MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence
Memo. No. 156

March 1968

Linear Decision and Learning Models

Marvin L. Minsky

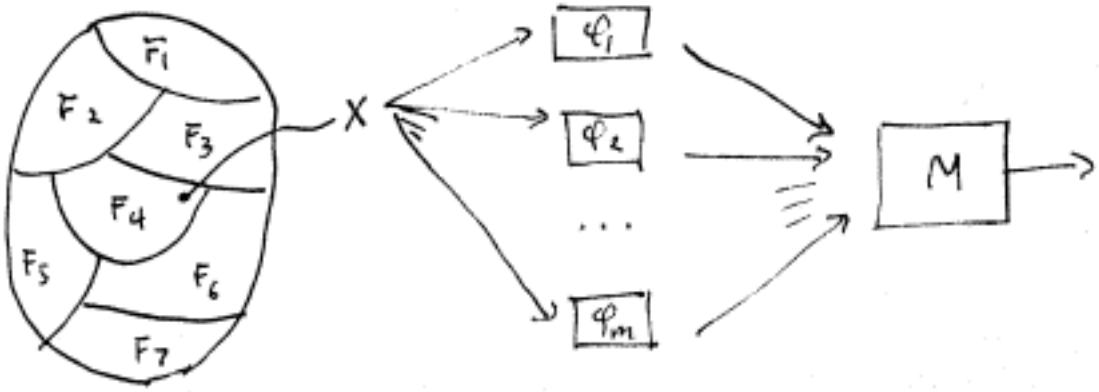
This Memorandum is a first draft of an essay on the simplest "learning" processes. Comments are invited. Subsequent sections will treat, among other things:

> The "stimulus-sampling" model of Estes
> Relations between Perceptron-type error reinforcement and Bayesian-type correlation reinforcement

and some other statistical methods viewed in the same way.

1.0 Decisions based on fixed set of tests

The events to be classified lie in a set $\{F_i\}$ of classes.



We are given also a set $\{\varphi_i(X)\}$ of "tests" or "experiments" that can be performed on the event that has just occurred. Define

$$\Phi(X) = (\varphi_1(X), \varphi_2(X), \ldots, \varphi_m(X))$$

to be the sequence of results of these tests, taken in some fixed order. Usually we will restrict each φ to be Boolean--that is, to have values 0 or 1.

> Thus the machine M cannot "see" the real events "directly," but only through the results of the experiments. There is really no alternative to this sort of indirectness, because only mystics believe in the possibility of the immediate and direct experience of reality--and whether or not this is in fact desirable, they are for some reason unable to transmit to others their grounds for believing this.

Given the outcome $\Phi(X)$ of a set of experiments, we would like to guess which class F_j contains the event X that has just occurred. In some situations we can be sure which F was responsible; for example, in the case that there is only one X that could have produced this particular value of Φ. (It is

convenient to think of $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_m)$ as a vector, so that we can talk about its value instead of the sequence of values of all the φ's.)

But in general, we cannot usually be sure which F_j was responsible. We shall consider the rather general case in which our knowledge about the situation can be summarized in the form of a table, or distribution, of probabilities:

$P(F_j/\Phi)$ = the probability that X is in F_j, given that the experiments have produced $\Phi = (\varphi_1, \ldots, \varphi_m)$.

Most of this chapter is devoted to discussing ways in which this sort of information could be acquired through experience. First, however, we will say a few words about how we could use this information if we had it!

1.1 Costs and decision criteria

If a particular $\Phi = \Phi_0$ occurs, and if we know that

$$\begin{cases} P(F_j/\Phi_0) = 1 \\ P(F_k/\Phi_0) = 0 \qquad (k \neq j) \end{cases}$$

there is no question that $X \in F_j$, and we can say so definitely. But if $0 < P < 1$ we have to guess. Usually, one will choose that j for which $P(F_j/\Phi_0)$ is largest. This guess will have the smallest chance of being wrong—the so-called "maximum likelihood criterion."

But in real situations, just trying to be right is not the only goal that may have to be considered. Different kinds of mistakes can have grossly different consequences. It can be more important to avoid the tiger than to acquire the lady. Let

$$c_{jk} = \text{the cost of guessing } F_k \text{ when X is really in } F_j.$$

Then the "expected" cost, if we guess F_k (given ϕ) is

$$C(k,\phi) = \sum_j c_{jk} P(F_j/\phi_0).$$

That is, this is the average price one will pay, over a large collection of experiments, if one has the policy of guessing F_k whenever one sees that particular ϕ . Clearly, our policy then should be to <u>choose that k for which C(k, ϕ) is smallest</u>.

Only in the most peculiar circumstances would it be that a correct guess c_{ii} would cost more than any incorrect guess c_{ij} $(i \neq j)$. But one might have a situation in which

$$\begin{cases} c_{ij} = 0 & \text{(all j)} \\ c_{jj} = 0 & \text{(all j)} \\ c_{ij} = 1 & (i \neq I, j \neq i) \end{cases}$$

so that one might as well always guess F_1 (regardless of the value of Φ), since there is no premium on being right otherwise. In the case in which all errors are equally costly:

$$\begin{cases} c_{ii} = 0 \\ c_{ij} = 1 \qquad (i \neq j) \end{cases}$$

we have

$$\sum_j c_{jk} P(F_j/\Phi) = 1 - P(F_k/\Phi)$$

(since $\sum_j P(F_j/\Phi) = 1$) and minimizing this is, as it should be, the same as maximizing $P(F_k/\Phi)$, i.e., the maximum likelihood strategy.

