STOCHASTIC LEARNING MODELS

FREDERICK MOSTELLER[1]
UNIVERSITY OF CHICAGO

1. Introduction

Four kinds of studies of learning might reasonably be discussed under the title of this paper: (1) mathematical research on the theory of neuron networks, (2) the design of self-organizing mechanisms such as robots or computing machines [16], (3) the parts of information and communication theory that fall in the field of statistical behavioristics [29], (4) stochastic learning models for simple psychological experiments. This paper deals with the fourth topic.

There is a small but growing body of literature on statistical models constructed to assist experimental psychologists in the design, analysis, and explanation of some comparatively simple trial-by-trial learning experiments carried out under highly controlled conditions. In these experiments the response is either classified categorically or given as a time measure.

Because these models emphasize both the step-by-step process of learning and its statistical features, problems of time dependence, statistical estimation, and occasionally problems in theoretical probability arise. Thus far, sufficiently little work both of an experimental and theoretical nature has been done on the models and their extensions that there is still considerable unity in the publications. Furthermore the notions involved are quite elementary.

In this brief discussion two general categories of mathematical learning models have been omitted. Thurstone [35] develops learning curves initially from an urn scheme, but turns from this probabilistic model to differential equations. Similarly Gulliksen [20] and Gulliksen and Wolfle [21] and many others before and since work from differential equations, rather than from the kind of trial-by-trial models that are principally discussed below. On the other hand, Hull's extensive work (for one example see [26]) has been omitted, though it is sometimes related to the models presented here, because his postulational system would require a review of its own.

Finally, Savage's theory of personal probability [32] can, as he points out (p. 44), be regarded as a device for giving expression to the phenomenon of learning by experience. He also notes that logic itself "can be interpreted as a crude but sometimes handy empirical psychological theory" (p. 20). Such theories are omitted on the grounds that they are oriented more normatively than empirically and would not be likely to describe well the kind of behavior emitted in the experiments discussed here.

2. Beginning notions

For simplicity a situation with two response classes is discussed first, together with the form of the operator used to change the response probabilities. In section 3, a generalization to more than two response classes is introduced. Discussion there helps explain the choice of the form of the operators used.

[1] On leave from Harvard University 1954–55.

In a particular experiment such as a T-maze, an organism may have only two possible responses—turning right or turning left. These responses are mutually exclusive and exhaustive (refusing to make a choice is disallowed). In general, let such a pair of alternatives be called A_1 and A_2. At any particular trial it is assumed that the probability of choosing alternative A_1 is p, while that of choosing A_2 is $q(= 1 - p)$.

After an alternative is chosen there is some outcome such as reward, nonreward, or punishment (examples: food, nothing, electric shock). The alternative chosen together with the outcome defines an event. Different events have associated with them operators which change the probabilities of the responses A_1 and A_2. It is not essential that the events be regarded as determined by alternatives and outcomes, but in the experimental situations considered thus far, it has been convenient to so regard them (sometimes one or the other is suppressed). Learning is achieved by changes in the probabilities of choosing alternatives. In the two-response situation it is adequate to describe the effect of an operator on the probability of A_1, because the sum of the probabilities of the alternatives must sum to unity. Two general forms for writing the operators are worth mentioning, the first for intuitive reasons, the second for computational convenience and everyday use.

The first, called the gain-loss form, is written

$$(1) \qquad Qp = p + a(1 - p) - bp, \qquad 0 \leqq a, b \leqq 1,$$

and the second, called the fixed-point form, is

$$(2) \qquad Qp = \alpha p + (1 - \alpha)\lambda, \qquad 0 \leqq \alpha \leqq 1, \qquad 0 \leqq \lambda \leqq 1.$$

The conditions on the parameters guarantee that Qp, the new probability of alternative A_1, has a value between 0 and 1. For reasons discussed below in the paragraph containing equation (8), negative values of α have been ruled out; consistency between the conditions of equations (1) and (2) could be achieved by adding the condition $a + b \leqq 1$ to equation (1).

The gain-loss form shows that the new probability is made up of the old probability plus a quantity proportional to its maximum possible gain and less a quantity proportional to its maximum possible loss. The fixed-point form shows the new probability as a weighted sum of the old probability and a value λ. The quantity λ is a fixed point of the operator because if p has the value λ, operating with Q gives $Q\lambda = \lambda$. In addition, λ is the limiting value obtained when the operator Q is applied repeatedly. This is easily shown because

$$(3) \qquad Q^n p = \alpha^n p + (1 - \alpha^n)\lambda.$$

If $|\alpha| < 1$, then as n tends to infinity the right-hand side of (3) tends to λ. In the special instance where α has the value unity, we have the identity operator because $Qp = p$. The quantity α measures the *ineffectiveness* of the operator. If $\alpha = 1$, the operator does not change the probability; if $\alpha = 0$, the probability leaps immediately to the limit point λ. Negative α's are not considered for reasons mentioned below.

Two operators do not ordinarily commute. It can be readily shown that

$$(4) \qquad (Q_1Q_2 - Q_2Q_1)p = (1 - \alpha_1)(1 - \alpha_2)(\lambda_1 - \lambda_2),$$

where the subscripts on the operators correspond to those on the parameters. Thus a pair of operators commute if and only if at least one is the identity operator, or if the operators have equal limit points. Commutativity does occur occasionally in experimental work. In Solomon and Wynne's escape-avoidance experiment, dogs jump over a barrier after a signal and before an electric shock (avoidance, A_1) or after both signal and shock have occurred (escape, A_2). In the analysis [6] of data from this experiment, it appeared that there were two operators, one associated with each alternative, and that both operators had limit points unity. The other condition for commutativity occurred in a rote learning experiment analyzed by Miller and McGill [30]. On each trial the same list of words was read (order scrambled between trials), the subject recorded all the words he could remember. Remembering a word improved its probability of recall after next reading, but forgetting a word did not seem to change its probability of being recalled on the following trial. Thus the operator associated with forgetting was—within measurement error—the identity operator. Commutativity is particularly convenient because it implies that the probability at the end of a sequence of trials depends only on the number of times each operator has been applied and not upon the particular order in which they were applied.

