AN EVALUATION OF THE PERCEPTRON THEORY

by

DALE RAYMOND LUMB

B. S., Kansas State University
of Agriculture and Applied Science, 1958

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
OF AGRICULTURE AND APPLIED SCIENCE

1959

TABLE OF CONTENTS

INTRODUCTION

As a consequence of the development of high-speed data processors, groups of investigators became aware of the difficulty of supplying data to high-speed digital computers with a speed comparable to that of the computer. Unfortunately, at the present stage of the art, all data collection must be done through the human channel which, while being amazingly flexible and complex, is inherently slow due to its low capacity. In this line very little has been done of any practical consequence. A few attempts to recognize time pattern can be listed, such as the work done at IBM, London University, and BTS on recognition of sound.

Most of the several approaches that have been investigated are of deterministic nature that would work well only in situations where the group of signals would be strictly constrained to be of a deterministic type. As a result of this situation, the various authors have been forced to introduce unbearable restrictions to the possibility of application of their methods. Typical examples of the failure to which these endeavors are doomed are the BTS digit recognizer that had to be regulated to a single speaker, and the recognizing machine developed at London University that systematically failed on some sound combinations.

In recent years a new point of view has been formulated. This point of view may be summarized by stating that to perform recognition, redundancy reduction, and noise elimination, one must deal with a cognitive system. In other words, only a system that is able to learn the probability distributions of the ensemble on which it is operating will have a fair chance to succeed.

Along this line of thought we find the contributions due to Allanson, at the University of Birmingham, Taylor, at London University, and Rosenblatt, at Cornell Aeronautical Laboratories.

The purpose of this paper is to investigate and evaluate the theory of operation of a machine or class of machines called the Perceptrons originated by Dr. Rosenblatt, of the Cornell Aeronautical Laboratories.

BASIC CONCEPTS OF STATISTICAL SEPARABILITY THEORY

The Theory of Statistical Separability is the theory of operation of a system called the Perceptron which operates according to certain statistical principles.

The system is so designed that it responds to a statistical bias. Information is stored on the basis of retaining that which is essential to the classification or discrimination of stimulus. Associative memory is employed rather than exact reproducibility of remembered materials.

For the Perceptron the idea of memory will be realizable in quite a different fashion as compared to the digital computer memory system. Representational memory employed in the computer is the logically translatable coding of desired information to be stored. If the Perceptron, for example, were to use digital computer memory devices, then for each retinal on--off cell there must be a corresponding storage for one bit of information. Then for a million on--off retinal cells a million storage units would be required. Although this requirement is realizable, the time required to select or identify the storage unit which corresponds most nearly to each new input would fall quite short of simulating any operation comparable to that of the human visual performance. Fortunately, this type of memory is not employed in the Perceptron. Rather an associative memory is used to identify or discriminate inputs. Although the organization of the Perceptron will be discussed later, it is sufficient at this point to say that the retinal cells are connected at random to a set of cells called association units. Thus any pattern of cells stimulated on the retina would activate a subset of these association cells. With associative type of memory the information content is contained in the connection patterns resulting from points of stimulation on the retina to cells of activity in the associations units.

In place of the idea of errorless retention, redundancy is employed in the use of the same associate units.

The system will occasionally make errors in identification of a pattern which has been correctly identified before, not because of malfunctioning of the electronic hardware, but because the system operates in a probabilistic manner. Since the nature of the system is statistical, the probability of correct recognition fluctuates with time. That is, the adapting of the system to its inputs is a function of time. Learning takes place, and then the system is said to adapt to its environment. Thus it follows that the statistical bias which determines the proper response will change with time.

The principle for connections is essentially random within limitations of the plan of organization. In an analogous manner the biological nervous system is assumed to have entire freedom in the details of connections.

According to biological nervous system theories, a system spontaneously adapts to its environment by two possible methods.

In one theory a system learns or adapts to its environment by change in network topology. As the nervous system adapts to its environment, neuron connections or branches of the neuron network continually change their topology.

The other theory assumes that a system adapts to its environment by changing a value function associated with the neurons. The network once established (upon birth) remains constant throughout the system's entire life and learning is accomplished by changes of some parameters of the neuron composition.

The latter theory of learning is the basis of Dr. Rosenblatt's Perceptron Theory.

ORGANIZATION OF THE PERCEPTRON

The basic organization of the Perceptron will consist of a sensory unit, two response units, R_1 and R_2, and their associated source sets, A_1 and A_2, respectively. The relatively simple model shown will be capable of a limited vocabulary; however, it will serve to illustrate the basic principles of the function and organization of the Perceptron. One method of pictorial representation of the basic organization is by use of the Venn diagram, Plate I, Fig. 1. The circles represent sets or classes of units, and the arrowed lines indicate directional excitatory connections of the various sets of units. The lines terminated by small circles indicate inhibitory connections. Figure 2, Plate I, is a schematic representation corresponding to Fig. 1, Plate I.

Now consider the laws or rules which govern the connections between the different sets of units of the Perceptron. The network of connections between S- and A-units is one of uniform random distribution. That is, any S-point may be connected to any A-unit with equal probability. Each S-point may be connected to several A-units distributed uniformly over the entire A-set. Each A-unit will have several S-points connected to it. These S-points are called the origin points of an A-unit. In the simplest Perceptron the origin points are uniformly distributed at random throughout the S-set. However, in order for the Perceptron to have sensitivity to contours and gradients, the origin points for a single A-unit must be concentrated in a small area such as an exponential distribution about a central point.

The A-units are connected to the R-units at random, similar to that of the S-points and A-unit connections. In general, this connection results in three A-subsets. One subset will be those A-units, denoted by A_1 set or R_1 source set, transmitting

EXPLANATION OF PLATE I

Fig. 1. A Venn diagram of the organization of a simple Perceptron.

Fig. 2. A schematic representation corresponding to Fig. 1.

PLATE I

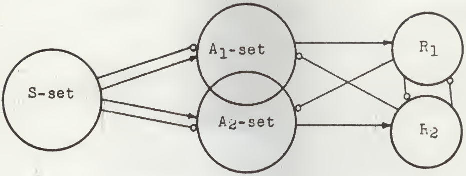


Fig. 1.

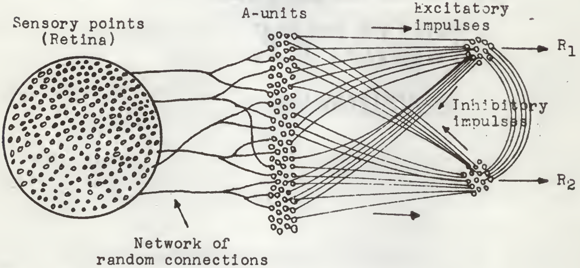


Fig. 2.

to R_1 response unit. The second subset consists of those A-units (A_2-set or R_2-source set) connected to the response unit R_2. The other is a small subset of overlapping A-units which is connected to both R_1 and R_2 response units.

In Fig. 1, the A_1-set is shown to be in the upper circle, the A_2-set the lower circle, and the intersection of the circles represents the overlapping set of A-units. There is no need for topographical segregation of the R-source sets in the actual system as the diagram was drawn in this manner for clarity.

As illustrated in the diagram of Fig. 1, the response units are mutually exclusive, that is, when R_1, for example, responds to a stimulus, it sends inhibitory impulses to A-units of A_2-set and to the other response unit R_2. Thus the rule of connection is that each R-unit inhibits the complement of its source set. When one R-unit has responded, it suppresses the other source-sets limiting the activity to the dominant A-set. The inhibitory impulses will prevent the non-dominant responses from being activated by impulses from the intersections of this source-set with the source-set of the dominant set.

Consider what happens upon presentation of the first stimulus to the sensory points of the Perceptron. A subset of uniformly distributed members of the A-system will respond. This set of points which is activated is the superset responding to the presented stimulus. At this point probably no R-unit will respond since the activated set of A-units is uniform in all source-sets. A response unit, say R_1, will be forced to respond by the experimenter. Then this response unit suppresses the other R-source sets. That is, the members of the original superset are inhibited except those of R_1 which become the dominant subset. The dominant R_1-subset consists of the active units of the R_1-source set which responds to a particular stimulus associated with R_1. Upon activation the active units of R_1 gain value with respect to the rival subsets. Since discrimination is based on the net value of the source-sets, then with increasing number of stimuli presented of this type the higher the probability that R_1 will respond autonomously. Similarly, with a different type of stimulus presented, R_2 may be forced to respond. When R_2 responds it inhibits the other source-sets; hence they are unable to gain value, and only the dominant R_2 subset (activated units of R_2) are allowed to gain value with respect to its complementary set.

The more presentations of the type of stimulus associated with R_2, the higher the value of R_2 due to stimulus S_{t_2} and the higher the probability of correct response of R_2.

In the biological nervous system there are three classes of cells: sensory, associative, and motor neurons. Corresponding to the biological system, the Perceptron has three elementary units which are the following: S-points (sensory points in a simulated retina), A-units (association units), and R-units (response units).

The sensory units receive the stimuli whatever they may be. For example, in the photoperceptron the stimulus will be proportional to the level of illumination. The response units may be considered the code center or a label of a particular class of stimulus.

The activity of the Perceptron upon the presentation of a stimulus will be divided into two classes, the predominant and the postdominant phases.

The predominant phase is only a transient phenomenon. When a stimulus has been shown to the sensory system, a certain number of A-units are activated. Some of these activated A-units will be members of both source-sets. The source-set, say R_1, for example, which contains the largest number of activated units will tend to have a higher net value than the other set. Thus R_1 will tend to respond. As R_1 responds, it suppresses the R_2-source set and the R_2-response unit. The above procedure takes place in a very short amount of time and is essentially a transient phenomenon.

Once a response unit has responded and the complement set has been suppressed, then the Perceptron is in the postdominant phase of its activity. During this phase the resulting unsuppressed activated A-units gain an increment of value, while the inactive A-units remain unchanged. It is evident that the next time the same stimulus is presented, the same reinforced A-units will be reactivated with a higher probability, and thus they will indicate the correct response. All of the above is a description of the reaction of the Perceptron to a presented stimulus.

The detailed analytical description of the predominant phase of Perceptron response was carried out by Dr. Rosenblatt and given in the Report on "A Theory of Statistical Separability in Cognitive Systems".

Two variables P_a, the expected proportion of A-units activated by a particular stimulus of a given size, and P_c, the expected proportion of A-units activated by one stimulus which are also activated by another stimulus, are sufficient to describe the predominant phase of the Perceptron.

The numerical evaluation of the equations for P_a and P_c were obtained by a Monte Carlo computation technique on the IBM 704 computer. The equations for P_a and P_c are essentially functions of the parameters of the Perceptron organization.

The connections of the system consist of random homogeneous distribution of connections between the S- and A-units. Each A-unit receives some excitatory connections and may, but not necessarily, receive some inhibitory connections from the sensory cells. The only restraint on the design of connections is that no two A-units are connected to identical sets of S-points. This restriction is placed so as to insure maximum difference in response of the system to different stimuli.

When any stimulus is presented to the sensory mosaic, a set of S-points is stimulated. The S-points are connected to the A-units by excitatory and inhibitory connections. If a sufficient number of net excitatory connections to an A-unit are excited by the stimulus, then the A-unit is activated. That is, if an A-unit receives a net amount of excitation greater than or equal to the threshold value, then that A-unit is to respond or become active. P_a and P_c are functions of the formulation of possible combination of excitatory and inhibitory connections and various levels of threshold values for the A-units. Hence the analysis concerning these quantities is essentially one of design possibilities of the Perceptron system.

This report will not be concerned with the aspect of the Perceptron analysis mentioned above, but it will be concerned with the feasibility of such a system for its intended purpose, that of learning its environment.

ALTERNATE PERCEPTRON MODELS

With the general statistical separability theory and the rules of organization given, several alternative Perceptrons are possible. On the basis of response unit discrimination there are two possible forms, the sum value and the mean value systems. In the sum value system, discrimination of the response units is based on a comparison of the total value of each A-subset (the set of active A-units per source-set).

Discrimination by the mean value system is the comparison of the mean value over the sets of active A-units. That is, an average is taken over each entire active subset, and the result is the mean value per active A-unit per subset, and the comparison for discrimination is made between the source-sets. The sum discriminating and the mean discriminating systems will be denoted by Σ-systems and μ-systems, respectively.

Three alternative Perceptron models will be considered with respect to the dynamics of the value change of each source set. One model is the uncompensated gain system called the Alpha

Perceptron. Each A-unit gains an increment of value per unit of time the cell is activated. When an A-unit is inactive or suppressed, it remains at a constant value which is determined by the number of reinforcements previously received. Thus the total value gain of a source-set per reinforcement is equal to the number of activated A-units per source-set. The mean value of the A-system increases with the number of reinforcements. This system has the advantage of being easy to design. However, it must operate under the restricted conditions that each response unit on the average is reinforced or becomes dominant with equal frequency. In the random environment the probability of correct response decreases to a random or chance expectancy of 0.5 when N_s, the number of stimuli presented to the system, becomes large enough. The system, under these conditions, becomes saturated.

The results of the analysis using mean discrimination for the Alpha system show that the performance is improved for higher values of n_s. In addition, the range of values of P_a for which the system operates satisfactorily is widened.

The second model is the constant-feed system which is called the Beta Perceptron. Independently of the number of reinforcements, a constant rate of value is fed to each source-set of the A-system. Hence the total value of all source-sets is always equal.

Within the source-set the active units take precedence over the inactive units; thus the value gain is distributed preferentially to the active units of each source set. The total value gain of a source-set per reinforcement is a constant, K, and the mean value of the A-system increases with time. An A-unit active for one unit of time gains K/N_{ar}, where N_{ar} is the number of active units in a source-set. The gain of an inactive A-unit outside the dominant set is K/N_{Ar}, where N_{Ar} equals the number of A-units connected to a response unit, while the gain of an inactive A-unit of the dominant set is zero.

The analysis of the Beta system has poorer performance than the Alpha system under all conditions, even with variation in n_{sr}, the number of stimuli associated to each response unit. The reason is the accumulation of value in the inactive units.

The parasitic gain system, or Gamma Perceptron, is the third Perceptron model that will be considered. The total value as well as the mean value of each source-set remains constant. Reinforcement produces only the effect of redistribution of the value among the A-units of a source-set. Within a source-set active A-units gain value at the expense of inactive A-units, which decrease in value.

Continuing the comparison of logical characteristics of the three systems, the total value gain of the source-set per reinforcement is, of course, zero. An A-unit active for one unit of time gains one increment of value. The inactive A-units outside of the dominant set gain zero value, while the inactive A-units of the dominant set loses $\frac{N_{ar}}{N_{Ar} - N_{ar}}$ increment of value.

ANALYSIS OF THE ALPHA PERCEPTRONS FOR IDEAL ENVIRONMENT

Response of the Alpha Systems with Uniform n_{s_r}

The performance of the Perceptron will be analyzed quantitatively on the basis of a hypothetical experiment. The experiment consists of a learning period and a testing period, during which time the capabilities of the machine will be evaluated.

During the learning period a specified number of stimuli, n_s, will be shown to the Perceptron. The experimenter will force each of these stimuli to become associated with one of the responses by forcing the desired response unit to respond. The stimuli for ideal environment each consists of a random collection of S-points to be stimulated. The stimuli will have the same measure, that is, each consists of the same number of S-points. It will be assumed that on the average an equal number of stimuli are associated to each response unit. In symbols n_{si} stimuli are associated during the learning period to response R_i.

In the testing period spontaneous reaction of the system to the previously reinforced stimulus s_t will be observed. Correct response is achieved if the testing and learning period responses are the same.

General considerations will now be given to the analysis of P_r, which is the probability of correct response during the testing period to stimuli previously reinforced during the learning period. Upon the presentation of a stimulus, the discrimination of the response units will be measured by the relative difference between the value of the source-sets. Thus the net bias B will be referred to as the net difference of value between R_1 and R_2 source-sets which results from a stimulus being presented. A convention to be used in this analysis is that only two response units will be assumed. However, the same analysis is valid for any number of response units. Under these conditions, if B is positive R_1 will be preferred, and if B is negative R_2 will be preferred.

The net bias B can be decomposed into two bias components--b, the controlled bias, and d, the random bias. The controlled bias b is the value gained by R_1 source-set due to stimulus S_t associated with R_1 when it was originally presented during the learning period. The random bias d is the net value between R_1 and R_2 source-sets due to all stimuli other than S_t.

Extensive use of statistical parameters will be made throughout this report and as each parameter is needed it will first be introduced in general statistical notation with the proper explanation. Then the application to the particular problem will be made.

