Lecture 7

QUEER USES FOR BAYES' THEOREM

I. J. Good (Good, 1950) has shown how we can use Bayes' theorem backwards to measure our own strengths of belief about propositions. For example, how strongly do you believe in extrasensory perception?

7.1. Extrasensory Perception.

What probability would you assign to the proposition that Mr. Smith has perfect extrasensory perception? He can guess right every time which number you are thinking of. Well now, to say zero--that, of course, is dogmatic. According to our theory, if you start out at $-\infty$ db, this means that you are never going to allow your mind to be changed by any amount of evidence, and you don't really mean that. But where *is* our strength of belief in a proposition like this? Our brains work pretty much the way this robot works, but *we* have an intuitive feeling for plausibility only when it's not too far from 0 db. We feel that something is more than likely to be so or less than likely to be so. We get fairly definite feelings about that. So the trick is to imagine an experiment. How much evidence would it take to bring my state of belief up to the place where I felt very perplexed and unsure about it? Not to the place where I believed it--that would overshoot the mark, and again we'd lose our resolving power. How much evidence would it take to bring you just up to the point where you were beginning to consider the possibility seriously?

We take this man who says he has extrasensory perception, and we will write down some numbers from 1 to 10 on a piece of paper and ask him to guess which numbers we've written down. We'll take the usual precautions to make sure against other ways of finding out. All right, if he guesses the first number correctly, of course we'll say "you're a very lucky person, but I don't believe it." And if he guesses two numbers correctly, we'll say "you're a very lucky person, but I don't believe it." By the time he's guessed four numbers correctly--well, I still wouldn't believe it. So my state of belief is certainly lower than -40 db. How many numbers would he have to guess correctly before you would really seriously consider the hypothesis that he has extrasensory perception? In my own case, I think somewhere around 10. My personal state of belief is, therefore, about -100 db. You could talk me into a $\pm$10 change fairly easily, and perhaps $\pm$20; but not much more than that.

7.2. Bayesian Jurisprudence.

It is interesting also to apply Bayes' theorem to various situations in which we can't really reduce it to numbers very well, but still it shows automatically what kind of information would be relevant to help us do plausible reasoning. Suppose someone in New York City has committed a murder, and you don't know at first who it is. Suppose there are 10 million people in New York City. On the basis of no knowledge but this, $e(\text{Guilty}|X) = -70$ db is the plausibility that any particular person is the guilty one.

How much positive evidence is necessary before we decide some man should be put away? Maybe +40 db, although your first reaction may be that this is not safe enough, and the figure ought to be higher. If we raise this figure, we give increased protection to the innocent, but at the cost of making it more difficult to convict the guilty; and at some point the

interests of society as a whole must take precedence over sentiment.

For example, if a thousand guilty men are set free, we know from only too much experience that two or three hundred of them will immediately proceed to inflict still more crimes upon society, and their escaping justice will encourage a hundred more to take up crime. So, I think it is clear that the damage done to society as a whole by allowing a thousand guilty men to go free, is far greater than that caused by falsely convicting one innocent man. If you have a sentimental reaction against this statement, I ask you to think: if you were a judge, would you rather face one man whom you had convicted falsely; or a hundred victims of crimes resulting from your lenience? Setting the threshold at +40 db will mean, crudely, that on the average not more than one conviction in ten thousand will be in error; a judge following this rule will probably not make one false conviction in a working lifetime on the bench. It seems to me that this is a reasonable figure that we can accept. Obviously, however, this matter ought to be researched much more carefully than we can do here.

So, if we took +40 db starting out from -70, this means that in order to get conviction you would have to produce about 110 db of evidence in favor of the guilt of this particular person.