Implicit in the discussion up to this point is the independence-of-path assumption. It has been assumed that the probability of a response on a particular trial depends on the value of p that has been achieved by successive application of the operators and not upon the detailed information about alternatives and outcomes that occurred on each trial. What "memory" the model has is contained in the value of p. Such an assumption is extremely drastic, and it is doubtful if it can long endure. By generalizing the notion of an operator to include more than the alternative and outcome of a single trial, it should be possible to widen the scope of the model. This has not been done thus far. The path-independence assumption makes the process under discussion Markovian, but the states of the system are not the alternatives, outcomes, or events, but the values of p assumed at a particular instant. With two operators, at the close of n trials there may be as many as 2^n different states available for the organism.

3. The general model

In general there are r mutually exclusive response classes or alternatives $A_1, A_2, \cdots, A_r$. The state of the organism at a given trial n is given by the vector $p(n) = (p_1, p_2, \cdots, p_r)$, $p_i \geqq 0$, $\Sigma p_i = 1$. An event E occurs which has associated with it the operator T. The new probability vector after the operator is applied is

$$p(n+1) = Tp(n) . \tag{5}$$

What should be the form of the operator? It has been reasoned as follows [8]. The identification between real classes of behavior and alternatives in the model is, to a certain extent, arbitrary. It would be desirable to have the description of the behavior by the model invariant under certain kinds of changes in classification. For example, in a bar-pressing experiment with a rat as subject, the rat may press the bar with his left paw, right paw, or with both paws. Yet an experimenter ordinarily does not distinguish among these classes of responses, rather he combines them into one response category for the record, and "treats" these responses in the same way, that is, all three classes have the same consequences for the rat. It would be desirable therefore to be able to combine such classes on any trial without altering the behavioral predictions. The form of the operator that makes this possible results from the *combining of classes condition.* Let us now be more specific.

We are clearly concerned only with *stochastic* operators, operators that send probability vectors into probability vectors, that is, in equation (5) both $p(n)$ and $p(n+1)$ are probability vectors. A ***projection*** of a probability vector $p = (p_1, p_2, \cdots, p_r)$ into a probability vector p' is obtained by choosing a nonempty subset σ of the elements of the vector and replacing all but one of these elements by zeros, and replacing the remaining one by the sum of the elements in the subset, while keeping unchanged the elements of the vector not in σ. [Example: (0.4, 0, 0, 0, 0.6) is one projection of (0.1, 0.1, 0.1, 0.1, 0.6), where the subset chosen for projection consists of the first four elements.] Call an operator that performs such a projection C. Then a stochastic operator T is said to satisfy the combining of classes condition if and only if

$$CTCp = CTp \tag{6}$$

for all projection operators C and all probability vectors p. Then we have the following [8]

THEOREM. *A stochastic operator T (for $r \geqq 3$) satisfies the combining of classes condition if and only if it has the form*

$$Tp = \alpha p + (1 - \alpha)\lambda , \tag{7}$$

where p and λ are probability vectors and α is a suitably chosen scalar.

Thus the innocuous sounding request that the model be such that an experimenter be able to combine response classes (which he treats in the same manner) either before or after he runs the experiment and still get the same result restricts the form of the operators tremendously, and this form is like that previously displayed in equation (2) except that p and λ are now vectors.

The spirit of the combining of classes condition is that the number of response classes is limited only by the imagination of the investigator. A lower bound on α can be obtained from an extension of the combining classes condition. The bound is

$$\frac{-1}{r-1} \leqq \alpha . \tag{8}$$

If we let r become arbitrarily large, then negative α's become inadmissible. This extension was suggested in conversation by L. J. Savage (see [7, section 1.8]). We have seen no experiments where negative values of α seemed appropriate. Negative α's lead to probabilities that oscillate above and below the limit point when the same operator is applied repeatedly.

It should be noted that the combining of classes condition is automatically satisfied for $r = 2$, and therefore that the theorem cannot be proved for that number of classes. But if we regard $r = 2$ as having arisen from the projection of some higher dimension in which the condition holds, the form of the operator required is that given in equation (2).

Estes [11] deduces equation (2) from set-theoretic considerations he regards as appropriate to conditioning experiments (stimulus sampling). His reasoning is essentially as follows. Let there be two response classes A_1 and A_2 and a stimulus set S formed of two disjoint subclasses C and $\sim C$. (Estes speaks of the numbers of elements in the various sets, we will speak of the measures of the elements, a minor change.) All elements in C are "conditioned to" A_1, those in $\sim C$ are conditioned to A_2. An organism perceives a subset of stimuli X such that the ratio of the measure of the elements in $X \cap C$ to the measure of C is the same as the ratio of the measure of C to the measure of S (homogeneity assumption). The value of this ratio is the probability of response A_1. After an event has occurred X breaks up into two disjoint subsets Y and Z. Whatever their previous state, all elements of Y are now "conditioned to" A_1 and all elements of Z to A_2. Assuming the measures of the elements are invariant, equation (1) can now be obtained after some algebra by letting $a = m(Y)/m(S)$ and $b = m(Z)/m(S)$, where $m(\ )$ stands for the measure of the set named in the parentheses.

By extending Estes' stimulus sampling analyses in the natural way to r classes and maintaining the corresponding kind of homogeneity assumption, it is possible to derive the operator obtained from the combining of classes condition [7, section 2.5]. Estes and Burke [14] also give attention to the problem of stimulus variability, that is, the measure of a stimulus element is its probability of occurring on a given trial. Different elements have independent probabilities of occurring on each trial.

The combining of classes condition has not had serious experimental testing. Most experiments analyzed thus far by the general model have employed two response classes. Some exceptions are the work of Neimark [31] who used three response classes, and the model behaved rather satisfactorily; and Flood [18] who used nine classes. Though Flood's analysis was not carried out in detail for the general model, it seems unlikely that the particular specialized models used would be satisfactory. It may well be that more strict experimental conditions are required as the number of response classes increases, if indeed the model holds. The obvious alternative to the combining of classes condition is that T should be a general $r \times r$ stochastic matrix. This means that each operator used in an experiment is required to have $r(r - 1)$ parameters estimated for it unless some method of reducing their number is employed. It must be recalled that the basic data are a sequence of responses (together with their outcomes) from which must be gleaned the estimates of these many parameters, scarcely a tasty task. The combining of classes condition reduces the number of parameters to r for each operator. But such remarks are not compelling, further tests are clearly in order.