The arithmetic mean of a distribution is the sum of the products of the values and their corresponding proportions. The arithmetic mean is also called the expected value of a member of the population to be chosen at random. If X is to denote a member of the set to be chosen at random and E denotes the expected value, then E(X) means the expected value of a member of the set chosen at random. The arithmetic mean is the center of gravity of a distribution since the sum of deviations from E(X) is zero. It should be noted that the expected value of a quantity X, for example, may be denoted by either E(X), or $\overline{X}$. Both notations will be used in this report.

Expressing the above two bias components as expected values plus their fluctuations, the following definitions result:

$\overline{b}$ = the expected bias gained by the R_1 source-set by the reinforcement due to the stimulus in question, s_t

$\overline{d}$ = the expected bias gained due to all reinforcements of the R_1 and R_2 source-sets, exclusive of s_t

Δb = the difference between the actual value of b from the expected value of b

Δd = the difference of the actual value of the random bias d, from the expected value of d.

In terms of the above components, the net bias may be expressed by

$$B = \overline{b} + \overline{d} + \Delta b + \Delta d \qquad (1)$$

For correct response of a particular stimulus, B must be positive for the stimulus. Therefore

$$\overline{b} + \overline{d} + \Delta b + \Delta d > 0$$

or

$$\overline{b} + \overline{d} > -(\Delta b + \Delta d)$$

which indicates that the sum of the expected bias must be greater than the fluctuation bias for correct response to occur.

The performance of the Perceptron systems may be measured by the correctness of response due to any particular stimulus in question. This is measured by P_r, the probability that when one stimulus of a class of stimuli associated with R_i-response unit is presented during the learning periods, this stimulus will be preferred over any particular response R_j in the testing phase.

From the previous considerations of the biases of the system, P_r would be directly proportional to the net expected bias and inversely proportional to the standard deviation of the bias components b and d. It is evident that P_r would be a function of the expected bias. However, this quantity must be normalized with respect to the standard deviation of the bias components denoted by $\sigma(b + d)$.

In notational form,

$$P_r = \int_{-\infty}^{\frac{\bar{b} + \bar{d}}{\sigma(b + d)}} f(X)\, dx \qquad (2)$$

where f is some suitable distribution function.

Δb and Δd, the error components of the bias, are not mutually independent because both components are functions of P_{Ai}, the actual proportion of A-units activated by the i^{th} stimulus.

Thus the standard deviation of (b and d) is difficult to evaluate. However, for a fixed value of Δb, $\sigma(d)$ could be calculated and the probability that the proper bias conditions would exist could be expressed by

$$\int_{-\infty}^{\frac{\bar{b} + \bar{d} + \Delta b}{\sigma_d}} \phi(Z)\, dZ$$

where $\emptyset$ is a suitable distribution function depending on Δb.

Now if the sum of all possible Δb were calculated, P_r may be written as follows:

$$P_r = \sum_{\Delta b} \left[\int_{-\infty}^{\frac{\bar{b} + \bar{d}}{\sigma_d}} \emptyset(Z) dZ \right] P(\Delta b) \qquad (3)$$

where $P(\Delta b)$ is the frequency of occurrence of Δb.

In order to simplify this expression, consider the quantity Δb. Δb is the error component of bias due only to the stimulus in question, the response of which is measured by P_r.

For the mean discriminating systems and the sum system with a large number of A-units, Δb is, in general, small compared to Δd. However, there is one condition which could make a critical difference in P_r if Δb was entirely neglected, and that is when $\Delta b = -b$. This indicates that when S_t is presented, then no A-unit in the R_1 source-set will be activated. For all other conditions, Δb can be neglected. The above sum reduces to one term which is

$$P_r = \left[\int_{-\infty}^{\frac{\bar{b} + \bar{d}}{\sigma_d}} \emptyset(Z) dZ \right] P(\Delta b \neq -b) \qquad (4)$$

The most logical choice for the distribution function, $\emptyset$, would be to assume a normal distribution function in view of the central limit theorem. Then P_r would be the normal distribution integral times the corrective factor $P(\Delta b \neq -b)$.

The expression for P_r becomes

$$P_r = P(\Delta b \neq -b) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

where

$$Z = \frac{\bar{b} + \bar{d}}{\sigma(d)} \qquad (5)$$

In the first analysis the study of the behavior of the system in an ideal environment will be carried out. Ideal environment is a simplification of the theoretical model, presented in order to simplify analysis, rather than an optimum. The important feature of ideal environment is that it simplifies the stimulus relationship associated to each response unit. Under this condition, each stimulus of the set of stimuli associated with a response unit has no correlation or relationship of any kind to any other stimulus of the same class. Another assumption is that all stimuli are of the same measure so that P_a will be identical for all stimuli.

The frequency of activation of the A-units will determine the bias of the source-sets and in turn the responses to be activated, thus determining correct recognition. With this in view very careful consideration must be made with respect to the details of the activity of the A-units during exposures.

Let the following notation be introduced:

P_i = the probability the i^{th} stimulus will be presented to the system

P_{Ai} = the probability that an A-unit will be activated by stimulus i.

The A-units are connected at random to the sensory system and the R-units. The expected value of P_{A_i} is the product of the values of P_{A_i} and their corresponding frequency of occurrence P_i.

In symbols, $$E(P_{A_i}) = \sum_i P_{A_i} P_i \qquad (6)$$

Since $E(P_{A_i})$ will be used quite frequently, the following notation is used: $P_a = E(P_{A_i})$. Several interpretations of P_a can now be projected, keeping in mind that P_a is an expected or mean value.

The most obvious meaning is that P_a is the probability that any randomly selected A-unit in the entire A-system will respond to a stimulus in question. It follows from this general definition that P_a is the proportion of A-units which will respond to a particular stimulus.

If a particular stimulus has activated an A-unit, it will gain an increment of value ΔV which has been set equal to unity. Then a final interpretation is that P_a is the expected value of a proportion of exposures on which an A-unit will gain an increment of value. In other words, P_a is the expected increment of value on the average that an A-unit will gain due to one exposure.

Many quantities in this analysis are expressible as a function of the random variable P_{A_i}. It is useful to measure the amount of variation in the value among the members of a population. One of the most frequently used measures of variability is variance, and its positive square root, the standard deviation denoted by σ^2 and σ, respectively.

In order to evaluate the variance of P_{A_i}, assume for $\sigma^2(P_{A_i})$ a series expansion of the type

$$\sigma^2(P_{A_i}) = \sum_{e=0}^{n} a_e \; P_a^{\,e} \qquad (7)$$

In practice, the powers of P_a higher than the second can be neglected since in most of the following considerations $P_a \ll 1$. Thus it can be assumed:

$$\sigma^2(P_{A_i}) = a_0 + a_1 P_a + a_2 P_a^{\,2} \qquad (8)$$

But noting that $P_a = 0$ if and only if $P_{A_i} = 0$ for all i, then

$$\sigma^2 \begin{pmatrix} P_{A_i} = 0 \\ P_a = 0 \end{pmatrix} = 0 = a_0 \qquad (9)$$

Therefore $\qquad \sigma^2(P_{A_i}) = a_1 P_a + a_2 P_a^{\,2} \qquad (10)$

But $P_a = 1$ if $P_{A_i} = i$ for all i. Thus

$$P_a = 1 \text{ implies } \sigma^2(P_{A_i}) = 0$$

or $\qquad a_1 + a_2 = 0, \; a_2 = -a_1 \qquad (11)$

and $\qquad \sigma^2(P_{A_i}) = a_1(P_a - P_a^{\,2}) \qquad (12)$

This being a variance of a population of probable numbers, its value cannot exceed 1; hence

$$\sigma^2(P_{A_i}) = P_a(1 - P_a) \qquad (13)$$

This coincides with the value given in reference (1) without justification. The above derivation indicates that this is the only feasible second order approximate of

$$\sigma^2(P_{A_i}) = f(P_a) \qquad (14)$$

Analysis for random environment is carried out because the analytical model used will serve as a basis of analysis for the modifications and extensions of more sophisticated Perceptrons.

For calculation of P_r as a function of P_a, the quantities which appear in the expression for P_r, namely, $\bar{b}$, $\bar{d}$, and σ_d, will now be calculated for the sum discriminating Alpha system.

Assuming non-overlapping source-sets, then the expected controlled bias $\overline{b}$ will be equal to the number of A-units activated in a source-set by a stimulus times the increment of value gained by an A-unit upon activation, which can be represented by $N_{a_r}\Delta V$. ΔV is assumed to be unity.

However, since overlap exists between source-sets, the effective value gained by a source-set is

$$\overline{b} = \overline{N}_{a_r} - \overline{N}_{a_c} \tag{15}$$

where $\overline{N}_{a_c}$ = expected number of common units activated by the stimulus in question.

Let $N_{a_{r_i}}$, N_{A_r}, and N_{a_r} be defined as follows:

$N_{a_{r_i}}$ = the number of active A-units per source-set when the i^{th} stimulus is presented

$N_{A_{r_j}}$ = the number of A-units connected to the response R_j, or in general

N_{A_r} = the number of A-units connected per response unit, since the variance of $N_{A_{r_j}}$ is considered negligible.

Then

$$N_{a_{r_i}} = N_{A_r} P_{A_i} \tag{16}$$

The expected value of N_{a_r} may now be calculated as follows:

$$E(N_{a_r}) = \sum_i N_{a_{r_i}} P_i = \sum_i N_{A_r} P_{A_i} P_i$$

$$= N_{A_r} \sum_i P_{A_i} P_i = N_{A_r} P_a \tag{17}$$

Similarly, $E(N_{a_c}) = P_a N_{A_C}$

where N_{A_C} = the number of A-units connected in common to R_i and R_j, a specified pair of response units

N_{a_c} = the number of A-units active in the N_{A_C} subset.

Substituting in the expression for the expected controlled bias yields

$$\bar{b} = P_a(N_{A_r} - N_{A_c})$$

Since $N_{A_r} - N_{A_c}$ is the number of effective A-units connected to a source-set, denoted by N_e, then $\bar{b} = P_a N_e$.

An experiment will be assumed in which the following conventions will be used. S_t has been selected to represent a known stimulus which will be used as a test stimulus. There is nothing special about this stimulus except that it has been chosen to represent any particular stimulus of the stimulus class associated to the R_1 source-set.

It is assumed that the number of stimuli associated to a response unit, n_{s_r}, are all equal. For the sake of calculation, the discrimination between stimuli belonging to response units R_1 and R_2 will be of concern, with S_t representative of the R_1 stimuli class.

The net bias is to measure the net value gained by the source-sets upon activation. The expected net bias is a measure of the net difference of value between the R_1 and R_2 source-sets due to stimuli reinforcements.

When any one stimulus is presented to the Perceptron in particular S_t, the expected value gained by the source-set which responds is equal to the number of effective units activated $\bar{N}_{a_r} - \bar{N}_{a_c}$, or $P_a N_e$. By definition, this value is the expected controlled bias $\bar{b}$. Since S_t will be assumed to be associated with R_1, then the $P_a N_e$ units activated by S_t form a set of units which will be called the $S_t R_1$ subset. Then stimuli other than

S_t associated to R_1 and R_2 presented during the learning period activate a portion of the S_tR_1 subset which tend to reinforce the S_tR_1 subset. Thus the result is to increase the probability for correct response of S_t during the test period. This overlapping bias reinforcement is measured by the random bias component, d.

The expected proportion of overlap of A-units to two stimuli is P_a for random environment. Then the expected bias $\bar{d}$ at the end of the learning period due to all stimuli belonging to R_1 and R_2 other than S_t, is equal to

$$\begin{aligned}\bar{d} = \bar{V}_1 - \bar{V}_2 &= P_a(P_aN_e)(n_{s_r} - 1) - P_a(P_aN_e)n_{s_r} \\ &= -P_a(P_aN_e) \qquad (18)\end{aligned}$$

where $\bar{V}_1 = P_a(P_aN_e)(n_{s_r} - 1)$ is the expected value of the R_1 source-set due to all stimuli associated with R_1 except S_t, and $\bar{V}_2 = P_a(P_aN_e)n_{s_r}$ is the expected value of the R_2 source-set due to all stimuli of the R_2 class.

The second quantity required for the P_r expression is the standard deviation of d, σ_d, which is defined as the positive square root of the variance of d, $\sigma_d{}^2$.

The error or random bias component, d, is given by

$$d = V_1 - V_2 \qquad (19)$$

where V_1 is the total value of the N_{a_1} units in the R_1 source-set and V_2 is the total value of the N_{a_2} units of the R_2 source-set at the time when S_t is presented. d is the net bias at the time when S_t is presented or the bias due to all other reinforcements of R_1 and R_2 other than S_t. V_1 is produced from $n_{s_r} - 1$ stimuli, other than S_t, associated with R_1, and V_2 is due to n_{s_r} stimuli which were associated with R_2 source-set.

The total value of either source-set R_1 or R_2 taken over all the active A-units which respond to S_t is given by

$$V_r = \sum_{j=1}^{N_{a_r}} v(a_j) \tag{20}$$

where $v(a_j)$ = the value of the a_j unit at the end of the learning period due to all stimulus other than S_t.

Before evaluating σ_d, several quantities will first be calculated.

The variance of N_{a_r} is

$$\sigma^2(N_{a_r}) = E(N_{a_r}{}^2) - \left[E(N_{a_r})\right]^2 \tag{21}$$

The expected value of N_{a_r} has been found to equal N_{A_r} P_a.

The expected value of $N_{a_r}{}^2$ can be found as follows:

$$E(N_{a_r}{}^2) = \sum N_{a_{r_i}}{}^2 P_i = \sum N_{A_r}{}^2 P_{A_i}{}^2 P_i = N_{A_r}{}^2 \sum P_{A_i}{}^2 P_i$$
$$= N_{A_r}{}^2\ P_a \text{ (see page 22 for } E(P_{A_i}{}^2)) \tag{22}$$

Substituting the above values in (1), the variance of N_{a_r} becomes:

$$\sigma^2(N_{a_r}) = N_{A_r}{}^2 P_a - (N_{A_r} P_a)^2$$
$$= N_{A_r}(1 - P_a)P_a \tag{23}$$

Each A-unit will be exposed n_{s_a} times with a probability of being activated upon each exposure of P_a. The value gained by the a_j unit upon the i^{th} exposure is $P_{A_i} \triangle V = P_{A_i}$, assuming the increment of value gained upon activation is unity.

The total value gained by the a_j unit upon n_{s_a} exposure can be represented by

$$v(a_j) = \sum_{i=1}^{n_{s_a}} P_{A_i} \tag{24}$$

The variance of $v(a_j)$ will be

$$\sigma^2\left[v(a_j)\right] = \sigma^2\left[\sum_{i=1}^{n_{s_a}} P_{A_i}\right] = (n_{s_a})\sigma^2(P_{A_i})$$

$$= (n_{s_a})(1 - P_a)P_a \qquad (25)$$

In order to evaluate the variance of the net value of a source-set $\sigma^2(v_r)$, which is a function of two random variables, consider the following derivation of the analogous expression.

The expression for d is a function of two random variables N_{a_r} and $v(a_j)$, so that additional considerations must be made as to the calculation of the variance of a quantity which is a function of two random variables.

In order to calculate the variance of $\sum_{j=1}^{N_{a_r}} v(a_j)$ of which N_{a_r} and $v(a_j)$ are random variables, the following derivation for the variance of summations of random variables is necessary.

Consider the summation

$$S = \sum_{i=1}^{n} x_i \qquad (26)$$

where x_i and n are random variables.

