PROBLEMS OF OPTIMIZATION IN LEARNING A LIST OF SIMPLE ITEMS

by

Patrick Suppes

TECHNICAL REPORT NO. 57

July 22, 1963

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
Applied Mathematics and Statistics Laboratories
STANFORD UNIVERSITY
Stanford, California

Problems of Optimization in Learning a List of Simple Items[1]

Patrick Suppes

1. Introduction.

The learning task I want to consider in this paper is, on the face of it, a very simple one. We want subjects to learn a list of items. The list is rather long, so we may want to consider the possibility of breaking up the list into blocks of varying size.

A number of applications of results on this kind of problem are easy to describe. Among the most interesting are those concerned with learning a second language. As a typical example, the items may be foreign words, and the problem is to learn their approximate meanings in English. In psychological terms, this means that to each word of the foreign language in the list the subject is asked to associate a unique English word. As we all know, in the learning of a second language, considerable variation of difficulty in single items is encountered. Some words in graphemic representation are almost identical to their English graphemic representation, but perhaps at the phoneme level considerable differences are present. For example, the word "application" is graphemically the same but phonetically different in English and French. In other cases, there is both graphemic and phonemic difference in the corresponding words of the two languages

---

[1] This research was supported by Contract SAE 9514 between Stanford University and the U. S. Office of Education, and was also partially supported by a grant from the Carnegie Corporation of New York.

but they have a common etymological root, and from this root stems both graphemic and phonemic similarity. An example is the English "number" and the French "nombre." In other cases, there is no apparent graphemic or phonemic similarity between corresponding words, and the association must be established in a pretty much arbitrary fashion, that is, without the use of common elements. A familiar example from standard mathematical vocabulary is the English word "set," the corresponding French word "ensemble" and the German word "Menge." Other kinds of cases that I have not mentioned arise when the graphemes of the two languages are written in different alphabets, as in the case of English and Russian. In this instance we may distinguish between words whose graphemic representations use common letters of the two alphabets, those that use letters unique to one of the alphabets, etc.

In the present paper, I want to ignore all these sources of differences in difficulty of learning items, and to assume that all items--rather, item pairs, meaning both the stimulus item and the correct response item--can be treated as paired associates. I do not want to defend the realism of this assumption in analyzing the learning of a foreign language. It is just that even with this highly simplifying assumption, we shall find that the problems besetting us are complex enough.

There are additional restrictions I want to impose on the structure of the learning task. In learning large lists of items, it is often the practice to have review sessions in which parts of the list are reviewed after they have been learned, presumably in order to maintain retention at a fairly high level. It will be an important and desirable

extension of the kind of results obtained in this paper to include the consideration of the size and nature of review sessions, but their omission is another useful simplification in this initial study.

In order to evaluate the proper sublist or block size it is first necessary to state explicitly the criterion with respect to which optimization is sought. One standard condition is to fix the total number of trials, with a definite number of the terminal trials allocated as test items. Performance on these test items is then used as the criterion in terms of which performance is evaluated. The optimization problem is then to decide how the remaining trials, apart from the test, should be allocated for training and learning purposes.

Another approach to choice of a criterion is to run each block of items to a behavioral criterion and then to evaluate different block sizes in terms of the total number of trials required to complete the learning. This method may also be combined with the test-trial procedure, although in this case some weighting has to be assigned to the two types of criteria.

For the purposes of the present paper, essentially the first criterion mentioned will be used. I assume there are a fixed number N of training trials; in particular, n training trials for each of m items, and thus $N = mn$. At the end of the N training trials each item i will have a mean error probability $E(q_i)$, where the mean is taken across subjects in the experiment. We then consider $E(q)$, the mean over items, i.e.,

$$E(q) = \frac{1}{m} \sum_{i=1}^{m} E(q_i) \ .$$

Our criterion of performance is simply $E(q)$, and our objective is to choose the block size k so as to minimize $E_k(q)$, (the subscript is used to indicate that the mean error probability at the end of training is a function of the block size k). Behavioral data on $E_k(q)$ may be obtained from test trials given at the end of training.

