Lecture 16

THE A_p DISTRIBUTION AND RULE OF SUCCESSION

Up to this point we have given our robot fairly general principles by which he can convert information into numerical values of prior probabilities, and convert posterior probabilities into definite final decisions; so he is now able to solve lots of problems. But he still operates in a rather inefficient way in one respect. When we give him new information and ask him to reason about it, he has to go back into his memory (this proposition X that involves everything that has ever happened to him). He must scan his entire memory storage reels for anything relevant to the problem before he can start reasoning on it. As the robot gets older this gets to be a more and more time-consuming process.

Now, human brains don't do this. We have some machinery built into us which summarizes our past conclusions, and allows us to forget the details which led us to those conclusions. We want to see whether it's possible to give the robot a definite mechanism by which he can store conclusions rather than isolated facts.

16.1. Memory Storage for Old Robots.

Let me point out another thing, which we will see is closely related to this problem. Suppose you have a penny and you are allowed to examine it carefully, convince yourself that it's an honest coin, has a head and tail, and center of gravity where it ought to be. Then, you're asked to give the

probability that this coin will come up heads on the first toss. I'm sure you'll say 1/2. Now, suppose you are asked to assign a probability to the proposition that there is life on Mars. Well, I don't know what your opinion is there, but on the basis of all the things that I have read on the subject, I would again say about 1/2 for the probability. But, even though I have assigned the same probability to them, I have a very different state of knowledge about those propositions. To see that, imagine the effect of getting new information. Suppose we tossed the coin five times and it comes up tails every time. You ask me what's my probability for heads on the next throw; I'll still say 1/2. But if you tell me one more fact about Mars, I'm ready to change my probability assignment completely. My state of belief has a great instability in the case of Mars, but there's something which makes it very stable in the case of the penny.

Now, it seemed to me for a long time that this was a fatal objection to Laplace's form of probability theory. We need to associate with a proposition not just a single number representing plausibility, but two numbers; one representing the plausibility, and the other how stable it is in the face of new evidence. And so, a kind of two-valued theory would have to be developed before it would make any sense. In the early 1950's, I even gave a talk at one of the Berkeley Statistical Symposiums, expounding this viewpoint. This is, furthermore, just what Carnap (1952) has done; his continuum of inductive methods consists of a class of probability functions $C_\lambda(h,e)$ in which λ is the "stability parameter."

But now, I think that there's a mechanism by which we can show that our present theory automatically contains all these things. So far, all the propositions we have asked the robot to think about are ones which had to be either true or false. Suppose we bring in new propositions of a different type. It doesn't make sense to say the proposition is either true or false,

but still we are going to say the robot assigns credibility to it. Now, these propositions are sometimes hard to state verbally, and I, at least, am never able to write a verbal statement that's unambiguous. But you noticed before that we can get around that very nicely by recognizing that if I state all probabilities conditional on X for a given problem, I've told you everything about X that's relevant to the problem. So, I want to introduce a new proposition A_p, defined by

$$(A|A_pE) \equiv p \tag{16-1}$$

where E is any additional evidence. If I had to render A_p as a verbal statement, it would come out something like this:

$A_p \equiv$ "Regardless of anything else you may have been told, the probability of A is p."

Now, A_p is a strange proposition, but if we allow the robot to reason with propositions of this sort, Bayes' theorem guarantees that there's nothing to prevent him from getting an A_p worked over onto the left side in his probabilities: $(A_p|E)$. Now, what are we doing here? We're talking about the "probability of a probability." I defined A_p by writing an equation. You ask me what it means, and I reply by writing more equations. So let's write the equations; if X says nothing about A except that it is possible for A to be true, and also possible for it to be false, then as we saw in the case of the "completely ignorant population" in Lecture 12,

$$(A_p|X) = 1, \qquad 0 \le p \le 1. \tag{16-2}$$

The transformation group arguments of Lecture 12 apply to this problem. As soon as we have this, we can use Bayes' theorem to get the probability (density) of A_p, conditional on other things. In particular,

$$(A_p|E) = (A_p|X)\,\frac{(E|A_p)}{(E|X)} = \frac{(E|A_p)}{(E|X)} \tag{16-3}$$

Now,

$$(A|E) = \int_0^1 (AA_p|E)\, dp \tag{16-4}$$

The propositions A_p are mutually exclusive and exhaustive (in fact, every A_p flatly and dogmatically contradicts every other A_q), so we can do this. We're just going to apply all of our mathematical rules with total disregard of the fact that A_p is a funny kind of proposition. We believe that these rules form a consistent way of manipulating propositions; their application cannot lead to contradictions. (Of course, we haven't really proved that they are consistent; we have proved only that if we represent degrees of plausibility by real numbers and require qualitative agreement with common sense, any other rules would be inconsistent.) But consistency is a purely structural property of the rules, which could not depend on the particular semantic meaning you or I might attach to a proposition. So now we can blow up the integrand of (16-4) by our Rule 1:

$$(A|E) = \int_0^1 (A|A_pE)(A_p|E)\, dp \tag{16-5}$$

But from the definition (16-1) of A_p, the first factor is just p, and so

$$(A|E) = \int_0^1 (A_p|E)\, p\, dp \tag{16-6}$$

The probability which our robot assigns to proposition A is just the <u>first moment</u> of the distribution of A_p. Therefore, the distribution of A_p should contain an awful lot more information about the robot's state of mind concerning A, than just the probability of A. I think the introduction of propositions of this sort solves both of the problems mentioned, and also gives us a powerful analytical tool for calculating probabilities.

