French, J.D. The Reticular Formation Scientific American, 1957, 196, No. 5, 54-60

23. Gamba, A. Optimum Performance of Learning Machines (awaiting publication)

24. Gibson, E.J. and Walk, R.D. The Visual Cliff Scientific American, 1960, 202, No. 4, 64-71

25. Gibson, J.J. Adaptation, After-Effect, and Contrast in the Perception of Curved Lines J. Experimental Psychol., 1933, 16, 1-31

26. Gibson, J.J. The Perception of the Visual World Houghton-Mifflin, New York, 1950

27. Gibson, J.J., Olum, P., and Rosenblatt, F. Parallax and Perspective during Aircraft Landings Amer. J. of Psychol., 1955, 68, 372-385

28. Greene, P.H. An Approach to Computers that Perceive, Learn, and Reason Proc. Western Joint Computer Conf., March 1959

29. Hay, J.C., Lynch, B.E., Smith, D.R., and Murray, A.E. Mark I Perceptron Operators' Manual Cornell Aeronautical Laboratory Report No. VG-1196-G-5, Buffalo, 1960

30. Hay, J.C., Martin, F.C., and Wightman, C.W. The Mark I Perceptron, Design and Performance, Record of IRE National Convention, Part 2, New York, 1960

31. Hay, J.C. and Wightman, C.W. The Mark I Perceptron, Research Trends, 1960, 8, No. 1, 1-4

32. Hayek, F.A. The Sensory Order University of Chicago Press Chicago, 1952

33. Hebb, D.O. The Organization of Behavior Wiley, New York, 1949

34. Hochberg, J.E. Nature and Nurture in Perception, in Challenge and Response in Psychology, Knopf (awaiting publication)

35. Householder, A.S. and Landahl, H.D. Mathematical Biophysics of the Central Nervous System Mathematical Biophysics Monograph Series, No. 1, Principia Press, Bloomington, Indiana, 1945

36. Jackson, J.H. Selected Writings of John Hughlings Jackson Hodder and Stoughton, London, 1931

37. Jahnke-Einde Tables of Functions Dover, New York

38. Jasper, H.H. Functional Properties of the Thalamic Reticular System, in Brain Mechanisms and Consciousness Blackwell, Oxford, 1954

39. John, E.R. Some Speculations of the Psychophysiology of Mind, in Scher, J. Toward a Definition of Mind Free Press, 1960

40. Joseph, R.D. On Predicting Perceptron Performance, Record of IRE National Convention, Part 2, New York, 1960

41. Joseph, R.D. Contributions to Perceptron Theory, Cornell Aeronautical Laboratory Report No. VG-1196-G-7, Buffalo, 1960

42. Joseph, R.D. Two Theorems on Error Correction (Project PARA, Technical Memorandum No. 17) Cornell Aeronautical Laboratory, Buffalo, 1960

43. Kleene, S.C. Representation of Events in Nerve Nets and Finite Automata, in Shannon and McCarthy (Ed.) Automata Studies Princeton University Press, 1956

44. Koffka, K. Principles of Gestalt Psychology Harcourt-Brace, New York, 1935

45. Köhler, W. Relational Determination in Perception, in Jeffress, Cerebral Mechanisms in Behavior (Hixon Symposium) Wiley, New York, 1951

46. Köhler, W. Discussion on McCulloch, W., Why the Mind is in the Head, in Jeffress, Cerebral Mechanisms in Behavior (Hixon Symposium) Wiley, New York, 1951

47. Landahl, H. D., McCulloch, W.S., and Pitts, W. A Statistical Consequence of the Logical Calculus of Nervous Nets Bull. Math. Biophysics, 1943, 5, 135-137

48. Lashley, K.S. The Relation between Mass Learning and Retention J. Comp. Neurol., 1926, 41, 1-58

49. Lashley, K.S. Brain Mechanisms and Intelligence University of Chicago Press, Chicago, 1929

50. Lashley, K.S. The Problem of Serial Order in Behavior, in Jeffress, Cerebral Mechanisms in Behavior (Hixon Symposium) Wiley, New York, 1951

51. Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H. What the Frog's Eye Tells the Frog's Brain Proc. IRE, 1959, 47, 1940-1951

52. Lorente de Nð, R. Cerebral Cortex: Architecture, in Fulton, J.F., Physiology of the Nervous System, Oxford University Press, New York, 1943

53. Lotka, A.J. Elements of Mathematical Biology, Dover, New York, 1956

54. MacKay, D.M. On Comparing the Brain with Machines American Scientist, 1954, 42, 261-68

55. MacKay, D.M. The Epistomological Problem for Automata, in Shannon and McCarthy, Automata Studies, Princeton University Press, Princeton, 1956