The main purpose of the present paper is to show that under fairly general assumptions, k should be chosen as large as possible when the learning process is faster than the forgetting process, i.e., the block size should be just the complete list of m items, or, in other words, we should choose $k = m$. And when the learning process is slower than the forgetting process, k should be chosen as small as possible, i.e., we should choose $k = 1$.

It should be emphasized that the results obtained here are all in terms of group means, but I trust that in most problems of application we are mainly interested in average group performance, for we cannot write separate curriculum material for the individual student.

It is natural to ask if there are any experimental studies that support the theoretical conclusion about optimal block size. A number of studies going back to 1924 are cited by Woodworth and Schlosberg (1954, p. 784), and they mainly support the conclusion that learning a list of simple items is faster when the block size is the whole list rather than some intermediate value. The 1932 experiment of Seibert, cited by Woodworth and Schlosberg, illustrates very well the empirical soundness of our main conclusion. In Seibert's

study 44 students were presented lists of 12 English-French equivalents in varying block size, to be learned so as to give the correct French word in response to the English word. Each pair was presented six times, and the block sizes were $k = 1, 4, 6, 12$. The average scores in terms of percent correct, after 50 minutes, and after two days, were as follows,

	Block size			
	1	4	6	12
After 50 minutes	35	39	44	49
After 2 days	31	33	34	47

The monotonic relation between percent correct and block size agrees with the theoretical results derived in the present paper. In terms of the other half of the theoretical results derived here, I have not found any data bearing on the choice of $k = 1$ when the rate of learning is slower than the rate of forgetting. The lack of relevant data seems to be due to the difficulty of analyzing from the standpoint of varying block size the usual studies cited in support of part over whole learning.

2. Stimulus sampling and linear models of learning and forgetting.

Because the results in this paper depend only on the mean learning and forgetting curves, they are compatible with a wide class of assumptions about the basic learning and forgetting processes. The conceptualization of learning underlying stimulus sampling models may be roughly described as follows.

The subject is presented with a sequence of trials, on each of which he makes a response that is one of several possible choices. In any particular setup it is assumed that there is a set of stimuli from which the subject draws a sample at the beginning of each trial. It is assumed that on each trial each stimulus is conditioned to at most one response. The probability of making a given response on any trial is postulated to be simply the proportion of sampled stimuli conditioned to that response, unless there are no conditioned stimuli in the sample, in which case there is a "guessing" probability for each response. Learning takes place by the following mechanism. At the end of the trial a reinforcing event occurs. The event identifies that one of the possible responses which was correct. With some fixed probability, the sampled stimuli become conditioned to this response if they are not already, and the organism begins another trial in a new state of conditioning.

It is not a very complicated matter to put these roughly stated assumptions in exact form and to derive from them the following recursion for the mean learning curve,

$$q_{n+1} = \alpha q_n \ . \tag{1}$$

Here q_n is the probability of an incorrect response on trial n, and the learning parameter is $\alpha = 1 - \frac{cs}{N}$, where c is the probability that a sample element becomes conditioned, N is the number of stimuli available for sampling in connection with a given item of the list, and s is the number of stimuli sampled on each trial, on the assumption that a fixed number is sampled throughout the experiment (modifications of this assumption of fixed sampling are easily introduced).

If we let the number of stimuli become large and require that s be approximately a fixed proportion of the population of stimuli, then it may be shown that in the limit, as the population of stimuli increases, we obtain the linear models, whose recursion for individual learning is precisely the same as the mean recursion given above (for a derivation of linear models from stimulus sampling models, see Estes and Suppes (1959).)

