Lecture 14

DECISION THEORY IN SIGNAL DETECTION

In this Lecture, I want to examine in detail one of the simplest applications of the general decision theory just formulated. As I pointed out in Lecture 6, the problem of detection of signals in noise is really exactly the same as Laplace's old problem of detecting the presence of unknown systematic influences in celestial mechanics, and Shewhart's (1931) more recent problem of detecting a systematic drift in machine characteristics, in industrial quality control. It is unfortunate that the basic identity of all these problems hasn't been more widely recognized, because it has forced workers in several different fields to rediscover the same things, with varying degrees of success, over and over again.

As you know by now, all we really have to do to solve this problem is to take the principles developed in Lectures 3, 10, and 12; and supplement them with the loss function criterion for converting final probabilities into decisions. However, the literature of this field has been largely created from the standpoint of the original decision theory before it was realized that it was mathematically identical with the original Laplace methods; or at least before the full implications of this fact had "sunk in." The existing literature therefore uses a different sort of vocabulary and set of concepts than I have been using up to now. Since it exists, we have no choice but to learn these terms and viewpoints if we want to read the literature of the field. So, I want to give you a very rapid, condensed review of the literature of the 1950's on these problems. My aim is to expose what is really essential, stripped of all unnecessary details. This material is also given in the papers of Middleton and van Meter (1955, 1956) and the treatise of Middleton (1960), in an enormously expanded form where a beginner can get lost for months without ever finding the real underlying principles. Just to have a complete, self-contained summary, I'll repeat a little bit from previous lectures.

14.1. Definitions and Preliminaries.

Notation:

$(A|B)$ = Conditional probability of A, given B

$(AB|CD)$ = Joint conditional probability of A and B, given D and C, . . ., etc.

For our purposes, everything follows from the single fundamental rule of calculation, which we have called Rule 1:

$$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{14-1}$$

If the propositions B, C are not mutually contradictory, this may be rearranged to give the rule of "learning by experience," Bayes' theorem:

$$(A|BC) = (A|C)\,\frac{(B|AC)}{(B|C)} = (A|B)\,\frac{(C|AB)}{(C|B)}\;. \tag{14-2}$$

If there are several mutually exclusive and exhaustive propositions B_i, then by summing (14-1) over them, we obtain the chain rule

$$(A|C) = \sum_i (A|B_iC)(B_i|C) \tag{14-3a}$$

or, for a simpler notation,

$$(A|C) = \sum_B (A|BC)(B|C) \tag{14-3b}$$

Now let

X = prior knowledge, of any kind whatsoever

S = signal

N = noise

$V = V(S,N)$ = observed voltage

D = decision about the nature of the signal

Any probabilities conditional on X alone are called prior probabilities. Thus we have

$(S|X)$ = prior probability of the particular signal S

$(N|X) = W(N)$ = prior probability of the particular sample of noise N.

In a linear system, $V = S + N$, and

$$(V|S) = (V|SX) = W(V - S) \quad . \qquad (14\text{-}4)$$

You may be disturbed by the absence of density functions, dS's, dN's, etc., which might be expected in the case of continuous S, N. Note, however, that our equations are homogeneous in these quantities, so they cancel out anyway. By $\sum_A$ I mean ordinary summation over some previously agreed set of possible values if A is discrete, integration with appropriate density functions if A is continuous.

A decision rule $(D_i|V_j)$, or for brevity just $(D|V)$, represents the process of drawing inferences about the signal from the observed voltage. If it is always made in a definite way, then $(D|V)$ has only the values 0, 1 for any choice of D and V; however we may also have a "randomized" decision rule according to which $(D|V)$ is a true probability distribution. Maintaining this more general view turns out to be a help in formulating the theory.

The essence of any decision rule, and in particular, any one which can be built into automatic equipment, is that the decision must be made on the basis of V alone; V is, by definition, the quantity which contains all the information actually used (in addition to the ever-present X) in arriving at the decision. Thus, if $Y \neq D$ is any other proposition, we have

$$(D|V) = (D|VY) \quad . \qquad (14\text{-}5)$$

An equivalent statement is that D depends on any proposition Y only through the intermediate influence of Y on V:

$$(D|Y) = \sum_V (D|V)(V|Y) \qquad (14\text{-}6)$$

14.2. Sufficiency and Information.

Equation (14-5) has interesting consequences; suppose we wish to judge the plausibility of some proposition Y, on the basis of knowledge of V and D. From (14-1),

$$(DY|V) = (Y|VD)(D|V) = (D|VY)(Y|V)$$

and using (14-5), this reduces to

$$(Y|VD) = (Y|V) \qquad (14\text{-}7)$$

Thus, if V is known, knowledge of D is redundant and cannot help us in estimating any other quantity. The reverse is not true, however; we could equally well use (17-1) in another way:

$$(VY|D) = (Y|VD)(V|D) = (Y|D)(V|YD).$$

Combining this with (14-7), there results the

Theorem: Let D be a possible decision, given V. Then $(V|D) \neq 0$, and

$$(Y|V) = (Y|D) \text{ if and only if } (V|D) = (V|YD) \qquad (14\text{-}8)$$

In words: knowledge of D is as good as knowledge of V for judgments about Y if and only if Y is irrelevant for judgments about V, given D. Stated differently: in the "environment" produced by knowledge of D, the propositions Y and V appear to be independent, i.e.

$$(YV|D) = (Y|D)(V|D) \qquad (14\text{-}9)$$

In this case, D is said to be a sufficient statistic for judgments about Y. In the next lecture, we will study the notion of sufficiency from a different point of view. Evidently, a decision rule which makes D a sufficient statistic for judgments about the signal S is in some sense superior to one without this property. However, such a rule does not necessarily exist. Equation

(14-9) is a very restrictive condition, since it must be satisfied for all values of Y, V, and all D for which $(D|V) \neq 0$.

As you might guess from this, the concept of sufficiency is closely related to that of information. The definition of sufficiency could equally well be stated as: D is a sufficient statistic for judgments about Y if it contains all the information about Y which V contains. Since D is determined from V, if it is not a sufficient statistic, it necessarily contains less information about Y than does V. In this statement, the term "information" was used in a loose, intuitive sense; does it remain true if we adopt Shannon's measure of information? Imagine that there are several mutually exclusive propositions Y_i, one of which must be true. For brevity we use, as above, the notation $\sum_Y f(Y) \equiv \sum_i f(Y_i)$. Then the entropy of Y with a specific value of D given is

$$H_D(Y) = - \sum_Y (Y|D) \log (Y|D) \qquad (14\text{-}10)$$

and its average over all values of D is

$$\overline{H}_D(Y) = \sum_D (D|X) \; H_D(Y) \qquad (14\text{-}11)$$

If

$$\overline{H}_C(Y) < \overline{H}_D(Y)$$

we say that C contains, on the average, more information about Y than does D. Note, however, that it may be otherwise for specific values of C and D.