The expected value of S and S^2 are found as follows:

$$ES = E_n\, E\left\{S \,\middle|\, n\right\} = E_n\left[nE(x_j)\right] = E(x_j)E(n) \qquad (27)$$

where in general $E_n\, f(n)$ is the expected value of f(n) taken over n.

$$ES^2 = E\left[\sum_{i=1}^{n} x_i\right]^2 = E\left[\sum_{i=1}^{n} x_i^2 + \sum_{i=j} x_i x_j\right]$$

$$= E_n\left[nE(x_i^2) + n(n - 1)(E(x_j))^2\right]$$

$$ES^2 = E(x_i^2)\, E(n) + E(n^2)(E(x_j))^2$$

$$- E(n)(E(x_j))^2 \qquad (28)$$

The variance is defined as the second moment minus the square of the first moment, or

$$\sigma^2 S = E(S^2) - (E(S))^2$$
$$= E(n)\left[E(x_i^2) - (E(x_j))^2\right]$$
$$+ (E(x_j))^2\left[E(n^2) - (E(n))^2\right]$$

$$\sigma^2\left(\sum_{i=1}^{n} x_i\right) = E(n)\sigma^2(x_j) + (E(x_j))^2\sigma^2(n)$$

Substituting in equation $N_{a_r} = n$ and $v(a_j) = x_j$,

then $$\sigma^2(v_r) = E(N_{a_r})\sigma^2(v(a_j)) + (Ev(a_j))^2\,\sigma^2(N_{a_r}) \qquad (30)$$
$$= (P_aN_{A_r})P_a(1 - P_a)n_{s_a} + (P_an_{s_a})^2\,P_a(1 - P_a)N_{A_r}$$
$$= P_a^2(1 - P_a)\,N_{A_r}\,n_{s_a}\;\;1 + P_an_{s_a}$$
$$\approx P_a^3(1 - P_a)N_{A_r}(n_{s_a})^2 \qquad (31)$$

It should be noted that those A-units which are in common to the two source-sets contribute equal value to both sets. Hence they do not affect the net difference in bias between the sets. Since only the number of effective units are under comparison, N_{A_r} may be replaced by N_e in the above expression.

From statistical theory it is known that the variance of a difference of non-correlated random variables is equal to the sum of the variance of each quantity, or

$$\sigma^2(V_1 - V_2) = \sigma^2(V_1) + \sigma^2(V_2) \qquad (32)$$

Now in this particular analysis the variance of each source-set is the same. Thus

$$\sigma^2(d_\Sigma) = \sigma^2(V_1) + \sigma^2(V_2) \qquad (33)$$
$$= 2\sigma(V_r)$$
$$\approx 2\,P_a^3(1 - P_a)\,N_e(n_{s_a})^2 \qquad (34)$$

The final quantity required in order to calculate P_r, the probability of correct response, is $P(\Delta b \neq -b)$, the probability that at least one A-unit will respond to the stimulus in question.

$P(\Delta b \neq -b)$ implies that $N_{a_r} - N_{a_c} > 0$; thus $P(\Delta b \neq -b)$ $= P(N_{a_r} - N_{a_c} > 0)$.

The probability that any particular A-unit on the average will not respond to a given stimulus is $(1 - P_a)$. In terms of the number of effective units N_e, the probability that no A-unit will respond to a given stimulus is

$$P(N_{a_r} - N_{a_c} = 0) = P(\Delta b = -b) = (1 - P_a)^{N_e} \qquad (35)$$

It follows that the complement of this probability is the probability that at least one unit will respond to a stimulus

$$P(\Delta b \neq -b) = 1 - (1 - P_a)^{N_e} \qquad (36)$$

From the above consideration the probability P_r is given by

$$P_r = \frac{1}{\sqrt{2\pi}} \left[1 - (1 - P_a)^{N_e}\right] \int_{-\infty}^{Z} \ell^{\frac{t^2}{2}} \, dt \qquad (37)$$

where in general $\quad Z = \dfrac{\bar{b} + \bar{d}}{\sigma_d}$

and in this particular case

$$Z = \sqrt{\frac{(1 - P_a)^{N_e}}{2\, P_a (n_{s_a})^2}} \qquad (38)$$

By use of the normal cumulative distribution tables, the above expression may be evaluated. Plates II, III, and IV show the results of such evaluations.

For the previous development for the equation of P_r and for those to follow, it should be noted that P_r represents the probability of correct responses for the stimulus S_t. However, no special constraints were placed on S_t as compared to other stimulus except the designation of S_t to a particular class of stimulus. Then S_t is a generic stimulus of its assigned class and all equations concerning S_t hold equally well for all members of the class of stimuli to which S_t belongs.

For the various graphs to follow, it will be useful to introduce the following relationships.

Let $\omega = \dfrac{N_{R_a}}{N_R}$ = proportion of R-units connected to an A-unit

$\omega_c = \dfrac{N_{A_c}}{N_{A_r}}$ = proportion of A-units connected in common to the R_i and R_j response units

$= P\left\{\text{unit } a_i \text{ belonging to the } R_i \text{ source-set is common to the } R_j \text{ source-set}\right\}$

$= \dfrac{\text{measure}\left\{R_i\right\}}{\text{measure}\left\{R'\right\}}$

where R' = the set of all R-units except R_i

From the definition of N_{R_a} it follows:

$$\text{Me}\left\{R_j\right\} = N_{R_a} - 1$$

and, of course $\text{Me}\left\{R'\right\} = N_R - 1$

Thus
$$\omega_c = \frac{N_{R_a} - 1}{N_R - 1} = \frac{\omega N_R - 1}{N_R - 1} \tag{39}$$

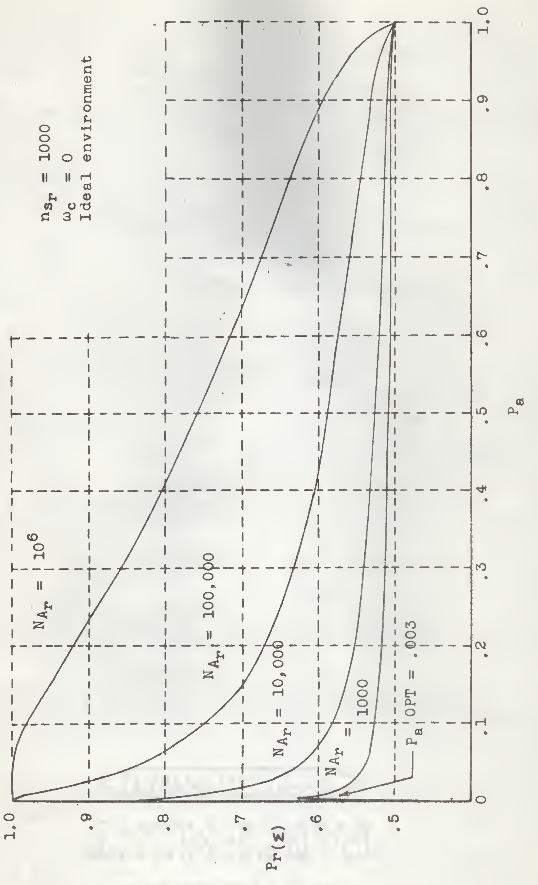
If N_R is large, ω_c approaches ω; and when $\omega_c = 0$, $\omega = \dfrac{1}{N_R}$.

The curves of Plate II show P_r as a function of P_a for several values of N_{A_r} with 1000 stimuli associated to each

EXPLANATION OF PLATE II

$P_r(\Sigma)$ at a function of P_a.

PLATE II


n_{s_r} = 1000
ω_c = 0
Ideal environment
N_{A_r} = 10^6
N_{A_r} = 100,000
N_{A_r} = 10,000
N_{A_r} = 1000
P_a OPT = .003
1.0
.9
.8
.7
.6
.5
$P_r(\Sigma)$
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
P_a

response unit and non-overlapping subsets. For a small number of A-units per source-set, P_a has a critically optimum point. Increasing the number of A-units increases the probability of correct response for a larger range of P_a. For $N_{A_r} = 10^6$, P_r is nearly unity for a range from $P_a = 0$ to $P_a = .05$.

With almost certainty that an A-unit will respond, that is, $P_a = 1$ for a given stimulus, then it is evident that P_r assumes chance expectancy ($P_r = .5$).

Plate III shows a set of curves for P_r as a function of n_{s_r}, the number of stimuli associated to each response. Parameters of the system consist of non-overlapping subsets and a fixed $P_a = .005$ which is rather an optimum value of P_a. From the curves it can be concluded that the number of stimuli which can be associated to a response unit for correct recognition increases with the number of A-units per subset.

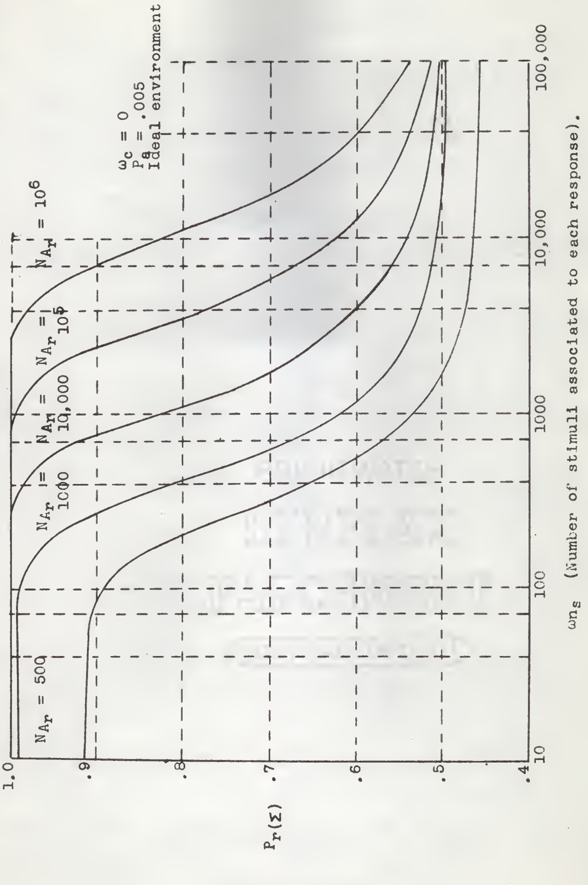
Plate IV gives the same sets of curves with the system parameters adjusted for $\omega = \omega_c = .5$, that is, the expected overlap among source-sets is 50 per cent.

Since these curves are for ideal environment, each stimulus of each class is independent or uncorrelated with any other stimulus. However, for any attempt to simulate this in a realistic environment, there would inevitably be a relationship between stimuli of a class. This would lead to mutual support between stimuli of a given category. Thus an increasing number of A-units would tend to be activated in common for stimuli of the same class, which would in turn increase the bias in the desired direction, making P_r higher under a given set of parameters.

EXPLANATION OF PLATE III

$P_r(\Sigma)$ as a function of n_{s_r}.

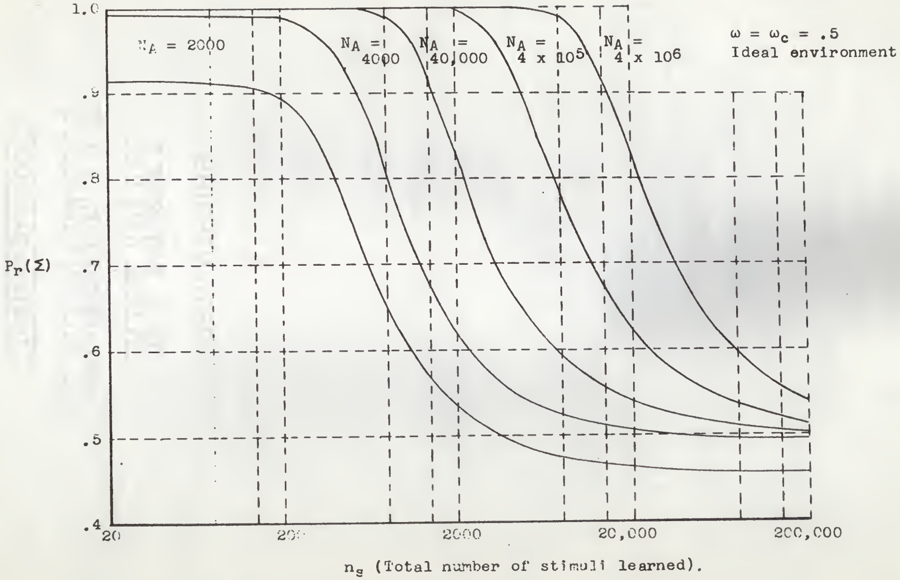
PLATE III


N_{A_r} = 500
N_{A_r} = 1000
N_{A_r} = 10,000
N_{A_r} = 10^5
N_{A_r} = 10^6
ω_c = 0
P_a = .005
Ideal environment
1.0
.9
.8
.7
.6
.5
.4
$P_r(\Sigma)$
10
100
1000
10,000
100,000
ωn_s (Number of stimuli associated to each response).

EXPLANATION OF PLATE IV

$P_r($ $)$ as a function of n_s.

PLATE IV



1.0
.9
.8
.7
.6
.5
.4
$P_r(\Sigma)$
N_A = 2000
N_A = 4000
N_A = 40,000
N_A = 4 x 10^5
N_A = 4 x 10^6
$\omega = \omega_c = .5$
Ideal environment
20
2000
20,000
200,000
n_s (Total number of stimuli learned).

In the following section the probability of correct response P_r will be calculated for the mean discriminating Alpha system. With mean discrimination the Perceptron responds to mean values of the active subsets of A-units rather than to sum values. In this system the component of variation of the controlled bias $\triangle b$ is zero, since it is due only to the variation in the number of A-units activated by the test stimulus. The mean value is measured over the entire number of A-units which are activated by S_t per source-set. Hence the expected bias $\overline{B}$ is the same as the bias in the sum system divided by the number of active effective units per source-set which can be represented by

$$\overline{B}_{\mu} = \overline{b}_{\mu} + \overline{d}_{\mu} = \frac{P_a N_e - P_a{}^2 N_e}{P_a\ N_e} = 1 - P_a \qquad (40)$$

As before, the variance of the value of an A-unit after n_{s_a} exposures is

$$\sigma^2 \left[v(a_j) \right] = P_a(1 - P_a)\ n_{s_a} \qquad (41)$$

For N_{a_r} active A-units per set, the variance of the mean value of the A-unit of one source-set is given by

$$\sigma^2 \left[\overline{v}(a_j) \right] = \frac{\sigma^2 \left[v(a_j) \right]}{N_{a_r}} \qquad (42)$$

Assuming disjunct sets (non-overlapping source-sets), the variance of the difference of the two means is

$$\sigma^2(d_{\mu}) = \sigma^2 \left[\overline{v}_1(a_j) \right] + \sigma^2 \left[\overline{v}_2(a_j) \right] \qquad (43)$$

and assuming the variance of both source-sets to be equal, the standard deviation of d, the positive square root of the variance, is

$$\sigma(d_{\mu}) = \sqrt{\frac{2\sigma^2\left[\bar{v}(a_j)\right]}{N_{a_r}}} \tag{44}$$

or substituting for $\sigma^2(v_r)$, N_{a_r}, and simplifying

$$\sigma(d_{\mu}) = \sqrt{\frac{2(1 - P_a)\, n_{s_a}}{N_{A_r}}} \tag{45}$$

Allowing for the correction of overlapping sets, N_{A_r} may be replaced by N_e (the effective number of A-units which contribute to the net bias between sets), and the above equation becomes

$$\sigma(d_{\mu}) = \sqrt{\frac{2\, n_{s_a}(1 - P_a)}{N_e}} \tag{46}$$

Correcting as before for the probability that no A-unit will respond, an analogous expression for P_r for the mean system can now be written.

$$P_{r(\mu)} = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

where
$$Z = \sqrt{\frac{(1 - P_a)^{N_2}}{2\, n_{s_a}}} \tag{47}$$

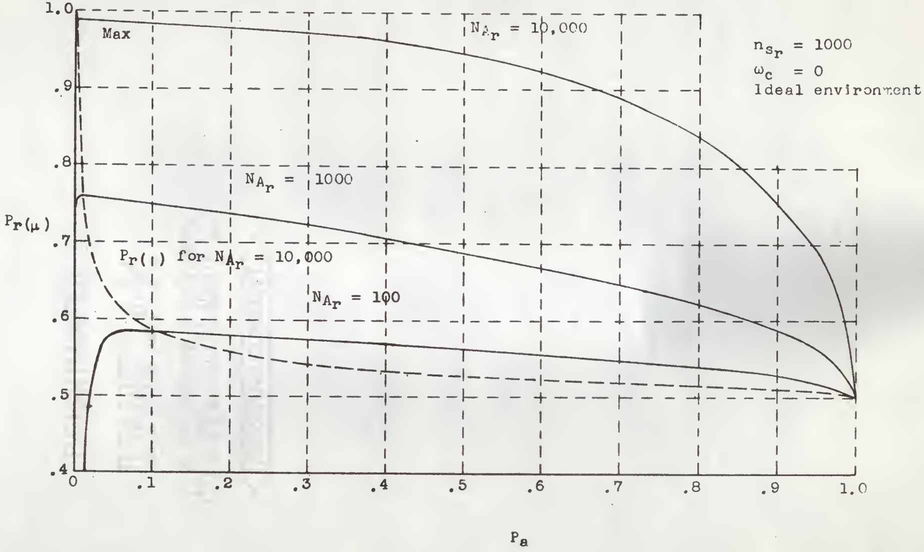
$P_{r(\mu)}$ as a function of P_a is illustrated by the set of curves in Plate V. The broken curve is given for $P_{r(\Sigma)}$ for $N_{A_r} = 10{,}000$ for comparison of the sum and the mean value systems. It is quite evident that under comparable conditions, the mean value system allows a much wider range of P_a for relatively good accuracy of correct recognition.