To see why, let's first note some lemmas about relevance. Suppose this evidence E consists of two parts; $E = E_aE_b$, where E_a is relevant to A and, given E_a, E_b is not relevant:

$$(A|E) = (A|E_aE_b) = (A|E_a) \tag{16-7}$$

By Bayes' theorem, it follows that, given E_a, A must also be irrelevant to E_b, for

$$(E_b|AE_a) = (E_b|E_a)\,\frac{(A|E_bE_a)}{(A|E_a)} = (E_b|E_a) \tag{16-8}$$

Let's call this property "weak irrelevance." Now does this imply that E_b is irrelevant to A_p? Evidently not, for (16-7) says only that the first moments of $(A_p|E_a)$ and $(A_p|E_aE_b)$ are the same. But suppose that for a given E_b, (16-7) holds independently of what E_a might be; call this "strong irrelevance." Then we have

$$(A|E) = \int_0^1 (A_p|E_aE_b)\; p\; dp = \int_0^1 (A_p|E_a)\; p\; dp. \tag{16-9}$$

If this is to hold for all $(A_p|E_a)$, the integrands must be the same

$$(A_p|E_aE_b) = (A_p|E_a) \tag{16-10}$$

and from Bayes' theorem it follows as in (16-8) that A_p is irrelevant to E_b:

$$(E_b|A_pE_a) = (E_b|E_a) \tag{16-11}$$

for all E_a.

Now, suppose our robot gets a new piece of evidence, F. How does this change his state of knowledge about A? We could expand directly by Bayes' theorem, which we have done before, but let's use our A_p this time,

$$(A|EF) = \int_0^1 (A_p|EF)\; p\; dp = \int_0^1 (A_p|E)\,\frac{(F|A_pE)}{(F|E)}\; p\; dp\;. \tag{16-12}$$

In this likelihood ratio, any part of E that is irrelevant to A_p can be struck out. Because, by Bayes' theorem, it is equal to

$$\frac{(F|A_pE_aE_b)}{(F|E_aE_b)} = \frac{(F|A_pE_a)\left[\dfrac{(E_b|FA_pE_a)}{(E_b|A_pE_a)}\right]}{(F|E_a)\left[\dfrac{(E_b|FE_a)}{(E_b|E_a)}\right]} = \frac{(F|A_pE_a)}{(F|E_a)} \tag{16-13}$$

where we have used (16-11). Now if E_a still contains a part irrelevant to A_p, we can repeat this process. Imagine this carried out as many times as

possible; the part E_{aa} of E that is left contains nothing at all that is irrelevant to A_p. E_{aa} must then be some statement only about A. But then by the definition (16-1) of A_p, we see that A_p automatically cancels out E_{aa} in the numerator: $(F|A_pE_{aa}) = (F|A_p)$. And so we have (16-12) reduced to

$$(A|EF) = \frac{1}{(F|E_{aa})} \int_0^1 (A_p|E)(F|A_p)\ p\ dp \qquad (16\text{-}14)$$

The weak point in this argument is that I haven't proved that it is possible to resolve E into a completely relevant part and completely irrelevant part. However, it is easy to show that in many applications it <u>is</u> possible. So, let's just say that the following results apply to the case where the prior information is "completely resolvable." We don't know whether it is the most general case; but we do know that it is not an empty one.

Now, $(F|E_{aa})$ is a troublesome thing which we would like to get rid of. It's really just a normalizing factor, and we can eliminate it the way we did in Equation (5-3); by calculating the odds on A instead of the probability. This is just

$$\frac{(A|EF)}{(a|FE)} = \frac{\int_0^1 (A_p|E)(F|A_p)\ p\ dp}{\int_0^1 (A_p|E)(F|A_p)\ (1\text{-}p)\ dp} = O(A|EF) \qquad (16\text{-}15)$$

The proposition E, which for this problem represents our prior evidence, now appears only in the combination $(A_p|E)$. This means that <u>the only property of E which the robot needs in order to reason out the effect of new information is this distribution</u> $(A_p|E)$. Everything that has ever happened to him which is relevant to this proposition A may consist of millions and millions of isolated separate facts. Whenever he receives new information, he does not have to go back and search his entire memory for every little detail of experience relevant to A. Everything he needs in order to reason about it is contained summarized in this one function, $(A_p|E)$. So, for each proposi-

tion about which he is going to have to reason, he can store a function like that in Figure (16.1). Whenever he receives new information, F, he will be well advised to calculate $(A_p|EF)$, and he then can erase his previous $(A_p|E)$ and for the future store only $(A_p|EF)$.

This shows that in a machine which does inductive reasoning, the memory storage problem is very much simpler than it is in a machine which does only deductive reasoning, like this one you have down at the end of the hall. This doesn't mean that the robot is able to throw away entirely all of his past experience, because there's always a possibility that some new proposition will come up which he has not had to reason about before. And whenever this happens, then, of course, he <u>will</u> have to go back to his original archives and search for every scrap of information he has relevant to this proposition.

With a little introspection, I think we would all agree that that's exactly what goes on in our minds. If you are asked how plausible you regard