Although stimulus sampling and linear models have not been applied very extensively to forgetting phenomena, and there is some reason to doubt that they will give an adequate detailed account of extinction of learning, it is a plausible assumption to suppose that the same kind of mean curve that holds for learning also holds approximately for forgetting, but in this case the roles of a probability of an error and a probability of a correct response are interchanged. Symmetric to the formulation of the learning process just given, we may characterize the forgetting process in terms of a fixed proportion of the population of stimuli becoming unconditioned with probability u on each trial on which the item had not appeared. Of course, from a physical standpoint, it would be natural to formulate the forgetting process in terms of a continuous

time parameter rather than in terms of the discrete trials on which the item is not presented, but it is convenient and natural in the context of discrete trial experiments to formulate the forgetting process in a manner exactly analagous to that of the formulation of the learning process. By the same sort of general arguments that may be used to derive (1) we may derive as the mean recursion for forgetting the equation

$$p_{n+1} = \beta p_n \tag{2}$$

where p_n is the probability of a correct response on trial n, and β is the forgetting parameter, which in terms of the formulation just given is equal to $1 - \frac{us}{N}$. Because the probability p of guessing a correct response when no stimulus elements are conditioned is often unequal to zero, equation (2) must be modified to:

$$p_{n+1} = \beta p_n + (1-\beta)p \ . \tag{3}$$

When we replace p_n by $1-q_n$ and p by $1-q$, we obtain the following recursion for forgetting in terms of the probability q_n of an error on trial n.

$$q_{n+1} = \beta q_n + (1-\beta)q \ . \tag{4}$$

It may be noted that the introduction of the probability $p = 1-q$ is also necessary for completely describing the learning process, as well as the forgetting process. In the case of the learning process the probability p characterizes the point at which learning begins. If, for example, the subject is responding by pressing one of several multiple choice keys, then it is certainly not correct to assume that

initially $p = 0$ but is, for example, something like $1/R$, where R is the number of response keys. It is important to be quite clear about the interpretation of equations (1) and (4). Equation (1) is the mean learning recursion that applies on trials on which an item is presented, and equation (4) applies on all other trials, i.e., on all trials on which other items are presented and thus there is an opportunity to forget the correct response to the item in question.

It is evident from equations (1) and (4) that the learning and forgetting processes, in terms of their mean recursions at least, are here formulated in terms of the three parameters, α, β and q.

3. Derivation of optimal block size.

To formulate a definite and reasonably manageable optimization problem, we shall assume the following:

A1. There are N total training trials.

A2. There are m items, and N/m is an integer equal to or greater than one.

A3. The block size is k (we want to optimize with respect to k).

A4. Items are arranged in a fixed order $1, \ldots, m$, and this order is not changed by changes in the block size.

A5. Each block of length k is run $n = N/m$ times in fixed order, and then the next block is run n times, and so on, until the entire list is completed.

A6. There is an initial uniform guessing probability q of a response error for all items.

As already remarked, we shall be concerned only with expectations. In particular, we shall compute first $E(q_i)$ for each item i, i.e., the expected error probability at the end of the N trials, and we shall then compute the mean $E(q)$ of the items.

As a first step toward computing $E(q_i)$, we want to find an expression for what happens to q_i during the course of training on the block to which i belongs. At the first appearance of item i, we have a reinforcement, followed by $k-1$ forgetting trials, followed by the second reinforcement, followed by $k-1$ forgetting trials, etc. It will be useful to write an expression, for what happens to q_i during one cycle, i.e., after one reinforcement trial and $k-1$ forgetting trials. Let us denote by ρ the reinforcement transformation and by φ the forgetting transformation. Then, at the end of one cycle,

$$\varphi(\rho(q_i)) = \varphi(\alpha q_i) = \beta^{k-1}\alpha q + (1-\beta^{k-1})q \ .$$

(On the right-hand side we write q rather than q_i, because we have assumed a uniform guessing probability q for all items.) Note that φ and ρ do not, of course, commute. Let us denote the transformation for the complete cycle by ψ. Thus

$$\psi(q_i) = \beta^{k-1}\alpha q + (1-\beta^{k-1})q \ .$$

At the end of the second cycle, we have

$$\psi^2(q_i) = \beta^{2(k-1)}\alpha^2 q + \beta^{k-1}\alpha(1-\beta^{k-1})q + (1-\beta^{k-1})q \ .$$