Acquisition of new information can never increase $\overline{H}$; let D, V, Y be, for the moment, any three quantities and form the expression

$$\overline{H}_V(Y) - \overline{H}_{DV}(Y) = \sum_{DVY} (DV|X)(Y|DV) \log (Y|DV)$$

$$- \sum_{VY} (V|X)(Y|V) \log (Y|V)$$

$$= \sum_{DVY} (DV|X)(Y|DV) \log [(Y|DV)/(Y|V)]$$

Using the by now familiar fact that $\log x \geq (1 - x^{-1})$, with equality if and only if $x = 1$, this becomes

$$\overline{H}_V(Y) - \overline{H}_{DV}(Y) \geq \sum_{DVY} (DV|X)[(Y|DV) - (Y|V)] = 0 \qquad (14\text{-}13)$$

Thus, $\overline{H}_{DV}(Y) \leq \overline{H}_V(Y)$, with equality if and only if Eq. (14-7) holds for all D, V, and Y for which $(DV|X) \neq 0$. Since (14-13) holds regardless of the meaning of D and V, we can equally well conclude that for all D, V, Y,

$$\overline{H}_D(Y) \geq \overline{H}_{DV}(Y) \leq \overline{H}_V(Y) \ .$$

Now letting D, V, Y resume their original meanings, we have in consequence of (14-7) $H_V(Y) = H_{DV}(Y)$, so that

$$\overline{H}_V(Y) \leq \overline{H}_D(Y) \qquad (14\text{-}14)$$

with equality if and only if Eq. (14-9) holds, i.e. if and only if D is a sufficient statistic. Thus, if by "information" we mean minus the average entropy of Y over the prior distribution of D or V, zero information loss in going from V to D is equivalent to sufficiency of D. Note that inequalities of the form (14-13) hold only for the averages $\overline{H}$, not for the H. Acquisition of a specific piece of information (that an event previously considered improbable had in fact occurred) may in some cases increase the entropy of Y. However, this is an improbable situation and on the average the entropy can only be lowered by additional information. This shows again that the term "information" is not a happy choice of word to describe entropy expressions. In spite of the entropy increase, the situation just described could hardly be called one of less information, but rather one of less certainty.

14.3. Loss Functions and Criteria of Optimum Performance.

In order to say that one decision rule is better than another, we need some specific criterion of what we want our detection system to accomplish. The criterion will vary with the application, and obviously no single decision rule can be best for all purposes. A very general type of criterion is obtained by assigning a loss function $L(D,S)$ which represents our judgment of how serious it is to make decision D when signal S is in fact present. In case there are only two possible signals; $S_o = 0$ (i.e. no signal), and $S_1 \neq 0$, and consequently two possible decisions D_o, D_1, there are two types of error, the false alarm $A = (D_1,S_o)$ and the false rest $R = (D_o,S_1)$. In some applications, one type of error might be much more serious than the other. Suppose that a false rest is considered ten times as serious as is a false alarm, while a correct decision of either type represents no "loss." We could then take $L(D_o,S_o) = L(D_1,S_1) = 0$, $L(D_o,S_1) = 10$, $L(D_1,S_o) = 1$. Whenever the possible signals and the possible decisions form discrete sets, the loss function becomes a loss matrix. In the above example,

$$L_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

The loss matrix plays approximately the same role in detection theory as does the payoff matrix in game theory. A player in a game may choose that strategy which maximizes his expected gain, and correspondingly we may choose that decision rule $(D|V)$ which minimizes the expected loss.

Instead of assigning arbitrarily a certain loss value to each possible type of detection error, we may consider information loss by the assignment $L(D,S) = -\log (S|D)$. This is somewhat more difficult to manipulate, because now $L(D,S)$ depends on the decision rule. A decision rule which minimizes information loss is one which makes the decision in some sense as close as possible to being a sufficient statistic for judgments about the signal. In exactly what sense seems never to have been clarified.

The conditional loss $L(S)$ is the average loss incurred when the specific signal S is present

$$L(S) = \sum_D L(D,S)(D|S) \qquad (14\text{-}15)$$

which may in turn be expressed in terms of the decision rule and the properties of the noise by using (14-6). The average loss is the expected value of this over all possible signals:

$$<L> = \sum_S L(S)(S|X) \tag{14-16}$$

Two different criteria of optimum performance now suggest themselves:

The Minimax Criterion. For a given decision rule $(D|V)$, consider the conditional loss $L(S)$ for all possible signals, and let $[L(S)]_{max}$ be the maximum value attained by $L(S)$. We seek that decision rule for which $[L(S)]_{max}$ is as small as possible. As we noted in the last lecture, this criterion concentrates attention on the worst possible case regardless of the probability of occurrence of this case, and it is thus in a sense the most conservative one. If the worst possible case is extremely unlikely to arise, one would call it too conservative. It has, however, the practical advantage that it does not involve the prior probabilities of the different signals, $(S|X)$, and therefore it can be applied in cases where the available information about the signal is of such an indefinite type that we do not know what prior probabilities to assign.

The Bayes Criterion. We seek that decision rule for which the average loss $<L>$ is minimized. In order to apply this, a prior distribution $(S|X)$ must be available.

Other criteria were proposed before the days of Decision Theory. In the Neyman-Pearson criterion, we fix the probability of occurrence of one type of error at some small value δ, and then minimize the probability of another type of error subject to this constraint. Siegert's "Ideal Observer" minimizes the total probability of error regardless of type. However, we will see below that these are both special cases of the Bayes criterion, for particular loss functions $L(D,S)$. The minimax criterion may also be considered a special case of the Bayes, in which we choose the worst possible $(S|X)$, after having found the decision rule which minimizes $\langle L\rangle$ for a given $(S|X)$. The basic identity of all these criteria came as quite a surprise to the early workers in this field.

Substituting in succession equations (14-15), (14-6), and (14-3) into (14-16), we obtain for the average loss

$$\langle L\rangle = \sum_{DV} \left[\sum_{S} L(D,S)(VS|X) \right] (D|V) \tag{14-17}$$

If $L(D,S)$ is a definite function independent of $(D|V)$ (this assumption excludes for the moment the information loss function), there is no function $(D|V)$ for which this expression is stationary in the sense of calculus of variations. We then minimize $\langle L\rangle$ merely by choosing for each possible V that decision $D_1(V)$ for which

$$K(D_1,V) \equiv \sum_{S} L(D_1,S)(VS|X) \tag{14-18}$$

is a minimum. Thus, we adopt the decision rule

$$(D|V) = \delta(D,D_1). \tag{14-19}$$

In general there will be only one such D_1, and the best decision rule is nonrandom. However, in case of "degeneracy," $K(D_1,V) = K(D_2,V)$, any randomized rule of the form

$$(D|V) = a\ \delta(D,D_1) + b\ \delta(D,D_2) \quad , \quad a + b = 1 \tag{14-20}$$

is just as good. This degeneracy occurs at "threshold" values of V, where we change from one decision to another.