Plate VI shows a definite advantage for the μ-system as compared to Plate III for the Σ-system. For instance, with $N_A = 10{,}000$ A-units per source-set, the value of $P_{r(\mu)}$ remains nearly unity for about 500 associations per response and slopes

EXPLANATION OF PLATE V

$P_r(\quad)$ as a function of P_a.

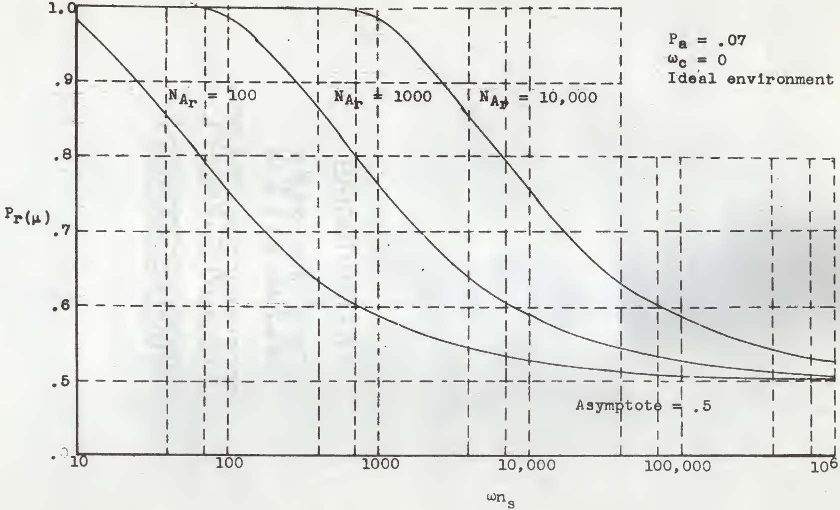
PLATE V


1.0
.9
.8
.7
.6
.5
.4
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
Max
N_{A_r} = 10,000
n_{s_r} = 1000
ω_c = 0
Ideal environment
N_{A_r} = 1000
$P_r(\mu)$
$P_r(\ )$ for N_{A_r} = 10,000
N_{A_r} = 100
P_a

EXPLANATION OF PLATE VI

$P_r($ $)$ as a function of ω_{n_s}.

PLATE VI



$P_a = .07$
$\omega_c = 0$
Ideal environment
$N_{A_r} = 100$
$N_{A_r} = 1000$
$N_{A_r} = 10,000$
Asymptote = .5
$P_r(\mu)$
1.0
.9
.8
.7
.6
.5
10
100
1000
10,000
100,000
10^6
ωn_s

off gradually. In addition, the graph shows that for a very large n_{s_r}, P_r reaches the chance expectancy of 0.5.

Effect on the Alpha Systems with Variation in n_{s_r}

In all of the previous analysis, n_{s_r}, the number of independent stimulus associated with each response unit, has been assumed to be equal for all response units. In reality, n_{s_r} may be considered as a random variable. In the ideal or random environment, it will be shown that the Alpha system will be incapable of efficient operation under the condition of random n_{s_r}.

First consider the case of non-overlapping subsets for the sum system of the Alpha Perceptron. Upon allowing n_{s_r} to vary, it will become quite evident that correct response is almost impossible. Consider the circumstances under which R_1 is associated with n_{s_1} stimuli and R_2 is associated with n_{s_2} stimuli. P_a may represent on the average the value gained per A-unit per stimulus. It follows that the value of the R_2 source set at the end of the learning period is $P_a\ N_{a_r}\ n_{s_2}$, the value of R_1 set before the presentation of the last stimulus is $(n_{s_1} - 1)\ P_a\ N_{a_r}$, and the value gained by the 1st stimulus is N_{a_r}. The expected net bias at the end of the learning period is

$$\overline{d} = \overline{v}_1 - \overline{v}_2 = (n_{s_1} - 1)P_aN_{a_r} + N_{a_r} - n_{s_2}P_aN_{a_r} \qquad (48)$$

As can readily be seen, if n_{s_1} is less than n_{s_2}, then the expected bias might easily be negative. Thus the probability of correct response for any stimulus of the n_{s_1} class which is presented in the test period is very low. In this case for correct response the value of the stimulus must not only be positive but it must be of sufficient magnitude to overcome the negative net bias of the system. If n_{s_2} is much greater than n_{s_1}, it is impossible to obtain correct response.

Now consider the condition in which the A-subsets are overlapping and the number of response units connected to each A-unit is large. n_{s_1}, n_{s_2}, ... will be picked from some distribution of n_{s_r}, not necessarily a normal distribution. The total number of stimuli presented to the system is not controlled; however, its expected value is $N_R n_{s_r}$. In this hypothetical experiment the variance of n_{s_r} will be considered to be large.

Under the above conditions a quantitative analysis of the mean discrimination Alpha Perceptron will be made. Before proceeding with the anlysis, the following relationships will be necessary.

If x is a random variable and f is any randomly varying function whose distribution depends on the value of x, then the expected value of f(x) is equal to the mean value taken over all x of the conditional mean value of the function f(x) relative to the hypothesis $x = \xi$, or in notational form

$$E\left[f(x)\right] = E_x\, E\left[f(x/x = \xi)\right] \tag{49}$$

Similarly, the expected value of the square of the same function is

$$E\left[(f^2(x)\right] = E_x\, E\left[f^2(x/x = \xi)\right] \tag{50}$$

The variance of such a random varying function in terms of the above equalities and inserting the second and third terms which sum to zero can be represented by

$$\sigma^2 f(x) = E_x E\left[f^2(x/x = \xi)\right] - E_x\left[E(f(x/x = \xi)\right]^2 + E_x\left[E(f(x/x = \xi)\right]^2 - E_x\left[E f(x/x = \xi)\right]^2 \tag{51}$$

$$\sigma^2 f(x) = E_x \sigma^2\left[f(x/x = \xi)\right] + \sigma^2\left[E f(x/x = \xi)\right] \tag{52}$$

Continuing with the experiment, let the test stimulus S_t activate n effective cells (non-overlapping units) in the R_1 source-set represented by $a_1, a_2, \ldots, a_n$. Furthermore, let $v(a_j)$ = the value of the a_j unit at the end of the learning series, except for the effect of S_t. The variance of the time conditional mean value of one source-set is represented by

$$\sigma^2\left(\frac{V_r}{n}\right) = \sigma^2\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right) \tag{53}$$

which is of the form of equation (52).

Substituting in equation (52),

$$\sigma^2\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right) = E_n\sigma^2\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right) + \sigma^2\left[E\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right)\right] \tag{54}$$

The expected value of $v(a_j) = P_a\, E\, N_{R_a}\, E_{n_r}$.

Consequently, $E\left(\frac{\sum_{1}^{n} v(a_j)}{n}\right)$ is independent of n. Therefore the second term of the above equation is zero, and it reduces

$$\sigma^2\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right) = E_n\,\sigma^2\left(\frac{\sum_{j=1}^{n} v(a_j)}{n}\right) \tag{55}$$

In order to express the total value of a source-set, let n, the number of non-common A-units reacting to S_t, be a fixed value. Then

$$\sum_{j=1}^{n} v(a_j) = \sum_{j=1}^{n} \sum_{k=1}^{n_{s_r}} x^r\,(a_j,k)$$
$$+ \sum_{j=1}^{n} \sum_{r=3}^{N_R} \sum_{k=1}^{n_{s_r}} x^r\,(a_j,k) \qquad (56)$$

$$\text{where } x^r(a_jk) = \begin{cases} 1 & \text{if the } K^{th} \text{ stimulus associated with response } r \text{ activates } a_j \\ 0 & \text{under all other conditions.} \end{cases}$$

The total value of R_1 source-set is the value gained in the set of units non-common to R_2 over n_{s_1} stimulations plus the value gained by those A-units which are common to other sets and which gain value due to other stimuli associated with other sets. This second term is a summation taken over all possible response units, $r = 3$ to N_R, and all n_{s_r}, the number of stimuli associated to response r.

For the sake of clarity for further calculation, assume that out of n cells $a_1 \ldots a_n$, m_r are in the r source-set. Then, of course, $m_1 = n$ and $m_2 = 0$.

The general term of the summation is independent for different values of r.

The variance of each term for different values of r takes on the form of equation (21).

For $r = 3 \ldots$

$$\sigma^2 \left[\sum_{k=1}^{n_{s_r}} \sum_{j=1}^{m_r} x^r(a_j,k)\right] = \sigma^2 \left[\sum_{k=1}^{n_{s_r}} m_r x^r(a_j,k)\right]$$

$$\sigma^2\left[m_r x^r(a_j,k)\right] \; E\, n_{s_r} + \left[E(m_r x^r(a_j,k)\right]^2 \times \sigma^2\,(n_{s_r}) \qquad (57)$$

$$= m_r P_a(1 - P_a)\, E\, n_{s_r} + (m_r P_a)^2\, \sigma^2\, (n_{s_r}) \tag{58}$$

For $r = 2$, the corresponding quantity is zero.

With $r = 1$, the variance is

$$\sigma^2 \left[\sum_{k=1}^{n_{s_r}} \sum_{j=1}^{n} x^r(a_j, k) \right] = n\, P_a(1 - P_a)\, E\, n_{s_r} + (n\, P_a)^2\, \sigma^2\, (n_{s_r}) \tag{59}$$

The total variance of $\sum_1^n V(a_i)$ is

$$\sigma^r \left[\sum_{j=1}^{n} v(a_j) \right] = P_a(1 - P_a)\, E\, n_{s_r} \sum_{r=1}^{N_R} m_r + (P_a n)^2\, \sigma^2\, (n_{s_r}) \sum_{r=1}^{N_R} m_r^2 \tag{60}$$

The summation $\sum_{r=1}^{N_R} m_r$ represents the total number of R connections originating from n cells, and is equal to nN_{R_a}. In order to compute $\sum_{r=1}^{N_R} m_r^2$, the variance in the intersections of different source-sets is neglected. Other than m_1 and m_2, $m_r = \omega_c' n$, where $\omega_c' = \dfrac{N_{R_a} - 1}{N_R - 2}$. ω_c' is found in the same manner as was ω_c on page 30, except that Me $\{R'\} = N_R - 2$ for ω_c'. Then

$$\sum_{r=1}^{N_R} m_r^2 = n^2 + \sum_{r=2}^{N_R} \omega_c'^2 n^2 = n^2 \left[1 + (N_R - 2)\omega_c'^2\right]$$

$$= n^2 \left[1 + \frac{(N_{R_a} - 1)^2}{N_R - 2} \right]$$

The variance required for equation (53) is then

$$\sigma^2 \left(\frac{\sum_1^n v(a_i)}{n} \right) = \frac{P_a(1 - P_a)}{n} \, E \, n_{s_r} \, N_{R_a} + P_a^2 \sigma^2 (n_{s_r}) \left[\frac{(N_{R_a} - 1)^2}{N_R - 2} + 1 \right] \quad (62)$$

Now the expected value of variance taken with respect to n, yields for the variance of the mean value in the R_1 source-set:

$$\sigma^2 \left(\frac{\sum_1^n v(a_i)}{n} \right) = E_n \left[\sigma^2 \left(\frac{\sum_1^n v(a_i)}{n} \right) \right]$$

$$= \frac{(1 - P_a)}{N_e} \, E \, n_{s_r} \, N_{R_a} + P_a^2 \sigma^2 (n_{s_r}) \left[\frac{(N_{R_a} - 1)^2}{N_R - 2} + 1 \right] \quad (63)$$

Since the above computation was general, the variance of R_2 set is the same.

Thus the total variance of the net random bias under the conditions of random n_{s_r} for the μ-discrimination of the Alpha Perceptron is given by twice the variance of the above, so that

$$\sigma^2 (d_\mu) = \frac{2(1 - P_a)}{N_r} \cdot E \, n_{s_r} \, N_{R_a} + 2 \, P_a^2 \sigma^2 (n_{s_r}) \left[\frac{(N_{R_a} - 1)^2}{(N_R - 2)} + 1 \right] \quad (64)$$

Again assuming a normal distribution for $P_r(\mu)$, the probability of correct response with random n_{s_r} for the Alpha Perceptron is given by the expression:

$$P_r(\mu) = \left[1 - (1 - P_a)^{N_e} \right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2} \, dt$$

where $$Z = \frac{1 - P_a}{\sigma_{d(\mu)}} \tag{66}$$

which reduces to equation (47) when $\sigma(n_{s_r}) = 0$.

Plate VII illustrates quite clearly the effects of permitting a large variation in n_{s_r}.

Several curves are plotted with P_r as a function of n_{s_r} for a system with 100 response units, and 10,000 A-units.

The broken curve represents the same system with no variation in n_{s_r}. A quite definite decrease in accuracy of performance of the system is indicated by allowing n_{s_r} to vary. The best operation results with disjunct sets. It should be kept in mind that the ideal environment condition is imposed in which each stimulus of each class is entirely independent from any other stimuli.

A similar situation exists when the size of the stimuli is allowed to vary. A qualitative examination will be sufficient to demonstrate this point.

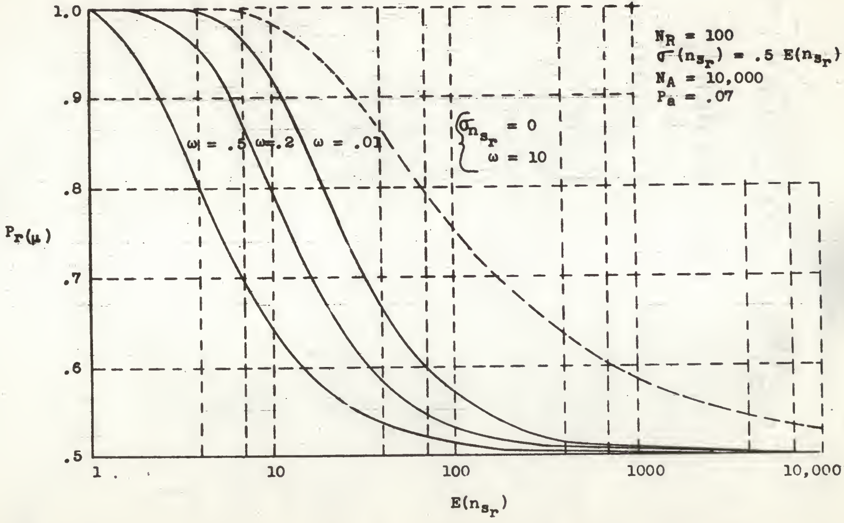
Consider the case where the stimuli of class R_2 were much larger than those stimuli associated to the R_1 source-set. P_a, the expected probability that any A-unit will be activated, will be greater for stimuli of class R_2, and the mean value of the R_2 set will grow faster than the mean value of the R_1 set. Then if the test stimulus S_t which has a disadvantage in measure is shown to the system, its reinforcement will probably not be sufficient to overcome the already favored bias toward R_2 set. Hence incorrect response will result.

From this example one can see that the effect of variation

EXPLANATION OF PLATE VII

$P_r(\mu)$ as a function of $E(n_{s_r})$.

PLATE VII


1.0
.9
.8
.7
.6
.5
$P_r(\mu)$
1
10
100
1000
10,000
$E(n_{s_r})$
$\omega = .5$
$\omega = .2$
$\omega = .01$
$\sigma_{n_{s_r}} = 0$
$\omega = 10$
$N_R = 100$
$\sigma(n_{s_r}) = .5\ E(n_{s_r})$
$N_A = 10,000$
$P_a = .07$

From this example one can see that the effect of variation in size of stimuli between the source-sets is to decrease the accuracy of the system.

The decrease in performance of a system due to stimuli size variation is less than the effect due to n_{s_r} variation, since P_a can be held reasonably constant over a wide range of retinal size variation.

THE GAMMA PERCEPTRONS FOR IDEAL ENVIRONMENT

Sum Discriminating Gamma System

The same logical analysis will be made for the Gamma Perceptron as was made for the Alpha system. Analysis will be carried out for both methods of discrimination.

The Gamma system will hold all sets at equal levels, and also it has the advantage of maintaining the mean value of the entire system constant. In terms of electronic simulation of this system, the above advantage would prevent the saturation of integrators and counters as would be found in the Alpha Perceptron. In a physiological system, this could mean that the cells are required to maintain an optimal range of sensitivity.