1.2 Inverting probabilities: Bayes' theorem

We have described decisions based upon knowledge of the probabilities $P(F_j/\Phi)$. Usually, one starts with access to a different set, $P(\Phi/F_j)$, of probabilities; $P(\Phi_0/F_j)$ is the probability that Φ_0 will occur if $X \in F_j$. For example, understanding of the mechanisms by which X's in each F_j affect the devices that compute the φ_i's theoretical calculation of what the $P(\Phi,F_j)$'s ought to be. Fortunately, one can calculate the $P(F_j/\Phi)$'s from these as follows: By the definition of the conditional probability symbol

$$P(B/A) = \frac{P(A \wedge B)}{P(A)}$$

where $P(B \wedge A)$ is the "joint" probability that both events B and A occur. But then, also

$$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$

because $P(A \wedge B) \equiv P(B \wedge A)$.

Therefore

$$P(F_j/\Phi) = \frac{P(\Phi/F_j) \cdot P(F_j)}{P(\Phi)} .$$

Since $P(\Phi) = \sum_k [P(F_k)P(\Phi/F_k)]$ is the probability that Φ will occur, regardless of which F_k occurs, we can write

$$P(F_j/\Phi) = \frac{P(\Phi/F_j) \cdot P(F_j)}{\sum_k [P(\Phi/F_k) \cdot P(F_k)]} \qquad (A)$$

showing that we can calculate the $P(F_j/\Phi)$'s if we know the $P(\Phi/F_j)$'s and also the "a priori" probabilities $P(F_j)$ of occurrences of X's of the different types. These are, of course, determined by the circumstances, and cannot be calculated from a theory of how the φ_j devices work.

The simple formula (A) is useful when one wants. so to speak, to invert the roles of "cause" and "effect" when one has either data or theory that goes in the wrong direction. We note that in many application one needs only to know the relative magnitudes of the different $P(F_j/\Phi)$'s, and since the denominators of (A) does not depend on j, one merely has to compute the value

of j for which

$$P(\Phi/F_j) \cdot P(F_j) \qquad (A')$$

is the largest.

1.3 Problems of implementing the probability approach

There are a variety of practical and theoretical objections to the use of formula (A). It is difficult, even in the simplest situations, to know enough about the $P(\Phi/F_j)$'s to use this method. One might be able to deduce them from a theoretical model of the situation, but then one would most likely be able to construct a more sophisticated decision method involving less computation. If there are many possible Φ's, then it becomes impractical to estimate the $P(\Phi,F)$'s on the basis of empirical observation. The amount of data would be too enormous. And, the system is completely unable to "guess" on Φ's it has not seen before. Finally, even if one had a complete catalog of the $P(\Phi,F)$'s, knowledge in this form is so free of structure that it would be very hard to adapt it to a similar, new situation. To be useful, knowledge has to be cast into structured models.

The simplest and most common kinds of models are the "parametric distributions." In this family of techniques--which include many standard methods of statistics, one has a procedure in which one

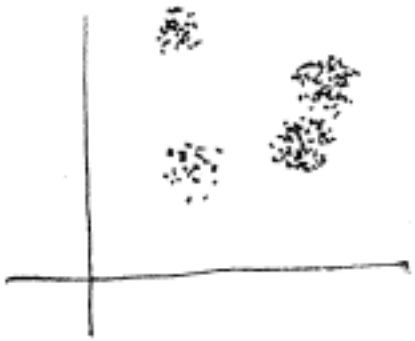
(a) assumes some form for the distributions of the(Φ/F)'s,

(b) fits the data to estimate the parameters of these distributions,

(c) designs a decision procedure based on the theory of the assumed distribution forms.

Example: One thinks of the $\Phi = (\varphi_1, \ldots, \varphi_m)$ as points in a vector space with the usual Cartesian distance metric.

Assume that each set $\{\Phi(X) \,|\, X \in F_j\}$ forms a "cluster" with (say) a symmetric normal distribution.



Each cluster has (say) the same variance (concentration) but different means (centers).

Then one can use the data to estimate the variance and means.

In this model, the decision of which F_j a given Φ should be assigned to can be made by finding which F-mean is closest to Φ. This, and many other related strategies are discussed at length in Nilsson's "Learning Machines," a book on the theory of a variety of statistical and threshold decision methods.

The trouble with the parametric models, and their relatives, is that even those that are the most sophisticated contain so little structure that they are usually thoroughly unsuitable for representing detailed knowledge about anything. (For example, they cannot satisfactorily represent finite-state processes.)

Nevertheless, we shall proceed to study the simplest such model, in which the φ's are statistically independent. Our conviction is that unless

this simple "linear" case is thoroughly understood, one can have little chance of making good "intuitive" judgements about more complicated systems.

1.4 Independence

We can evade some of the problems mentioned in §1.3 if we can assume that the tests $\varphi_i(X)$ are statistically independent for each F_j. Mathematically this means that for any sequence $\Phi(X) = (\varphi_1(X),\ldots,\varphi_m(X))$ of values of Φ we can assert that