14.4. A Discrete Example.

Consider the case already mentioned, where there are two possible signals S_0, S_1, and a loss matrix

$$L_{ij} = \begin{pmatrix} L_{00} & L_{01} \\ L_{10} & L_{11} \end{pmatrix} = \begin{pmatrix} 0 & L_r \\ L_a & 0 \end{pmatrix}$$

where L_a, L_r are the losses incurred by a false alarm and a false rest, respectively. Then

$$\begin{aligned} K(D_0,V) &= L_{01}(VS_1|X) = L_r(VS_1|X) \\ K(D_1,V) &= L_{10}(VS_0|X) = L_a(VS_0|X) \end{aligned} \tag{14-21}$$

and the decision rule that minimizes <L> is

$$\left.\begin{array}{l} \text{Choose } D_1 \text{ if } \dfrac{(VS_1|X)}{(VS_0|X)} > \dfrac{L_a}{L_r} \\[2ex] \text{Choose } D_0 \text{ if } \dfrac{(VS_1|X)}{(VS_0|X)} < \dfrac{L_a}{L_r} \\[2ex] \text{Choose either at random in case of equality.} \end{array}\right\} \tag{14-22}$$

In words: if the prior probability that the observed voltage is due to the signal exceeds the probability that it is due to noise alone by a factor greater than the ratio of false alarm loss to false rest loss, we decide that the signal is present. If the prior probabilities of signal and no signal are

$$(S_1|X) = p, \qquad (S_0|X) = q = 1 - p \tag{14-23}$$

respectively, we have $(VS_1|X) = (V|S_1)(S_1|X) = p(V|S_1)$, etc., and the decision rule becomes

$$\text{Choose } D_1 \text{ if } \frac{(V|S_1)}{(V|S_0)} > \frac{qL_a}{pL_r}\text{ , etc.} \tag{14-24}$$

The left-hand side of (14-24) is called a likelihood ratio. It depends only on the statistical properties of the noise, and is the quantity which should be computed by the optimum receiver according to the Bayes criterion. The same quantity is the essential one regardless of the assumed loss function and regardless of the probability of occurrence of the signal; these affect only the threshold of detection. Furthermore, if the receiver merely computes this likelihood ratio and delivers it at the output without making any decision, it provides us with all the information we need to make optimum decisions in the Bayes sense. Note particularly the generality of this result, which is one of the most important ones for our applications; no assumptions are needed as to the type of signal, linearity of the system, or statistical properties of the noise.

We now work out, for purposes of illustration, the decision rules and their degree of reliability, for several of the above criteria, in the simplest possible problem that I mentioned back in Lecture 4, to illustrate the principle of maximum likelihood. We have a linear system in which the voltage is observed at a single instant, and we are to decide whether a signal, which can have only amplitude S_1, is present in noise, which is gaussian with mean square value $<N^2>$:

$$W(N) = \frac{1}{\sqrt{2\pi<N^2>}} \exp\left[- \frac{N^2}{2<N^2>} \right] \tag{14-25}$$

The likelihood ratio in (14-24) then becomes

$$\frac{(V|S_1)}{(V|S_0)} = \frac{W(V-S_1)}{W(V)} = \exp\left[\frac{2VS_1-S_1{}^2}{2<N^2>} \right] \tag{14-26}$$

and since this is a monotonic function of V, the decision rule can be written as

$$\text{choose} \begin{Bmatrix} D_1 \\ D_0 \end{Bmatrix} \quad \text{when } V \begin{Bmatrix} > \\ < \end{Bmatrix} V_b \tag{14-27}$$

with

$$\frac{V_b}{\sqrt{<N^2>}} = \frac{1}{2s} \; [2 \; \log \left(\frac{qL_a}{pL_r} \right) \; + s^2] = v_b \tag{14-28}$$

in which

$s \equiv \frac{S_1}{\sqrt{<N^2>}}$ is the voltage signal-to-noise ratio, and

$v \equiv \frac{V}{\sqrt{<N^2>}}$ is the normalized voltage.

Now we find for the probability of a false rest:

$$(R|X) = (D_0S_1|X) = p \sum_V (D_0|V)(V|S_1) = p \int_{-\infty}^{V_b} dV\, W(V-S_1)$$

$$= p\,\Phi(v_b - s\ ) \tag{14-29}$$

and for a false alarm,

$$(A|X) = (D_1S_0|X) = q \sum_V (D_1|V)(V|S_0) = q \int_{V_b}^{\infty} dV\, W(V)$$

$$= q[1 - \Phi(v_b)] \quad . \tag{14-30}$$

Here $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{14-31}$$

numerical values of which are given in most mathematical tables. For $x > 2$, a good approximation is

$$1 - \Phi(x) \simeq \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \quad . \tag{14-32}$$

As a numerical example, if $L_r = 10\ L_a$, $q = 10\ p$, these expressions reduce to

$$(A|X) = 10\ (R|X) = \frac{10}{11}\ [1 - \Phi\left(\tfrac{1}{2}\ s\right)] \tag{14-33}$$

The probability of a false alarm is less than 0.027, and of a false rest less than 0.0027 for $s > 4$. For $s > 6$, these numbers become 1.48×10^{-3}, 1.48×10^{-4} respectively.

Let us see what the minimax criterion would give in this problem. The conditional losses are

$$\begin{aligned} L(S_0) &= L_a \sum_V (D_1|V)(V|S_0) = L_a \int_{-\infty}^{\infty} (D_1|V)\ W(V)\ dV \\ L(S_1) &= L_r \sum_V (D_0|V)(V|S_1) = L_r \int_{-\infty}^{\infty} (D_0|V)\ W(V-S_1)\ dV \end{aligned} \tag{14-34}$$

Writing $f(V) \equiv (D_1|V) = 1 - (D_0|V)$, the only restriction on $f(V)$ is $0 \le f(V) \le 1$. Since L_a, L_r, and $W(V)$ are all positive, a change $\delta f(V)$ in the neighborhood of any given point V will always increase one of the quantities (14-34) and decrease the other. Thus when the maximum L(S) has been made as small as possible, we will certainly have $L(S_0) = L(S_1)$, and the problem is thus to minimize $L(S_0)$ subject to this constraint. Suppose that for some particular $(S|X)$ the Bayes decision rule happened to give $L(S_0) = L(S_1)$. Then this particular solution must be identical with the minimax solution, for with the above constraint, $\langle L\rangle = [L(S)]_{max}$, and if the Bayes solution minimizes $\langle L\rangle$ with respect to all admissible variations $\delta f(V)$ in the decision rule, it a fortiori minimizes it with respect to the smaller class of variations which keep $L(S_0) = L(S_1)$. Therefore our optimum decision rule will have the same form as before: There is some threshold V_m such that

$$f(V) = \begin{cases} 0, & V < V_m \\ 1, & V > V_m \end{cases} \qquad (14\text{-}36)$$

Any change in V_m from the value which makes $L(S_0) = L(S_1)$ necessarily increases one or the other of these quantities. The equation determining V_m is therefore

$$L_a \int_{V_m}^{\infty} W(V)\ dV = L_r \int_{-\infty}^{V_m} W(V-S_1)\ dV$$

or, in terms of normalized quantities,

$$L_a[1 - \Phi(v_m)] = L_r\ \Phi(v_m - s) \qquad (14\text{-}37)$$

Note that (14-30), (14-31) give the conditional probabilities of false rest and false alarm for any decision rule of type (14-36), regardless of whether the threshold was determined from (14-28) or not; for the arbitrary threshold V_0

$$\begin{aligned} (R|S_1) &= (V < V_0|S_1) = \Phi(v_0 - s) \\ (A|S_0) &= (V > V_0|S_0) = \frac{1}{2}\,[1 - \Phi(v_0)] \end{aligned} \qquad (14\text{-}38)$$

From (14-28) we see that there is always a particular ratio (p/q) which makes the Bayes threshold V_b equal to the minimax threshold V_m. For values of (p/q) other than this worst value, the Bayes criterion gives a lower average loss than does the minimax, although one of the conditional losses $L(S_0)$,

$L(S_1)$ will be greater than the minimax value.