"The Gamma system can be thought of, physiologically, as involving a constant chemical or nutrient distribution rate, which is normally just sufficient to balance the expected rate of utilization."[1]

Now we will proceed with the analysis of the Gamma system.

For the sum discrimination, the expected controlled bias $\bar{b}$ is the same as in the Alpha system, since $\bar{b}$ is the value gained by the R_1 source set due to the presentation of S_t. Therefore $\bar{b} = P_a N_e$. n_{s_1} and n_{s_2} are the number of stimuli associated with R_1 and R_2 sets, respectively. If the test stimulus is deleted from the learning series, then $n_{s_1} - 1$ stimuli are associated to R_1 and n_{s_2} stimuli are associated to R_2.

The value gained by the A-units active for one unit of time is $(n_{s_1} - 1) P_a^2 N_e$, but the value lost from inactive A-units of the dominant set is

$$(n_{s_1} - 1) P_a N_e (1 - P_a) E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right)$$

Therefore

$$\bar{V}_1 = (n_{s_1} - 1) P_a N_e \left[P_a - (1 - P_a) E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right) \right] \qquad (66)$$

Similarly, for R_2 source-set

$$\bar{V}_2 = n_{s_2} P_a N_e \left[P_a - (1 - P_a) E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right) \right] \qquad (67)$$

Then the expected net bias due to all stimuli associated with R_1 and R_2 except S_t it

$$\bar{d} = \bar{V}_1 - \bar{V}_2 = (n_{s_1} - 1 - n_{s_2}) \left[P_a - (1 - P_a) E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right) \right] \qquad (68)$$

---

[1] Rosenblatt, Frank. "The Perceptron--A theory of statistical separability in cognitive systems." Cornell Aeronautical Laboratory, Inc. Report No. VG-1196-G-1, January, 1958.

$E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right)$ will now be evaluated.

$$E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right) = E\left(\frac{N_{a_r} P_a}{N_{A_r}P_a - N_{a_r}P_a}\right) = E\left(\frac{N_{a_r} P_a}{N_{a_r} - N_{a_r}P_a}\right) = \frac{P_a}{1 - P_a} \qquad (69)$$

Substituting this quantity into the expression for $\bar{d}$ yields

$$\bar{d} = (n_{s_1} - 1 - n_{s_2}) \left[P_a - (1 - P_a) \frac{P_a}{1 - P_a}\right] = 0 \qquad (70)$$

The expected total net bias is then

$$\bar{B} = \bar{b} + \bar{d} = P_a N_e \qquad (71)$$

Now let a stimulus activate m_1 units $(a_1 \ldots a_{m_1})$ in the R_1 source-set (exclusive of common units), and m_2 units $(b_1 \ldots b_{m_2})$ in the R_2 source-set.

For fixed n_{s_r} stimuli per source-set, then the R_1 source-set component d_1 given m_1 is

$$d_1/m_1 = \sum_{r=1}^{N_R} \sum_{n=1}^{n_{s_r}} \sum_{i=1}^{m_1} x^r(a_i,h) \qquad (72)$$

where

$$x^r(a_i, h) \begin{cases} 1 = \text{when } a_i \text{ of the r source-set is activated by the h th stimulus associated to response r} \\ = -E\left(\frac{N_{a_r}}{N_{A_r} - N_{a_r}}\right) = -\frac{P_a}{1 - P_a} \\ \quad \text{when } a_i \text{ of the r source-set but is not activated by the h th stimulus} \\ 0 = \text{for all other conditions} \end{cases}$$

For fixed n_{s_r}, the variance of d_1 given m_1 is

$$\sigma^2 (d_1/t) = \sum_{r=1}^{N_R} E\, n_{s_r}\, \sigma^2 \left(\sum_{i=1}^{m_1} x^r(a_i, h)\right)$$

$$= E\, n_{s1}\, \sigma^2 \left(\sum_{i=1}^{m_1} x'(a_i,h)\right)$$

$$+ E\, n_{s2}\, \sigma^2 \left(\sum_{i=1}^{m_1} x^2(a_i,h)\right)$$

$$+ \sum_{r=3}^{N_R} \sigma^2 \left(\sum_{i=1}^{m_1} x^r(a_i,h)\right) \qquad (73)$$

By definition of m_1, the second term in the above equation is zero. In this formulation the variation measured by the variance of one source-set due to all stimulus associated to all response units is to be calculated.

In the Gamma system the h th stimulus associated with R_1 activates a certain number of non-common units, m_1, and the increment gained by the active units is 1 and the value lost by an inactive unit in the R_1 set is $P_a/1 - P_a$. Assuming the average taken over all stimuli, the net value of a unit per stimulus is

$$P_a - \frac{P_a}{(1 - P_a)}(1 - P_a) = 0$$

Previously it was shown that the expected value of a source-set was zero. The variance is equal to the second moment in this case, and the calculation of the variance proceeds as follows.

$$\sigma^2 \left(\sum_{i=1}^{m_1} x^r(a_i,\ h)\right) = m_1\, \sigma^2 \left[x^r(a_i,\ h)\right]$$

$$= m_1\, E\left[x^r(a_i,\ h)\right]^2$$

$$= m_1 \left[P_a - (1 - P_a)\left(\frac{P_a}{1 - P_a}\right)^2\right]$$

$$= m_1 \left(\frac{P_a}{1 - P_a}\right) \qquad (74)$$

Consequently the variance of d given m_1 is

$$\sigma^2 (d_1/m_1) = E\, n_{s_1} \frac{P_a}{(1 - P_a)} m_1$$

$$+ (N_{R_a} - 1)\, E\, n_{s_r} \frac{P_a}{1 - P_a} m_1$$

$$= E\, n_{s_r}\, m_1 \frac{P_a}{1 - P_a} N_{R_a} \tag{75}$$

The variance of d_2/m_2 is given by the previous calculation with m_1 replaced by m_2. Then the total variance of both R_1 and R_2 source-set is

$$\sigma^2 (d/m_1, m_2) = \sigma^2 (d_1/m_1) + \sigma^2 (d_2/m_2)$$

$$= E\, n_{s_r} \frac{P_a}{(1 - P_a)} N_{R_a} (m_1 + m_2) \tag{76}$$

On the average, then, the number of units in a source-set activated by any stimulus is equal to $P_a N_e$. The variance of d in the sum system is

$$\sigma^2 (d_\Sigma) = 2\, E\, n_{s_r}\, P_a (1 - P_a)^{-1} N_{R_a} (P_a N_e) \tag{77}$$

From the above equation it may be noted that the variance of n_{s_r} does not enter into the final variance of d in any way. This indicates that in the Gamma Perceptron the restriction that n_{s_r} must be uniform is removed.

The expression for the probability of correct response in terms of the system parameters for the Gamma system with sum discrimination and varying n_{s_r} is

$$P_r(\Sigma) = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

$$\text{where} \quad Z = \frac{P_a N_e}{\sqrt{2 \, E n_r \, P_a^2 \, (1 - P_a)^{-1} \, N_{R_a} N_e}} = \sqrt{\frac{N_e(1 - P_a)}{2 \, N_{R_a} \, E \, n_{s_r}}} \tag{78}$$

Mean Discriminating Gamma System

A quite similar development follows for the mean discrimination system. The expected bias is the same as in the previous case divided by the average number of active units $P_a N_e$. Therefore $\bar{b} = 1$ and $\bar{d} = 0$; hence $\bar{B} = 1$.

As before, $E(d_1/m_1) = 0$, and $E(d_2/m_2) = 0$

The variance of the mean value of d_1/m_1 is

$$\frac{\sigma^2 \, (d_1/m_1)}{m_1^2} = E \, n_{s_r} \, \frac{P_a(1 - P_a)^{-1} \, N_{R_a}}{m_1} \tag{79}$$

The total variance of d may be expressed as

$$\sigma^2 \, d_{(\mu)} = \frac{\sigma^2 \, (d_1/m_1)}{m_1^2} + \frac{\sigma^2 \, (d_2/m_2)}{m_2^2} \tag{80}$$

$$= 2 \, E \, n_{s_r} \, \frac{P_a \, N_{R_a}}{(1 - P_a) \, P_a N_e}$$

$$= 2 \, E \, n_{s_r} \, \frac{N_{R_a}}{N_e(1 - P_a)} \tag{81}$$

The expression for Z may be written

$$Z = \sqrt{\frac{N_e(1 - P_a)}{2 \, E \, n_{s_r} \, N_{R_a}}} \tag{82}$$

which is the same as Z for $P_{r(\Sigma)}$.

Thus under ideal environment conditions for the Gamma system

$$P_{r(\mu)} = P_{r(\Sigma)} \tag{83}$$

Comparison of performance for the Alpha and Gamma systems is illustrated by Plate VIII which shows $P_{r(\mu)}$ versus $E\, n_{n_{s_r}}$. The graph is for ideal environment conditions and for the assumption that the variance of n_{s_r} is equal to half of its expected value. The Gamma system has a definite advantage under these conditions.

ANALYSIS OF THE ALPHA SYSTEMS FOR DIFFERENTIATED ENVIRONMENT

Alpha Perceptron for Sum Discrimination

All analysis up to this point to determine the performance of the various Perceptron systems has been for an experiment under the assumption of ideal environment. It has been assumed that each stimulus belonging to a particular class was chosen to be a random collection of points on the retinal area of the sensory cells. With this random environment there was no correlation among any stimuli within any class. Likewise, the stimuli for the different classes were chosen at random. Then the correlation of stimuli between classes is also zero. The only restriction was that the measure of the stimuli be uniform.

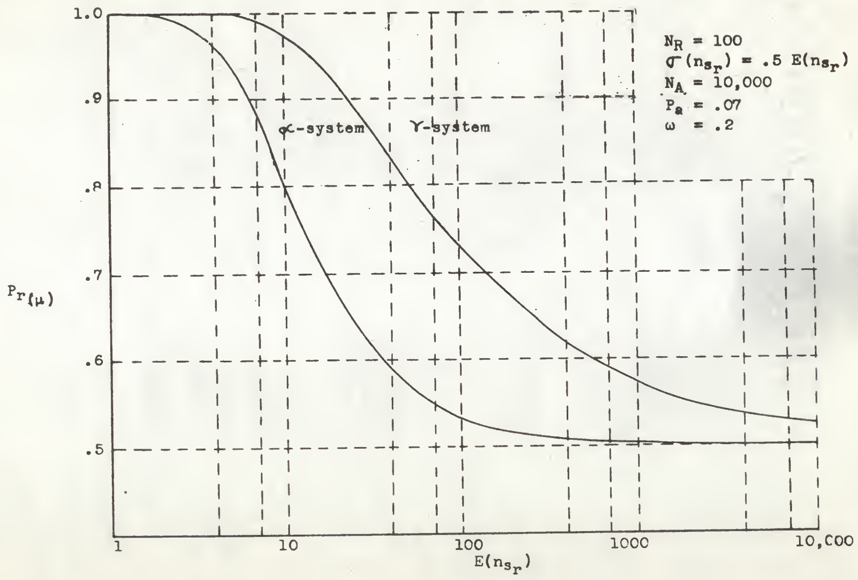
It has been shown that the performance of all Perceptron systems decay to a chance expectance for correct response under random environment conditions. This result, of course, could have been predicted.

However, the previous analysis was for the purpose of comparison of the possible Perceptron systems and to serve as an

EXPLANATION OF PLATE VIII

Comparison of the Alpha and Gamma systems, for variable n_{s_r}.

PLATE VIII


1.0
.9
.8
.7
.6
.5
$P_{r(\mu)}$
1
10
100
1000
10,000
$E(n_{s_r})$
α-system
γ-system
N_R = 100
σ(n_{s_r}) = .5 E(n_{s_r})
N_A = 10,000
P_a = .07
ω = .2

analytical model.

Now it is of importance to determine the reaction of the Perceptron systems to classes of stimuli with some kind of relationship and correlation among the stimuli of a class.

In the remaining analysis of this report the performance of the Perceptron systems will be evaluated under an experiment in which non-random or differential environment conditions exist. Differential environment means that the stimuli of any particular class have some correlation in their characteristics. For example, one class might be circles with different locations within a defined retinal region, and the other class might be a set of squares with various locations within the same specified region. Under these conditions it will be shown that the recognition performance of the Perceptron can be made to approach an asymptotic level different from chance expectancy with increasing number of stimuli.

Before proceeding to the analysis with differentiated environment, several new symbols and concepts need discussion.

In the ideal environment case it was assumed that since there was no correlation between stimuli, the expected portion of overlap of A-units between S_t and stimuli of class 1 or 2 was equal to P_a. In the present case there exists a relationship between stimuli of the same class which will be measured by various forms of P_c. In general, P_c may be defined as the conditional probability that an A-unit activated by one stimulus S_1 will also be activated by another stimulus S_2.

Let $P_{c_{xy}}$ represent the expected value of P_c for two stimuli from classes x and y. $P_{c_{tx}}$ is the expected value of P_c between stimulus S_t and all stimuli belonging to classes other than 1 and 2. The probability P_c represents a mean or expected value as did P_a. P_c for finite number of stimuli exposures will be represented by $P_{c_{t1(s)}}$ and $P_{c_{t2(s)}}$. $P_{c_{t1}}$ is a measure of relationship between the test stimulus S_t and another stimulus of class 1. If $P_{c_{t1(j)}}$ denotes the expected value of $P_{c_{t1}}$ for the j^{th} unit, there is a resulting distribution of $P_{c_{t1(j)}}$ over the A-units.

Stimulus S_t associated with R_1 will activate P_aN_e A-units. These units may be thought of as a particular subset of R_1 source-set. Suppose a stimulus S_1 associated with R_1 is shown to the Perceptron. Then P_{c11} is the expected proportion of these units in the S_tR_1 subset that will be activated. These units will gain an increment of value (by convention $\triangle V = 1$).

P_{c11} may also be interpreted as the expected value on the average that an A-unit of this S_tR_1 subset will gain upon an exposure by S_1. P_{c11} represents the expected probability that an A-unit which is activated by a particular stimulus of class R_1 will also be activated by any other stimulus of class R_1.

In an analogous manner, P_{c12} is the expected probability that an A-unit will respond to a stimulus of class R_2 given that it responds to a particular stimulus of class R_1.

In view of these interpretations of these symbols, then the expected bias $\overline{d}$ due to all stimuli other than S_t may be calculated. Consider n_{s_r} to be equal for both sets. Each stimulus associated with R_1 will add an increment of value to the S_tR_1

subset equal to the number of units in the S_tR_1 subset, P_aN_e, times the average value that each unit gains due to any other stimuli of class R_1.

The increment of value gained by this S_tR_1 set due to a stimulus of class R_2 is equal to $P_aN_eP_{c12}$.

Hence the expected bias $\bar{d}$ due to all stimuli other than S_t may be expressed by

$$\bar{d} = \bar{V}_1 - \bar{V}_2 = P_aN_e\,(n_{s_r} - 1)P_{c_{11}} - P_aN_e\,P_{c_{12}}\,n_{s_r} \qquad (84)$$

All stimuli associated to response units other than R_1 and R_2 contribute an increment of value $P_aN_eP_{c_{1x}}$ to both R_1 and R_2 sets. Hence the net value added to R_1 or R_2 effectively cancels out. Then the above expression for $\bar{d}$ is general and independent of overlapping among source-sets. $\bar{d}$ may be written in the following form.

$$\bar{d}_{\Sigma} = P_aN_e(n_{s_r} - 1)(P_{c_{11}} - P_{c_{12}}) - P_aN_e\,P_{c_{12}} \qquad (85)$$

It is evident that $\bar{d}$ will not be a small fraction of $\bar{b}$ as was the case in the ideal environment, but that $\bar{d}$ will be proportional to n_{s_r}, depending on the difference of $P_{c_{11}}$ and $P_{c_{12}}$.

If $P_{c_{11}} > P_{c_{12}}$, that is, for classes of stimuli sufficiently dissimilar, then correct response will almost always occur, provided n_{s_r} is the same for both R-sets.

Assuming non-uniform n_{s_r}, then for the Alpha system in differentiated environment

$$\bar{d}_{\Sigma} = P_aN_e\,(n_{s_1}\,P_{c_{11}} - n_{s_2}\,P_{c_{12}}) \qquad (86)$$

In the following analysis P_r will be evaluated in terms of the Alpha system parameters for the Alpha system with uniform n_{s_r} for all responses in a differentiated environment.