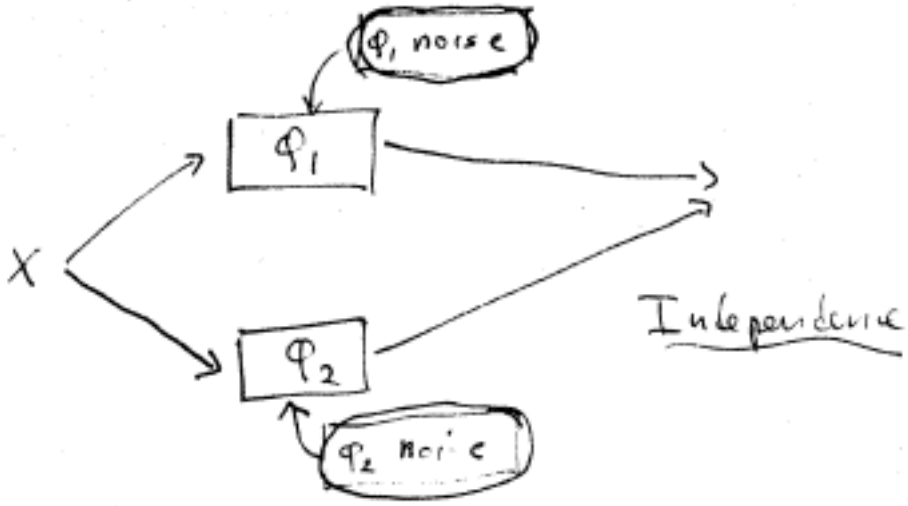
$$P(\varphi_1(X) \wedge \cdots \wedge \varphi_m(X)/F_j) = P(\varphi_1(X)/F_j) \cdot \ldots \cdot P(\varphi_m(X)/F_j)$$

for each j. More compactly we say

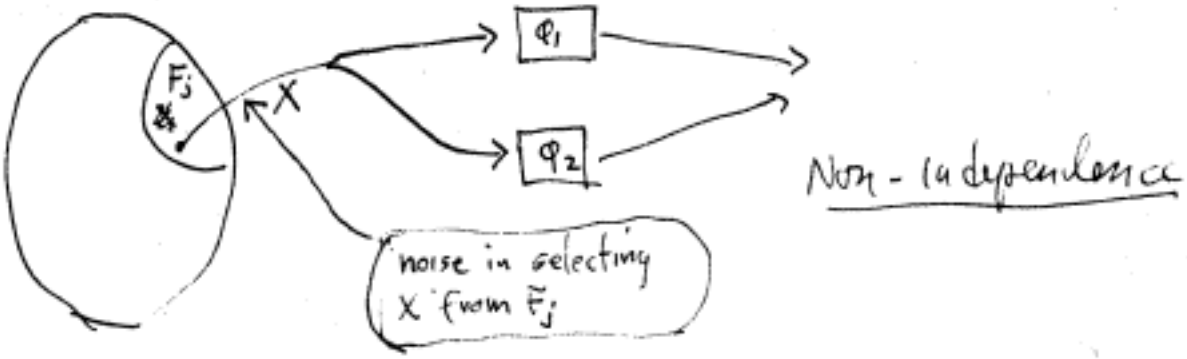
$$P(\Phi/F_j) = \prod_k P(\varphi_k/F_j). \qquad (B)$$

Informally, given that F_j has occurred, $P(\varphi_k/F_j)$ gives the distribution of each of the φ's. If one is further told the values of some of the φ's, independence means that this gives absolutely no further information about the values of the remaining φ's! We want to emphasize that this is a most stringent conditio

Experimentally one might get independence when there are variations in responses of φ's because of "noise" or measurement uncertainties within the individual φ-mechanisms. To the extent that this is why the φ's fluctuate, one would not then expect the values of some φ's to help predict those of others.



But where the variations in a φ, given F_j, are due to selection of different X's from the same F-class, one would not ordinarily assume independence, since the value of one of the φ's tells something about which X in F has occurred, and hence could help at least partly to predict how another φ will behave.



An extreme example of non-independence is the following, in which there are two functions, φ_1 and φ_2, and two classes, F_1 and F_2.

$\varphi_1(X)$ = a pure random variable with $P(\varphi_1(X) = 1) = \frac{1}{2}$.
Its value is determined by tossing a coin, not by X.

$$\varphi_2(X) = \begin{cases} \varphi_1(X) \text{ if } X \epsilon F_1 \\ 1 - \varphi_1(X) \text{ if } X \epsilon F_2. \end{cases}$$

Then $P(\varphi_1 \wedge \varphi_2 / F_1) = \frac{1}{2}$.

But $P(\varphi_1 / F_1) = P(\varphi_2 / F_1) = \frac{1}{2}$

hence $P(\varphi_1 / F_1) \cdot P(\varphi_2 / F_1) = \frac{1}{4}$.

Notice that <u>neither</u> φ_1 <u>nor</u> φ_2 <u>taken alone give any information whatever about F!</u> Each appears to be a random coin toss. But from both one can determine perfectly which F has occurred, for

while $\varphi_1 = \varphi_2 \Longrightarrow F_1$,
$\varphi_1 \neq \varphi_2 \Longrightarrow F_2$

with absolute certainity.

Remark: We will assume only independence within each class F_j. If X ranges over several F's then knowing one φ-value can help guess another. For example, suppose that

$$\varphi_1 = \varphi_2 = 0 \text{ if } X \epsilon F_1$$
$$\varphi_1 = \varphi_2 = 1 \text{ if } X \epsilon F_2.$$

The two φ's can (in fact are) independent on each F. But if we did not know that $X \epsilon F_1$ and we then told that $\varphi_1 = 0$, we could indeed then predict that $\varphi_2 = 0$ also, without this violating our independence assumption. If we had been told that $X \epsilon F_1$, we could have already predicted the value of φ_2, and in that case learning the value of φ_1 would not have helped!

2.0 The linear maximum likelihood estimator

We will assume that the φ's are independent (for each F_j). Suppose that we have just observed a case in which $\Phi = (\varphi_1, \ldots, \varphi_m)$ and we want to know which F_j has the greatest probability: Which is largest of

$$P(F_1/\Phi), \ldots, P(F_n/\Phi)?$$

Now, using formula (A) of §1.2, this is equivalent to choosing the j for which $P(\Phi/F_j) \cdot P(F_j)$ is largest. For brevity, define

$$p_{ij} \equiv P(\varphi_i = 1/F_j)$$

and

$$p_j \equiv P(F_j).$$

We will discuss only Boolean φ's; that is, each φ has value 0 or 1.