Suppose now we learn that this person had a motive. What does that do to the plausibility of his guilt? Well, Bayes' theorem says

$$e(\text{Guilty}|\text{Motive}) = e(\text{Guilty}|X) + 10 \log_{10} \frac{(\text{Motive}|\text{Guilty})}{(\text{Motive}|\text{Not Guilty})} \qquad (7\text{-}1)$$

$$\simeq -70 - 10 \log_{10} (\text{Motive}|\text{Not Guilty})$$

since $(\text{Motive}|\text{Guilty}) \simeq 1$; i.e., we consider it quite unlikely that the crime had no motive at all. Thus, the significance of learning that the person had a motive depends almost entirely on the probability $(\text{Motive}|\text{Not Guilty})$ that an innocent person would also have a motive. This evidently

agrees exactly with our common sense; if the deceased were kind and loved by all, hardly anyone would have had a motive to do him in. Learning that, nevertheless, our suspect did have a motive, would then be very significant information. If the victim had been an unsavory character, who took great delight in all sorts of foul deeds, then a great many people would have a motive, and learning that our suspect was one of them, is not so significant. The point of this is that we don't really know what to make of the information that our suspect had a motive, unless we also know something about the character of the deceased. But how many members of juries would realize that, unless it was specifically pointed out to them?

Suppose that a very enlightened judge, with powers not given to judges under present law, had perceived this fact and, when testimony about the motive was introduced, he directed his assistants to obtain for the jury the most reliable data possible on the number of people in New York City who had a motive. This number was N_m. Then

$$(\text{Motive}|\text{Not Guilty}) = \frac{N_m - 1}{(\text{number of people in New York}) - 1} \simeq 10^{-7}(N_m - 1)$$

and equation (7-1) reduces, for all practical purposes, to

$$e(\text{Guilty}|\text{Motive}) \simeq -70 + 10 \log [10^7/(N_m - 1)] = -10 \log (N_m - 1). \quad (7\text{-}2)$$

You see that the population of New York has cancelled out of the equation; as soon as we know the number of people who had a motive, then it doesn't matter any more how large the city was.

Well, you can go on this way for a long time, and I think you will find it both enlightening and entertaining to do so. For example, we now learn that the suspect had bought a gun the day before the crime. Or that he was seen at the scene of the crime shortly before. If you have ever been told not to trust Bayes' theorem, you should follow a few examples like this a good deal further, and see how infallibly it tells you what information

would be relevant, what irrelevant, in plausible reasoning. Even in situations where we would be quite unable to say what numerical values should be used, it still reproduces qualitatively just what your common sense (after perhaps a little meditation) tells you.

7.3. Testing Scientific Theories.

Another class of applications of Bayes' theorem, which has been discussed vigorously by philosophers of science for over a century, concerns the reasoning process of a scientist, by which he accepts or rejects his theories in the light of the observed facts. I mentioned in the second lecture that this consists largely of the use of two forms of syllogism,

$$\text{one strong: } \left\{\begin{array}{l}\text{If A, then B}\\ \underline{\text{\quad B false \quad}}\\ \text{A false}\end{array}\right\}, \text{ and one weak: } \left\{\begin{array}{l}\text{If A, then B}\\ \underline{\text{\quad B true \quad}}\\ \text{A more plausible}\end{array}\right\}$$

We see that these correspond to the use of Bayes' theorem in the forms

$$(A|b) = (A|X)\,\frac{(b|A)}{(b|X)}\,, \qquad (A|B) = (A|X)\,\frac{(B|A)}{(B|X)}$$

respectively. It is at once obvious that Bayes' theorem accounts for the strong syllogism; for if $(B|A) = 1$, Bayes' theorem gives $(A|b) = 0$; our rules for plausible reasoning include those of deductive reasoning as a special case.

Interest here centers on the question whether the second form of Bayes' theorem gives a satisfactory quantitative version of the weak syllogism. Let us consider a specific example given by Professor George Polya [Polya, 1954; Vol. II, pp. 130-132]. The planet Uranus was discovered by Herschel in 1781. Within a few decades (i.e. by the time Uranus had traversed about one third of its orbit), it was clear that it was not following exactly the path prescribed for it by the Newtonian theory (laws of mechanics and gravitation). At this point, a naive application of the strong syllogism might lead one to conclude that the Newtonian theory was demolished. However,

its many other successes had established the Newtonian theory so firmly that to the French astronomer Leverrier, an alternative hypothesis was rendered more plausible: there must be still another planet beyond Uranus, whose gravitational pull is causing the discrepancy.