First consider the sum-discriminating system.

$$\text{Let } x^r(a_j,K) = \begin{cases} 1 & \text{if the } K^{th} \text{ stimulus of the } r^{th} \text{ class, and a member of the } r^{th} \text{ source-set activates the } a_j \text{ cell under the condition that the } a_j \text{ cell is activated by the test stimulus } S_t. \\ 0 & \text{under all other conditions.} \end{cases}$$

Then the bias component d_1 due to the R_1 source-set is

$$d_1 = \sum_{j=1}^{N_e} \sum_{k=1}^{n_{sr}} \sum_{r=1}^{N_R} x^r(a_j,K)$$

$$= \sum_{j=1}^{N_e} \sum_{k=1}^{n_{sr}} \left[x'(a_j,K) + \sum_{r(a_j)} x^r(a_j,K) \right] \qquad (87)$$

where $r(a_j)$ goes through all response units to which a_j is connected, except R_1 and R_2.

For simplicity in the following analytical development, let Y_j represent the value gained by unit a_j from all stimuli (the sum over K). Then

$$Y_j = \sum_{k=1}^{n_{sr}} \left[x'(a_j,K) + \sum_{r(a_j)} x^r(a_j,K) \right] \qquad (88)$$

The variance of d_1 given S_t, a fixed member of class R_1, and, assuming that the values of different A-units are independent, is

$$\sigma^2(d_1/t) = \sum_{j=1}^{N_e} \sigma^2 \sum_{k=1}^{n_{sr}} \left[x'(a_j,K) + \sum_{r(a_j)} x^r(a_j,K) \right]$$

$$= \sum_{j=1}^{N_e} \sigma^2 \, (Y_j) \qquad (89)$$

$Y_j = 0$ with a probability of $1 - P_a$, that is, if the unit a_j is not activated by S_t. $Y_j = 1$ with a probability of P_a. The condition expectation of Y_j given that S_t activates a_j is denoted by $E(Y_j/t)$, the conditional second moment is $E(Y_j^2/t)$, and the conditional variance $\sigma^2(Y_j/t)$.

The first and second moments and total variance are

$$E\,Y_j = P_a\,E(Y_j/t)$$
$$E\,Y_j^2 = P_a\,E(Y_j^2/t)$$
$$\sigma^2\,(Y_j) = E\,Y_j^2 - (E\,Y_j)^2$$

Substituting in terms of the conditional expectations of Y_j and introducing the second and third terms which sum to zero yields:

$$\begin{aligned}\sigma^2\,(Y_j) &= P_a\,E(Y_j^2/t) - P_a(E\,Y_j/t)^2 + P_a(E\,Y_j/t)^2 \\ &\quad - \left[P_a(E\,Y_j/t)\right]^2 \\ &= P_a\,\sigma^2\,(Y_j/t) + P_a(1 - P_a)\,E(Y_j/t) \qquad (90)\end{aligned}$$

In order to simplify notation, let the conditional value of Y_j given t be represented by $\widetilde{Y}_j$, then $E(Y_j/t) = E(\widetilde{Y}_j)$, $E(Y_j^2/t) = E(\widetilde{Y}_j^2)$, etc.

For different exposures the contribution made to the conditional Y_j (given that a_j is activated by S_t) is independent. Then the sum over K is independent of the variance of Y_j, assuming n_{s_r} to be fixed, then the variance of Y_j is

$$\begin{aligned}\sigma^2\,(Y_j/t) &= n_{s_r}\,\sigma^2\left[x'(a_j,1) + \sum_{r(a_j)} x^r(a_j,1)\right] \\ &= n_{s_r}\,\sigma^2\left[x'(a_j,1)\right] + \sigma^2\left[\sum_{r(a_j)} x^r(a_j,1)\right] \qquad (91)\end{aligned}$$

Before actually evaluating $\sigma^2(d_1/t)$, the above variances must be evaluated.

The expected value that unit a_j will gain on the average due to a stimulus of class one given that a_j is activated by S_t, is $P_{ct1(j)}$. The same type of reasoning as was used in deriving $E(P_{ai}^2)$ can be used for the second moment of $P_{ct1(j)}$ which results with $E(P^2_{ct1(j)i}) = P_{ct1(j)}$ (see page 22). Then it follows that

$$\sigma^2\left[x'(a_j,1)\right] = E\left[x'(a_j,1)\right]^2 - \left[E\, x'(a_j,1)\right]^2$$
$$= P_{ct1(j)} - P_{ct1(j)}^2 \qquad (92)$$

and
$$\sigma^2\left[\sum_{r(a_j)} x^r(a_j,1)\right] = \sum_{r(a_j)} \sigma^2\left[x^r(a_j,1)\right]$$
$$= \sum\left[P_{ct1(j)} - P_{ct1(j)}^2\right] \qquad (93)$$

Making the proper substitutions, the variance of d_1 given t is

$$\sigma^2(d_1/t) = \sum_{j=1}^{N_e}\left\{P_a\, n_{sr}\sigma^2\left[x'(a_j,1) + \sum_{r(a_j)} x^r(a_j,1)\right]\right.$$
$$\left. + P_a(1 - P_a)\, n_{sr}^2\left(E\left[x'(a_j,1) + \sum_{r(a_j)} x^r(a_j,1)\right]\right)^2\right\}$$
$$= \sum_{j=1}^{N_e}\left\{P_a\, n_{sr}\left[P_{ct1(j)} - P_{ct1(j)}^2 + \sum_{r(a_j)}\left(P_{ct1(j)} - P_{ct1(j)}^2\right)\right]\right.$$
$$+ P_a(1 - P_a)\, n_{sr}^2\left[P_{ct1(j)} + \left(\sum_{r(a_j)} P_{ctr(j)}\right)^2\right.$$
$$\left.\left. + 2\, P_{ct1(j)}\sum_{r(a_j)} P_{ctr(j)}\right]\right\} \qquad (94)$$

Now assuming that $\dfrac{\sum_{r(a_j)} P_{ctr(j)}^2}{N_{R_a} - 1} = P_{ctx(j)}^2$ where $P_{ctx(j)}$ is the mean value of P_c measured for unit a_j between stimulus S_t and all stimuli of classes other than 1 and 2.

$$\sigma^2(d_1/t) = P_a\, n_{sr}\, N_e\left[P_{ct1} - P_{ct1}^2 - \sigma_j^2(P_{ct1})\right.$$
$$\left. + (N_{R_a} - 1)\left[P_{ctx} - P_{ctx}^2 - \sigma_j^2(P_{ctx})\right]\right]$$

$$+ P_a(1 - P_a)\, n_{s_r}^2\, N_e \left[P_{ct1}^2 + \sigma_j^2(P_{ct1}) + (N_{R_a} - 1)^2\right]$$
$$x\, (P_{ctx}^2 + \sigma_j^2(P_{ctx}) + 2(N_{R_a} - 1)(P_{ct1}\, P_{ctx} + \epsilon) \qquad (95)$$

where the cross product term ϵ is assumed to be negligible. $P_{ct1(j)}$, the probability of the a_j unit responding to both S_t and S_1, will constitute a distribution over j whose variance is $\sigma_j^2\,(P_{ct1})$. Similarly, $\sigma_j^2\, P_{ctx}$ is the variance associated with the distribution of $P_{ctx(j)}$ over the set of A-units. The variance $\sigma_j^2\,(P_{ct1})$ and $\sigma_j^2\,(P_{ctx})$ will be considered as empirical values which are to be measured for any particular case in question, since they are not, as yet, yielded to an analytic approach.

The values of P_c result in a crude approximation to a normal distribution. An estimate of the standard deviation is that it would be equal to half the expected value of the variable. The results of an experiment conducted by Dr. Rosenblatt resulted in showing that the above was a conservative estimate.

Now consider S_t to be any stimulus of the first class. The expected value of d_1 given S_t under these conditions results in the modification of the previous $E(d_1/t)$ by a factor of P_a, the probability that S_t will be activated, in the following manner.

$$E(d_1/t) = n_{s_r} \sum_{j=1}^{N_e} \left[P_a\, P_{ct1(j)} + \sum_{r(a_j)} P_a\, P_{ct1(j)}\right]$$
$$= P_a\, n_{s_r}\, N_e \left[P_{ct1} + (N_{R_a} - 1)\, P_{ctx}\right] \qquad (96)$$

From formula (52), the total variance of d_1 in terms of the conditional variance is

$$
\begin{aligned}
\sigma^2(d_1) &= E_t \sigma^2(d_1/t) + \sigma_t^2 (E\, d_1/t) \\
&= P_a\, n_{s_r}\, N_e \Big[P_{c11} - P_{c11}^2 - \sigma_s^2(P_{c11}) - \sigma_j^2 (P_{c11}) \\
&\quad + (N_{R_a} - 1) \left(P_{c1x} - P_{c1x}^2 - \sigma_s^2(P_{c1x}) - \sigma_j^2(P_{c1x}) \right) \Big] \\
&\quad + P_a(1 - P_a)\, n_{s_r}^2\, N_e \Big[P_{c11}^2 + \sigma_s^2(P_{c11}) \\
&\quad + \sigma_j^2(P_{c11}) + (N_{R_a} - 1)^2 (P_{c1x}^2 + \sigma_s^2(P_{c1x}) \\
&\quad + \sigma_j^2(P_{c1x}) + 2(N_{R_a} - 1)(P_{c11}\, P_{c1x}) \Big] \\
&\quad + P_a^2\, n_{s_r}^2\, N_e^2 \Big[\sigma_s^2 (P_{c11}) + (N_{R_a} - 1)^2 \sigma_s^2(P_{c1x}) \\
&\quad + 2(N_{R_a} - 1)\epsilon \Big]
\end{aligned} \tag{97}
$$

where $\sigma_s^2(P_{c11})$ and $\sigma_s^2(P_{c1x})$ represent the variances of P_{c11} and P_{c1x}, respectively, taken over all test stimuli S_t of the set, and ϵ is the covariance of $P_{c11}P_{c1x}$ which will be assumed to be negligible. The variances with subscript of S may be considered as empirical variables to be measured for the case in question. This variance of P_{ct1} depends on the shape of the stimuli of class one. If the stimuli of this class are all the same and uniformly distributed over the infinite retina of the sensory system, then P_{ct1} will be identical for any stimuli of the class chosen as test stimulus, and its variance is zero. However, if the stimulus of the given class varies widely in shape and its distribution on the retina, the variance of P_{ct1} may be considerable. The variance of the bias component d_2 of the R_2 source-set will be equal to that of the R_1 set given by equation (71) with P_{c11} replaced by P_{c12}.

Now the probability for correct response of the Alpha Perceptron with sum discrimination under differentiated environment can be written as follows:

$$P_r(\Sigma) = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \ell^{-t^2/2}\, dt$$

$$\text{where} \quad Z = \frac{P_a N_e \left[1 - (n_{s_r} - 1)(Pc_{11} - Pc_{12}) - Pc_{12}\right]}{\sqrt{\sigma^2(d_1) + \sigma^2(d_2)}} \qquad (98)$$

Now examine Z of the above expression. The numerator of Z is proportional to n_{s_r}, and the square of the denominator contains two components each of which contains two additive components, one proportional to n_{s_r} and the other proportional to $n_{s_r}{}^2$ for a given Perceptron.

Thus Z takes on the form of $\frac{C_1 + C_2\, n_{s_r}}{\sqrt{C_3\, n_{s_r} + C_4\, n_{s_r}{}^2}}$ which can be written $\frac{\frac{C_1}{n_{s_r}} + C_2}{\sqrt{\frac{C_3}{n_{s_r}} + C^4}}$, where C's are constants.

Consequently Z will approach a limit of $\frac{C_2}{\sqrt{C_4}}$ as the number of stimuli associated to each response, n_{s_r}, increases. The importance of this is that the Perceptron will approach a better than chance limit for probability of correct response with increasing experience.

Alpha Perceptron for Mean Discrimination

Before studying the results of $P_{r(\Sigma)}$, the probability of correct response for mean discrimination will be considered. The expected net bias $\overline{B}$ for the mean system is equal to $\overline{B}$ for the sum system divided by the number of units activated on the average in any source-set by any stimulus, namely, $P_a N_e$. Then

$$\overline{B} = \left[1 - (n_{s_r} - 1)(P_{c_{11}} - P_{c_{12}}) - P_{c_{12}}\right] \tag{99}$$

The variance of d_1 given S_t for this system is

$$\sigma^2\,(d_1/t) = \sigma^2 \left| \frac{\sum_{j=1}^{N_e} Y_j}{N_{a_1(t)}} \right| \tag{100}$$

where $N_{a_r(t)}$ = the number of A-units activated by S_t in the r source-set, and

$$E(d_1/t) = E \left| \frac{\sum_{j=1}^{N_e} Y_j}{N_{a_1(t)}} \right|$$

$$= \sum_{\mathcal{G}} P_a^{K}\,(1 - P_a)^{N_e - K} \sum_{\ell=1}^{K} \frac{E\,\widetilde{Y}_\ell}{K} \tag{101}$$

where $P_a^K(1 - P_a)^{N_e - K}$ = probability that the particular combination occurs in which only K out of N_e units are activated by S_t.

$\sum_{\ell=1}^{K} \frac{E\,\widetilde{Y}_\ell}{K}$ = the average value an A-unit in the K set gains by activation from S_t. For a given partition K, N_e - K, there are $\binom{N_e - 1}{K - 1} = \frac{(N_e - 1)!}{(K - 1)!\,(N_e - K)!}$ = number of different ways that K - 1 active units can be selected from N_e - 1 units.

$\mathcal{G}$ = the class of all possible partitions of the K and N_e into K, N_e - K.

The possibility that K = 0, that is, no unit responds to S_t, has been excluded, so that k can range from 1 to N_e. Then

$$E(d_1/t) = \sum_{K=1}^{N_e} \binom{N_e - 1}{K - 1} P_a^K(1 - P_a)^{N_e - K} \sum_{\ell=1}^{K} \frac{E\,\widetilde{Y}_\ell}{K}$$

$$= \sum_{\ell=1}^{K} E\,\widetilde{Y}_\ell \sum_{K=1}^{N_e} \binom{N_e - 1}{K - 1} \frac{P_a{}^K(1 - P_a)^{N_e-K}}{K}$$

$$= \sum_{\ell=1}^{N_e} \frac{E\,\widetilde{Y}_\ell}{N_e} \sum_{K=1}^{N_e} P_a{}^K(1 - P_a)^{N_e-K} \binom{N_e}{K} \qquad (102)$$

Now consider the second summation. The probability that the particular combination in which $K = 0$ is $(1 - P_a)^{N_e}$. If $K = 0$ were admitted the sum that all combinations would come up is $= 1$. Thus the above sum is

$$\sum_{K=1}^{N_e} P_a{}^K(1 - P_a)^{N_e-K} \binom{N_e}{K} = 1 - (1 - P_a)^{N_e}$$

Since N_e is large, then the sum is approximately $= 1$, in which case

$$E(d_1/t) = \sum_{\ell=1}^{N_e} \frac{E\,\widetilde{Y}_j}{N_e} = \frac{1}{N_e} \sum_{j=1}^{N_e} E\,\widetilde{Y}_j \qquad (103)$$

and the ℓ-index may be replaced by the j-index since the sum reduces to the average value of an average set.