Then, using (B) of §1.4, we can define

$$p_j \cdot \prod_{\varphi_i=1} p_{ij} \cdot \prod_{\varphi_i=0} (1-p_{ij})$$

as the quantity to be maximized. It is formally convenient to multiply and divide by $\prod_{\varphi_i=1} (1-p_{ij})$ to obtain

$$p_j \cdot \prod_{\varphi_i=1} \left(\frac{p_{ij}}{1-p_{ij}}\right) \cdot \prod_{\text{all } i} (1-p_{ij})$$

for then we have only to compute the j that maximizes

$$B_j \cdot \prod_{\varphi_i=1} \left(\frac{p_{ij}}{1-p_{ij}}\right) \qquad \text{(C)}$$

where B_j is a constant that does not depend on $\underline{\varphi}$ so that the influence of the actual experiment is concentrated in the product expression. The terms

$$\left(\frac{p_{ij}}{1-p_{ij}}\right)$$

are the "odds" or "likelihood ratios" of getting $\varphi_i = 1$ if $X \in F_j$.

It is formally convenient now to replace (C) by its logarithm because <u>sums</u> are more familiar than products. This changes only the form, not the content; since $\log(x)$ increases when x does, we still select the maximum of

$$b_j + \sum_{\varphi_i=1} \log \left(\frac{p_{ij}}{1-p_{ij}}\right) \qquad (D)$$

where $b_j = \log \beta_j$. The term

$$w_{ij} \equiv \log \left(\frac{p_{ij}}{1-p_{ij}}\right)$$

is aptly called* the "weight of the evidence" of φ_i in favor of F_j. Now we can write

$$b_j + \sum_{\varphi_i=1} w_{ij}$$

$$= b_j + \sum_{\text{all } i} w_{ij}\varphi_i$$

where all that does not depend on the outcome of the actual experiment Φ is absorbed into the constant b_j which is therefore a sort of "a priori" weight for F_j (as opposed to the $\sum_j w_{ij}\varphi_i$ term which is "a posteriori").

The important conclusion is that the decision can be made upon the basis of a <u>linear</u> combination of the terms. This is due directly to the assumption we made about the independence of the p_{ij}'s for each j.

There are slight assymmetries in our formula that come from treating $\varphi_i=1$ as an occurrence of an event and $\varphi_i=0$ as a non-occurrence of the same event. This is quite arbitrary, and makes

---

* by I. J. Good

it hard to return to the more general situation in which the epxeriments $\{\varphi_i\}$ could each have many values. Besides, our algebra introduces a quite unnecessary risk of dividing by zero! But on the whole, we gain an heuristically clearer picture and, with this insight, it would be easy enough to return to the original formulae for repairs.

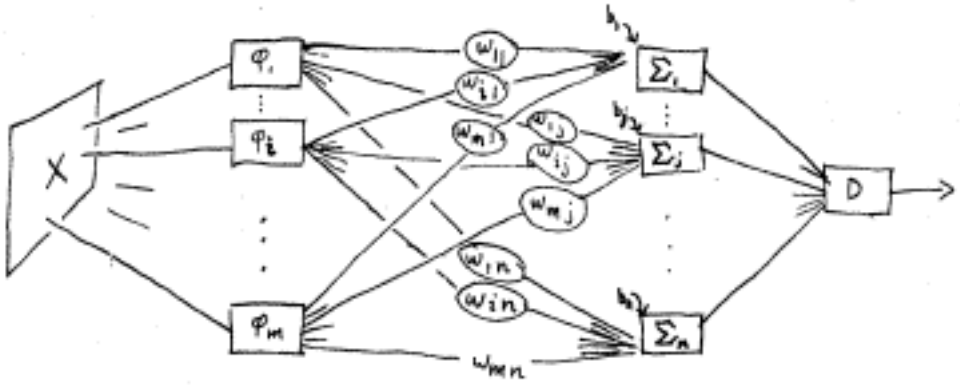
Formally, one can go slightly further in simplifying the formulae. Create a new "experiment" φ_0 wh. value is always 1. Define $w_{0j} = b_j$. Then if we re-define the vector Φ to be $(\varphi_0, \varphi_1, \ldots, \varphi_m)$ and define vectors $W_j = (w_{0j}, w_{1j}, \ldots, w_{mj})$ our procedure is: choose that j for which

$$W_j \cdot \Phi$$

is maximal. Later we will give a more-or-less meaningful geometric interpretation to this vector product formalism, but right here it is not very illuminating.

2.1 Layer-Machines

Formula (D) of §2.0 suggests the design of a machine for making our decision:

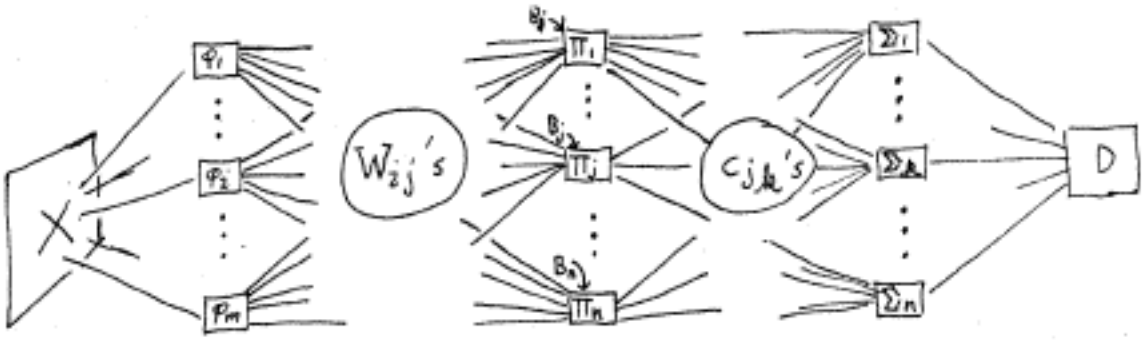


where D is a device that simply decides which of its inputs is the largest. Each φ-device emits a standard-sized pulse (if $\varphi(X) = 1$) when X is presented. The pulses are multiplied by the w_{ij} quantities as indicated, and summed at the Σ-boxes.