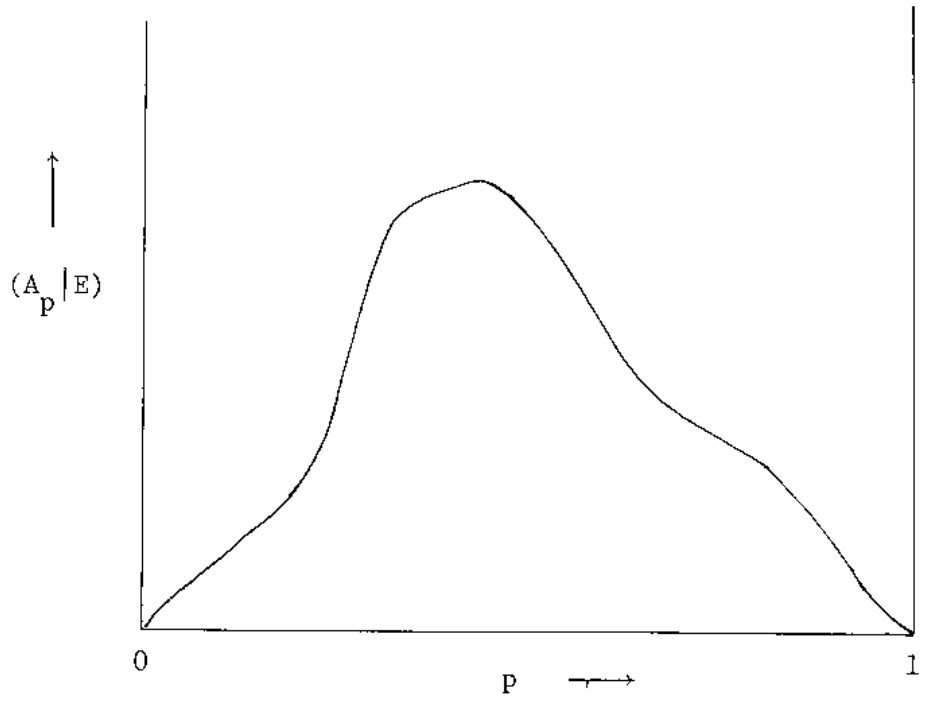

Figure 16.1

some proposition, you don't go back and recall all the details of everything that you ever learned about this proposition. You recall your previous state of mind about it. How many of us can still remember the argument that first convinced us that

$$\frac{d \sin x}{dx} = \cos x \quad ?$$

Let's look once more at Equation (16-14). If the new information F is to make any appreciable change in the probability of A, we can see from this integral what has to happen. If the distribution of $(A_p|E)$ was already very sharply peaked at one particular value of p, then $(F|A_p)$ will have to be even more sharply peaked at some other value of p, if we are going to get any appreciable change in the probability. On the other hand, if the distribution $(A_p|E)$ is a very broad one, then, of course, almost any small amount of slope in $(F|A_p)$ can make a big change in the probability which the robot assigns to A. So, the stability of the robot's state of mind is essentially the <u>width</u> of the distribution $(A_p|E)$. I don't think there's any single number which fully describes this stability. On the other hand, whenever he has accumulated enough evidence so that $(A_p|E)$ is fairly well sharply peaked at some value of p, then the variance of that distribution becomes a pretty good measure of how stable his state of mind is. The greater amount of previous information he has collected, the narrower his A_p-distribution will be, and therefore the harder it will be for any new evidence to change that state of mind.

Now we can see the difference between the penny and Mars. In the case of the penny, my distribution $(A_p|E)$, based on my prior knowledge, is represented by a curve something like Figure (16.2a). In the case of the question of life on Mars, my state of knowledge is described by an $(A_p|E)$ distribution something like Figure (16.2b), qualitatively. The first moment is the same

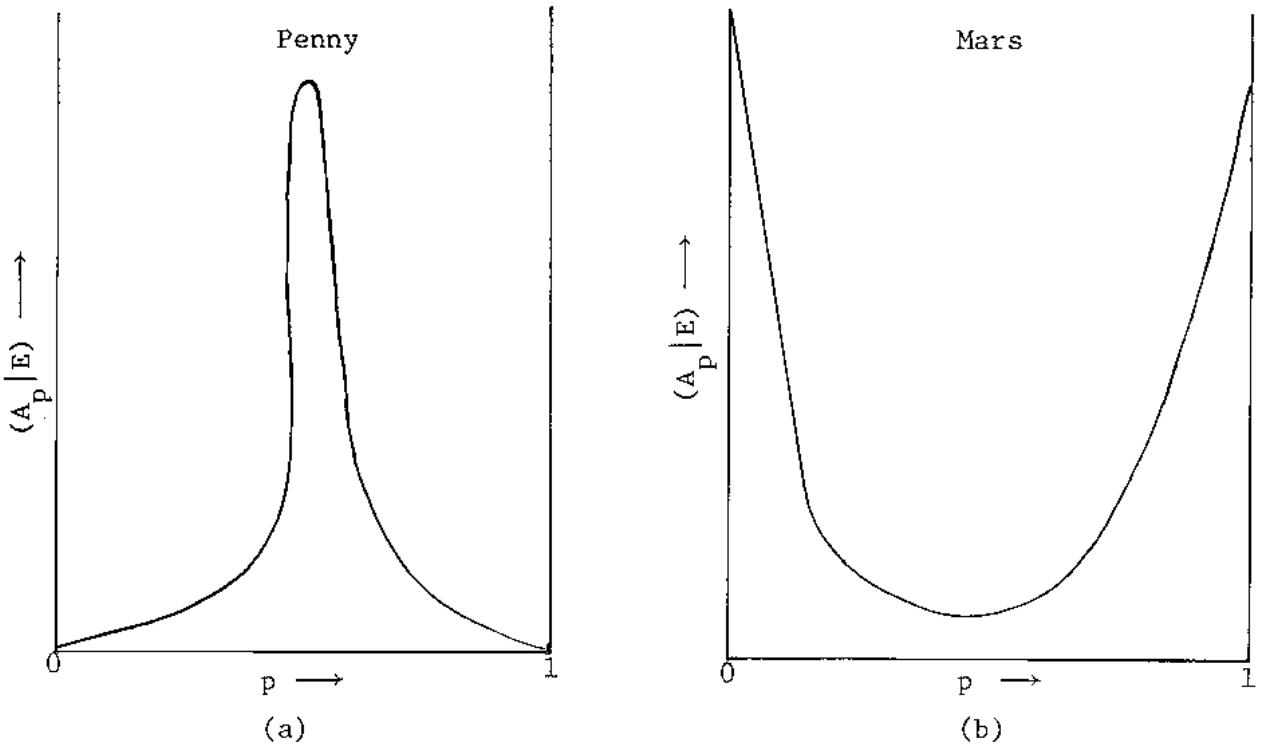

Figure 16.2

in the two cases. So, I assign probability 1/2 to either one; nevertheless, there's all the difference in the world between my state of knowledge about those two propositions, and this difference is represented in the distribution of $(A_p|E)$.