Similarly, the second moment of d_1 given that a_j is activated by S_t will be

$$E(d_1{}^2/t) = E\left[\frac{(\sum_{j=1}^{N_e} Y_j)^2}{N_{a_1(t)}{}^2}\right]$$

$$= E\left[\frac{\sum_{j=1}^{N_e} Y_j{}^2}{N_{a_1(t)}{}^2}\right] + E\left[\frac{\sum_{\substack{i,j=1\\ i\neq j}}^{N_e} Y_i Y_j}{N_{a_1(t)}{}^2}\right] \qquad (104)$$

$$\text{where}\quad E\left[\frac{\sum_{j=1}^{N_e} Y_j^2}{N_{a_1(t)}}\right] = \sum_{\mathcal{G}} P_a^2(1-P_a)^{N_e-K}\,\frac{\sum_{\ell=1}^{K} E\,\widetilde{Y}_\ell^2}{K^2}$$

$$E\left[\frac{\sum_{\substack{i,j=1\\ i\neq j}}^{N_e} Y_i Y_j}{N_{a_1(t)}^2}\right] = \sum_{\mathcal{G}} P_a^K(1-P_a)^{N_e-K} \sum_{\substack{\ell,i=1\\ \ell\neq r}}^{K} \frac{E\,\widetilde{Y}_\ell \widetilde{Y}_r}{K^2}$$

To simplify notation, let

$$P_a^K\,(1-P_a)^{N_e-K} = P'$$

Then the conditional second moment will be

$$E(d_1^2/t) = \sum_{\mathcal{G}} \frac{P'}{K^2}\left(\sum_{\ell=1}^{K} E\,\widetilde{Y}_\ell^2 + \sum_{\substack{\ell,r=1\\ \ell\neq r}}^{K} E\,\widetilde{Y}_\ell \widetilde{Y}_r\right) \tag{105}$$

Introducing a second and third term which sum to zero yields:

$$E(d_1^2/t) = \sum_{\mathcal{G}} \frac{P'}{K^2}\left[\sum_{\ell=1}^{K} E\,\widetilde{Y}_\ell^2 - \sum_{\ell=1}^{K} (E\,\widetilde{Y}_\ell)^2 + \sum_{\ell=1}^{K} (E\,Y_\ell)^2 + \sum_{\substack{\ell,r\\ \ell\neq r}}^{K} E\,\widetilde{Y}_\ell\, E\,\widetilde{Y}_r\right]$$

$$= \sum_{\mathcal{G}} \frac{P'}{K^2}\left[\sum_{\ell=1}^{K}\sigma^2(\widetilde{Y}_\ell) + \sum_{\ell,i=1}^{K} E\,\widetilde{Y}_\ell\, E\,\widetilde{Y}_r\right]$$

$$= \sum_{\mathcal{G}} \frac{P'}{K^2}\left[\sum_{\ell=1}^{K}\sigma^2(\widetilde{Y}_\ell) + \left(\sum_{\ell=1}^{K} E\,\widetilde{Y}_\ell\right)^2\right] \tag{106}$$

Evaluating the sum over $\mathcal{G}$ gives

$$E(d_1^2/t) = \sum_{K=1}^{N_e} \frac{P'}{K^2}\binom{N_e-1}{K-1}\sum_{\ell=1}^{N_e}\sigma^2(\widetilde{Y}_\ell)$$

$$+ \sum_{K=1}^{N_e} \frac{P'}{K^2}\binom{N_e-2}{K-2}\left[\left(\sum_{j=1}^{N_e} E\,\widetilde{Y}_j\right)^2 - \sum_{j=1}^{N_e}(E\,\widetilde{Y}_j)^2\right]$$

$$+ \sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 1}{K - 1} \left[\sum_{j=1}^{N_e} (E\ \tilde{Y}_j)^2 \right] \tag{107}$$

where the above quantity in brackets of the second term represents the terms of the cross products resulting from the second term of the previous expression.

Due to the definition of $E\ \tilde{Y}_j$, a_j must be active and cross product terms specify that $a_j \neq a_i$, so that a_i must be active also. Thus at least two units must be active. Then the number of different combinations is $\binom{N_e - 2}{K - 2}$.

For the squared term $(E\ \tilde{Y}_j)^2$, only a_j must be active so that the corresponding number of different combinations for this term is $\binom{N_e - 1}{K - 1}$.

The conditional second moment of d_1 may be written as follows:

$$E(d_1{}^2/t) = \sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 1}{K - 1} \sum_{\ell=1}^{N_e} \sigma^2\ (\tilde{Y}_\ell)$$

$$+ \left(\sum_{j=1}^{N_e} E\ \tilde{Y}_j\right)^2 \sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 2}{N_e - 2}$$

$$+ \sum_{j=1}^{N_e} (E\ \tilde{Y}_j)^2 \left[\sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 1}{K - 1} - \sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 2}{K - 2} \right] \tag{108}$$

Simplification of the coefficients of the terms proceeds as follows:

$$\sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 1}{K - 1} = \sum_{K=1}^{N_e} \frac{P'}{K^2} \frac{(N_e - 1)!}{(K - 1)!(N_e - K)!}$$

$$\sum_{K=1}^{N_e} \frac{P'(N_e)!}{N_e K(K)!(N_e - K!)} = \frac{1}{N_e} \sum_{K=1}^{N_e} \frac{P'}{K} \binom{N_e}{K} \tag{109}$$

Let $$Q = \sum_{K=1}^{N_e} \frac{P'}{K} \binom{N_e}{K} = \sum_{K=1}^{N_e} \frac{P_a{}^K (1 - P_a)^{N_e - K}}{K} \binom{N_e}{K}$$

Therefore $$\sum_{K=1}^{N_e} \frac{P'}{K^2} \binom{N_e - 1}{K - 1} = \frac{1}{N_e} Q$$

where $$Q \approx \sum_{K=1}^{N_e} P_a (1 - P_a)^{N_e - K} \binom{N_e}{K}$$

$$\approx \frac{1}{P_a N_e}$$

For the coefficient of the second term:

$$\sum_{K=2}^{N_e} \frac{P'}{K^2} \binom{N_e - 2}{K - 2} = \sum_{K=2}^{N_e} \frac{P'}{K^2} \frac{(N_e - 2)!}{(K - 2)!(N_e - K)!}$$

$$= \sum_{K=2}^{N_e} \frac{P'(K - 1)}{K\, N_e(N_e - 1)} \binom{N_e}{K}$$

$$= \frac{1}{N_e(N_e - 1)} \sum_{K=2}^{N_e} \left(1 - \frac{1}{K}\right) P' \binom{N_e}{K}$$

$$= \frac{1}{N_e(N_e - 1)} (1 - Q) \tag{110}$$

A combination of the above two terms yields the coefficient of the third term:

$$\sum_{K=1}^{N_e} \frac{P_a{}^K (1 - P_a)^{N_e - K}}{K^2} \binom{N_e - 1}{K - 1} - \sum_{K=2}^{N_e} \frac{P_a{}^K (1 - P_a)^{N_e - K}}{K^2} \binom{N_e - 2}{K - 2}$$

$$= \frac{1}{N_e} Q - \frac{1}{N_e(N_e - 1)} (1 - Q) = \frac{Q\, N_e - 1}{N_e(N_e - 1)} \tag{111}$$

Substituting these coefficients in the expression (108) for the second moment of d_1/t gives

$$E(d_1^2/t) \approx \sum_{j=1}^{N_e} \sigma^2 (\widetilde{Y}_j) \frac{Q}{N_e} + (\sum_{j=1}^{N_e} E\,\widetilde{Y}_j)^2 \left(\frac{1 - Q}{N_e(N_e - 1)}\right)$$

$$+ \sum_{j=1}^{N_e} (E\,\widetilde{Y}_j)^2 \left[\frac{Q\,N_e - 1}{N_e(N_e - 1)}\right] \qquad (112)$$

Now the total variance of d_1/t may be calculated.

$$\sigma^2 (d_1/t) = E(d_1^2/t) - \left[E(d_1/t)\right]^2$$

$$= \frac{Q}{N_e} \sum_{j=1}^{N_e} \sigma^2 (\widetilde{Y}_j) + \left(\frac{1 - Q}{N_e(N_e - 1)}\right) (\sum_{j=1}^{N_e} E\,\widetilde{Y}_j)^2$$

$$+ \left(\frac{Q\,N_e - 1}{N_e(N_e - 1)}\right) \sum_{j=1}^{N_e} (E\,\widetilde{Y}_j)^2 - \left(\frac{1}{N_e}\right)^2 (\sum_{j=1}^{N_e} E\,\widetilde{Y}_j)^2$$

Combining terms:

$$\sigma^2(d_1/t) = \frac{Q}{N_e} \sum_{j=1}^{N_e} \sigma^2(\widetilde{Y}_j) + \frac{Q\,N_e - 1}{N_e(N_e - 1)} \sum_{j=1}^{N_e} (E\,\widetilde{Y}_j)^2$$

$$+ \frac{1 - Q\,N_e}{N_e^2(N_e - 1)} (\sum_{j=1}^{N_e} E\,\widetilde{Y}_j)^2 \qquad (113)$$

Substituting in the required expressions which were calculated for the sum system results in the following:

$$\sigma^2 (d_1/t) = Q\, n_{s_r} \Big[P_{ct1} - P_{ct1}^2 - \sigma_j^2 (P_{ct1})$$

$$+ (N_{R_a} - 1) \left(P_{ctx} - P_{ctx}^2 - \sigma_j^2 (P_{ctx})\right)\Big]$$

$$+ \frac{N_e\,Q - 1}{N_e - 1} n_{s_r}^2 \Big[P_{ct1}^2 + \sigma_j^2 (P_{ct1})$$

$$+ (N_{R_a} - 1)^2 \left(P_{ctx}^2 + \sigma_j^2 (P_{ctx})\right)$$

$$+ 2(N_{R_a} - 1)\, P_{ct1}\, P_{ctx}\Big]$$

$$- \frac{N_e\,Q - 1}{N - 1} n_{s_r}^2 \left[P_{ct1} + (N_{R_a} - 1)\, P_{ctx}\right]^2$$

$$= Q\, n_{s_r} \Big[P_{ct1} - P_{ct1}^2 - \sigma_j^2 (P_{ct1}) + (N_{R_a} - 1)\big(P_{ctx} - P_{ctx}^2 - \sigma_j^2 (P_{ctx})\big)\Big] + \frac{N_e Q - 1}{N_e - 1} n_{s_r}^2 \Big[\sigma_j^2 (P_{ct1}) + (N_{R_a} - 1)^2 \sigma_j^2 (P_{ctx})\Big] \qquad (114)$$

Letting S_t by any stimulus of the class one the conditional expectation of d_1/t and its variance are respectively:

$$E(d_1/t) = n_{s_r} \Big[P_{ct1} + (N_{R_a} - 1) P_{ctx} \Big]$$

$$\sigma_t^2 (E\, d_1/t) = n_{s_r}^2 \Big[\sigma_s^2 (P_{c11}) + (N_{R_a} - 1)^2 \sigma_s^2 (P_{c1x}) + 2 (N_{R_a} - 1) \epsilon \Big] \qquad (115)$$

where as before ϵ represents the covariance of P_{c11} P_{c1x}, which will be assumed negligible. Making the proper substitutions from the previously derived expressions and assuming the approximation of $Q \approx \frac{1}{P_a N_e}$, the total variance of d_1 is given by the general equation

$$\sigma^2 (d_1) = E_t \sigma^2 (d_1/t) + \sigma_t^2 (E\, d_1/t)$$

$$\approx \frac{n_{s_r}}{P_a N_e} \Big[(P_{c11} - P_{c11}^2 - \sigma_s^2 (P_{c11}) - \sigma_j^2 (P_{c11}) + (N_{R_a} - 1)\big(P_{c1x} - P_{c1x}^2 - \ \ _s^2 (P_{c1x}) - \sigma_j^2 (P_{c1x})\big)\Big] + \left[\frac{1}{P_a(N_e - 1)} - \frac{1}{N_e - 1} \right] n_{s_r}^2 \Big[\sigma_j^2 (P_{c11}) + (N_{R_a} - 1)\, \sigma_j^2 (P_{c1x}) \Big] + n_{s_r}^2 \Big[\sigma_s^2 (P_{c11}) + (N_{R_a} - 1)^2 \sigma_s^2 (P_{c1x}) \Big] \qquad (116)$$

The standard deviation of d for the mean system is equal to

$$\sigma_{d(\mu)} = \sqrt{\sigma^2 (d_1) + \sigma^2 (d_2)}$$

where σ^2 (d_2) is equal to σ^2 (d_1) with $P_{c_{11}}$ replaced with $P_{c_{12}}$.

Then the probability of correct response for the mean system is given by

$$P_{r(\mu)} = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

$$\text{where} \quad Z = \frac{(n_{s_r} - 1)(P_{c_{11}} - P_{c_{12}}) + (1 - P_{c_{12}})}{\sqrt{\sigma^2 (d_1) + \sigma^2 (d_2)}} \tag{117}$$

To further study the capacity and capabilities of the Alpha Perceptron P_d, the probability that two stimuli associated to two different response units during the learning period will be correctly discriminated in the test period.

The equations for P_r gave an analytical indication of the correctness of response for one test stimulus, while P_d indicates the correctness of discrimination of stimuli.

If symbols were redefined, then the correctness of response to S_{t_2}, a test stimulus associated with R_2, could be determined. Assuming the P_r's to be independent, then the probability of correct discrimination, P_d, would be equal to the product of the P_r's for S_{t_1} and S_{t_2}.

With this idea in mind, let some symbols be examined closely. By convention, the S_tR_1 subset is the set of units in the R_1 source-set which are activated by S_t. Then in this subset each unit gains one unit of value due to one exposure of S_t.

The expected net bias $\bar{d}$ was due to all stimuli other than S_t which (when assuming n_{s_r} to be uniform) means that $(n_{s_r} - 1)$

stimuli were associated with R_1 and n_{s_r} stimuli were associated to R_2. Then the resultant mean value per unit due to the unbalanced association toward R_2 is equal to $P_{c_{12}}$. This can be seen since $P_{c_{12}}$ may be interpreted as the value gained on the average by a unit of the S_tR_1 set by a stimulus associated with R_2.

It follows that the net expected reinforcement bias due to S_t is equal to $1 - P_{c_{12}}$.

By slight modification various degrees of relationship can be obtained between the two known stimuli S_{t_1} and S_{t_2}, where S_{t_1} denotes the test stimulus associated with R_1, and S_{t_2} represents the test stimulus associated with R_2.

If the unbalanced reinforcement bias toward R_2 $P_{c_{12}}$ is replaced by $P_{c_{t_1t_2}}$ (the expected value of P_c between S_{t_1} and S_{t_2}), then the resulting equation for correct response will be correct for assuming that S_{t_2} corresponds to one stimuli associated with R_2. Now another equation for correct response will assume that S_{t_1} is the R_1 test stimulus and S_{t_2} the oppositely associated R_2 stimulus. This equation will be denoted by $P_{r(t1)}$. $P_{r(t2)}$ will represent the corresponding equation in which S_{t_2} is the test stimuli of R_2, and S_{t_1} is the oppositely associated stimulus of R_2.

Assuming the P_r's to be independent, the probability that both known stimuli are associated correctly is the product of individual probabilities of correct response which is equal to the probability of correct discrimination, P_d.

$$P_{r(t1)}\, P_{r(t2)} = P_d \tag{118}$$

Thus when S_{t_1} and S_{t_2} have a specified difference measured by

$P_{ct_1t_2}$ and $P_{ct_2t_1}$, P_d represents the correctness of response of both known stimuli during the test period.

Now if $P_{ct_1t_2} = P_{ct_2t_1}$, and both stimuli are the same size, then the equation for $P_{r(t2)}$ is the same as $P_{r(t1)}$ with P_{c11} replaced by P_{c22} in the Z expression. Subject to the restrictions of uniform stimuli size and uniform n_{s_r}, P_d can be written for the sum system as

$$P_{d(\Sigma)} = \frac{\left[1 - (1 - P_a)^{N_e}\right]^2}{2\pi} \int_{-\infty}^{Z_1} \ell^{-t^2/2}\,dt \int_{-\infty}^{Z_2} \ell^{-t^2/2}\,dt$$

$$\text{where}\quad Z_r = \left[\frac{P_a N_e\,(n_{s_r} - 1)(P_{c_{rr}} - P_{c12}) + (1 - P_{ct_1t_2})}{\sigma_{d(\Sigma)}}\right] \qquad (119)$$

and $\sigma_{d(\Sigma)_1}$ = positive square root of equation (97)

$\sigma_{d(\Sigma)_2}$ = positive square root of equation (97) with P_{c11} replaced by P_{c12}.

Similarly, for the mean discriminating system:

$$P_{d(\mu)} = \left[\frac{1 - (1 - P_a)^{N_e}}{\sqrt{2\pi}}\right]^2 \int_{-\infty}^{Z_1} \ell^{-t^2/2}\,dt \int_{-\infty}^{Z_2} \ell^{-t^2/2}\,dt$$

$$\text{where}\quad Z_r = \left[\frac{(n_{s_r} - 1)(P_{c_{rr}} - P_{c12}) + (1 - P_{ct_1t_2})}{\sigma_{d(\mu)_r}}\right] \qquad (120)$$

and $\sigma_{d(\mu)_1}$ = positive square root of equation (116), and

$\sigma_{d(\mu)_2}$ = positive square root of equation (116) with P_{c11} replaced by P_{c22}.

It may be noted that if $P_{ct_1t_2}$ becomes increasingly large $1 - P_{ct_1t_2}$, the expected known reinforcement bias approaches zero. Provided there were no other stimuli of other classes for reinforcement, the expression for Z would approach an asymptote chance response.