Returning for the moment to <u>costs</u>, if we combine the observations of §1.1 and §2.0 we will want to minimize (for k)

$$\sum_j c_{jk} \cdot B_j \cdot \prod_{\varphi_i=1} W_{ij}$$

where $W_{ij} = p_{ij}/(1-p_{ij})$. It is interesting that this more complicated procedure also lends itself to the layer-structure



2.1 The two-class case

If there are only two classes F_1 and F_2 the decision formula becomes very simple: we choose F_1 if

$$c_1 + \sum_i w_{i1}\varphi_i > c_2 + \sum_i w_{i2},$$

i.e., when

$$\sum_i (w_{i1} - w_{i2})\,\varphi_i > c_2 - c_1 \qquad (A)$$

which has the form

$$\sum_i \alpha_i\,\varphi_i > \theta. \qquad (B)$$

A decision function of this sort is called a "linear threshold function" because the choice depends upon whether the linear function $\Sigma\alpha_i\varphi_i$ exceeds the "threshold" θ. It is certainly of the simplest ways to make a decision that is not entirely trivial.

Notice that we derived the formula on the basis of some very strong assumptions--notably the independence of the p_{ij}'s. These will not be true in general, and indeed it will be exceptional that the conditions will be close enough to make the formula (A) useful. The formula (B) is somewhat more general--and therefore somewhat more widely useful in that, while it retains the linear threshold form, it does not require that α_{ij} be precisely

$$\log\left(\frac{p_{i1}}{p_{i2}}\frac{(1-p_{i2})}{(1-p_{i1})}\right),$$

and suggests that even if the statistical assumptions of independence do not hold, there might be other values of α_{ij} that could be used. This is sometimes true.

3.0 Estimating the p_{ij}'s

3.1 Laplace estimation

The most obvious way to estimate p_{ij} is to present some events in F_j and record the value of φ_i. If N events occur and we observe H occurrences of $\varphi_i = 1$ we can estimate that

$$p \cong \frac{H}{N}.$$

Naturally, this estimate will give different values for different samples; it has a statistical distribution. In fact, it will have a binomial distribution about the true mean, with variance

$$\sigma^2 = \frac{p(1-p)}{N}.$$

We will not derive this well-known result, but most readers will recognize that it is plausible because

(i) if n=1 the distribution has two points whose weights and distances from p give the variance

$$\sigma^2 = (1-p)p^2 + p(1-p)^2$$
$$= p(1-p)$$

(ii) one is used to the uncertainty decreasing with the square root of sample size,

(iii) the binomial distribution is so simple and fundamental that one would not expect any other factor to enter.

(end of mathematical joke)

To use this we have to keep a record of N. To "update" the estimate after each observation denote the N-th estimate of p by $p^{[N]}$:

If $\varphi = 1$ then

$$p^{[N+1]} = \frac{H+1}{N+1}$$

$$= \left(1 - \frac{1}{N+1}\right)\frac{H}{N} + \frac{1}{N+1}$$

while if $\varphi = 0$,

$$p^{[N+1]} = \left(1 - \frac{1}{N+1}\right)\frac{H}{N}.$$

Both can be summarized by

$$p^{[N]} = (1 - \tfrac{1}{N})p^{[N-1]} + \tfrac{1}{N}\varphi^{[N]}$$

where $\varphi^{[N]}$ is the value of φ at the N-th observation.
This suggests still another way to estimate p_{ij}! What would happen if we replaced the multipliers $(1 - \frac{1}{N})$ and $\frac{1}{N}$ by constants that don't depend on N--the number of trials? It is desirable to make these constants still add to unity, just as $\frac{1}{N}$ and $\left(1 - \frac{1}{N}\right)$ do, so that the estimated "probabilities" will also add to unity.

3.2 Reinforcement estimation

We are estimating the probability that $\varphi(X) = 1$. After each event $\varphi^{[t]}$ we revise our t-th estimate $p^{[t]}$ by the formula

$$p^{[t+1]} = \theta\, p^{[t]} + (1 - \theta)\, \varphi^{[t]} \qquad (R)$$

where θ is a constant between 0 and 1. For simplicity we begin by setting $p^{[0]} = \varphi^{[0]}$. By applying the formula repeatedly we obtain

$$\begin{aligned}
p^{[0]} &= \varphi^{[0]} \\
p^{[1]} &= \theta\varphi^{[0]} + (1 - \theta)\varphi^{[1]} \\
p^{[2]} &= \theta^2\varphi^{[0]} + (1 - \theta)[\varphi^{[2]} + \theta\varphi^{[1]}] \\
&\ldots \\
p^{[n]} &= \theta^n\varphi^{[0]} + (1 - \theta)[\varphi^{[n]} + \theta\varphi^{[n-1]} + \theta^2\varphi^{[n-2]} + \ldots + \theta^{n-1}\varphi^{[1]}]
\end{aligned}$$