Working backwards, Leverrier computed the mass and orbit of a planet which could produce the observed deviation and predicted where the new planet would be found. An observatory received Leverrier's prediction on September 23, 1846, and on the evening of the same day, the new planet (Neptune) was discovered within one degree of the predicted position!

Instinctively, we feel that the plausibility of the Newtonian theory was increased by this little drama. The question is, how much? The attempt to apply Bayes' theorem to this problem will give us a good example of the complexity of actual situations faced by scientists, and also of the caution which must be exercised in reading the rather confused literature on these problems.

Following Polya's notation, let T stand for the Newtonian theory, N for the part of Leverrier's prediction that was verified. Then Bayes' theorem gives for the posterior probability of T,

$$(T|N) = (T|X)\,\frac{(N|TX)}{(N|X)}\,. \tag{7-3}$$

Suppose we try to evaluate $(N|X)$. This is the prior probability of N, regardless of whether T is true or not. Since $N = N(T+t) = NT + Nt$, we have, by applying Rule 3, then Rule 1,

$$\begin{aligned}(N|X) &= (NT + Nt|X) = (NT|X) + (Nt|X)\\ &= (N|TX)(T|X) + (N|tX)(t|X)\end{aligned} \tag{7-4}$$

and you see that $(N|tX)$ has intruded itself into the problem. But in the problem as stated this quantity is not defined; the statement t ≡ "Newton's theory is false" has no definite implications until we specify what alterna-

tive we have to put in place of Newton's theory.

For example, if there were only a single possible alternative according to which there could be no planets beyond Uranus, then $(N|tX) = 0$, and Bayes' theorem would again reduce to deductive reasoning, giving $(T|N) = 1$, independently of the prior probability $(T|X)$. On the other hand, if Einstein's theory were the only possible alternative, its predictions do not differ appreciably from those of Newton's theory for this phenomenon, and we would have $(N|tX) = (T|X)$, whereupon Bayes' theorem reduces to $(T|N) = (T|X)$. Verification of Leverrier's prediction might elevate the Newtonian theory to certainty, or it might have no effect at all on its plausibility! It depends entirely on this: Against which specific alternatives are we testing Newton's theory?

Now to a scientist who is judging his theories, this conclusion is the most obvious exercise of common sense. Yet statisticians have developed criteria for accepting or rejecting theories (Chi-squared test, etc.) which make no reference to any alternatives. A practical difficulty of this was pointed out forcefully by Sir Harold Jeffreys (Jeffreys, 1939); there is not the slightest use in rejecting any hypothesis H unless we can do it in favor of some definite alternative H' which better fits the facts.* Bayes' theorem tells us much more than this: unless the observed facts are absolutely impossible on hypothesis H, it is meaningless to ask how much those facts tend "in themselves" to confirm or refute H. Not only the mathematics, but also our common sense (if we think about it for a minute) tells us that we have not asked any definite, well-posed question

*I don't mean to argue against the use of the Chi-squared test itself; later in these lectures, when we take up significance tests, we will see that in some cases it is very nearly the right test to answer a different question, namely: "Within a certain specified class of alternatives H', do any exist which better fit the facts, and how much improvement in fit is possible?"

until we specify the possible alternatives to H.

Of course, as the observed facts approach impossibility on hypothesis H, we are led to worry more and more about H; but mere improbability, however great, cannot in itself be the reason for doubting H. For example, if I toss a coin 1000 times, then no matter what the result is, the specific observed sequence of heads and tails has a probability of only 2^{-1000}, or minus 3000 decibels, on the hypothesis that the coin is honest. If, after having tossed it 1000 times, I still believe that the coin is honest, it can be only because the observed sequence is even more improbable on any alternative hypothesis that I am willing to consider seriously. This situation will be analyzed more deeply later on, where it will lead to a general formulation of significance tests.