For simplicity of notation, we let

$$y = \beta^{k-1} .$$

At the end of the $(n-1)$st cycle, we have

$$\psi^{n-1}(q_i) = [y^{n-1}\alpha^{n-1} + y^{n-2}\alpha^{n-2}(1-y) + \ldots + y\alpha(1-y) + (1-y)]q ,$$

$$= [\alpha^{n-1}y^{n-1} + \frac{(1-y)(1-\alpha^{n-1}y^{n-1})}{1-\alpha y}]q .$$

As is apparent from the definition of ψ, this result is independent of i, for we have thus far considered only a block of trials duplicated for each item. We have one more learning trial to add, and then may express the remaining trials after the last reinforcement, as forgetting trials. There are N total trials. The question is, how many of these N trials have occurred when the n^{th} reinforcement, the last reinforcement, for item i occurs. To find in which block item i occurs, we may express i as

$$i = bk + j ,$$

for integers b and j, with $j \leq k$. Before the first reinforcement of item i, bkn trials were used by the preceding b blocks, and $j-1$ trials by the predecessors of i in its own block. Then $k(n-1)$ trials are used up by the trials starting with the first reinforcement of item i and ending just before the last reinforcement. We may express this last reinforcement by multiplying $\psi^{n-1}(q_i)$ by α to have:

$$\alpha\psi^{n-1}(q_i) .$$

The remaining $N - (bkn+j-1+k(n-1)+1)$ trials are forgetting trials. We thus have

$$E(q_i) = \beta^{N-((b+1)kn+j-k)}(\alpha\psi^{n-1}(q_i))$$

$$+(1-\beta^{N-((b+1)kn+j-k)})q$$

or, in more convenient form for the present,

$$E(q_i) = q-(q-\alpha\psi^{n-1}(q_i))\beta^{N-((b+1)kn+j-k)} .$$

Under our homogeneity assumption for items, only the exponent of β depends on i, and so to obtain

$$E(q) = E(E(q_i)) = \frac{1}{m}\sum_{i=1}^{m} E(q_i) ,$$

we need to sum this final term of our expression for $E(q_i)$. For convenience of notation, let $\gamma = \frac{1}{\beta}$. Then, when $i = bk+j$, and $h =$ the number of blocks

$$\sum_{b=0}^{h-1}\sum_{j=1}^{k} \gamma^{(b+1)kn+j-k} = \gamma^{k(n-1)+1}+\gamma^{k(n-1)+2}+ \ldots +\gamma^{k(n-1)+k}$$

$$+ \gamma^{kn+k(n-1)+1}+ \ldots +\gamma^{kn+k(n-1)+k}$$

$$+ \gamma^{2kn+k(n-1)+1}+ \ldots +\gamma^{2kn+k(n-1)+k}$$

$$+ \ldots +\gamma^{(h-1)kn+k(n-1)+1}+ \ldots +\gamma^{(h-1)kn+k(n-1)+k}$$

$$= \gamma^{k(n-1)+1}\{[1+\gamma+ \ldots +\gamma^{k-1}]+\gamma^{kn}[1+\gamma+ \ldots +\gamma^{k-1}]$$

$$+ \ldots +\gamma^{(h-1)kn}[1+\gamma+ \ldots +\gamma^{k-1}]\}$$

$$= \gamma^{k(n-1)+1}\left(\frac{1-\gamma^{k}}{1-\gamma}\right)[1+\gamma^{kn}+ \ldots +\gamma^{(h-1)kn}]$$

$$= \gamma^{k(n-1)+1}\left(\frac{1-\gamma^{k}}{1-\gamma}\right)\left[\frac{1-\gamma^{hkn}}{1-\gamma^{kn}}\right]$$