These relations and several previous remarks are illustrated in Figure (14.1), in which we plot the conditional losses $L(S_0)$, $L(S_1)$ and the average loss $\langle L\rangle$ as functions of the threshold V_0, for the case $L_a = \frac{3}{2} L_r$, $p = q = \frac{1}{2}$. The minimax threshold is at the common crossing-point of these curves, while the Bayes threshold occurs at the lowest point of the $\langle L\rangle$ curve. One sees how the Bayes threshold moves as the ratio (p/q) is varied, and in particular that the value of (p/q) which makes $V_b = V_m$ also leads to the maximum values of the $\langle L\rangle_{min}$ obtained by the Bayes criterion. Thus we could also define a "maximin" criterion; first find the Bayes decision rule which gives minimum $\langle L\rangle$ for a given $(S|X)$, then vary the prior probabilities $(S|X)$ until the maximum value of $\langle L\rangle_{min}$ is attained. This is the worst possible (in the Bayes sense) prior probability, and the decision rule thus obtained is identical with the one resulting from the minimax criterion.

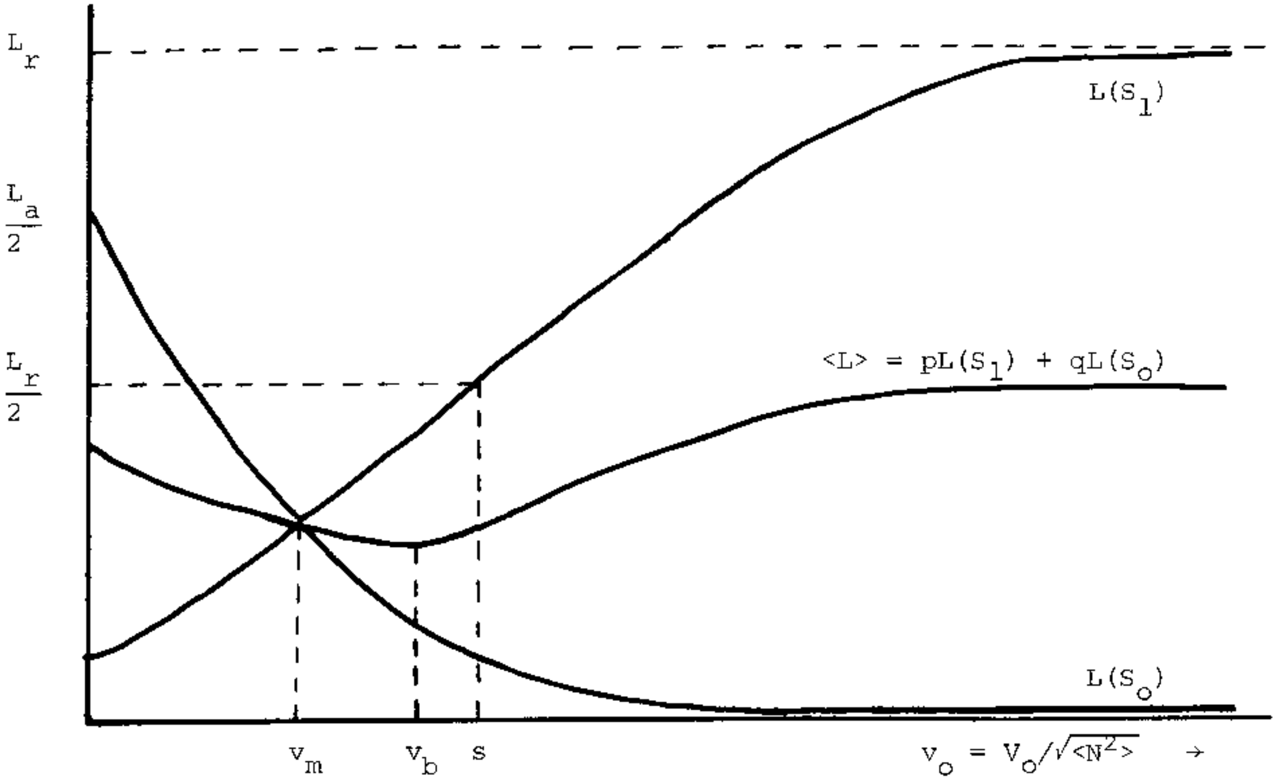

Figure 14.1. Conditional and Average Losses as functions of the detection threshold V_o. The $L(S_1)$ curve is symmetric about the point $\{s, L_r/2\}$.

The Neyman-Pearson criterion is easily discussed in this example: Suppose the conditional probability of a false alarm $(D_1|S_0)$ is held fixed at some small value ε, and we wish to minimize the conditional probability $(D_0|S_1)$ of a false rest, subject to this constraint. Now the Bayes criterion minimizes the average loss

$$\langle L\rangle = pL_r(D_0|S_1) + qL_a(D_1|S_0)$$

with respect to any admissible variation $\delta(D|V)$ in the decision rule. In particular, therefore, it minimizes it with respect to the smaller class of variations which hold $(D_1|S_0)$ constant at the value finally obtained. Thus it minimizes $(D_0|S_1)$ with respect to these variations and solves the Neyman-Pearson problem; we need only choose the particular value of the ratio (qL_a/pL_r) which results in the assumed value of ε according to equations (14-28), (14-30).

We find for the Neyman-Pearson threshold, from (14-38)

$$\Phi(v_{np}) = 1 - \varepsilon \tag{14-39}$$

and the conditional probability of detection is

$$(D_1|S_1) = 1 - (D_0|S_1) = \Phi(s - v_{np}) \tag{14-40}$$

This is the cumulative normal distribution, plotted in Appendix . First finding from the graph, v_{np} for given ε, we find that if $\varepsilon = 10^{-3}$, a detection probability of 99 per cent or better is attained for $s > 6$.

It is important to note that these numerical examples depend critically on our assumption of gaussian noise. If the noise is not gaussian, the actual situation may be either more or less favorable than indicated by the above relations. It is well known that in one sense gaussian noise is the worse possible kind; because of its maximum entropy properties, gaussian noise can obscure a weak signal more completely than can any other noise of the same average power. On the other hand, gaussian noise is a very favorable kind from which to extract a fairly strong signal, because the probability that the noise will exceed a few times the RMS value $\sqrt{<N^2>}$ becomes vanishingly small. Consequently, the probability of making an incorrect decision on the presence or absence of a signal goes to zero very rapidly as the signal strength is increased. The high reliability of operation found above for $s > 6$ would not be found for noise possessing a probability distribution with wider "tails".

The type of noise distribution to be expected in any particular case depends, of course, on the physical mechanism which gives rise to the noise. When the noise is the resultant of a large number of small, independent effects, the central limit theorem of probability theory tells us that the gaussian distribution will be our best bet regardless of the nature of the individual sources.

Well, as the BBC announcers say, that is the end of my summary. All of these apparently different criteria lead, when worked out, to a probability ratio test. In the case of a binary decision, it took the simple form (14-22). Of course, any decision process can be broken down into successive binary decisions, so this case really has the whole story in it. All the different criteria amounted, in the final analysis, only to different philosophies about how you choose the threshold value at which you change your decision.

14.5. How Would Our Robot Do It?

Now let's see how this problem appears from the viewpoint of our robot. The rather long arguments we had to go through above (and even they are very highly condensed, I assure you!) to get the result are due only to the orthodox view which insists on looking at the problem backwards, i.e. on concentrating attention on the final decision rather than on the inductive reasoning process which logically has to precede it. To the robot, if our job is to make the best possible decision as to whether the signal is present, the obvious first thing we must do is calculate the probability that the signal is present.