which we can write as

$$\begin{aligned}
p^n &= (1 - \theta)\left[\varphi^{[n]} + \theta^1\varphi^{[n-1]} + \ldots + \theta^{n-1}\varphi^{[1]} + \frac{\theta^n}{1-\theta}\,\varphi^{[0]}\right] \\
&= \frac{1\,\varphi^{[n]} + \theta\varphi^{[n-1]} + \ldots + \theta^{n-1}\,\varphi^{[1]} + \frac{\theta^n}{1-\theta}\,\varphi^{[0]}}{1 + \theta + \ldots + \theta^{n-1} + \frac{\theta^n}{1-\theta}} \\
&= \frac{\sum_j c_j \varphi^{[j]}}{\sum_j c_j}
\end{aligned}$$

where the denominator is equal to $1/(1 - \theta)$. The last formula is written to show more clearly that $p^{[n]}$ can be expressed as a simple weighted sum, i.e., average of the $\varphi^{[t]}$'s. We have two reasons for writing the sum in this form:

(1) Exponential decay of "memory"

The formula shows that the value of α at time n is

(a) a weighted average of the $\varphi^{[t]}$'s

(b) the weights, i.e., the influence of older events falls off exponentially. For we can define

$$(1-\theta)\ \alpha^{[n]} \cong \sum_{0}^{\infty} \theta^{t}\ \varphi^{[n-t]}$$

and we can write

$$(1-\theta)\ \frac{\partial \alpha^{[n]}}{\partial \varphi^{[t]}} = \theta^{n-t}$$

where the derivative can be interpreted as showing how much $\alpha^{[n]}$ depends on the φ that occurred t moments before. The first term, $\varphi^{[0]}$, always retains slightly more weight (by a factor of $\frac{1}{1-\theta}$) but it, too, is subject to the same decay.

(2) The procedure is an estimator for p

This follows from the general theorem that if $x_1,\ldots,x_n$ are independent random variables and $E\langle X\rangle$ is the mean or "expected" value of a randon variable, then

$$E\langle \Sigma a_i x_i \rangle = \Sigma a_i E\langle x_i \rangle$$

for any set $\{a_i\}$ of constants. Since each φ_i had $E\langle\varphi\rangle = p$ we get

$$E\langle \alpha^n \rangle = \frac{\sum_j c_j \cdot p}{\sum_j c_j} = p.$$

The exponential form has advantages in some situations:

(1) We do not have to store the number N of trials, and

(2) The estimator lends itself to "adaptation" for changing situations.

On the other hand, it does not optimally use the experience in non-changing situations.

We can appraise the efficiency of (R) in using its data by computing its variance--the mean square error about p.

3.3 The variance of the reinforcement estimator (R).

Suppose that X has the distribution f(X) with mean μ and is subject to the transformation

$$X' = \theta X + (1 - \theta)\varphi$$

where prob(φ = 1) = p. Then the distribution of X' is

$$f'(X) = p\alpha(X) + (1 - p)\beta X$$

where

$$\alpha(X) = \frac{1}{\theta} f(1 - \frac{1-X}{\theta}) \text{ and } \beta(X) = \frac{1}{\theta} f(\frac{1}{\theta}). \qquad (I)$$

Then the variance of f'(X) is (as shown below)

$$\sigma_{f'}^2 = p\,\sigma_\alpha^2 + (1-p)\,\sigma_\beta^2 + p(1-p)(\mu_\alpha-\mu_\beta)^2. \qquad \text{(II)}$$

If the variance of f is σ_f^2 then from (I.) we have

$$\begin{cases} \sigma_\alpha^2 = \theta^2\sigma_f \\ \mu_\alpha = 1-\theta+\theta\mu_f \end{cases} \qquad \begin{cases} \sigma_\beta^2 = \theta^2\sigma_f \\ \mu_\beta = \theta\mu_f \end{cases}$$

hence

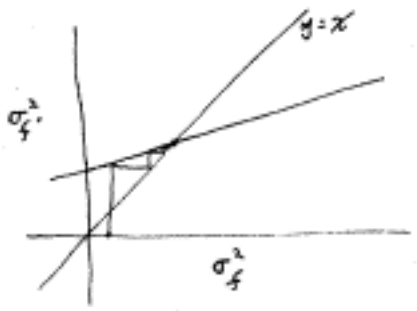
$$\sigma_{f'}^2 = \theta^2\sigma_f^2 + p(1-p)(1-\theta)^2.$$

This is linear in σ_f^2 with slope $\theta^2 < 1$, hence iteration must lead to a limit ("fixed") point for the variance. At this limit we must have

$$\sigma^2 = \theta^2\sigma^2 + p(1-p)(1-\theta)^2$$

so

$$\sigma_{\text{lim}}^2 = p(1-p)\,\frac{1-\theta}{1+\theta}.$$



Proof of (II):

$$\sigma_{f'}^2 = \int f'(X)(X-\mu_{f'})^2 = \int f'(X)X^2 - \mu_{f'}^2,$$

$$= p\int\alpha(X)X^2 + (1-p)\int\beta(X)X^2 - [p\mu_\alpha+(1-p)\mu_\beta]^2$$

$$= p[\int\alpha(X)X^2 - \mu_\alpha^2] + p\mu_\alpha^2 - p^2\mu_\alpha^2$$

$$+ (1-p)[\int\beta(X)X^2-\mu_\beta^2] + (1-p)\mu_\beta^2 - (1-p)^2\mu_\beta^2$$

$$- 2p(1-p)\mu_\alpha\mu_\beta$$

$$= p\sigma_\alpha^2 + (1-p)\sigma_\beta^2 + p(1-p)(\mu_\alpha-\mu_\beta)^2.$$

3.4 The role of θ as a memory time-constant

When the reinforcement process $\overset{(R)}{\wedge}$ has been in operation for a very long time, the variance of its expected value does not approach zero (as it does for the $\overset{a}{\wedge}$ Laplace estimate when N increases beyond bound). This is because it depends more upon the recent past than upon the remote past; indeed the very most recent term can alter the current value by more than $1-\theta$, a constant. (In the uniform weight process the effect is less than $\frac{1}{n}$ which becomes arbitrarily small.) If we "equate" the variances