Now, incidentally, I might mention an amusing thing. While I was first working some of this out, a newspaper story showed up from which I would like to read you a few sentences. This is from the Associated Press, December 14, 1957, entitled, "Brain Stockpiles Man's Most Inner Thoughts." It starts out: "Everything you have ever thought, done, or said--a complete record of every conscious moment--is logged in the comprehensive computer of your brain. You will never be able to recall more than the tiniest fraction of it to memory, but you'll never lose it either. These are the findings of Dr. Wilder Penfield, Director of the Montreal Neurological Institute, and a leading Neurosurgeon. The brain's ability to store experiences, many

lying below consciousness, has been recognized for some time, but the extent of this function is recorded by Dr. Penfield."

Now there are several examples given, of experiments on patients suffering from epilepsy. Stimulation of a definite location in the brain recalled a definite experience from the past, which the patient had not been previously able to recall to memory. This has happened many times. I'm sure you have all read about these things. Here are the concluding sentences of this article. Dr. Penfield now says, "This is not memory as we usually use the word, although it may have a relation to it. No man can recall by voluntary effort such a wealth of detail. A man may learn a song so he can sing it perfectly, but he cannot recall in detail any one of the many times he heard it. Most things that a man is able to recall to memory are generalizations and summaries. If it were not so, we might find ourselves confused by too great a richness of detail."

16.2. An Application.

Now let's imagine that a "random" experiment is being performed. From the results of the experiment in the past, we want to do the best job we can of predicting results in the future. To make the problem a definite one, introduce the propositions:

X ≡ "For each trial we admit two prior hypotheses: A true, and A false. The underlying 'causal mechanism' is assumed the same at every trial. This means, for example, that (1) the probability assigned to A at the n'th trial does not depend on n, and (2) evidence concerning the results of past trials retains its relevance for all time; thus for predicting the outcome of trial 1,000, knowledge of the result of trial 1 is just as

relevant as knowledge of the result of trial 999.

There is no other prior evidence.

$N_n \equiv$ "A true n times in N trials in the past."

$M_m \equiv$ "A true m times in M trials in the future."

The verbal statement of X suffers from just the same ambiguities that we have found before, and which have caused so much trouble and controversy in the past. One of the important points I want to put across in these talks is that you have not given any precise description of the prior information until you have given, not verbal statements, but equations, which specify the prior probabilities to be used. In the present problem, this more precise statement of X is, as before

$$(A_p|X) = 1 \quad , \qquad 0 \le p \le 1 \tag{16-16}$$

with the additional understanding that the same A_p-distribution is to be used for calculations pertaining to all trials. What we are after is $(M_m|N_n)$. First, note that by many repetitions of our Rule 1 and Rule 2, in the same way that we found Equation (5-34), we have the binomial distributions

$$\begin{aligned} (N_n|A_p) &= \binom{N}{n} p^n (1-p)^{N-n} \\ (M_m|A_p) &= \binom{M}{m} p^m (1-p)^{M-m} \quad . \end{aligned} \tag{16-17}$$

Note that, although A_p sounds like an awfully dogmatic and indefensible statement to us the way we've introduced it, this is actually the way in which probability is introduced in almost all present testbooks. One postulates that an event possesses some intrinsic, "absolute" or "physical" probability, whose numerical value we can never determine exactly. Nevertheless, no one questions that such an "absolute" probability exists. Cramér (1946, p. 154), for example, takes it as his fundamental axiom. That is just as dogmatic a statement as our A_p; and I think it is, in fact, just our A_p. The equations you see in current textbooks are all like the two I have just

written; whenever p appears as a given number, there's an A_p hiding in the right-hand of your probability symbols.

Mathematically, the only difference between what we're doing here and what is done in current textbooks is that we recognize the existence of that right-hand side for all probabilities, and we are not afraid to use Bayes' theorem to work any proposition whatsoever back and forth from one side of our symbols to the other. I think that in refusing to make free use of Bayes' theorem, modern writers are depriving themselves of the most powerful single principle in probability theory. When a problem of statistical inference is studied long enough, sometimes for decades, one is always forced eventually to a conclusion that could have been derived in three lines from Bayes' theorem. We saw this in the quality-control example and in the case of decision theory; and we'll see several more examples in the remainder of these talks.

Now, we need to find the prior probability $(N_n|X)$. This is already determined from $(A_p|X)$, for our trick of resolving a proposition into mutually exclusive alternatives gives us

$$(N_n|X) = \int_0^1 (N_n A_p|X)\, dp = \int_0^1 (N_n|A_p)(A_p|X)\, dp = \binom{N}{n} \int_0^1 p^n (1-p)^{N-n}\, dp \quad .$$

The integral we have to evaluate is the complete Beta-function:

$$\int_0^1 x^r (1-x)^s\, dx = \frac{r!\, s!}{(r+s+1)!} \tag{16-18}$$

Thus, we have

$$(N_n|X) = \begin{cases} \frac{1}{N+1}, & 0 \le n \le N \\ 0, & N < n \end{cases} \tag{16-19}$$

i.e., just the uniform distribution of maximum entropy. $(M_m|X)$ is similarly found. Now we can turn (16-17) around by Bayes' theorem:

$$(A_p|N_n) = (A_p|X)\, \frac{(N_n|A_p)}{(N_n|X)} = (N+1)(N_n|A_p) \tag{16-20}$$

and so finally the desired probability is

$$(M_m|N_n) = \int_0^1 (M_m A_p|N_n)\, dp = \int_0^1 (M_m|A_p N_n)(A_p|N_n)\, dp \quad . \qquad (16\text{-}21)$$

Since $(M_m|A_p N_n) = (M_m|A_p)$ by the definition of A_p, we have everything in the integrand on the board. Substituting into (16-21), we have again an Eulerian integral, and our result is

$$(M_m|N_n) = \frac{\binom{n+m}{n}\binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} \qquad (16\text{-}22)$$