Flood [17] has tried out models closely related to the ones presented here to see how they perform in learning to play the von Neumann-Morgenstern types of games. By assigning parameter values and applying the appropriate operators after every trial, one can by use of a random number table compute the behavior of what might be called a stat-rat, a mathematical organism that learns in exact accordance with the rules of the particular model. When the model is used for such game-learning it becomes a self-organizing system. Flood used parameter estimates based on experiments with real rats, so with that slender tie to real learning, and because of its intrinsic interest to statisticians, we quote from his summary (pp. 156–157):

"The experimental results consist of Monte Carlo computations for the stat-rat, contests between stat-rat and a human subject, and comparisons of performance of stat-rat and a human subject when playing the same static game. Very limited data indicate that

(a) The stat-rat usually learns a good strategy when a constant mixed-strategy is played against him. In Morra [a 9 × 9 symmetric game] and the other games played the stat-rat seemed to settle on essentially the best strategy within 200 trials or so.

(b) A person proficient at games would win against the stat-rat in Morra.

(c) The stat-rat does reasonably well in a static game, in comparison with the human subject, but a statistician would certainly defeat the stat-rat." He goes on to say that it seems unlikely that a Markov process will be adequate to describe human learning. We do not disagree. As we have said, the models are developed for quite simple situations—for example, Estes and Burke [14, p. 279] say of one of their models "for experimental verification of the present theory we shall look to experiments involving responses no more complex than flexing a limb, depressing a bar, or moving a key."

4. Experimental models and estimation

In this section some models for specific experiments are described, together with some remarks about estimation, and occasionally the results of experiments.

Bush and Mosteller [7] classify models ordinarily used for choice experiments into three categories: experimenter controlled, subject controlled, and experimenter-subject controlled.

Experimenter-controlled models can well be illustrated by a prediction experiment like that of Humphreys [27]. On a proportion π of the trials a lamp lights, but on $1 - \pi$ of the trials it does not. After a signal that the trial is to begin, the subject predicts whether the lamp will or will not light on the ensuing trial. There are thus two outcomes O_1 and O_2 corresponding to lamp "on" or "off." Let p_n be the probability of the prediction "on" (A_1) for the nth trial. In this problem two operators have been used:

Event	Operator	Probability of Application
O_1	$p_{n+1} = Q_1 p_n = \alpha_1 p_n + (1 - \alpha_1)\lambda_1$	π_1
O_2	$p_{n+1} = Q_2 p_n = \alpha_2 p_n + (1 - \alpha_2)\lambda_2$	$\pi_2 (= 1 - \pi_1)$

For the experiment described here the limit points are ordinarily chosen as $\lambda_1 = 1$, $\lambda_2 = 0$, the intuitive argument being that if O_1 always occurred, the subject would ultimately be practically certain to predict A_1, and if O_2 always appeared the subject would be practically certain to predict A_2 after many trials.

If there are many subjects, each starting with the initial probability p for an A_1 response, then at the nth trial there is a distribution of values of p. Let this distribution have mean $V_{1,n}$. Then it can be shown [7, section 4.3] that the mean is given by

$$V_{1,n} = V_{1,\infty} - (V_{1,\infty} - p)\bar{\alpha}^n , \tag{9}$$

where

$$a_i = (1 - \alpha_i)\lambda_i , \qquad i = 1, 2 ,$$

$$\bar{a} = \pi_1 a_1 + \pi_2 a_2 ,$$

$$\bar{\alpha} = \pi_1 \alpha_1 + \pi_2 \alpha_2 ,$$

$$V_{1,\infty} = \frac{\bar{a}}{1 - \bar{\alpha}} .$$

Additional moments of the distribution can be deduced.

In the special light experiment described above, the symmetry between "on" and "off" suggests $\alpha_1 = \alpha_2$. This assumption together with the values of the limits mentioned earlier yields $V_{1,\infty} = \pi_1$. Thus with these special assumptions the model states that asymptotically the subject will predict "on" for the lamp the same proportion of times it comes on. Human subjects do this! The experiment has been performed often, with different stimuli, different instructions, and in different laboratories. Statisticians and mathematicians notice that the subject is not correct as often as he could be if he would choose the prediction corresponding to max (π_1, π_2) and stick to it. (This experiment has not, so far as the author is aware, been performed with lower animals as subjects.) References and discussion bearing on this experiment can be found in [7, chapter 13], [8], Estes and Straughan [15], Estes [13], Flood [18]. The experiment has been extended to three predictions by Neimark [31]. Further, Burke, Estes, and Hellyer [3] used an ingenious variation of this experiment to study, in association with a variation of the model, the effect of stimulus variability upon rate of verbal conditioning. The stimulus sampling approach was made operational in such a manner that from the results of two experiments, the quantitative results of a third could be forecast.

This type of model is called experimenter-controlled (or noncontingent) because the application of the operators is entirely in the hands of the experimenter.

In a subject-controlled experiment, the application of the operators is determined by the response given by the subject. The Solomon-Wynne experiment [33] briefly described earlier exemplifies this (p_n is the probability of avoidance on trial n):

Event	Operator	Probability of Application
A_1 (avoidance before shock)	$p_{n+1} = Q_1 p_n = \alpha_1 p_n + (1 - \alpha_1)\lambda_1$	p_n
A_2 (escape after shock)	$p_{n+1} = Q_2 p_n = \alpha_2 p_n + (1 - \alpha_2)\lambda_2$	$1 - p_n$