$$\frac{p(1-p)}{n} \cong p(1-p)\,\frac{1-\theta}{1+\theta}$$

we get

$$n \sim \frac{1+\theta}{1-\theta}\,.$$

For small values of θ this is a little more than unity, showing (correctly) that the estimate is based almost entirely upon the last event. In fact, in this case

$$p^{[t]} = \theta p^{[t]} + (1-\theta)\varphi^{[t]} \sim \varphi^{[t]}.$$

For large θ, i.e., close to unity, we have

$$n \sim \frac{2}{1-\theta}$$

so that if $\theta = 1 - \frac{1}{m}$ then the variance is about that one would obtain by simple averaging of the last 2m samples! Thus one can think of the quantity $\frac{1}{1-\theta}$ as corresponding roughly to a time constant for "forgetting."

3.5 The Samuel compromise

In his classical paper about "Some Studies in Machine Learning using the Game of Checkers," Arthur L. Samuel uses an ingenious combination of probability estimation methods. In his application it occasionally happens that a new evidence term φ_i is introduced (and an old one is abandoned because it has not been of much value in the decision process). When this happens there is a problem of preventing violent fluctuations, because after one or a few trials the term's probability estimate will have a large variance as compared with older terms that have better statistical records. Samuel uses the following algorithm to "stabilize" his system: he sets $p^{[0]} = \frac{1}{2}$ and

$$p^{[t+1]} = (1 - \frac{1}{N}) p^{[t]} + \frac{1}{N} \varphi^{[t]}$$

where

$N = 16$	if	$t < 32$
$N = 2^n$	if	$2^n \leq t < 2^{n+1}$
$N = 256$	if	$256 \leq t$.

Thus, in the beginning the estimate is made as though the probability had already been estimated to be $\frac{1}{2}$ on the basis of several, i.e., the order of 16, trials. Then, in the "middle" period, the algorithm approximates the uniform wieghting procedure. Finally (when $t \sim 256$) the procedure changes to the exponential decay mode, with fixed N, so that recent experience can outweigh earlier results. The use of powers of two represents a convenient computer-program technique for dong this.

In Samuel's system, the terms actually used have the form

$$C^{[t]} = 2p^{[t]} - 1$$

so that the "estimator" ranges in the interval $-1 \leq C_n \leq +1$ and can be treated as a "correlation coefficient." I mention this here only to justify Samuel's initial setting of $p^{[0]}$ to $\frac{1}{2}$, i.e., to $C^{[0]}$ to 0. In his context, this setting makes perfect sense, whereas in our interpretation the setting of $p^{[0]}$ to $\frac{1}{2}$ would be arbitrary.

3.6 Variants of the reinforcement estimator

Consider the "reinforcement process"

$$\boxed{\alpha' = \theta\alpha + (1-\theta)\varphi} \qquad (R_1)$$

If the distribution of α has mean μ then the distribution of α' will be

$$\mu' = (1-p)\theta\mu + p(\theta\mu + (1-\theta))$$

because there is probability $(1-p)$ that $\varphi=0$ and probability p that $\varphi=1$. Then

$$\mu' = \theta\mu + (1-\theta)p.$$

Applying this again we get for the mean of $(\alpha')'$,

$$\mu'' = \theta(\theta\mu + (1-\theta)p) + (1-\theta)p$$
$$= \theta^2\mu + (1-\theta)p(1+\theta)$$
$$= \theta^2\mu + (1-\theta^2)p$$

and similarily, if we apply R_1 n times, we get

$$\mu^{[n]} = \theta^n\mu + (1-\theta^n)\cdot p$$

Clearly as $n \to \infty$, $\mu^{[n]} \to p$.

This analysis can be replaced by a more general method, by using the following two simple observations:

Lemma 1: If $\alpha' = f(\alpha,\varphi)$ is linear in φ then if $\mu(\alpha) = m$ and $p(\varphi=1)$ then

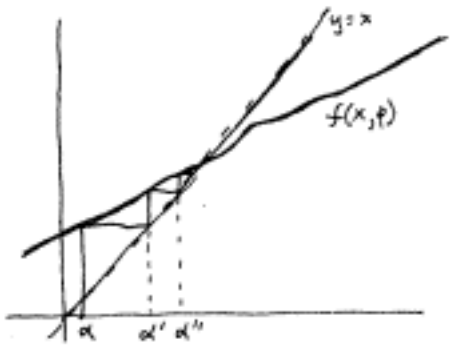
$$\mu(\alpha') = f(m,p).$$

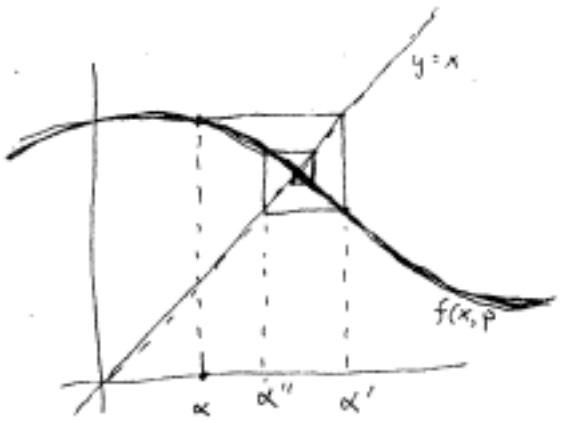
Proof: $\mu(\alpha') = (1-p)f(m,0) + p\cdot f(m,1)$
$= f(m,0) + p[f(m,1) - f(m,0)]$
$= f(m,p)$, because one can "interpolate" a linear function