Note that this is <u>not</u> the same as the hypergeometric distribution (5-23) of sampling theory. Let's look at this result first in the special case M = m = 1. It will then reduce to the probability of A being true in the next trial, given that it had been true n times in the previous N trials. The result is

$$(A|N_n) = \frac{n+1}{N+2} \quad . \qquad (16\text{-}23)$$

This is <u>Laplace's rule of succession</u>. It occupies a supreme position in probability theory; it has been easily the most misunderstood and misapplied rule in the theory, from the time Laplace first gave it in 1774. In almost any book on probability you'll find this rule mentioned very briefly, mainly in order to warn the reader not to use it. But we've got to take the trouble to understand it because in our design of this robot, Laplace's rule of succession is, like Bayes' theorem, one of the most important rules we have. It is a new rule for converting raw information into numerical values of probabilities, and it gives us one of the most important connections between probability and frequency.

<u>16.3. Laplace's Rule of Succession</u>.

Poor old Laplace has been lampooned for generations because he illustrated use of this rule by calculating the probability that the sun will rise tomorrow, given that it has risen every day for the past 5,000 years. One gets a

rather large factor in favor of the sun rising again tomorrow, of course. With no exceptions at all as far as I know, modern writers on probability have considered this a pure absurdity. Even Jeffreys and Carnap find fault with the rule of succession.

I have to confess to you that I am unable to see anything at all absurd about the rule of succession. I recommend very strongly that you do a little literature searching, and read some of the objections various writers have to it. I think you will see that in every case the same thing has happened. First, Laplace was quoted out of context, and secondly, in order to demonstrate the absurdity of the rule of succession, the author applies it to a case where it was never intended to be applied, because there is additional prior information which was not taken into account.

If you go back and read Laplace (1819) himself, you will see that in the very next sentence after this sunrise episode, he points out to the reader that this is the probability based only on the information that the event has occurred n times in N trials, and that our knowledge of celestial mechanics represents a great deal of additional information. Of course, if you have additional information beyond the numbers n and N, then you ought to take it into account. You are then considering a different problem, the rule of succession no longer applies, and you can get an entirely different answer. This theory gives the results of consistent plausible reasoning on the basis of the information which was put into it.

Let me give you three famous examples of the kind of objections to the rule of succession which you find in the literature. (1) Suppose the solidification of hydrogen to have been once accomplished. According to the rule of succession, the probability that it will solidify again if the experiment is repeated is 2/3. This does not in the least represent the state of belief of any scientist. (2) A boy is 10 years old today. According to the rule

of succession, he has the probability 11/12 of living one more year. His grandfather is 70; and so according to this rule he has the probability 71/72 of living one more year. The rule violates qualitative common sense! (3) Consider the case N = n = 0. It then says that any conjecture without any verification has the probability 1/2. Thus there is probability 1/2 that there are exactly 137 elephants on Mars. Also there is probability 1/2 that there are 138 elephants on Mars. Therefore, it is certain that there are at least 137 elephants on Mars. But the rule says also that there is probability 1/2 that there are no elephants on Mars. The rule is logically self-contradictory!

The trouble with examples (1) and (2) is obvious in view of our earlier remarks; in each case, an enormous amount of highly relevant prior information, known to all of us, was simply ignored, producing a flagrant misuse of the rule of succession. But let's look a little more closely at example (3). Wasn't the law applied correctly here? I certainly can't claim that we had prior information about elephants on Mars which was ignored, can I? And even if I could, that still wouldn't account for the self-contradiction. Evidently, if the rule of succession is going to survive example (3), there must be some very basic points about the use of probability theory which we still have to learn.

Well, now, what do we mean when we say that there's no evidence for a proposition? The question is not what you or I might mean colloquially by such a statement. The question is, what does it mean to the robot? What does it mean in terms of probability theory?

The prior information we used in derivation of the rule of succession was that the robot is told that there are only two possibilities: A true, and A false. His entire "universe of discourse" consists of only two propositions. In the case N = 0, we could solve the problem also by direct appli-

cation of the principle of indifference, Rule 4; and this will of course give the same answer $(A|X) = 1/2$, that we got from the rule of succession. But just by noting this, we see what is wrong. Merely by admitting the possibility of three different propositions being true, instead of only two, we have already specified prior information different from that used in deriving the rule of succession.

If the robot is told to consider 137 different ways in which A could be false, and only one way in which it could be true, then the prior probability of A is 1/138, not 1/2. So, we see that the example of the elephants on Mars was, again, a gross misapplication of the rule of succession.

Moral: Probability theory, like any other mathematical theory, cannot give us a definite answer unless we ask it a definite question. We should always start a problem with an explicit enumeration of the different propositions we're going to consider. That is part of the "boundary conditions" which must be specified before we have a uniquely defined mathematical problem. If you say, "I don't know what the possible propositions are," that is mathematically equivalent to saying, "I don't know what problem I want to solve." This is just the point that I have already belabored back in Lecture 7.

In this connection we have to remember that probability theory never solves problems of actual practice, because all such problems are infinitely complicated. We solve only idealizations of the real problem, and the solution is useful to the extent that the idealization is a good one. In the example of the solidification of hydrogen, the prior information which our common sense uses so easily, is actually so complicated that nobody knows how to convert it into a prior probability assignment. I don't think there is any reason to doubt that probability theory is, in principle, competent to deal with such problems; but we have not yet learned how to translate them into

mathematical language without oversimplifying so much that the solution is useless.

Laplace's rule of succession provides a definite solution to a definite problem. Everybody denounces it as nonsense because it is not also the solution to some other problem. The case where the problem can be reasonably idealized to one with only two hypotheses to be considered, a belief in a constant "causal mechanism," and no other prior information, is the only case where it applies. You can, of course, generalize it to any number of hypotheses, and let me just give you the result of doing this.