In this experiment $\lambda_1 = \lambda_2 = 1$, and $p_0 = 0$. The problem is to estimate the values of the learning rate parameters α_1 and α_2. Data from this experiment form a matrix of 1's and 0's corresponding to avoidance and escape respectively for 30 dogs for some 20 trials. As a first approximation, it can be assumed that all dogs have identical parameters (the experimenters take pains that the shock is adjusted to the dog, and that the barrier to be jumped is adjusted to shoulder height for each dog). In principle, the likelihood of the entire matrix can be written down and maximized for the two parameters under discussion. The maximization would be a matter of exploring the α_1, α_2 unit square, not a small job. The procedure actually used in [6], [7] is to obtain an estimate for α_2, based on a maximum likelihood procedure using only trials to the first avoidance (this yielded a value 0.92). Then this estimate was regarded as a fixed known value in the further estimation of α_1. The estimation of α_1 can be sketched briefly. The probability of escape after one previous avoidance is

$$q_{n,1} = \alpha_1 \alpha_2^{n-1} . \tag{10}$$

This can be estimated from $1 - x_{n,1}/N_{n,1}$, where $N_{n,1}$ is the number of dogs on trial n that have just one previous avoidance, and $x_{n,1}$ is the number of these dogs that avoided on trial n. Since α_2 is regarded as known, α_1 can now be estimated. Estimates like this can be obtained for each number n, and for each number k of previous avoidances. These estimates may then be pooled (the value obtained was $\alpha_1 = 0.80$). It is an open question what the optimum method of estimating parameters is in all these applications. Furthermore, it would be valuable to know among those methods leading to estimates at reasonable computing costs which are to be preferred. To compare the behavior of the model with that of the real dogs the experiment was replicated in a Monte Carlo manner with 30 stat-dogs each of whom had the parameters mentioned above. A few of the comparisons given between the stat-dogs and real dogs are shown in table I. The observed standard deviations of the statistics are more for the purpose of direct comparison than for significance testing, because the model is designed to reproduce variability as well as averages. Only half of the comparisons given in [7] are shown here, the author arbitrarily chose every other one.

One result of the model is to show that one avoidance trial is worth about the same as 2.7 escape trials in learning to avoid [because $(0.92)^{2.7} \cong 0.80$].

The Miller-McGill rote-learning study [30] is another illustration of subject-controlled events. The data again form a matrix of 1's and 0's (corresponding to recall and nonrecall), rows corresponding to words, columns corresponding to trials. The operators take the form

TABLE I

COMPARISONS OF THE STAT-DOG DATA AND THE REAL DOG DATA OF THE SOLOMON-WYNNE EXPERIMENT, 30 DOGS IN EACH SET

STATISTIC	STAT-DOGS		REAL DOGS	
	Mean	Standard Deviation	Mean	Standard Deviation
Trials before first avoidance	4.13	2.08	4.50	2.25
Total shocks	7.60	2.27	7.80	2.52
Alternations	5.87	2.11	5.47	2.72
Trials before first run of four or more avoidances	9.47	3.48	9.70	4.14

$$\begin{aligned} Q_1 p &= \alpha_1 p + (1 - \alpha_1)\lambda_1 && \text{(recall)} \\ Q_2 p &= p && \text{(nonrecall)} \end{aligned} \tag{11}$$

where preliminary investigation indicates that approximately an identity operator is associated with nonrecall. Here the problem is to estimate simultaneously p_0, the initial probability of recall, α_1, and λ_1. One proposal for such estimation [7, section 10.7] is to let x_ν be the number of words recalled at least $\nu + 1$ times, and let N_ν be the number of word-trials required for those x_ν words to be recalled $\nu + 1$ times after each has been recalled ν times. Formally this is equivalent to inverse binomial sampling, where sampling from a binomial population continues until c successes are observed [here x_ν plays the role of c, and $p_\nu = \alpha_1^\nu p_0 + (1 - \alpha_1^\nu)\lambda_1$ plays the role of the fixed binomial probability]. As an estimate of p_ν we could use

$$\hat{p}_\nu = \frac{x_\nu - 1}{N_\nu - 1}, \tag{12}$$

which has a variance roughly approximated by

$$\sigma^2(\hat{p}_\nu) \cong \frac{p_\nu^2(1 - p_\nu)}{x_\nu}. \tag{13}$$

This suggests the possibility of minimizing a χ^2-like quantity

$$S = \sum_{\nu=0}^{\Omega} \frac{(p_\nu - \hat{p}_\nu)^2}{\sigma^2(\hat{p}_\nu)}, \tag{14}$$

where Ω is the largest number of recalls used in the estimation procedure. It is possible to write the three minimizing equations and solve them approximately by iterative methods in a couple of days using a hand calculator. It may be of interest that in an experiment with 32 words (scrambled between presentations), it was estimated that $p_0 = 0.23$, $\alpha_1 = 0.85$, $\lambda_1 = 0.98$ (not far from the intuitive estimate of unity). Roughly speaking the subject recalls about a quarter of the words on the first trial and on each successive trial adds to that number 15 per cent of the remaining words. For those who have learned lists presented in the same serial order on each trial it should be mentioned that it is rather more difficult to learn a scrambled list unless some specific preparations are taken.

Turning now to experimenter-subject controlled events we take the example of the T-maze, or the two-armed bandit. The description sounds similar to that of the Humphreys experiment, but it is not identical. In the two-armed bandit experiment, the subject is seated before an apparatus with two buttons, a left-hand button and a right-hand button. After a light signals that the bandit is activated the subject may press one or the other button. When the right-hand button is pressed a poker chip is delivered on π_1 of the trials. When the left-hand button is pressed a poker chip is delivered on π_2 of the trials. The pay-off proportions need not add to unity; for example, in one experiment $\pi_1 = 1$, $\pi_2 = 0.50$, in another $\pi_1 = 0.50$, $\pi_2 = 0$, and in still another $\pi_1 = 0.75$, $\pi_2 = 0.25$. The sequence of decisions actually made forms the information for study. Four operators seem to be appropriate to this experiment corresponding to four events formed from the two responses (right or left) and the two outcomes (reward or nonreward). The symmetry of the two sides suggests that reward following a right-hand choice should yield the same improvement in the probability of choosing the right-hand button that reward following a left-hand choice gives in improving the probability of choosing the left-hand button (that is, the operators should have the same value of α). Similar remarks apply for nonreward.

Four models have been suggested thus far for this type of experiment. These models differ only in the restrictions placed on the parameter values. It is convenient to give each in the form of a 2 × 2 table whose entries are the result of a single application of the operator to p, the probability of choosing the right-hand side (for convenience in discussion we assume $\pi_1 > \pi_2$, so that the right-hand side is the favorable side).

REINFORCEMENT-EXTINCTION MODEL

	Left	Right
Reinforcement	$\alpha_1 p$	$\alpha_1 p + 1 - \alpha_1$
Nonreinforcement	$\alpha_2 p + 1 - \alpha_2$	$\alpha_2 p$

Reinforcement increases the probability of choosing the rewarded side toward unity, while nonreinforcement reduces the probability of choosing that side toward zero.