$$= \gamma^{k(n-1)+1}\left(\frac{1-\gamma^{k}}{1-\gamma}\right)\left[\frac{1-\gamma^{N}}{1-\gamma^{kn}}\right]$$

because $N = hkn$. Replacing now γ by $1/\beta$, we obtain

$$\sum_{b}\sum_{j}\beta^{-[(b+1)kn+j-k]} = \frac{1}{\beta^{k(n-1)+1}} \cdot \frac{\frac{\beta^{k}-1}{\beta^{k}}}{\frac{\beta-1}{\beta}}\left[\frac{\frac{\beta^{N}-1}{\beta^{N}}}{\frac{\beta^{kn}-1}{\beta^{kn}}}\right]$$

$$= \frac{\beta^{k}-1}{\beta-1}\left[\frac{\beta^{N}-1}{\beta^{N}(\beta^{kn}-1)}\right]$$

$$= \frac{1-x}{1-\beta}\left[\frac{1-\beta^{N}}{\beta^{N}(1-x^{n})}\right] ,$$

where $x = \beta^{k}$. Combining results, and noting that $x = \beta y$, we then have as a final expression

$$E(q) = \frac{1}{m}\sum_{i=1}^{m} E(q_i)$$

$$= q - (q-\alpha\psi^{n-1}(q))\,\frac{\beta^{N}}{m}\sum_{b=0}^{h-1}\sum_{j=1}^{k}\beta^{-[(b+1)kn+j-k]}$$

$$= q - (q-\alpha\psi^{n-1}(q))\,\frac{1-\beta^{N}}{(1-\beta)m}\left(\frac{1-x}{1-x^{n}}\right)$$

$$= q - \left[q - \alpha\left[\frac{\alpha^{n-1}}{\beta^{n-1}} x^{n-1} + \frac{(1-\frac{x}{\beta})(1-\frac{\alpha^{n-1}x^{n-1}}{\beta^{n-1}})}{1-\frac{\alpha x}{\beta}}\right]q\right] \left(\frac{1-\beta^N}{(1-\beta)m}\right)\left(\frac{1-x}{1-x^n}\right)$$

$$= \left\{1 - \left[\frac{(1-\epsilon x) - \alpha(1-\epsilon x)\epsilon^{n-1}x^{n-1} - (\alpha-\epsilon x)(1-\epsilon^{n-1}x^{n-1})}{1-\epsilon x}\right]\left(\frac{1-\beta^N}{(1-\beta)m}\right)\left(\frac{1-x}{1-x^n}\right)\right\} q \ ,$$

where $\epsilon = \frac{\alpha}{\beta}$. We may then simplify this last expression to obtain a reasonably simple result.

$$E(q) = \left\{1 - \frac{(1-\beta^N)(1-\alpha)}{(1-\beta)m}\left(\frac{1-\epsilon^n x^n}{1-\epsilon x}\right)\left(\frac{1-x}{1-x^n}\right)\right\} q \tag{5}$$

To determine how the mean probability of an error at the end of training will vary with the selection of the block size k , we note first that x is a function of k , in particular , $x = \beta^k$. Whence as k increases, x decreases, and vice versa, because β is between 0 and 1. If we look at the derivative of (5) with respect to x , the whole matter hinges on the sign of the derivative of

$$\left(\frac{1-\epsilon^n x^n}{1-\epsilon x}\right)\left(\frac{1-x}{1-x^n}\right), \tag{6}$$

because

$$\frac{(1-\beta^N)(1-\alpha)}{(1-\beta)m}$$

is positive. However, it is easily shown that the derivative of (6) is positive if $\epsilon > 1$, negative if $\epsilon < 1$, and, of course, zero if $\epsilon = 1$, for the range of x we are considering, namely, $0 < x < 1$.