If there are to be only two possibilities, S_0, S_1, taken into account, then after we have seen voltage V, the odds are from (5-5)

$$O(S_1|VX) = O(S_1|X)\,\frac{(V|S_1)}{(V|S_0)} \qquad (14\text{-}41)$$

If we give the robot the loss function (14-21) and ask him to make the decision which minimizes the expected loss, he will evidently use the decision rule

$$\text{choose } D_1 \text{ if } \; O(S_1|V) \equiv \frac{(S_1|V)}{(S_0|V)} > \frac{L_a}{L_r} \qquad (14\text{-}42)$$

etc. But from Rule 1, $(VS_1|X) = (S_1|V)(V|X)$, $(VS_0|X) = (S_0|V)(V|X)$, and (14-42) is identical with (14-22). So, just from looking at this problem the other way around, our robot derives the same final result in exactly two lines!

You see that all this discussion of strategies, admissibility, conditional losses, etc., was completely unnecessary. Except for the introduction of the loss function at the end, there's nothing in decision theory that isn't already contained in basic probability theory, if we can only free ourselves from the dogma that "probability statements can be made only about random variables," and use the theory in the full generality given to it by Laplace.

This comparison shows why the development of decision theory has, more than any other single factor, led to this revolution in statistical thought. For about thirty years, Jeffreys tried valiantly to explain the Laplace methods to statisticians, and his efforts met only with a steady torrent of denials and ridicule. The quotation about Bayes' theorem applied to quality-control testing that I gave you back in Lecture 5 is a relatively mild example; if you have a taste for such things, you can find, particularly in the works of Fisher and von Mises, some attacks on the viewpoint of Laplace and Jeffreys which make my polemics seem rather tame. It is really astonishing how much emotional fervor can be generated by something that outsiders might consider a rather dry and dull branch of mathematics.

It is real poetic justice that the work of one of the most respected of the "orthodox" statisticians, which was hailed, very properly, as perhaps the greatest advance in statistical practice yet produced, turned out to give, after very long and complicated arguments, exactly the same final results that the despised Laplace methods give you immediately. The only proper conclusion, it seems to me, is that the supposed distinction between statistical inference and probability theory was entirely artificial--a tragic error of judgment which has wasted perhaps a thousand man-years of our best mathematical talent in the pursuit of false goals. There is no longer any justification for trying to make this non-existent distinction.

Suppose that, in the above case of a linear system with gaussian noise, we apply Bayes' theorem in the logarithmic form of Lecture 5. If now we let S_0 and S_1 stand for numerical values giving the amplitudes of the two possible signals, the evidence for the signal is increased by

$$\log \frac{(V|S_1)}{(V|S_0)} = \frac{(V - S_0)^2 - (V - S_1)^2}{2<N^2>}$$

$$= \text{const.} + \frac{S_1 - S_0}{<N^2>} V \qquad (14\text{-}43)$$

so, the observed voltage is just a linear function of the number of db evidence for S_1.

A funny thing happened in the history of this subject. You know that electrical engineers started out not knowing anything whatsoever about statistics. They knew about signal to noise ratios. Receiver input circuits were designed for many years on the basis that signal to noise ratio was maximized. More specifically, it turned out that if you take the ratio of (peak signal)^2 to mean square noise, and find the design of input stages of the receiver which will maximize this quantity, this turned out to be a very useful thing. This leads to the solution which is now called the classical matched filter. It has been discovered independently by at least a dozen people. I believe the first person to work out this matched filter theory was the late Professor W. W. Hansen, in about 1941. I was working with him, beginning in 1942, on problems of radar detection. He circulated a little memorandum at the time in which he gave this solution for the design of the optimum response curve of an IF strip. Years later I was thinking about an entirely different problem (an optimum antenna pattern), and when I finally got the solution, I recognized it as exactly the same thing that Bill Hansen had worked out many years before. I'll give you this theory in a later lecture. Since then I see, almost every time I open a journal concerned with these problems, that somebody else has a paper with the same solution in it.

Now, in the 1950's, people got more sophisticated about the way they handled their detection problems, and they started using this wonderful new tool, statistical decision theory, to see if there were still better ways of handling these design problems. The strange thing happened that in the case of a linear system with gaussian noise, the optimum solution which decision theory leads you to, turns out to be exactly the same old classical matched filter. When I first saw this, I was very surprised that two approaches so entirely different should lead to the same solution. But, note that our robot represents a viewpoint from which it is not at all surprising that the two lines of argument would have to give the same result. The best statistical analysis you can make of the problem will always be one in which you calculate the <u>probability</u> that the various signals are present by means of Bayes' theorem. But, in the case of a linear system with gaussian noise, the observed voltage is itself just a linear function of the posterior probability measured in db. So, they are essentially just two different ways of formulating the same problem.

The different approaches to the theory simply amount to different philosophies of how you choose that value of probability at which you will change your decision. Because of the fact that they all lead to the same probability ratio test, they must necessarily all be derivable from Bayes' theorem, in agreement with out robot's prediction back in Lecture 4.

The problem just examined by several different decision criteria is, of course, the simplest possible one. In a more realistic problem we will observe the voltage v(t) as a function of time, perhaps several voltages $v_1(t)$, $v_2(t)$, ... in several different channels. We may have many different possible signals $S_a(t)$, $S_b(t)$ ... to distinguish, or we may need not only to decide whether a given signal is present, but also to make the best estimates of one or more signal parameters (such as intensity, starting time, frequency, phase, rate of frequency modulation, etc.). Therefore, just as in the problem of quality control discussed in Lectures 5, 6, the details can become arbitrarily complicated. But these extensions are, from the Bayesian viewpoint, straightforward in that they require no new principles beyond those already given.

I want to come back to some of these more complicated problems of detection and filtering toward the end of these lectures; but for now let's look at another elementary kind of decision problem. In the ones discussed so far, we used Bayes' theorem, but not maximum entropy. Now I want to show you a kind of problem where we need maximum entropy. but not Bayes' theorem.

14.6. The Widget Problem.

This problem was first propounded at a symposium held at Purdue University in November, 1960--at which time, however, the full solution was not known. This was worked out later (Jaynes, 1963c), and some numerical approximations were improved in the computer work of Tribus and Fitts (1968).

The widget problem has proved to be interesting in more respects than originally realized. It is a decision problem in which there is no occasion to use Bayes' theorem, because no "new" information is acquired. Thus it would be termed a "no data" decision problem in the sense of Chernoff and Moses (1959). However, at successive stages of the problem we have more and more prior information; and digesting it by maximum entropy leads to a sequence of prior probability assignments, which lead to different decisions. Thus it is an example of the "pure" use of maximum entropy, as in statistical mechanics. It is hard to see how the problem could be formulated mathematically at all without use of maximum entropy, or some other device [like the one considered in Lecture 10 (Sec. 10.8)] which turns out in the end to be mathematically equivalent to maximum entropy.

The problem is interesting also in that we can see a continuous gradation from decision problems so simple that common sense tells us the answer instantly with no need for any mathematical theory, through problems more and more involved so that common sense has more and more difficulty in making a decision, until finally we reach a point where nobody has yet claimed to be able to see the right decision intuitively, and we require the mathematics to tell us what to do.