The above equations assume that S_{t_1} can be any stimulus of the class 1 set and S_{t_2} may likewise be any stimuli of class 2. That is to say, all the stimuli within a class have a high correlation with each other. With this consideration $P_{c_{11}}$, $P_{c_{12}}$, $P_{c_{22}}$, etc., are used in the numerator for Z rather than the specific values $P_{ct_{11}}$, $P_{ct_{12}}$, $P_{ct_{22}}$, etc. However, when $P_{c_{11}}$ is not typically the expected value $P_{ct_{11}}$, then the more specific values must be used to be representative of the particular situation. For example, class 1 might consist of a set of circles and class 2 a set of ellipses. If a typical circle of class 1 was chosen as S_{t_1} and an ellipse of nearly circular form was chosen to be S_{t_2}, the value of $P_{ct_1t_2}$ is close to unity. If such a selection is picked for computation, $P_{ct_{11}}$, $P_{ct_{1x}}$, etc., should replace $P_{c_{11}}$, $P_{c_{1x}}$, etc., and in addition the conditional variance t would replace $\sigma^2(d)$. These stimuli are no longer typically selected from the class in which they are members.

In classes such as squares versus circles, the maximum value of $P_{c_{12}}$ was .63 in which the centers of the two figures were identical, according to a simulation experiment performed by Dr. Rosenblatt. In this type of class any square or any circle may be taken for discrimination and there is no need to modify equation (120).

Thus far it has been shown that the Alpha Perceptron can perform correct discrimination between known previously reinforced stimulus with increasing number of n_{s_r}.

Consider the effect of the known reinforcement bias $(1 - P_{c_{12}})$ with large n_{s_r}. Unless the term $(P_{c_{11}} - P_{c_{12}})$ is extremely small, the known reinforcement bias will have negligible effect. As n_{s_r} becomes large enough regardless of the size of $(P_{c_{11}} - P_{c_{12}})$, the known reinforcement bias for any particular stimulus during the learning period becomes negligible. This means that the system will respond just as accurately to a test stimulus which has never before been presented or reinforced. This demonstrates that the system approaches a condition for which P_r is better than a chance level, even for a stimulus of zero known bias. Therefore the ability of the Perceptron to form perceptual generalizations has been shown.

Expressions for P_g, the probability of correct generalization, are obtained from the equations for P_r, with zero reinforcement bias of the test stimulus. Then under the conditions of uniform n_{s_r} and fixed stimulus size, expressions for P_g are the following.

$$P_{g(\Sigma)} = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

$$\text{where} \quad Z = \frac{P_a\, N_e\, n_{s_r}\, (P_{c_{11}} - P_{c_{12}})}{\sigma_{d(\Sigma)}} \qquad (121)$$

$$P_{g(\mu)} = \left[1 - (1 - P_a)^{N_e}\right] \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-t^2/2}\, dt$$

$$\text{where} \quad Z = \frac{n_{s_r}\, (P_{c_{11}} - P_{c_{12}})}{\sigma_{d(\mu)}} \qquad (122)$$

and $\sigma_{d(\Sigma)}$ and $\sigma_{d(\mu)}$ are the same as equations (97) and (116) respectively.

Plate IX illustrates the results of the performance equations of P_r and P_g for the mean-discriminating Alpha system. The parameters of the system such as N_e, ω, etc., were modeled after the results of a particular square-circle discrimination simulation experiment. P_g may be interpreted as the probability that any circle or square placed at random within the bounds of the experiment is correctly recognized.

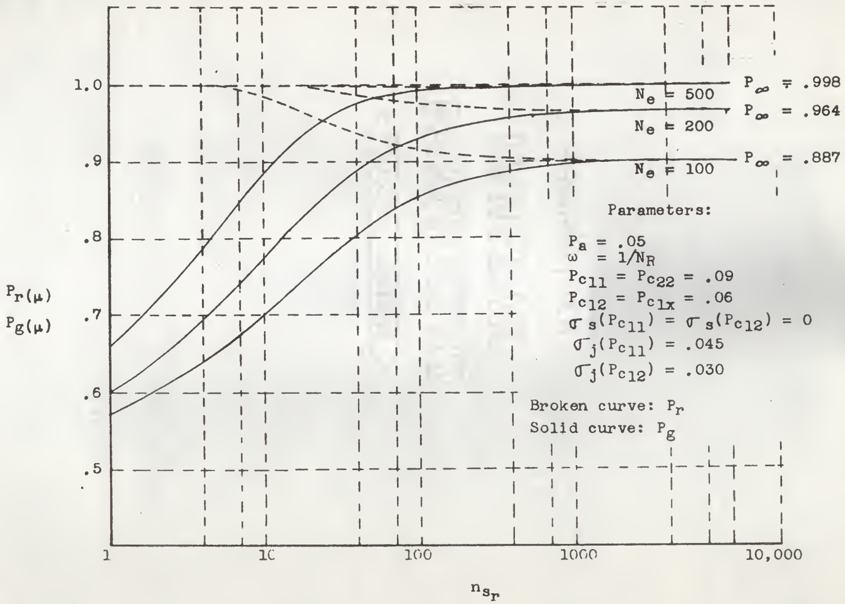
Three pairs of curves are given in Plate IX. One pair of curves (P_r and P_g versus n_{s_r}) is for a system with $N_e = 100$ units. The other two pairs of curves are with $N_e = 200$ and $N_e = 500$. In all cases P_g starts slightly above a 0.5 level for n_{s_r} small and approaches an upper asymptote. For small n_{s_r} the known reinforcement is zero for P_g. However, as the number of stimuli increases, this term has negligible effect so that P_g approaches its upper asymptote. The curves for P_r with n_{s_r} small are nearly unity since the known reinforcement will have little interference from bias due to other associations. As n_{s_r} increases, P_r approaches the P_g asymptote which can be made close to unity by increasing the number of effective A-units. P_r approaching P_g indicates that the specific reinforcement bias becomes increasingly negligible in comparison to the steadily increasing bias due to the difference in P_c's. Both P_r and P_g, in the limit, converge to the same asymptote.

Plate X shows three pairs of curves with the probability of error $1 - P_g$ versus N_A. The solid curves represent $\omega = 0.5$ and the broken curves represent a system with disjunct source-sets. From the curves it is evident that as the number of

EXPLANATION OF PLATE IX

P_r and P_g as function of n_{s_r} parameters based on square-circle discrimination.

PLATE IX

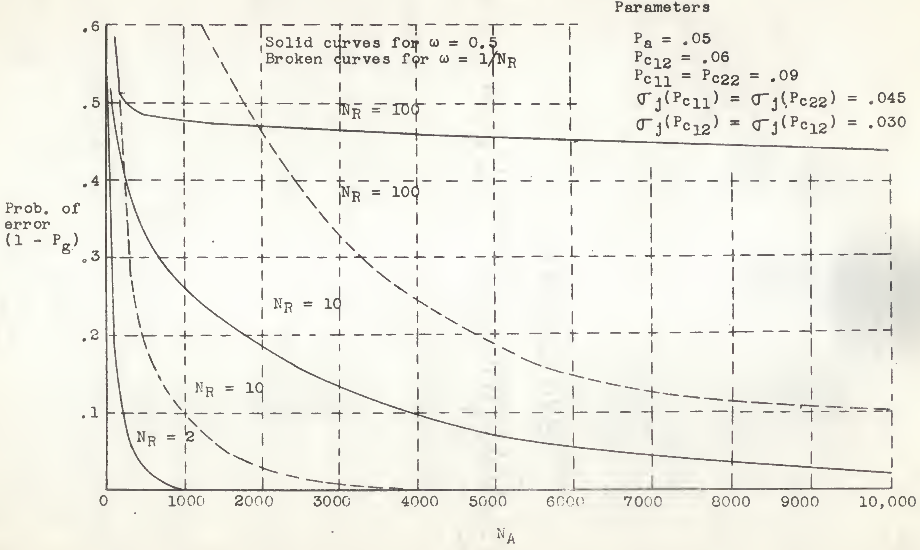


1.0
.9
.8
.7
.6
.5
$P_r(\mu)$
$P_g(\mu)$
$N_e = 500$
$N_e = 200$
$N_e = 100$
$P_\infty = .998$
$P_\infty = .964$
$P_\infty = .887$
Parameters:
$P_a = .05$
$\omega = 1/N_R$
$P_{c11} = P_{c22} = .09$
$P_{c12} = P_{c1x} = .06$
$\sigma_s(P_{c11}) = \sigma_s(P_{c12}) = 0$
$\sigma_j(P_{c11}) = .045$
$\sigma_j(P_{c12}) = .030$
Broken curve: P_r
Solid curve: P_g
1
10
100
1000
10,000
n_{s_r}

EXPLANATION OF PLATE X

1 - $P_g(\mu)$ as function of N_A.

PLATE X



Parameters
$P_a = .05$
$P_{c12} = .06$
$P_{c11} = P_{c22} = .09$
$\sigma_j(P_{c11}) = \sigma_j(P_{c22}) = .045$
$\sigma_j(P_{c12}) = \sigma_j(P_{c12}) = .030$
Solid curves for $\omega = 0.5$
Broken curves for $\omega = 1/N_R$
$N_R = 100$
$N_R = 100$
$N_R = 10$
$N_R = 10$
$N_R = 2$
Prob. of error $(1 - P_g)$
.6
.5
.4
.3
.2
.1
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10,000
N_A

response units increases, the size of the system in terms of the number of A-units increases rapidly to attain a given probability criterion.

The variance for large systems increases with N_R^2 for a fixed ω. For small systems, with disjunct sets, the variance increases with N_R. As also can be seen from the curves overlapping source-sets are desirable for small systems, and disjunct source-sets are desirable for large systems.

CONCLUSION

The first analysis of this report was concerned with the performance of the Alpha and Gamma Perceptrons under ideal environment conditions. Although the major goal of using the random stimulus constraints was to achieve an analytical model for further analysis, several characteristics of the Perceptron resulted. It was found that the Alpha systems learned to respond with better-than-chance accuracy for previously reinforced stimuli. The probability of correct response decreased to a chance level with increasing number of independent stimuli associated with each response unit. Correct response for the Gamma system was independent of the variation in the number of stimuli associated to each response unit.

Mean discrimination was superior for the Alpha Perceptron. For the Gamma Perceptron the probability of correct response was identical for both methods of discrimination. Of course, with ideal environment there was no basis for generalization in recognition of non-previously reinforced stimuli since no relationship among stimuli existed.

With differentiated or non-random environment, the performance of the Alpha Perceptron with both methods of discrimination was investigated. Stimuli within a class were correlated, of which P_c was a measure of stimuli relationship.

Under these conditions the probability of correct response for stimuli of a class approached a better-than-chance asymptote with increasing number of stimuli associated to a response unit. This asymptote approached one for large enough number of A-units.

If the Perceptron was actually to indicate that it could adapt to its environment, then it must be capable of generalization. That is, after sufficient learning, it should be able to recognize stimuli of a class even though they had never been presented before to the system. With stimuli within the classes being correlated, generalization was not only possible but also the probability of correct generalization converged to the same asymptote as P_r. In other words, it could be concluded that after considerable experience the Alpha Perceptron performed just as accurately to the recognition of stimuli which had never before been shown to the system as to stimuli which had been reinforced previously.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Dr. Ugo Gagliardi for many helpful suggestions in the preparation of this thesis.

REFERENCES

1. Allanson, J. T.
   Some properties of a randomly connected neutral network. Third London Symposium on Information Theory. 1955.

2. Ashby, W. Ross.
   Design for a brain. New York: John Wiley & Sons; London: Chapman & Hall, 1952.

3. Cramer, Harald.
   Mathematical methods of statistics. Princeton: Princeton University Press, 1946.

4. Davis, K. H., R. Riddulph, and S. Balashek.
   Automatic recognition of spoken digits. Second London Symposium on Communication Theory. 1952.

5. Feller, William
   Introduction to probability theory. New York: John Wiley & Sons, 1950.

6. Hilgard, E. R.
   Theories of learning. New York: 1948.

7. Hull, C. L., et al.
   Mathematico-deductive theory of rate learning. Yale, 1940.

8. ________________.
   The principles of behavior. New York. 1943.

9. Kendall, M. G.
   The advanced theory of statistics, vol. I, p. 178. London, 1952.

10. Kleene, S. C.
    Representation of events in nerve nets and finite automata. In Shannon and McCarthy, Automata studies. Princeton: Princeton University Press, 1956.

11. Lashley, K. S.
    Brain mechanisms and intelligence. Chicago: University of Chicago Press, 1929.

12. MacKay, D. M.
    In Symposium Information Theory, M.O.S., London, 1950.

13. Minsky, Marvin L.
    Some universal elements for finite automata. In Shannon and McCarthy, Automata studies. Princeton: Princeton University Press, 1956.

14. Mood, Alexander M.
Introduction to the theory of statistics. New York: McGraw-Hill, 1950.

15. Rochester, N., et al.
Tests on a cell assembly theory of the action of the brain, using a large digital computer. I.R.E. Trans. on Information Theory. IT-2, 80-93. 1956.

16. Rosenblatt, Frank.
The Perceptron--A theory of statistical separability in cognitive systems. Cornell Aeronautical Laboratory, Inc. Report No. VG-1196-G-1, January, 1958.

17. ________________
Two theorems of statistical separability in the Perceptron. Cornell Aeronautical Laboratory, Inc. Report No. VG-1196-G-2, Sept., 1958.

18. Selfridge, Oliver G.
Pattern recognition and learning. Third London Symposium on Information Theory. 1955.

19. Shannon, C. E.
A mathematical theory of communication. Bell System Tech. Jrnl. 27(1948) 379, 623.

20. Shannon, C. E. and W. Weaver
The mathematical theory of communication. Urbana: University of Illinois Press, 1949.

21. Taylor, W. K.
Electrical simulation of some nervous system functional activities. Third London Symposium on Information Theory. 1955.

22. Uttley, A. M.
Conditional probability machines and conditioned reflexes. In Shannon and McCarthy, Automata studies. Princeton: Princeton University Press, 1956.

23. Von Neumann, John.
Probabilistic logics and the synthesis of reliable organisms from unreliable components. In Shannon and McCarthy, Automata studies. Princeton: Princeton University Press, 1956.

24. ________________.
The general and logical theory of automata. In cerebral mechanisms in behavior (The Hixon Symposium). New York: John Wiley & Sons, 1951.

25. Wall, P. D., et al.
Factors limiting the maximum impulse transmitting ability of an afferent system of nerve fibres. Third London Symposium on Information Theory. 1956.

26. Woodward, P. M.
Probability and information theory. New York: McGraw-Hill, 1953.

AN EVALUATION OF THE PERCEPTRON THEORY

by

DALE RAYMOND LUMB

B. S., Kansas State University
of Agriculture and Applied Science, 1958

AN ABSTRACT OF
A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE UNIVERSITY
OF AGRICULTURE AND APPLIED SCIENCE

1959

This report was written to investigate and evaluate the statistical analysis of the Perceptron proposed by Dr. Rosenblatt, of Cornell Aeronautical Laboratories. It was desired to determine the feasibility of the self-adaptive cognitive system presented in reference 16.

The evaluation was carried out in coordination with electronic signal recognition research, Project 264 of the Engineering Experiment Station, Kansas State University. Project 264 shares the same basic idea of Dr. Rosenblatt's work on the Perceptron. This is, both projects deal with a system capable of learning the statistical characteristics of the input ensembles. However, the two projects are quite different in their mechanisms necessary to accomplish their goal.

A statistical analysis was employed in order to determine the characteristics and performance properties of several Perceptron models. The expressions representing the accuracy of recognition with various sets of system parameters specified were illustrated by the graphs given at the completion of each analysis.

Under the ideal environment conditions the Perceptron systems investigated in this report were capable of associating a specific number of stimuli to specific response units. However, these associations could not be retained as the number of stimuli presented to the system increased. In other words, under these conditions the Perceptron systems were not capable of self-adapting to the environment of uncorrelated signal ensembles.

With uncorrelated signal ensembles there was no basis for generalization. Mean discrimination resulted in better performance than sum discrimination for the Alpha Perceptron.

The Gamma system proved to be capable of performance independent of the stimuli measure. Correctness of response was the same for both methods of discrimination of the Gamma Perceptron.

In differentiated environment where the stimuli within classes were correlated, self adapting to the environment was possible. In fact, the probability of correct response approached a better-than-chance asymptote with increasing number of stimuli associated to a response unit. This asymptote approached unity for a large enough number of A-units.

The Perceptron was capable of generalization so that self adaptation to its environment was realized. With increasing experience, the probability of correct generalization converged to the same asymptote as P_r.