In psychological work the phenomenon of secondary reinforcement often affects results. A verbal description runs that the presence of stimuli associated with reinforcement can itself be reinforcing. Such reasoning would suggest that choosing a side that had previously been reinforced might improve the probability that that side is chosen, though presumably the improvement would not be as great as when primary reinforcement is present. This reasoning suggests the alternative secondary reinforcement model.

SECONDARY REINFORCEMENT MODEL

	Left	Right
Reinforcement	$\alpha_1 p$	$\alpha_1 p + 1 - \alpha_1$
Nonreinforcement	$\alpha_2 p$	$\alpha_2 p + 1 - \alpha_2$

The third model is obtained from the reinforcement-extinction model by equating the α's, it might be regarded as an information type of model because reinforcement and nonreinforcement have equivalent though opposite effects. In the literature it has been called the equal-alpha model rather than the information model.

The fourth model assumes that nonreinforcement does not change the probabilities. This can be achieved by setting $\alpha_2 = 1$ in either of the first two models, but the titles of the tables then lose their meaning. This has been called the identity-operator model.

If the reinforcement-extinction model were appropriate, $\alpha_2 \neq 1$, $\pi_1 \neq 1$, then an organism does not ultimately learn to go to the favorable side with probability unity. Even if he were certain to go to the right, on some trial he would not be rewarded, this would reduce his probability of going to the right, and he has a finite probability of going to the left on a later trial. If α_2 is nearly unity, however, this effect can be small. If we think of a large number of identical organisms undergoing this experiment, then after a large number of trials we anticipate a distribution of values of p for these organisms which does not degenerate to a point distribution. In a reasonable sense the probabilities 0 and 1 of going to the right-hand side are reflecting barriers. When this model is specialized to the equal-alpha model a more definite prediction can be computed [13], [8], namely that the asymptotic mean of the distribution of values of p is

$$V_{1,\infty} = \frac{1-\pi_2}{2-\pi_1-\pi_2}. \tag{15}$$

For example, if $\pi_1 = 0.50$, $\pi_2 = 0$, $V_{1,\infty} = \frac{2}{3}$. Generally speaking this result has not been observed in T-maze or two-armed bandit experiments, but Detambel [10] did observe this result in an experiment with flashing lights. The subject was told that he had the choice of two telegraph keys to push, that one key was "correct" and the other key "incorrect" on each trial (contrary to fact for the group of subjects of principal interest), and that whenever he chose the correct key a light flash would appear. For three sets of π's with values 1.00:0; 0.50:0; 0.50:0.50, the asymptotic results were in accord with equation (15). The principal support for the equal-alpha model comes from the 0.50:0 group. The data are readily available in [13].

The secondary reinforcement model gives a rather different prediction. The values 0 and 1 in that model act as absorbing barriers. Sooner or later the subject stabilizes on one alternative. Furthermore, the alternative need not be the favorable one. By chance alone the animal may ultimately stabilize on the wrong side. This phenomenon has not occurred in any of the two-armed bandit experiments with human subjects reported in [7, chapter 13], but in conversation Solomon Weinstock [37] reported that 2 of about 30 rats in a 0.75:0.25 T-maze experiment appeared to have stabilized on the unfavorable side after a large number of trials.

Wilson and Bush [38] in a T-maze experiment with paradise fish attempted to reproduce some of the conditions of the experimenter-controlled event model by running an experimental group in a maze that had a transparent divider so that the fish could see that the caviar (reinforcement) was placed in the other side of the maze when he had made a choice that did not yield reinforcement. One and only one side was reinforced on each trial. In the control group an opaque divider was used to correspond to the ordinary two-armed bandit conditions. For the last 50 of 140 trials run under 0.75:0.25 conditions the number of fish in various class intervals is given in table II.

TABLE II

DISTRIBUTION OF NUMBER OF TURNS TO FAVORABLE SIDE IN LAST 50 TRIALS OF 140 TRIAL T-MAZE EXPERIMENT WITH PARADISE FISH (75:25). (WILSON AND BUSH)

TURNS TO FAVORABLE SIDE	CONTROL GROUP (Opaque Divider)	EXPERIMENTAL GROUP (Transparent Divider)
0–5	1	4
6–10	0	2
11–15	0	1
16–20	1	0
21–25	2	0
26–30	0	0
31–35	3	3
36–40	2	0
41–45	7	2
46–50	11	10
Number of fish	27	22

The roughly U-shaped distribution for the experimental group would be in good accord with the secondary reinforcement model or the identity operator model, but it seems not to agree well with the equal-alpha model. From considerations of order of magnitude one would not expect to find so many fish near the extremes in the equal-alpha model, nor does one expect so many fish stabilizing on the unfavorable side in the reinforcement-extinction model if α_2 has a reasonable effect.

No very good suggestions other than maximum likelihood have been proposed for estimating the learning parameters in such experiments (unless equal-alpha or identity operator models are assumed). The author suggests the following possibility as an approximate method until a better one is proposed. From preliminary work with T-maze and similar experiments, it would appear that both α_1 and α_2 are rather close to unity, say 0.95 or greater, in most experiments. Initially the value of p is near 0.50. In this neighborhood the effect of an operator is to add or subtract about $\frac{1}{2}(1 - \alpha)$ to the value of p (more precisely $(1 - \alpha)(1 - p)$ or $(1 - \alpha)p$). It is proposed to ignore the fact that not all operators commute and replace the model by an additive model for the early trials as indicated by the additive operators.