We see here that, even when the application is only qualitative, classical probability theory is still useful to us in a normative sense; it is the test by which we can detect inconsistencies in our own reasoning. Some authors have argued strongly against the use of Bayes' theorem for testing hypotheses. But when we take the trouble to learn what it actually says, we find that Bayes' theorem tells immediately what is needed before we have any rational criterion for testing hypotheses.

This brings us to some comparisons with the literature. In Polya's discussion of Bayes' theorem applied to the status of Newton's theory before and after Leverrier's feat, no specific alternative to Newton's theory is stated; but from the numerical values used (loc. cit., p. 131) we can infer that the alternative H' was one according to which it was known that one more planet existed beyond Uranus, but all directions on the celestial sphere were considered equally likely. Unfortunately, in the calculation no distinction was made between $(N|X)$ and $(N|tX)$; and consequently the quantity which Polya interprets as the ratio of posterior to prior probabi-

lities of Newton's theory, is actually the ratio of posterior to prior odds. This is, in our notation, $(N|TX)/(N|tX) = (N|TX)/(N|H'X) \simeq 13{,}000$.

The conclusions are much more satisfactory when we notice this. Whatever prior probability $(T|X)$ we imagine Newton's theory to have, if H' is the only alternative considered, then verification of Leverrier's prediction increased the evidence for Newton's theory by $10 \log_{10}(13{,}000) \simeq 41$ decibels. Actually, if there were a new planet, it would be reasonable to adopt a different alternative hypothesis H", according to which its orbit would lie in the plane of the ecliptic, as Polya points out. If, on hypothesis H", all values of longitude are considered equally likely, we might reduce this figure to about $10 \log_{10}[(N|TX)/(N|H"X)] = 10 \log_{10}(180) \simeq 23$ decibels. In view of the great uncertainty as to just what the alternative is, it seems to me any value between these extremes is more or less reasonable.

There was a difficulty (which Polya interpreted as revealing an inconsistency in Bayes' theorem), that if the probability of Newton's theory were increased by a factor of 13,000, then the prior probability was necessarily lower than (1/13,000); but this contradicts common sense, because Newton's theory was already very well established before Leverrier was born. Recognition that we are, in the above numbers, dealing with odds rather than probabilities, completely removes this objection and makes Bayes' theorem appear quite satisfactory in describing the inductive reasoning of a scientist. This is a good example of the way in which objections to the Bayes-Laplace methods which you find in the literature, disappear when you look at the problem more carefully.

But the example also shows clearly that in practice the situation faced by the scientist is so complicated that there is little hope of applying Bayes' theorem to give quantitative results about the relative status of theories. Also there is no need to do this, because the real difficulty

of the scientist is not in the reasoning process itself; his common sense is quite adequate for that. The real difficulty is in learning how to formulate new alternatives which better fit the facts. Usually, when one succeeds in doing this, the evidence for the new theory soon becomes so overwhelming that no one needs probability theory to tell him what conclusions to draw. So, I would say that in principle the application of Bayes' theorem in the above way is perfectly legitimate; but in practice it is of very little use to a scientist.

7.4. Different Views on Probability Theory.

Professor L. J. Savage (Savage, 1954) has written an excellent survey of the foundations of statistics, in which he clearly recognizes, and gives a rigorous discussion of, many of the points that I am trying to put across here in a more informal way. He gives a broad classification of attitudes toward probability theory into three different camps:

(a) Objectivistic. Probability has nothing whatsoever to do with "degree of reasonable belief" or inductive reasoning. By "probability" we must mean only observable frequencies in independent repetitions of a random experiment.

(b) Personalistic. Probability can be used legitimately to describe the degree of confidence that a particular individual has in the truth of a proposition, but probability assignments are not unique; two individuals having the same prior evidence may assign different probabilities without either being unreasonable.

(c) "Necessary views hold that probability measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another. They are generally regarded by their holders as extensions of logic, which

tells when one set of propositions necessitates the truth of another."

Here I have merely summarized Savage's description of objectivistic and personalistic views, but quoted his statement about "necessary" views in full. This is the view which he imputes to Laplace; or more accurately, Laplace's view is described (p. 278) as a "naive necessary one."