Lemma 2: If $\left|\frac{\partial f(\alpha,\varphi}{\partial\alpha}\right| < 1$ then the limit of $\alpha, f(\alpha), f(f(\alpha)), \ldots f^{(n)}(\alpha)$ exists and is the (unique) solution of the equation

$$y = f(y,p).$$

Proof: This is a "fixed point theorem" and the diagrams below show why it is true:





Now we can apply these to the formula R_1, and have only to solve

$$y = \theta y + (1-\theta)p$$

or

$$y = \frac{1-\theta}{1-\theta}\, p = p.$$

which, as we already know is the mean limit of the means of the iterated process.

But with the lemmas, we can also analyse some other systems. Another interesting one is

$$\boxed{\alpha' = \theta(\alpha + \varphi)} \qquad (R_2)$$

and we have

$$y = \theta(y + p) \qquad \text{or} \qquad y = \frac{\theta}{1-\theta}\, p$$

so that this, too, is an estimator for p which has to be corrected by the constant factor $\frac{\theta}{1-\theta}$.

Another, somewhat different reinforcer is

$$\alpha' = \begin{cases} \alpha+1 & \text{if} \quad \varphi=1 \\ \theta\alpha & \text{if} \quad \varphi=0 \end{cases} \qquad (R_3)$$

which can be written as

$$\alpha' = \varphi + \alpha(\varphi + \theta - \varphi\theta).$$

This satisfies the conditions of the Lemmas, since

$$\left|\frac{\partial\alpha'}{\partial\alpha}\right| = |\varphi+\theta-\varphi\theta| = |1-(1-\varphi)(1-\theta)| < 1$$

so we have

$$y = p + y(p + \theta - p\theta)$$

or

$$y = \left(\frac{1}{1-\theta}\right)\left(\frac{p}{1-p}\right).$$

This is an estimator (with the $\frac{1}{1-\theta}$ correction factor) of the likelihood ratio $\left(\frac{p}{1-p}\right)$. It is interesting that this is so easily obtained by a reinforcement process as simple as: "if $\varphi = 1$ occurs, <u>add</u> 1 to α; if $\varphi = 0$ occurs, <u>multiply</u> α by θ"!

Another simple form is

$$\alpha' = \theta\alpha + \varphi \qquad (R_4)$$

which leads to the estimate

$$y = \frac{1}{1-\theta} \cdot p$$

Finally, one might consider the very simple form

$$\alpha' = \alpha + \varphi. \qquad (R_5)$$

This "diverges," i.e., the α's grow beyond bound (and do not satisfy the condition for Lemma 2). Still, if one is making decisions by comparing different p_{ij}'s, one can use the ratios of these simple "scores" as likelihood ratios, to obtain the "uniform weighting" type of behavior. We include R_5 only to indicate its heuristic similarity to the others.

3.7 A simple "synaptic" reinforcer theory

Let us make a simple "neuronal model." The model is to "account" for the following phenomena:

1. There is to be a quantity α that estimates $p_{ij} = P(\varphi_i/F_j)$;
2. The only information available are the occurrence of $\varphi_i = 1$ and $X \epsilon F_j$.

The bag B_1 contains a very high and constant concentration of a substance A. When φ_i or F_j occur--or "fire"--the walls of the corresponding bags B_i and/or C_j become "permeable" to A for a moment. If φ_i alone occurs, nothing really changes, because B_i is surrounded by the impermeable C_j. If F_j alone fires, it loses some A by diffusion to the outside; in fact, if α is the amount of A in C_j it may be assumed (by the usual laws of diffusion and concentration) to lose some fraction $(1-\theta)$ of α:

$$\alpha' = \theta\alpha \quad \text{if} \begin{cases} F_j \text{ occurs and} \\ \varphi_i = 0 \end{cases}$$

If <u>both</u> φ_i and F_j are active then approximately the same loss will occur from C_j. But an essentially constant amount <u>b</u> will be "injected" from B_i to C_j. So

$$\alpha' = \theta\alpha + b \quad \text{if} \begin{cases} F_j \text{ occurs and} \\ \varphi_i = 1 \end{cases}$$

We can assume that <u>b</u> is constant because the concentration of A is very high in B_i compared to that in C_j. Or one can invent any number of similar variations. In any case we get

$$\boxed{\alpha' = \theta\alpha + \varphi b}$$

so that in the limit the mean of α will approach

$$\frac{b}{1-\theta}\, p$$

which is proportional to, and hence an estimator of, $p_{ij} = \text{Prob}(\varphi_i / F_j)$.

Thus the simple geometry together with the idea of a membrane becoming permeable briefly following a nerve impulse gives us a quantity that is an estimator of the appropriate probability.

How could this representation of probability be translated into a useful neuronal mechanism? One could imagine all sorts of schemes: ionic concentrations--or rather, their logarithms!--could become membrane potentials, or conductivities, or even probabilities of occurrences of other chemical events. The "anatomy" and "physiology" of our model could easily be modified to obtain P processes and their attendant "likelihood ratios." Indeed, it is so easy to imagine variants--the idea is so insensitive to details--that I don't propose it to be considered seriously, except as a family of simple yet intriguing models that a neural theorist should have available.