ADDITIVE APPROXIMATION MODEL

	Left	Right
Reinforcement	$p - \epsilon$	$p + \epsilon$
Nonreinforcement	$p - \delta$	$p + \delta$

The probability on trial n would then be approximately

$$p_n = p_0 + a_n\epsilon + b_n\delta \tag{16}$$

where a_n and b_n are the net numbers of times ϵ and δ have previously been added (where $\epsilon \doteq \frac{1}{2}(1 - \alpha_1)$, $\delta \doteq \frac{1}{2}(1 - \alpha_2)$). Maximum likelihood would suggest

$$P \doteq \prod_n p_n^{x_n} q_n^{1-x_n}\,, \tag{17}$$

where $x_n = 1$ or 0 according as right or left is chosen on the nth trial. Taking logarithms and differentiating gives for the likelihood equations

$$\begin{aligned} \sum \frac{x_n - p_n}{p_n(1-p_n)}\, a_n &= 0\,, \\ \sum \frac{x_n - p_n}{p_n(1-p_n)}\, b_n &= 0\,. \end{aligned} \tag{18}$$

Though these are tiresome to solve because p_n depends on ϵ and δ, it will be recalled that p_n is in the neighborhood of 0.50. Therefore the denominators $p_n(1 - p_n)$ might well be ignored, at least for the preliminary calculations. If this is done the equations (18) reduce to

$$\begin{aligned} \epsilon \Sigma a_n^2 + \delta \Sigma a_n b_n &= \Sigma a_n x_n - \tfrac{1}{2}\Sigma a_n \\ \epsilon \Sigma a_n b_n + \delta \Sigma b_n^2 &= \Sigma b_n x_n - \tfrac{1}{2}\Sigma b_n\,. \end{aligned} \tag{19}$$

This approximation might be obtained more directly by minimizing

$$\Sigma (x_n - p_n)^2\,. \tag{20}$$

After the first approximation is obtained, it might be possible to improve the estimate slightly by taking some account of the value of p obtaining when an operator is applied. One advantage of this additive model is that the four models described earlier can be discriminated on the basis of the single estimation procedure. If $\epsilon > 0$, $\delta < 0$, the reinforcement-extinction model applies; if $\epsilon > 0$, $\delta > 0$, the secondary reinforcement model applies; if $\delta = 0$ the identity operator model applies; and if $\epsilon = -\delta$ the equal-alpha model applies. Preliminary trials of this method with stat-rat data suggest it may be useful. From calculations based on the first 40 trials, the secondary reinforcement model appears to be appropriate to the fish data. The general idea, of course, is not restricted to T-maze problems.

A few results from T-maze and two-armed bandit experiments assuming the identity operator for nonreward may give some notion of the relative difficulty of discrimination between the more favorable and less favorable side. For convenience, in table III we give $\theta = 1 - \alpha$, a measure of the effectiveness of reinforcement in improving the probability associated with the reinforced alternative. Stanley's experiments [34] were performed with hungry rats, reinforcement was food, nonreinforcement was lack of food. Goodnow's data replicate Stanley's on Harvard students. In the "play free" situation reinforcement was a poker chip exchangeable for one cent, nonreinforcement was absence of chip. For her "pay to play" condition reinforcement was as in "play free," but nonreinforcement represented loss of one chip. Robillard used Harvard freshmen with chips worth zero, one, and five cents. An additional experiment using a playing card version of the two-armed bandit was performed by Bush, Davis, and Thompson with reinforcement five cents, nonreinforcement zero cents. Further discussion of these and other experiments is available in [7].

5. Other applications

Contrary to the impression that might be created by previous sections of this discussion, early papers on this general model by Estes [11], [12] and Bush and Mosteller [4] were oriented toward applications to problems of latency and rate. In the runway experiment of Graham and Gagné [19], the times (latency or running time) taken by rats to traverse a runway measured the learning. In a Skinner-box the rate at which a rat presses a bar or at which a pigeon pecks a key under various schedules of reinforcement measures performance. To bring the model to bear on such time problems requires a tie-up between time, the operators, and probability. Without going into details here, the general approach has been to quantize time into small increments h, and to regard each interval h as a trial offering an opportunity for the application of an operator. Then in rate problems instead of sticking closely to the stated model one replaces the results of the detailed step-wise application of the operators by an average value obtained by weighting the possible outcomes by their probabilities of occurrence. One does still further violence to the model by then turning from the approximate difference equation to a differential equation approximation. The general technique is routine for the physicist. In the case of latencies, the resulting curves usually fit the sequences of average latencies for several animals quite well. But the support for the model is not strong, because the types of curves generated have the general shape a curve-fitter would have chosen by eye and a reasonable number of parameters are available for the fitting. One feels the model should do more than fit curves of means. In the case of rate problems the model could in principle obtain a little support from its differential prediction [4] for various kinds of partial reinforcement schedules, but technical difficulties about eating time, recovery from eating, and activity level make it difficult to carry out the opera-

TABLE III

Summary of Estimates of $\theta = 1 - a$ for Various T-Maze and Two-Armed Bandit Experiments. The Given Value of n Represents the Number of Subjects in Each Group

Stanley: Rats in T-Maze ($n = 7$)

Schedule	100:0	50:0	75:25
Estimated parameter	0.039*	0.038	0.036

* An underestimate because of the special manner in which the experiment was performed.

Goodnow: Harvard Students Using Two-Armed Bandit ($n = 5$)

	100:0	50:0	100:50	75:25
Pay to play	0.112	0.038	0.072	0.033
Play free	0.049	0.021		

Robillard: Harvard Freshmen Using Two-Armed Bandit ($n = 10$)

	50:0	30:0	80:0	80:40	60:30
$.00	0.042	0.033	0.057	0.025	0.025
$.01	0.031				
$.05	0.027				

Bush, Davis, Thompson: High School and College Science Students Using Playing Cards ($n = 10$)

	50:0
High School	0.042
College Science	0.066

tions. Furthermore with bar-pressing and key-pecking the response itself has a staccato nature, and there is a question whether the model should attempt to describe the individual pecks or presses or rather bursts of these. Thus in spite of the attention given it thus far, the Skinner-box experiments cannot be said to be satisfactorily handled by the model—at least the evidence suggests only that by various approximate devices it generates a class of curves that can be used for fitting, but the kinds of further consequences we wish a model to generate remain untested and in some instances ungenerated. For latency in the runway, a somewhat more general model has been developed [7] for Weinstock's runway data [36] which attempts not only to fit mean curves, but also to generate the variances and percentage points of the distributions associated with latencies on each trial. This attempt represents an improvement in the direction desired, but the model still does not account for the extremely large latencies observed. The model referred to visualizes the rat's progress down the runway as composed of forward movements and intervals of standing still. A fixed number of forward movements is required to complete the run. A forward movement requires a reaction time plus a random interval of time, while standing still uses only a random interval. Random intervals are drawn from gamma-type distributions. The notion of an experimental trial is retained, with p_n being the probability of a forward movement throughout trial n, and $1 - p_n$ is the probability of standing still. A reinforcement operator adjusts p_n after each experimental trial. One possible improvement in this model to help account for large latencies is to admit the fact (observed in experiments) that some activities cancel previous forward movements. This complication has not been added to the latency model.