I want to say something about each of these adjectives, because I am expounding a viewpoint which I believe is the same as Laplace's (although from this distance in time, there is no way to be sure of that in every detail). Since the term "necessary" was coined by Savage, we have to accept its definition as given above; but we can still ask whether the definition properly describes Laplace's view (or the one I am developing, if there is any difference). Now in order to answer this, it would clearly be absurd to try to consult every statement about probability made by, or in the name of, Laplace. We have to distinguish clearly between probability theory and things that have been said about probability theory; too often, they are entirely different. The only way to find out what Laplace's form of probability theory "really says" about some question is to look at the equations Laplace gave us, in some specific case where the question comes up.

Now, where is an equation which says that probability measures the extent to which one set of propositions, out of logical necessity, confirms the truth of another? Where, indeed, is the relation in logic which tells when one set of propositions necessitates the truth of another? The relations of logic are of the form, "If A implies B, and if B implies C, then ...." There is nothing in logic which tells us whether A does in fact imply B. In other words, the relations of logic are only rules for the consistent manipulation of implications; they do not tell us whether some proposed implications are correct, but only whether they are mutually consistent.

It is exactly the same in probability theory. The basic equations are simply,

$$(AB|C) = (A|BC)(B|C)$$

$$(A|B) + (a|B) = 1$$

These, you see, are again statements of the form, "If C implies B to the extent $(B|C)$, and if BC implies A to the extent $(A|BC)$, then ...." There is nothing which tells whether C does in fact imply B to the extent $(B|C)$. In other words, the relations of probability theory are only rules for the consistent manipulation of partial implications; they do not tell us whether some proposed probability assignments are correct, but only whether they are mutually consistent.

If, on meditation, I decide that my personal probabilities are $(B|C) = 3/4$, $(A|BC) = 4/5$, $(AB|C) = 1/2$, then probability theory tells me that I am reasoning inconsistently. It does not tell me how to resolve that inconsistency.

But we can, in the case of probability theory, make a much stronger statement. What did we just learn? How much did verification of Leverrier's prediction N, out of logical necessity, confirm the truth of Newton's theory T? Bayes' theorem not only did not answer this, but it explicitly stated the opposite of the "necessary" view: Unless N is absolutely impossible on hypothesis T, it is meaningless to ask how much N, in itself, confirms the truth of T.

How about Rule 4? Isn't that an equation that tells us that one proposition does, out of logical necessity, confirm the truth of another to a definite extent? No, it isn't. Mathematically, the rule asserts one thing, and one thing only: if the sum of N equal numbers is unity, then each of the numbers must be N^{-1}. Rule 4 assigns definite numerical values to probabilities only after we have arbitrarily specified the set of propositions $A_1 \ldots A_N$ that

we're going to consider. Nothing in probability theory tells us that this specific set of propositions was the right set to introduce.

Consider two different problems; in problem (1) we have N different propositions, $A_1 \ldots A_N$. In problem (2) we have one more proposition A_{N+1} that must be taken into account. In general, for a given specific piece of evidence E, the probability $(A_1|E)$ will be different in the two problems. We saw this in detail when we studied multiple hypothesis testing in Lecture 6; addition of hypothesis D to the problem completely changed the numerical value of $(A|E)$.

Probability theory not only does not say that evidence E confirms the truth of A to some definite extent; it explicitly denies that any such relation exists. The probability $(A|E)$ does not depend only on A and E; it depends also on which alternatives to A we are considering, and it is mathematically indeterminate until those alternatives have been specified.

So, I think we have to plead "not guilty" to any charge that Laplace's formulation of probability theory is a "necessary" one. Indeed, if anyone is guilty of supposing that one proposition confirms the truth of another to any unique extent, it is the "objectivist" who teaches his students how to accept or reject hypotheses without considering the alternatives. Laplace's theory will not allow us to commit that error of reasoning.