56. MacKay, D.M. Operational Aspects of Intellect, in Mechanization of Thought Processes (Vol. I) H.M. Stationery Office, London, 1959

57. McCulloch, W.S. and Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity Bull. Math. Biophysics, 1943, 5, 115-133

58. Milner, P.M. The Cell Assembly: Mark II Psch. Rev., 1957, 64, 242-252

59. Minsky, M.L. Neural-Analog Networks and the Brain Model Problem (Thesis) Princeton University, 1954

60. Minsky, M.L. Some Methods of Artificial Intelligence and Heuristic Programming, in Mechanization of Thought Processes (Vol. I) H.M. Stationery Office, London, 1959

61. National Physical Laboratory, Mechanization of Thought Processes (Proceedings of Symposium No. 10), H.M. Stationery Office, London, 1959

62. Newell, A., Shaw, J.C., and Simon, H.A. Elements of a Theory of Human Problem Solving Psych. Rev., 1958, 65, 151-166

63. Newell, A., Shaw, J.C., and Simon, H.A. A Variety of Intelligent Learning in a General Problem Solver, in Yovits and Cameron (Ed.) Self-Organizing Systems Pergamon Press, New York, 1960

64. Olds, J. and Milner, P. Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain J. Comp. Physiol. Psych., 1954, 47, 419-427

65. Olds, J. A Preliminary Mapping of Electrical Reinforcing Effects in the Rat Brain J. Comp. Physiol. Psych., 1956, 49, 507-512

66. Olds, J. Adaptive Functions in Paleocortical and Related Structures, in Harlow and Woolsey, Biological and Biochemical Bases of Behavior, University of Wisconsin Press, Madison, 1958

67. Papert, S. Some Mathematical Models of Learning, Proceedings of Fourth London Symposium on Information Theory, 1960 (awaiting publication)

68. Penfield, W., and Rasmussen, T. The Cerebral Cortex of Man, Macmillan, New York, 1950

69. Pitts, W. The Linear Theory of Neuron Networks: The Static Problem, Bull. Math. Biophysics, 1942, 4, 169-175

70. Pitts, W. The Linear Theory of Neuron Networks: The Dynamic Problem, Bull. Math. Biophysics, 1943, 5, 23-31

71. Pitts, W. and McCulloch, W.S. How We Know Universals: The Perception of Auditory and Visual Forms, Bull. Math. Biophysics, 1947, 9, 127-147

72. Pribram, K.H. A Review of Theory in Physiological Psychology, Annual Review of Psychology, Vol. 11, Annual Reviews, Inc., Palo Alto, California, 1960

73. Rashevsky, N. Mathematical Biophysics, University of Chicago Press, Chicago, 1938

74. Rashevsky, N. Mathematical Biophysics: Physico-Mathematical Foundations of Biology (Vol. II), Dover, New York, 1960

75. Riesen, A.H. The Development of Visual Perception in Man and Chimpanzee, Science, 1947, 106, 107-108

76. Roberts, L.G. Pattern Recognition with an Adaptive Network, Record of IRE National Convention (Part 2), New York, 1960

77. Rochester, N., Holland, J.H., Haibt, L.H., and Duda, W.L. Tests on a Cell Assembly Theory of the Action of the Brain, Using a Large Digital Computer, IRE Transactions on Information Theory, IT-2, 1956, 80-93

78. Rosenblatt, F. The Perceptron, A Perceiving and Recognizing Automaton (Project PARA), Cornell Aeronautical Laboratory Report No. 85-460-1, January 1957

79. Rosenblatt, F The Perceptron: A Theory of Statistical Separability in Cognitive Systems, Cornell Aeronautical Laboratory Report No. VG-1196-G-1, January 1958

80. Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psych. Rev., 1958, 65, 386-408

81. Rosenblatt, F. The Design of an Intelligent Automaton, Research Reviews, Office of Naval Research, Washington, October 1958, 5-13

82. Rosenblatt, F. Two Theorems of Statistical Separability in the Perceptron, in Mechanization of Thought Processes (Vol. I), H.M. Stationery Office, London, 1959

83. Rosenblatt, F. A Conjecture on the Biochemistry of Memory Mechanisms (Project PARA Technical Memorandum No. 10), Cornell Aeronautical Laboratory, Buffalo, 1959

84. Rosenblatt, F. Perceptron Simulation Experiments, Proc. IRE, 1960, 48, 301-309

85. Rosenblatt, F. Perceptual Generalization over Transformation Groups, in Yovits and Cameron (Ed.), Self-Organizing Systems, Pergamon Press, New York, 1960