Making use of Estes' set-theoretic formulation for conditioning, Bush and Mosteller [5] have attempted a model for generalization and discrimination, the notions of which have been used further for discussions of psychoanalytic displacements [2], [9].

One direction in which the application of these models may well move is that suggested by problems of social interaction among the participants of a group in group decision. Bales, Flood, and Householder [1] have tried to use such a model to explain who says what and to whom in a discussion, where, of course, the actual content of the discussion is very grossly categorized. Hays and Bush [25] have regroomed the old war horse, the Humphreys-type experiment, to study decision making in groups of three. As before, following a signal, a light was turned on or not turned on (on in 75 per cent of the trials). The group of three individuals was required to make a single guess whether the light would or would not come on. Two models are proposed for test: (1) the group-actor model in which a group as a whole behaves as does the single individual in the Humphreys experiment described at the beginning of section 4; (2) the voting model in which it is assumed that the individuals behave in accord with the prediction for Humphreys experiment, but that the group decision is by majority rule. Model 1 would say the group predicts "on" 75 per cent of the time after many trials, and model 2 predicts (approximately) that the asymptote is $P = \pi^3 + 2\pi^2(1 - \pi) = 3\pi^2 - 2\pi^3$ or about 0.84. The data, with the malice data often have, refused to choose between these models after 100 trials. Both fit quite decently. If the experiment can be run for more trials without fatigue factors, perhaps a decision can be reached.

Though time measures remain an outstanding problem other measures of intensity seem even less tractable; no one has tried to bring amount of salivation, angle of knee kick, pull on a leash, galvanic skin response, or blood pressure into this framework. Very likely such measures require a more complicated model for the organism.

6. Probability considerations

Problems in theoretical probability raised by these models have encouraged some authors [22], [23], [24], [28] to investigate the nature of the distributions of values of p generated by them, as well as to consider more general problems suggested by them.

Harris [24] considers a process generated by a starting probability row vector $\xi^{(0)} = (Z_1^{(0)}, Z_2^{(0)}, Z_3^{(0)})$, $Z_i^{(0)} \geqq 0$, $\Sigma Z_i = 1$, together with three stochastic matrices:

$$(21)\qquad A_1 = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1-\alpha & 0 \\ \beta & 0 & 1-\beta \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1-\rho & \rho & 0 \\ 0 & 1 & 0 \\ 0 & \sigma & 1-\sigma \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $0 < \alpha, \beta, \rho, \sigma < 1$. (Note that A_1 and A_2 do not satisfy the combining of classes condition.) One matrix A_i is selected, $Z_i^{(0)}$ being the probability that A_i is selected, and a new vector

$$(22)\qquad \xi^{(1)} = \xi^{(0)}B_1$$

is computed where B_1 is the one matrix selected from the three. Now another A_i (named for convenience B_2) is selected, this time with probabilities given by the components of $\xi^{(1)}$, and B_2 postmultiplies $\xi^{(1)}$ to yield $\xi^{(2)}$. At the kth stage an A_i is selected (B_{k+1}) in accordance with the components of $\xi^{(k)}$ and premultiplied by $\xi^{(k)}$ to yield $\xi^{(k+1)}$. (The matrix A_3 corresponds to "no change" or standing still.) A sequence $B_1, B_2, \cdots$ is said to *conclude* (A_i) if all B_j for j greater than some integer are equal to the same A_i. A sequence *concludes* if it concludes (A_1) or concludes (A_2). Harris proves the

THEOREM. *For any* $\xi^{(0)} \neq (0, 0, 1)$ *the sequence* $B_1, B_2, \cdots$ *concludes with probability* 1. *Let* $\pi(\xi)$ *be the probability that the sequence concludes* (A_1) *if* $\xi^{(0)} = \xi$. *Then for* $\xi \neq (0, 0, 1)$, $\pi(\xi)$ *is a continuous function of* ξ *which is* 0 *when* $\xi = (0, 1, 0)$, *and* 1 *when* $\xi = (1, 0, 0)$. *Also* $0 < \pi(\xi) < 1$ *for* ξ *not equal* $(1, 0, 0)$, $(0, 1, 0)$ *or* $(0, 0, 1)$; *and* $\pi(\xi)$ *satisfies the functional equation*

$$(23)\qquad \pi(\xi) = Z_1\pi(\xi A_1) + Z_2\pi(\xi A_2) + Z_3\pi(\xi)$$

where $\xi = (Z_1, Z_2, Z_3)$.

Bellman [24] considers the functional equation

$$(24)\qquad f(x) = xf[a + (1 - a)x] + (1 - x)f(\sigma x) ,$$

where his x can be regarded as our p in the case of two subject-controlled events, $a + (1 - a)x$ corresponds to an operator $Q_1p = \alpha_1p + (1 - \alpha_1)\lambda_1$ with our $\alpha_1 = 1 - a$, $\lambda_1 = 1$, and his σx corresponds to Q_2p with $\alpha_2 = \sigma$, $\lambda_2 = 0$. He shows that a limiting continuous solution exists, is monotone in x, concave if $a + \sigma \leqq 1$, convex if $a + \sigma \geqq 1$, analytic, and unique. He extends the discussion to the two-dimensional case and states that no new problems arise for still higher dimensions.

Shapiro [24] studies methods of solving the functional equation (24), and shows that the method of successive approximations converges uniformly to the continuous solution if and only if the initial approximation $f_0(x)$, $f_0(0) = 0$, $f_0(1) = 1$, is bounded in the interval [0, 1] and continuous at 0 and 1 (a very mild restriction, since $f_0(x) = x$ is a natural initial approximation). The rate of convergence is studied. A necessary and sufficient condition for x to be the continuous solution is that

$$(25)\qquad \sigma = 1 - a + o(a^2) .$$

Karlin [28] introduces a different method of attack for random walk problems leading to functional equations like (24). He is able to deduce many of the results of Bellman and Shapiro, as well as additional ones for the case of two reflecting barriers. The latter is obtained when the coefficients on the right-hand side of equation (24) are interchanged. Among other results is the extension to the three response experimenter-subject controlled event situation where one response is that of standing still: (1) $x \rightarrow \sigma x$ with probability $\pi_1(1 - x)$; (2) $x \rightarrow \alpha x + 1 - \alpha$ with probability $\pi_2 x$; and (3) $x \rightarrow x$ with probability $(1 - \pi_1)(1 - x) + (1 - \pi_2)x$.