There are K different hypotheses, $\{A_1, A_2, \ldots, A_K\}$, a belief that the "causal mechanism" is constant, and no other prior information. We perform a random experiment N times, and observe A_1 true n_1 times, A_2 true n_2 times, etc. Of course, $\sum_i n_i = N$. On the basis of this evidence, what is the probability that in the next $M = \sum_i m_i$ repetitions of the experiment, A_i will be true exactly m_i times? To find the distribution $(m_1 \ldots m_K | n_1 \ldots n_K)$ that answers this, define the prior knowledge by a K-dimensional uniform prior distribution

$$(A_{p_1 \cdots p_K} | X) = C\, \delta(p_1 + \ldots + p_K - 1) \quad , \quad p_i \geq 0 \tag{16-24}$$

To find the normalization constant C, we set

$$\int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; (A_{p_1 \cdots p_K} | X) = 1 = C\, I(1) \tag{16-25}$$

where

$$I(r) \equiv \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; \delta(p_1 + \ldots + p_K - r) \tag{16-26}$$

Direct evaluation of this would be rather messy, so let's use the following trick. First, take the Laplace transform of (16-26)

$$\int_0^\infty e^{-\alpha r} I(r)\, dr = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; e^{-\alpha(p_1 + \ldots + p_K)} = \frac{1}{\alpha^K} \; . \tag{16-27}$$

Or, inverting the Laplace transform,

$$I(r) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \frac{e^{\alpha r}}{\alpha^K}\, d\alpha = \frac{1}{(K-1)!}\frac{d^{K-1}}{d\alpha^{K-1}} e^{\alpha r}\Big|_{\alpha=0}$$

$$= \frac{r^{K-1}}{(K-1)!} \tag{16-28}$$

Thus,

$$C = \frac{1}{I(1)} = (K-1)! \tag{16-29}$$

By this device, we avoided having to consider complicated details about different ranges of integration over the different p_i, that would come up if we tried to evaluate (16-26) directly.

The prior distribution $(n_1 \ldots n_K | X)$ is then, using the same trick,

$$(n_1 \ldots n_K | X) = \frac{N!}{n_1! \ldots n_K!}\int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, (A_{p_1 \ldots p_K} | X)$$

$$= \frac{N!\ (K-1)!}{n_1! \ldots n_K!}\, J(1) \tag{16-30}$$

where

$$J(r) \equiv \int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, \delta(p_1 + \ldots + p_K - r) \tag{16-31}$$

which we evaluate as before by taking the Laplace transform:

$$\int_0^\infty e^{-\alpha r}\, J(r)\, dr = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, e^{-\alpha(p_1 + \ldots + p_K)}$$

$$= \prod_{i=1}^K \frac{n_i!}{\alpha^{n_i+1}} \tag{16-32}$$

So, as in (16-28), we have

$$J(r) = \frac{n_1! \ldots n_K!}{2\pi i}\int_{-i\infty}^{i\infty} \frac{e^{\alpha r}\, d\alpha}{\alpha^{N+K}} = \frac{n_1! \ldots n_K!}{(N+K-1)!}\, r^{N+K-1} \tag{16-33}$$

and

$$(n_1 \ldots n_K | X) = \frac{N!(K-1)!}{(N+K-1)!}, \qquad n_i \geq 0,\ n_1+\ldots+n_K = N \tag{16-34}$$

Therefore, by Bayes' theorem

$$(A_{p_1 \ldots p_K} | n_1 \ldots n_K) = (A_{p_1 \ldots p_K} | X) \frac{(n_1 \ldots n_K | A_{p_1 \ldots p_K})}{(n_1 \ldots n_K | X)}$$

$$= \frac{(N+K-1)!}{n_1! \ldots n_K!} p_1^{n_1} \ldots p_K^{n_K} \delta(p_1+\ldots+p_K-1) \tag{16-35}$$

and finally

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \, (m_1 \ldots m_K | A_{p_1 \ldots p_K}) (A_{p_1 \ldots p_K} | n_1 \ldots n_K)$$

$$= \frac{M!}{m_1! \ldots m_K!} \frac{(N+K-1)!}{n_1! \ldots n_K!} \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \, p_1^{n_1+m_1} \ldots p_K^{n_K+m_K} \delta(p_1+\ldots+p_K-1) \tag{16-36}$$

The integral is the same as J(1) except for the replacement $n_i \to n_i+m_i$. So, from (16-33),

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \frac{M!}{m_1! \ldots m_K!} \frac{(N+K-1)!}{n_1! \ldots n_K!} \frac{(n_1+m_1)! \ldots (n_K+m_K)!}{(N+M+K-1)!} \tag{16-37}$$

or, reorganizing into binomial coefficients,

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \frac{\binom{n_1+m_1}{n_1} \ldots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}} . \tag{16-38}$$

In the case where we want just the probability that A_1 will be true on the next trial, we need this formula with $M = m_1 = 1$, all other $m_i = 0$. The result is the generalized law of succession:

$$(A_1 | n_1, N, K) = \frac{n_1 + 1}{N + K} . \tag{16-39}$$

You see that in the case $N = n_1 = 0$, this reduces to the answer provided by the principle of indifference, Rule 4, which it therefore contains as a

special case. If K is a power of 2, this is the same as a method of inductive reasoning proposed by Carnap in 1945, which he denotes as c*(h,e) in his "Continuum of Inductive Methods."

Now, use of the rule of succession in cases where N is very small is rather foolish, of course. Not really wrong; just foolish. Because if we have no prior evidence about A, and we make such a small number of observations that we get practically no evidence; well, that's just not a very promising basis on which to do plausible reasoning. We can't expect to get anything useful out of it. We do, of course, get definite numerical values for the probabilities, but these values are very "soft," i.e., very unstable, because the A_p distribution is still very broad for small N. Our common sense tells us that the evidence N_n for small N provides no reliable basis for further predictions, and we'll see in the next lecture that this conclusion also follows as a consequence of the theory we're developing here.