Finally, it turned out to be very close to an important real problem faced by oil prospectors. The details of the real problem are shrouded in proprietary caution; but I'm not giving away any secrets if I tell you that, a few years ago, I spent a week at the research laboratories of one of our large oil companies, lecturing for over 20 hours on the widget problem. They made me go through every part of the calculation in excruciating detail--much more than we have time for here--with a room full of engineers armed with slide-rules, checking up on every stage of the numerical work. I've often wondered since how far they have extended the theory beyond the original problem, and how much it helped them; but I don't expect to find out.

Well, here is the problem. Mr. A is in charge of a Widget factory, which proudly advertises that it can make delivery in 24 hours on any size order. This, of course, is not really true, and Mr. A's job is to protect, as best he can, the Advertising Manager's reputation for veracity. This means that each morning he must decide whether the day's run of 200 Widgets will be painted red, yellow, or green. (For complex technological reasons, not relevant to the present problem, only one color can be produced per day.) We follow his problem of decision through several stages of increasing knowledge.

Stage 1. When he arrives at work, Mr. A checks with the stock room and finds that they now have in stock 100 red widgets, 150 yellow, and 50 green. His ignorance lies in the fact that he does not know how many orders for each type will come in during the day. Clearly, in this state of ignorance, Mr. A will attach the highest significance to any tiny scrap of information about orders likely to come in today; and if no such scraps are to be had, we do not envy Mr. A his job. Still, if a decision has to be made on no more information than this, his common sense will probably tell him that he had better build up that stock of green widgets.

Stage 2. Mr. A, feeling the need for more information, calls up the front office and asks, "Can you give me some idea of how many orders for red, yellow, and green widgets are likely to come in today?" They reply, "Well, we don't have the breakdown of what has been happening each day, and it would take us a week to compile that information from our files. But we do have a summary of total sales last year. Over the last year, we sold a total of 13,000 red, 26,000 yellow, and 2600 green. Figuring 260 working days, this means that last year we sold an average of 50 red, 100 yellow, 10 green each day." If Mr. A ponders this new information for a few seconds, I think he will change his mind, and decide to make yellow ones today.

Stage 3. The man in the front office calls Mr. A back to say, "It just occurred to me that we do have a little more information that might possibly help you. We have at hand not only the total number of widgets sold last year, but also the total number of orders we processed. Last year we got a total of 173 orders for red, 2600 for yellow, and 130 for green. This means that customers who use red widgets ordered, on the average, 13000/173 = 75 widgets per order, while the average orders for yellow and green were 26000/2600 = 10, and 2600/130 = 20 respectively." This new data doesn't change the expected daily demand; but if Mr. A is very shrewd and ponders it very hard, I think he may change his mind again, and decide to make red ones today.

Stage 4. Mr. A is just about to give the order to make red ones when the front office calls him again to say, "We just got word that a messenger is on his way here with an emergency order for 40 green widgets." Now, what should he do? Up to this point, Mr. A's decision problem has been simple enough so that reasonably good common sense will tell him what to do. But now, I think he is in trouble; qualitative common sense is just not powerful enough to solve his problem, and he needs a mathematical theory to determine a definite optimum decision.

Let's summarize all the above data in a table:

	R	Y	G	Decision
1. In Stock	100	150	50	G
2. Avg. Daily Order Total	50	100	10	Y
3. Avg. Individual Order	75	10	20	R
4. Specific Order			40	?

Table 14.1. Summary of four stages of the Widget Problem.

In the last column I give the decisions that seemed, to me, to be the best ones before I had worked out the mathematics. Do other people agree with this intuitive judgment? Professor Myron Tribus has put this to the test by giving talks about this problem, and taking votes from the audience before the solution is given. Let me quote his findings as given in their paper (M. Tribus and G. Fitts, 1968). They use D_1, D_2, D_3, D_4 to stand for the optimum decisions in stages 1, 2, 3, 4 respectively:

"Before taking up the formal solution, it may be reported that Jaynes' widget problem has been presented to many gatherings of engineers who have been asked to vote on D_1, D_2, D_3, and D_4. There is almost unanimous agreement about D_1. There is about 85 percent agreement on D_2. There is about 70 percent agreement on D_3, and almost no agreement on D_4. One conclusion stands out from these informal tests; the average engineer has remarkably good intuition in problems of this kind. The majority vote for D_1, D_2, and D_3 has always been in agreement with the formal mathematical solution. However, there has been almost universal disagreement over how to defend the intuitive solution. That is, while many engineers could agree on the best course of action, they were much less in agreement on why that course was the best one."

14.7. Solution For Stage 2.

Now, how are we to set up this problem mathematically? In a real life situation, evidently, the problem would be a little more complicated than indicated so far, because what Mr. A does today also affects how serious his problem will be tomorrow. Mr. A's decision each day should not depend only on orders expected for that day; they should be based on his best estimates of orders likely to come in for all future days, and on the consequences of failure to meet all orders not only today but also in the future. That would get us into the subject of dynamic programming. But for now, just to keep the problem simple, let's solve only the truncated problem in which he makes decisions on a day to day basis with no thought of tomorrow.

We have just to carry out the steps enumerated under "General Decision Theory" at the end of the last lecture. Since Stage 1 is almost too trivial to work with, consider the problem of stage 2. First, enumerate the possible "states of nature" θ_j. These correspond to all possible order situations that could arise; if Mr. A knew in advance exactly how many red, yellow, and green widgets would be ordered today, his decision problem would be trivial. Let $n_1 = 0, 1, 2, \ldots$ be the number of red widgets that will be ordered today, and similarly n_2, n_3 for yellow and green respectively. Then any conceivable order situation is given by specifying three non-negative integers $\{n_1,n_2,n_3\}$. Conversely, every ordered triple of non-negative integers represents a conceivable order situation.

Next, we are to assign prior probabilities $(\theta_j|X) = (n_1n_2n_3|X)$ to the states of nature, which maximize the entropy of the distribution subject to the constraints of our prior knowledge. We solved this problem generally in Lecture 10, Equations (10-26)--(10-32); and so we just have to translate the result into our present notation. The index i on x_i in Lecture 10 now corresponds to the three integers n_1, n_2, n_3; the functions $f_k(x_i)$ also correspond to the n_i, since the prior information at this stage is that the expectations $\langle n_1\rangle$, $\langle n_2\rangle$, $\langle n_3\rangle$ of orders for red, yellow, and green widgets are given as 50, 100, 10 respectively. With three average values given, we will have three Lagrange multipliers λ_1, λ_2, λ_3, and the partition function (10-30) becomes

$$Z(\lambda_1,\lambda_2,\lambda_3) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_3=0}^{\infty} \exp(-\lambda_1 n_1 - \lambda_2 n_2 - \lambda_3 n_3)$$

$$= \Pi_{i=1}^{3} (1 - e^{-\lambda_i})^{-1} \qquad (14\text{-}44)$$

The λ_i are determined from (10-32):

$$\langle n_i \rangle = - \frac{\partial}{\partial \lambda_i} \log Z$$

$$= \left(\frac{1}{e^{\lambda_i} - 1} \right) \qquad (14\text{-}45)$$

The maximum-entropy probability assignment (10-28) for the states of nature $\theta_j = \{n_1 \; n_2 \; n_3\}$ therefore factors:

$$p(n_1 n_2 n_3) = p_1(n_1) \; p_2(n_2) \; p_3(n_3) \qquad (14\text{-}46)$$

with

$$p_i(n_i) = (1 - e^{-\lambda_i}) \; e^{-\lambda_i n_i} \; , \quad n_i = 0, 1, 2, \ldots$$

$$= \frac{1}{\langle n_i \rangle + 1} \left[\frac{\langle n_i \rangle}{\langle n_i \rangle + 1} \right]^{n_i} \qquad (14\text{-}47)$$