Putting these results together, and remembering that $\epsilon = \frac{\alpha}{\beta}$, we then have:

(7) $\begin{cases} \text{If } \alpha < \beta \text{ then } E(q) \text{ is a monotonically decreasing function of } k . \\ \text{If } \alpha > \beta \text{ then } E(q) \text{ is a monotonically increasing function of } k . \end{cases}$

Now if $\alpha < \beta$, then the rate of learning is faster than the rate of forgetting. And if $\alpha > \beta$ learning is slower than forgetting, for if $\alpha = 1$ there is no learning, or if $\beta = 1$ there is no forgetting. Hence we reach the conclusion from (7)

(8) $\begin{cases} \text{If the rate of learning is faster than the rate of forgetting,} \\ \text{choose the block size as large as possible, i.e., choose } k = m , \\ \text{in order to minimize } E(q) . \\ \text{If the rate of learning is slower than the rate of forgetting} \\ \text{choose the block size as small as possible, i.e., choose } k = 1 , \\ \text{in order to minimize } E(q) . \end{cases}$

To give some sense of the numerical character of these results, let us take the simplest list of interest, namely, $m = 2$, and examine the difference between $E_1(q)$ and $E_2(q)$, where the subscript indicates the block size of 1 or 2. The two extreme reinforcement schedules are $n = 2$ and $n = \infty$, i.e., we give in one case two training trials on each of the two items, and in the other, an infinite number of training trials.

When $n = 2$

$$E_1(q) = [1 - \frac{(1-\alpha^2)(1+\beta^2)}{2}]q$$

$$E_2(q) = [1 - \frac{(1-\alpha)(1+\alpha\beta)(1+\beta)}{2}]q .$$

To take a fairly extreme numerical case, if $\alpha = .1$, $\beta = .9$ and $q = 1.0$, then $E_1(q) = .104$ and $E_2(q) = .024$. The absolute difference between $E_1(q)$ and $E_2(q)$ is not large here, but the percentage difference certainly is, i.e., $E_1(q)$ is more than four hundred percent larger than $E_2(q)$.

When we consider $n = \infty$ with the same values of α, β and q the absolute difference is also large. Here obviously $E_1(q) = .500$, because the first item is completely forgotten and the second item completely learned.

For $n = \infty$

$$E_k(q) = \left\{ 1 - \frac{(1-\alpha)(1-\beta^k)}{1-\alpha\beta^{k-1}} \quad \frac{1}{(1-\beta)m} \right\} q ,$$

and $E_2(q) = .060$, which is a much smaller mean error probability than .500.

Although the numerical computations given here do not arise from any experimental data, it should be clear that the theoretical results obtained here are easily applied to data. In general only the parameters α and β need be estimated. If several block sizes are studied it will be easy enough to see how well the results derived here stand up. On data like those of Seibert cited at the beginning, it is apparent the theory will work out fairly well. The first critical and interesting point is to extend the block size considerably to see if the monotonicity is preserved. Fairly large scale experiments using Russian-English word pairs are now underway at Stanford, jointly with E. J. Crothers and Ruth Weir. We hope to report on these matters soon.

The results derived for presentation of the items in a fixed order can be extended to paired-associate schedules in which the order of items in a block is randomized on each presentation. Unfortunately the expressions become very cumbersome because of the varying number of trials between reinforcements.

Another direction of greater conceptual interest is toward consideration of other schedules than those of fixed block size. Given the mean recursions (1) and (4) for learning and forgetting, m items and n training trials on each item, what is the efficiency of intermittent review sessions of varying block size? More generally, what is the optimal arrangement of training trials, for given α and β, without restriction to fixed block size? I hope in the near future to be able to report on some of these additional questions.

References

Estes, W. K. and Suppes, P. Foundations of statistical learning theory, II. The stimulus sampling model, Technical Report No. 26, Institute for Mathematical Studies in the Social Sciences, Stanford University, October 1959, 141 pp.

Woodworth, R. S. and Schlosberg, Harold. Experimental Psychology, Revised edition, New York: 1954.