The real reason for introducing the rule of succession lies in the cases where we <u>do</u> get a significant amount of information from the random experiment; i.e., when N is a large number. In this case, fortunately, we can pretty much forget about these fine points concerning prior evidence. The particular initial assignment $(A_p|X)$ will no longer have much influence on the results, for the same reason as in the particle-counter problem. This remains true for the generalized case leading to (16-38). You see from (16-39) that as soon as the number of observations N is large compared to the number of hypotheses K, then the probability assigned to any particular hypothesis depends for all practical purposes, only on what we have observed, and not on how many prior hypotheses there are. If you contemplate this for ten seconds, I think your common sense will tell you that the criterion $N \gg K$ is exactly the right one for this to be so.

16.4. Confirmation and Weight of Evidence.

Now, I'd like to introduce a few new ideas which are suggested by our calculations involving A_p. We saw that the stability of probability assignment in the face of new evidence is essentially determined by the width of the A_p distribution. If E is prior evidence and F is new evidence, then

$$(A|EF) = \int_0^1 (A_p|EF)\ p\ dp = \frac{\int_0^1 (A_p|F)(A_p|E)\ p\ dp}{\int_0^1 (A_p|F)(A_p|E)\ dp} \tag{16-40}$$

We'll say that F is compatible with E, as far as A is concerned, if having the new evidence, F, doesn't make any appreciable change in the probability of A; i.e.,

$$(A|EF) = (A|E) \tag{16-41}$$

The new evidence can make an enormous change in the distribution of A_p without changing the first moment. It might sharpen it up very much, or broaden it. We could become either more certain or more uncertain about A, but if F doesn't change the center of gravity of the A_p distribution, we still end up assigning the same probability to A.

Now, the stronger property: the new evidence F confirms the previous probability assignment, if F is compatible with it, and at the same time, gives us more confidence in it. In other words, we exclude one of these possibilities, and with new evidence F the A_p distribution narrows. Suppose F consists of performing some random experiment and observing the frequency with which A is true. In this case $F = N_n$, and our previous result, Eq. (16-20), gives

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!}\, p^n (1-p)^{N-n}$$

$$\simeq (\text{constant}) \cdot \exp\left[-\frac{(p-f)^2}{2\sigma^2}\right] \tag{16-42}$$

where

$$\sigma^2 = \frac{f(1-f)}{N} \tag{16-43}$$

and f = (n/N) is the observed frequency of A. The approximation is derived by expanding $\log(A_p|N_n)$ in a Taylor series about its peak value, and is valid when n>>1 and (N-n)>> 1. If these conditions are satisfied, then $(A_p|N_n)$ is very nearly symmetric about its peak value. Then, if the observed frequency f is close to the prior probability (A|E), the new evidence N_n will not affect the first moment of the A_p distribution, but will sharpen it up, and that will constitute a confirmation as I defined it. This shows one more connection between probability and frequency. I defined the "confirmation" of a probability assignment according to entirely different ideas than are usually used to define it. I defined it in a way that agrees with our intuitive notion of confirmation of a previous state of mind. But it turned out that the same experimental evidence would constitute confirmation on either the frequency theory or our theory.

Now, from this we can see another useful notion; which I'll call weight of evidence.

Let's consider A_p, given two different pieces of evidence, E and F.

$$(A_p|EF) = (\text{constant})(A_p|E)(A_p|F) \tag{16-44}$$

If the distribution $(A_p|F)$ was very much sharper than the distribution $(A_p|E)$, then the product of the two would still have its peak at practically the value determined by F. In this case, we would say that the evidence F carries much greater "weight" than the evidence E. If we have F, it doesn't really matter much whether we take E into account or not. On the other hand, if we don't have F, then whatever evidence E may represent will be extremely significant, because it will represent the best we are able to do. So, acquiring one piece of evidence which carries a great amount of weight can

make it, for all practical purposes, unnecessary to continue keeping track of other pieces of evidence which carry only a small weight.

Of course, this is exactly the way our minds operate. When we receive one very significant piece of evidence, we no longer pay so much attention to vague evidence. In so doing, we are not being very inconsistent, because it wouldn't make much difference anyway. So, our intuitive notion of weight of evidence is bound up with the sharpness of this A_p distribution. Evidence concerning A that we consider very significant is not necessarily evidence that makes a big change in the probability of A. It is evidence that makes a big change in this distribution of A_p. Now seeing this, we can get a little more insight into the principle of indifference, Rule 4, and also make contact between this theory and Carnap's methods of inductive reasoning.

Before we can use the principle of indifference to assign numerical values of probabilities, there are two different conditions that have to be satisfied: (1) we have to be able to analyze the situation into mutually exclusive, exhaustive possibilities; (2) having done this, we must then find that the available information gives us no reason to prefer any of the possibilities to any other. In practice, these conditions are hardly ever met unless there's some evident element of symmetry in the problem. But there are two entirely different ways in which condition (2) might be satisfied. It might be satisfied as a result of ignorance, or it might be satisfied as a result of positive knowledge about the situation.

To illustrate this, let's suppose that a person who is known to be very dishonest is going to toss a coin and there are two people watching him. Mr. A is allowed to examine the coin. He has all the facilities of the National Bureau of Standards at his disposal. He performs thousands of experiments with scales and calipers and magnetometers and microscopes, X-rays, and neutron beams, and so on. Finally, he is convinced that the coin <u>is</u> perfectly

honest. Mr. B is not allowed to do this. All he knows is that a coin is being tossed by a shady character. He suspects the coin is biased, but he has no idea in which direction.