Thus in stage 2, Mr. A's state of knowledge about today's orders is given by three exponential distributions:

$$p_1(n_1) = \frac{1}{51} \left(\frac{50}{51} \right)^{n_1}$$

$$p_2(n_2) = \frac{1}{101} \left(\frac{100}{101} \right)^{n_2}$$

$$p_3(n_3) = \frac{1}{11} \left(\frac{10}{11} \right)^{n_3} \qquad (14\text{-}48)$$

which completes step 2. Step 3, application of Bayes' theorem to digest new evidence E, is absent because there is no new evidence. Therefore, the decision must be made directly from the prior probabilities (14-48), as is always the case in statistical mechanics. So, we now proceed to step 4, enumerate the possible decisions. These are $D_1 \equiv$ make red ones today, $D_2 \equiv$ make yellow ones, $D_3 \equiv$ make green ones. In step 5, we are to introduce a loss function $L(D_i, \theta_j)$. Mr. A's judgment is that there is no loss if all orders are filled today; otherwise the loss will be proportional to--and in view of the invariance of the decision rule under proper linear transformations that we noted at the end of Lecture 13, we may as well take it equal to--the total number of unfilled orders.

The present stock of red, yellow, and green widgets is $S_1 = 100$, $S_2 = 150$, $S_3 = 50$ respectively. On decision D_1 (make red widgets) the available stock S_1 will be increased by the day's run of 200 widgets, and the loss will be

$$L(D_1;n_1n_2n_3) = g(n_1 - S_1 - 200) + g(n_2 - S_2) + g(n_3 - S_3) \quad (14\text{-}49)$$

where $g(x)$ is the ramp function

$$g(x) \equiv \begin{cases} x, & x \geq 0 \\ 0, & x \leq 0 \end{cases} \quad (14\text{-}50)$$

Likewise, on decision D_2, D_3 the loss will be

$$L(D_2;\ n_1n_2n_3) = g(n_1 - S_1) + g(n_2 - S_2 - 200) + g(n_3 - S_3) \quad (14\text{-}51)$$

$$L(D_3;\ n_1n_2n_3) = g(n_1 - S_1) + g(n_2 - S_2) + g(n_3 - S_3 - 200) \quad (14\text{-}52)$$

So, if decision D_1 is made, the expected loss will be

$$\begin{aligned} \langle L\rangle_1 &= \sum_{n_i} p(n_1n_2n_3)\ L(D_1;n_1n_2n_3) \\ &= \sum_{n_1=0}^{\infty} p_1(n_1)\ g(n_1-S_1-200) + \sum_{n_2=0}^{\infty} p_2(n_2)\ g(n_2-S_2) \\ &\qquad + \sum_{n_3=0}^{\infty} p_3(n_3)\ g(n_3-S_3) \end{aligned} \quad (14\text{-}53)$$

and similarly for D_2, D_3. The summations are elementary, giving

$$\langle L\rangle_1 = \langle n_1\rangle e^{-\lambda_1(S_1+200)} + \langle n_2\rangle e^{-\lambda_2 S_2} + \langle n_3\rangle e^{-\lambda_3 S_3}$$

$$\langle L\rangle_2 = \langle n_1\rangle e^{-\lambda_1 S_1} + \langle n_2\rangle e^{-\lambda_2(S_2+200)} + \langle n_3\rangle e^{-\lambda_3 S_3}$$

$$\langle L\rangle_3 = \langle n_1\rangle e^{-\lambda_1 S_1} + \langle n_2\rangle e^{-\lambda_2 S_2} + \langle n_3\rangle e^{-\lambda_3(S_3+200)} \quad (14\text{-}54)$$

or, inserting numerical values

$$\langle L\rangle_1 = 0.131 + 22.480 + 0.085 = 22.70$$

$$\langle L\rangle_2 = 6.902 + 3.073 + 0.085 = 10.06$$

$$\langle L\rangle_3 = 6.902 + 22.480 + 4 \times 10^{-10} = 29.38 \qquad (14\text{-}55)$$

showing a strong preference for decision $D_2 \equiv$ "make yellow ones today," as common sense had already anticipated.

You will recognize that Stage 2 of Mr. A's decision problem is mathematically the same as the theory of the harmonic oscillator in quantum statistical mechanics. There is still another engineering application of the harmonic oscillator equations, in some problems of message encoding, that we'll see when we take up communication theory. I'm trying to emphasize the generality of this theory, which is mathematically quite old and well known, but which has been applied in the past only in some specialized problems in physics. This general applicability can be seen only after we are emancipated from the orthodox view that all probability distributions must be justified in the frequency sense. Historically, this made it appear to most workers in statistical mechanics that the methods of Gibbs could be justified only via unproved "ergodic hypotheses" (in spite of the fact that Gibbs himself never mentioned them). But if we interpret Gibbs' equations not as assertions about frequencies but as examples of inductive reasoning based on the principle of maximum entropy, it is clear that the reasoning doesn't depend on ergodic properties or any other aspect of the laws of physics--*ergo*, the canonical ensemble formalism of Gibbs can be applied to any problem of inductive reasoning where the given information can be stated in the form of mean values.

14.8. Solution For Stage 3.

In Stage 3 of Mr. A's problem we have some additional pieces of information giving the average *individual* orders for red, yellow, and green widgets.

To take account of this new information, we need to set up a more detailed enumeration of the states of nature, in which we take into account not only the total orders for each type, but also the breakdown into individual orders. We could have done this also in stage 2, but since at that stage there was no information available bearing on this breakdown, it would have added nothing to the problem. However, in the interest of checking the consistency of this theory, you may find it amusing to retrace stage 2 on this basis and see how it would have led to exactly the same results given above.

In stage 3, a possible state of nature can be described as follows. We receive u_1 individual orders for 1 red widget each, u_2 orders for 2 red widgets each, ..., u_r individual orders for r red widgets each. Also, we receive v_y orders for y yellow widgets each, and w_g orders for g green widgets each. Thus a state of nature is specified by an infinite number of non-negative integers

$$\theta = \{u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots\} \qquad (14\text{-}56)$$

and conversely every such set of integers represents a conceivable state of nature, to which we assign a probability $p(u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots)$.

Today's total demand for red, yellow and green widgets is, respectively

$$n_1 = \sum_{r=1}^{\infty} r\, u_r$$

$$n_2 = \sum_{y=1}^{\infty} y\, v_y$$

$$n_3 = \sum_{g=1}^{\infty} g\, w_g \qquad (14\text{-}57)$$

the expectations of which were given in stage 2 as $\langle n_1 \rangle = 50$, $\langle n_2 \rangle = 100$, $\langle n_3 \rangle = 10$. The total number of individual orders for red, yellow, and green widgets are respectively

$$m_1 = \sum_{r=1}^{\infty} u_r$$

$$m_2 = \sum_{y=1}^{\infty} v_y$$

$$m_3 = \sum_{g=1}^{\infty} w_g \qquad (14\text{-}58)$$

And the new feature of stage 3 is that $\langle m_1\rangle$, $\langle m_2\rangle$, $\langle m_3\rangle$ are also known. For example, the statement that the average individual order for red widgets is 75 means that $\langle n_1\rangle = 75\langle m_1\rangle$.