86. Rosenblatt, F. On the Convergence of Reinforcement Procedures in Simple Perceptrons, Cornell Aeronautical Laboratory Report No. VG-1196-G-4, Buffalo, February 1960

87. Rosenblatt, F. Tables of Q-Functions for Two Perceptron Models, Cornell Aeronautical Laboratory Report No. VG-1196-G-6, Buffalo, May 1960

88. Ruch, T.C. Sensory Mechanisms, in Stevens, S.S., Handbook of Experimental Psychology, Wiley, New York, 1951

89. Ruch, T.C. Motor Systems, in Stevens, S.S., Handbook of Experimental Psychology, Wiley, New York, 1951

90. Sauer, E.G.F. Celestial Navigation by Birds, Scientific American, 1958, 199, No. 2, 42-47

91. Shannon, C.E. and McCarthy, J. Automata Studies, Princeton University Press, Princeton, 1956

92. Shimbel, A. and Rapoport, A. A Statistical Approach to the Theory of the Central Nervous System, Bull. Math. Biophysics, 1948, 10, 41-55

93. Sholl, D.A. The Organization of the Cerebral Cortex, Wiley, New York, 1956

94. Sperry, R.W. Mechanisms of Neural Maturation, in Stevens, S.S., Handbook of Experimental Psychology, Wiley, New York, 1951

95. Sperry, R.W. Physiological Plasticity and Brain Circuit Theory, in Harlow and Woolsey, Biological and Biochemical Bases of Behavior, University of Wisconsin Press, Madison, 1958

96. Stark, L. and Baker, F. Stability and Oscillations in a Neurological Servomechanism, J. of Neurophysiol., 1959, 22, 156-164

97. Sutherland, N.S. Stimulus Analyzing Mechanisms, in Mechanization of Thought Processes, H.M. Stationery Office, London, 1959

98. Sutherland, N.S. Theories of Shape Discrimination in Octopus, Nature, 1960, 186, 840-844

99. Taylor, W.K. Electrical Simulation of Some Nervous System Functional Activities, Proc. of Third London Symposium on Information Theory, Butterworths, London, 1955

100. Turing, A.M. On Computable Numbers, with an Application to the Entscheidungs-problem, Proceedings of the London Mathematical Society, Ser. 2, Vol. 2 (1936-37), 230-265

101. Uttley, A.M. Conditional Probability Machines and Conditioned Reflexes, in Shannon and McCarthy, Automata Studies, Princeton University Press, Princeton, 1956

102. Uttley, A.M. Conditional Probability Computing in a Nervous System, in Mechanization of Thought Processes, H.M. Stationery Office, London, 1959

103. VonNeumann, J. The General and Logical Theory of Automata, in Jeffress, Cerebral Mechanisms in Behavior (Hixon Symposium), Wiley, New York, 1951

104. VonNeumann, J. Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components, in Shannon and McCarthy, Automata Studies, Princeton University Press, Princeton, 1956

105. VonNeumann, J. The Computer and the Brain, Yale University Press, New Haven, 1958

106. VonSenden, M. Raum und Gestagtauffassung bei operierten Blindgeborenen vor und nach der Operation, Barth, Leipzig, 1932

107. Wallach, H. Memory Effects in Perception, Acta Psychologica, 1955, 11, 180

108. Woolsey, C.N. Organization of Somatic Sensory and Motor Areas of the Cerebral Cortex, in Harlow and Woolsey, Biological and Biochemical Bases of Behavior, University of Wisconsin Press, Madison, 1958

109. Yovits, M.C. and Cameron, S. (Editors) Self-Organizing Systems (Proceedings of an Interdisciplinary Conference), Pergamon Press, New York, 1960

110. Bremermann, H.J. The Evolution of Intelligence: The Nervous System as a Model of its Environment Technical Report No. 1, Contract Nonr 477(17), University of Washington, July 1958

111. Pontrjagin, L. Topological Groups, Princeton University Press, Princeton, 1939

112. Stevens, S.S. Handbook of Experimental Psychology, Wiley, New York, 1951

113. Hubel, D.H. and Wiesel, T.N. Receptive fields of single neurons in the cat's striate cortex Journal of Physiology, 1959, 148, 574-591

114. Shoulders, K.R. Research in Microelectronics Using Electron-Beam Activated Machining Techniques Stanford Research Institute, Menlo Park, California, September 1960

115. Kesler, C. Preliminary Experiments on Perceptron Applications to Bubble Chamber Event Recognition, in Cognitive Systems Research Program Report No. 1, Cornell University, January 1961

116. Murray, A. Phase 1 Interim Report: Perceptron Applicability to Photointerpretation Cornell Aeronautical Laboratory Report No. VE-1446-G-1, November 1960