Condition (2) is satisfied equally well for both of these people. Each of them would start out by assigning probability one-half to each face. The same probability assignment can described a condition of complete ignorance or a condition of very great knowledge. Now, this sort of situation has seemed paradoxical for a long time. Why doesn't Mr. A's extra knowledge make any difference? Well, of course, it does make a difference. It makes a very important difference, but one that doesn't show up until we start performing this random experiment. The difference is not in the probability of A, but in the distribution of A_p.

Suppose the first toss is heads. To Mr. B, that constitutes evidence that the coin is biased to favor heads. And so, on the next toss, he would assign new probabilities to take that into account. But to Mr. A, the evidence that the coin is honest carries overwhelmingly greater weight than the evidence of one throw, and he'll continue to assign a probability of 1/2.

Well, now, you see what's going to happen. To Mr. B, every toss of the coin represents new evidence about its bias. Every time it's tossed, he will revise his assignments for the next toss; but after several tosses his assignments will get more and more stable, and in the limit $N \to \infty$ they will tend to the observed frequency of heads. To observer A, the evidence of symmetry continues to carry greater weight than the evidence of almost any number of throws, and he persists in assigning probability 1/2. Each has done consistent plausible reasoning on the basis of the information available to him, and our theory accounts for the behavior of each.

If you assumed that Mr. A had perfect knowledge of symmetry, you might conclude that his A_p distribution is a true δ-function. In that case, his

mind could never be changed by any amount of new data from the random experiment. Of course, that's a limiting case that's never reached in practice. Not even the Bureau of Standards can give us evidence that good.

16.5. Carnap's Inductive Methods.

Carnap (1952) gives an infinite family of possible "inductive methods," by which one can convert prior information and frequency data into a probability assignment and an estimate of frequencies for the future. His principle is that the final probability assignment $(A|N_nX)$ should be a weighted average of the prior probability $(A|X)$ and the observed frequency, $f = n/N$. Assigning a weight N to the "empirical factor" f, and an arbitrary weight λ to the "logical factor" $(A|X)$ leads to the method which Carnap denotes by $c_\lambda(h,e)$. Introduction of the A_p distribution accounts for this in more detail; the theory developed here includes all of Carnap's methods as special cases corresponding to different prior distributions $(A_p|X)$, and leads us to reinterpret λ as the weight of prior evidence. Thus, in the case of two hypotheses, the Carnap λ-method is the one you can calculate from the prior distribution $(A_p|X) = (\text{const.})\cdot[p(1-p)]^r$, with $2r = \lambda - 2$. The result is

$$(A|N_nX) = \frac{2n + \lambda}{2N + 2\lambda} = \frac{(n+r) + 1}{(N+2r) + 2} \quad . \qquad (16\text{-}45)$$

Greater λ thus corresponds to a more sharply peaked $(A_p|X)$ distribution.

In our coin-tossing example, the gentleman from the Bureau of Standards reasons according to a Carnap method with λ of the order of, perhaps, thousands to millions; while Mr. B, with much less prior knowledge about the coin, would use a λ of perhaps 5 or 6. (The case $\lambda = 2$, which gives Laplace's rule of succession, is much too broad to be realistic for coin tossing; for Mr. B surely knows that the center of gravity of a coin can't be moved by more than half its thickness from the geometrical center. Actually, as we will see in Lecture 19, this analysis isn't always applicable to tossing of real coins,

for reasons having to do with the laws of physics.)

From the second way I wrote Equation (16-45), you see that the Carnap λ-method corresponds to a weight of prior evidence which would be given by $(\lambda-2)$ trials, in exactly half of which A was observed to be true. Can we understand why the weighting of prior evidence is λ = (number of prior trials + 2), while that of the new evidence N_n is only (number of new trials) = N? Well, look at it this way. The appearance of the (+2) is the robot's way of telling us this: prior knowledge that it is possible for A to be either true or false, is equivalent to knowledge that A has been true at least once, and false at least once. This is hardly a derivation; but I think it makes excellent common sense.

But let's pursue this line of reasoning a step further. We started with the statement X: it is possible for A to be either true or false at any trial; but that is still a somewhat vague statement. Suppose we interpret it as meaning that A has been observed true exactly once, and false exactly once. If we grant that this state of knowledge is correctly described by Laplace's assignment $(A_p \mid X) = 1$, then what was the "pre-prior" state of knowledge before we had the data X? To answer this, we need only apply Bayes' theorem backwards, as we did at the beginning of Lecture 7. The result is: our "pre-prior" A_p-distribution must have been

$$(A_p \mid\ )\ dp = (\text{const.})\ \frac{dp}{p(1-p)} \tag{16-46}$$

which is the quasi-distribution representing "complete ignorance," or the "basic measure" of our parameter space, that we found by transformation groups in Lecture 12. So, here is another line of thought that could have led us to this measure.

It appears, then, that if we have definite prior evidence that it is possible for A to be either true or false on any one trial, then Laplace's

rule $(A_p|X) = 1$ is the appropriate one to use. But if initially we are so completely uncertain that we're not even sure whether it is possible for A to be true on some trials and false on others, then we should use the prior (16-46).

How different are the numerical results which the pre-prior assignment (16-46) gives us? Repeating the derivation of (16-20) with this pre-prior assignment we find that, provided n is not zero or N,

$$(A_p|N_n)' = \frac{(N-1)!}{(n-1)!\,(N-n-1)!}\,p^{n-1}(1-p)^{N-n-1} \tag{16-47}$$

which leads, instead of to Laplace's rule of succession, to the mean-value estimate of p:

$$(A|N_n)' = \int_0^1 (A_p|N_n)'\; p\; dp = \frac{n}{N} \tag{16-48}$$

equal to the observed frequency, and identical with the maximum-likelihood estimate of p. Likewise, provided $0 < n < N$, we find instead of (16-22) the formula

$$(M_m|N_n)' = \frac{\binom{m+n-1}{m}\binom{M-m+N-n-1}{M-m}}{\binom{N+M-1}{M}} \tag{16-49}$$

All of these results correspond to having observed one less success and one less failure.