REFERENCES

[1] R. F. Bales, M. M. Flood and A. S. Householder, "Some group interaction models," The RAND Corporation, RM-953, 1952.

[2] F. R. Brush, R. R. Bush, W. O. Jenkins, *et al.*, "Stimulus generalization after extinction and punishment: an experimental study of displacement," *Jour. Abnor. and Soc. Psych.*, Vol. 47 (1952), pp. 633–640.

[3] C. J. Burke, W. K. Estes, and S. Hellyer, "Rate of verbal conditioning in relation to stimulus variability," *Jour. Exp. Psych.*, Vol. 48 (1954), pp. 153–161.

[4] R. R. Bush and F. Mosteller, "A mathematical model for simple learning," *Psych. Review*, Vol. 58 (1951), pp. 313–323.

[5] ———, "A model for stimulus generalization and discrimination," *Psych. Review*, Vol. 58 (1951), pp. 413–423.

[6] ———, "A stochastic model with applications to learning," *Annals of Math. Stat.*, Vol. 24 (1953), pp. 559–585.

[7] ———, *Stochastic Models for Learning*, New York, John Wiley and Sons, 1955.

[8] R. R. Bush, F. Mosteller, and G. L. Thompson, "A formal structure for multiple-choice situations," *Decision Processes*, New York, John Wiley and Sons, 1954, pp. 99–126.

[9] R. R. Bush and J. W. M. Whiting, "On the theory of psychoanalytic displacement," *Jour. Abnor. and Soc. Psych.*, Vol. 48 (1953), pp. 261–272.

[10] M. H. Detambel, "A re-analysis of Humphreys' 'Acquisition and extinction of verbal expectations.'" Unpublished M. A. thesis, Indiana University, 1950.

[11] W. K. Estes, "Toward a statistical theory of learning," *Psych. Review*, Vol. 57 (1950), pp. 94–107.

[12] ———, "Effects of competing reactions on the conditiong curve for bar pressing," *Jour. Exp. Psych.*, Vol. 40 (1950), pp. 200–205.

[13] ———, "Individual behavior in uncertain situations: an interpretation in terms of statistical association theory," *Decision Processes*, New York, John Wiley and Sons, 1954, pp. 127–137.

[14] W. K. Estes and C. J. Burke, "A theory of stimulus variability in learning," *Psych. Review*, Vol. 60 (1953), pp. 276–286.

[15] W. K. Estes and J. H. Straughan, "Analysis of a verbal conditioning situation in terms of statistical learning theory," *Jour. Exp. Psych.*, Vol. 47 (1954), pp. 225–234.

[16] B. G. Farley and W. A. Clarke, "Simulation of self-organizing systems by digital computer," *Trans. Inst. of Radio Engineers, Professional Group on Information Theory*, PGIT-4 (Sept. 1954), pp. 76–84.

[17] M. M. Flood, "On game-learning theory and some decision-making experiments," *Decision Processes*, New York, John Wiley and Sons, 1954, pp. 139–158.

[18] ———, "Environmental nonstationarity in a sequential decision-making experiment," *Decision Processes*, New York, John Wiley and Sons, 1954, pp. 287–299.

[19] C. H. Graham and R. M. Gagné, "The acquisition, extinction, and spontaneous recovery of a conditioned operant response," *Jour. Exp. Psych.*, Vol. 26 (1940), pp. 251–280.

[20] H. Gulliksen, "A rational equation of the learning curve based on Thorndike's law of effect," *Jour. General Psych.*, Vol. 11 (1934), pp. 395–434.

[21] H. Gulliksen and D. L. Wolfle, "A theory of learning and transfer: I, II," *Psychometrika*, Vol. 3 (1938), pp. 127–149 and 225–251.

[22] T. E. Harris, "A method for limit theorems in Markov chains" (abstract), *Annals of Math. Stat.*, Vol. 23 (1952), p. 141.

[23] ———, "On chains of infinite order," to be published.

[24] T. E. Harris, R. Bellman and H. N. Shapiro, "Studies in functional equations occurring in decision processes," The RAND Corporation, P-382, 1953.

[25] D. G. Hays and R. R. Bush, "A study of group action," *Amer. Sociological Rev.*, Vol. 19 (1954), pp. 693–701.

[26] C. L. Hull, C. I. Hovland, R. T. Ross, *et al.*, *Mathematico-deductive Theory of Rote Learning*, New Haven, Yale University Press, 1940.

[27] L. G. Humphreys, "Acquistion and extinction of verbal expectations in a situation analogous to conditioning," *Jour. Exp. Psych.*, Vol. 25 (1939), pp. 294–301.

[28] S. Karlin, "Some random walks arising in learning models I," *Pacific Jour. Math.*, Vol. 3 (1953), pp. 725–756.

[29] G. A. Miller and F. C. Frick, "Statistical behavioristics and sequences of responses," *Psych. Review*, Vol. 56 (1949), pp. 311–324.

[30] G. A. Miller and W. J. McGill, "A statistical description of verbal learning," *Psychometrika*, Vol. 17 (1952), pp. 369–396.

[31] E. D. Neimark, "Effects of type of non-reinforcement and number of alternative responses in two verbal conditioning situations." Unpublished Ph.D. thesis, Indiana University, 1953.

[32] L. J. Savage, *The Foundations of Statistics*, New York, John Wiley and Sons, 1954.

[33] R. L. Solomon and L. C. Wynne, "Traumatic avoidance learning: acquisition in normal dogs," *Psych. Monographs*, Vol. 67 (1954), No. 354.

[34] J. C. Stanley, Jr., "The differential effects of partial and continuous reward upon the acquisition and elimination of a running response in a two-choice situation." Unpublished Ed. D. thesis, Harvard University, 1950.

[35] L. L. Thurstone, "The learning function," *Jour. General Psych.*, Vol. 3 (1930), pp. 469–494.

[36] S. Weinstock, "Resistance to extinction of a running response following partial reinforcement under widely spaced trials," *Jour. Compar. and Physiological Psych.*, Vol. 47 (1954), pp. 318–322.

[37] ———, personal communication.

[38] T. R. Wilson and R. R. Bush, personal communication.