Why have I answered this objection at such great, and repetitious, length? For several decades, authors of works on probability and statistics have been repeating the charge that Laplace's theory is nonsense <u>because</u> it supposes that for any two propositions A, B, there is a definite numerical value of $(A|B)$. The most casual glance at Laplace's <u>equations</u> shows that this is simply not true.

I think the trouble comes ultimately from some unfortunate historical accidents. After Laplace's death, some nineteenth-century philosophers made

ridiculous misapplications of probability theory, asserted that their nonsensical conclusions were "mathematically proved," and invoked the authority of Laplace to back them up. No man's reputation ever suffered more from the antics of enthusiastic but uncritical friends. The rise of the "objectivist" viewpoint in the twentieth century is an understandable, but misdirected, reaction against this lunacy. Instead of analyzing the transgressions and learning how to avoid such mistakes in the future, it was much easier to attack Laplace.

On the other hand, isn't it perfectly obvious that probability theory is an extension of logic, in exactly the sense alluded to by Savage? Probability theory fills in the gap between logical proof and disproof and shows us how to reason consistently in the intermediate region where, of necessity, virtually all of our actual reasoning takes place. It clearly includes deductive logic as a special case. I am continually amazed at the caution with which mathematicians approach this issue, and at their extreme reluctance to take the problem of inductive reasoning seriously. One gets the impression that an extension of logic is some enormously difficult, and probably impossible, problem which ordinary mortals had better leave alone.

Part of our communication gap here lies in the fact that no one has ever given an explicit answer to this question: What is it that we should prove about a proposed extension of logic before mathematicians will take it seriously? What are the tests that it has to pass? If you demand a proof that Laplace's theory is "correct," then I'm afraid I don't know what the question means. If you want to see a proof that it is the only possible extension of logic, then I would reply that it is surely not unique. But I think we have given fairly convincing arguments for the view that it is the only possible extension of logic which is internally consistent and represents degrees of plausibility by real numbers. You can, of course, hope to see

more rigorous and more general arguments than I have given; and I hope that you will. In this connection, let me just mention that the book of Savage (Savage, 1954) contains a great deal of this more refined analysis using measure theory, which is applicable to our problem.

How about other kinds of extensions of logic, in which we don't represent plausibility by real numbers? The possibilities of such "lattice theories" seem endless, and I want to say a little more about them in the last lecture. However, before dashing off to explore them, one should realize this: unless and until some specific failure of Laplace's theory is discovered, we have no rational basis for saying that a different theory is any better than the one we already have, and no clue to tell us in what way we should want another theory to be any different.

So, I would like to propose this as a working procedure. Let's take the good points of Savage's definition of "personalistic" and "necessary" views and combine them into a single definition; and above all, let's acknowledge their proper source:

(d) Laplace's Theory. Probability theory is an extension of logic which describes the consistent inductive reasoning of an idealized being who represents degrees of plausibility by real numbers. The numerical value of any probability $(A|B)$ will in general depend not only on A and B, but also on the entire background of other propositions that this being is taking into account. A probability assignment is "subjective" in the sense that it describes a state of knowledge rather than anything which could be measured in an experiment; but it is completely "objective" in the sense that it is independent of the personality of the user; two beings with the same total background of knowledge must assign the same probabilities.

Now for that other adjective, "naive". This is more difficult to discuss, because it is vague. A dictionary definition of naive is: "of unaffected simplicity." To call any mathematical theory naive in that sense is, I think, very great praise; and praise of which Laplace's theory is fully deserving. But I don't think Savage meant it in that way. I think he meant that Laplace did not hesitate to apply probability theory in all sorts of problems where a modern statistician would fear to tread. Our little excursion into jurisprudence is, no doubt, a good example. But, of course, if probability theory really is an extension of logic, there shouldn't be any restriction on the kind of problem treated; in principle, we ought to be able to apply it to any situation where plausible inference is needed. The only way of judging whether this is so, is simply to apply Laplace's theory to many specific situations, particularly those where the objectivists have warned us not to use it, and see for ourselves just how naive the results are, and whether the objectivist can produce any better results. We have already done some of this in the last three lectures, and many more examples will come up in later ones.