With six average values given, we will have six Lagrange multipliers $\{\lambda_1\mu_1;\lambda_2\mu_2;\lambda_3\mu_3\}$. The maximum-entropy probability assignment will have the form

$$p(u_1u_2\ldots;v_1v_2\ldots;w_1w_2\ldots) = \exp(-\lambda_0-\lambda_1n_1-\mu_1m_1-\lambda_2n_2-\mu_2m_2-\lambda_3n_3-\mu_3m_3)$$

which factors:

$$p(u_1u_2\ldots;v_1v_2\ldots;w_1w_2\ldots) = p_1(u_1u_2\ldots)p_2(v_1v_2\ldots)p_3(w_1w_2\ldots) \qquad (14\text{-}59)$$

The partition function also factors:

$$Z = Z_1(\lambda_1\mu_1)\ Z_2(\lambda_2\mu_2)\ Z_3(\lambda_3\mu_3) \qquad (14\text{-}60)$$

with

$$Z_1(\lambda_1\mu_1) = \sum_{u_1=1}^{\infty}\sum_{u_2=1}^{\infty}\ldots \exp[-\lambda_1(u_1+2u_2+3u_3+\ldots)-\mu_1(u_1+u_2+u_3+\ldots)]$$

$$= \Pi_{r=1}^{\infty}\frac{1}{1-e^{-r\lambda_1-\mu_1}} \qquad (14\text{-}61)$$

with similar expressions for Z_2, Z_3. To find λ_1, μ_1 we apply the general rule, Equation (10-32):

$$\langle n_1\rangle = \frac{\partial}{\partial\lambda_1}\sum_{r=1}^{\infty}\log(1-e^{-r\lambda_1-\mu_1}) = \sum_{r=1}^{\infty}\frac{r}{e^{r\lambda_1+\mu_1}-1} \qquad (14\text{-}62)$$

$$\langle m_1\rangle = \frac{\partial}{\partial\mu_1}\sum_{r=1}^{\infty}\log(1-e^{-r\lambda_1-\mu_1}) = \sum_{r=1}^{\infty}\frac{1}{e^{r\lambda_1+\mu_1}-1} \qquad (14\text{-}63)$$

Comparing with equations (14-57), (14-58), we see that

$$\langle u_r \rangle = \frac{1}{e^{r\lambda_1+\mu_1} - 1} \tag{14-64}$$

and now the secret is out--Stage 3 of Mr. A's decision problem is just the theory of an ideal Bose-Einstein gas in quantum statistical mechanics!

If we treat the ideal Bose-Einstein gas by the method of the grand canonical ensemble, we obtain just these equations, in which the number r corresponds to the r'th single-particle energy level, u_r to the number of particles in the r'th state, λ_1 and μ_1 to the temperature and chemical potential.

In the present problem it is clear that for all r, $\langle u_r \rangle \ll 1$, and that $\langle u_r \rangle$ cannot decrease appreciably below $\langle u_1 \rangle$ until r is of the order of 75, the average individual order. Therefore, μ_1 will be numerically large, and λ_1 numerically small, compared to unity. This means that the series (14-62), (14-63) converge very slowly and are useless for numerical work unless you have a big computer. However, we can transform them into rapidly converging sums as follows:

$$\sum_{r=1}^{\infty} \frac{1}{e^{\lambda r+\mu} - 1} = \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} e^{-n(\lambda r+\mu)}$$

$$= \sum_{n=1}^{\infty} \frac{e^{-n\mu}}{1 - e^{-n\lambda}} \tag{14-65}$$

The first term is already an excellent approximation. Similarly,

$$\sum_{r=1}^{\infty} \frac{r}{e^{\lambda r+\mu} - 1} = \sum_{n=1}^{\infty} \frac{e^{-n(\lambda+\mu)}}{(1 - e^{-n\lambda})^2} \tag{14-66}$$

and so (14-62) and (14-63) become

$$\langle n_1 \rangle \simeq \frac{e^{-\mu_1}}{\lambda_1^2} \tag{14-67}$$

$$\langle m_1 \rangle \simeq \frac{e^{-\mu_1}}{\lambda_1} \tag{14-68}$$

or

$$\lambda_1 \simeq \frac{<m_1>}{<n_1>} = \frac{1}{75} = 0.0133 \tag{14-69}$$

$$e^{\mu_1} \simeq \frac{<n_1>}{<m_1>^2} = 112.5 \tag{14-70}$$

$$\mu_1 = 4.722 \tag{14-71}$$

Tribus and Fitts, evaluating the sums exactly by computer, get $\lambda_1 = 0.0131$, $\mu_1 = 4.727$; so our approximations (14-67), (14-68) are very good, at least in the case of red widgets.

The probability that u_r has a particular value is, from (14-59) or (14-61),

$$p(u_r) = (1 - e^{-r\lambda_1 - \mu_1})\, e^{-(r\lambda_1+\mu_1)u_r} \tag{14-72}$$

which has the mean value (14-64) and the variance

$$\mathrm{var}(u_r) = <u_r^2>-<u_r>^2 = \frac{e^{r\lambda_1+\mu_1}}{(e^{r\lambda_1+\mu_1} - 1)^2} \tag{14-73}$$

The total demand for red widgets

$$n_1 = \sum_{r=1}^{\infty} r u_r \tag{14-74}$$

is expressed as the sum of a large number of independent "random variables". The probability distribution for n_1 will have the mean value (14-67) and the variance

$$\mathrm{var}(n_1) = \sum_{r=1}^{\infty} r^2\, \mathrm{var}(u_r) = \sum_{r=1}^{\infty} \frac{r^2\, e^{r\lambda_1+\mu_1}}{(e^{r\lambda_1+\mu_1} - 1)^2} \tag{14-75}$$

which we convert into the rapidly convergent sum

$$\sum_{r,n=1}^{\infty} nr^2\, e^{-n(r\lambda+\mu)} = \sum_{n=1}^{\infty} n\, \frac{e^{-n(\lambda+\mu)} + e^{-n(2\lambda+\mu)}}{(1 - e^{-n\lambda})^3} \tag{14-76}$$

or, approximately,

$$\mathrm{var}(n_1) \simeq \frac{2e^{-\mu_1}}{\lambda_1^{\,3}} = \frac{2}{\lambda_1} \langle n_1 \rangle \ . \qquad (14\text{-}77)$$

At this point I have to anticipate some mathematical facts concerning the Central Limit Theorem, that we'll study later. Because n_1 is the sum of a large number of small terms, the probability distribution for n_1 will be very nearly gaussian:

$$p(n_1) \simeq A \exp\left\{-\frac{\lambda_1(n_1-\langle n_1\rangle)^2}{4\langle n_1\rangle}\right\} \qquad (14\text{-}78)$$

for those values of n_1 which can arise in many different ways. For example, the case $n_1 = 2$ can arise in only two ways: $u_1 = 2$, or $u_2 = 1$, all other u_k being zero. On the other hand, the case $n_1 = 150$ can arise in an enormous number of different ways, and the "smoothing" mechanism of the central limit theorem can operate. Thus, Equation (14-78) will be a good enough approximation for the large values of n_1 of interest to us, but it may not be for small n_1.