Classroom Notes

Math 394B Summer Quarter 2005

Week 1

Probability is very young among the branches of Mathematics. Its mathematical foundations didn't reach a sound stable level until the 1940s, when N.N.Kolmogorov laid down the modern form for its axiomatics.

Being relatively young, it may not come as a surprise that it has been plagued until very recently by discussions about its foundations that find little parallel in other branches of Mathematics. After all, since the first half of the 20th Century saw a spectacular development in Probability Theory, based on a sometimes fuzzy mathematical framework, so it is understandable that some of the pioneers would try to keep the fuzziness they used against the new framework.

We will look at two aspects that have inflamed discussions at a religious-like level. The first regards the "countable additivity" axiom, and is very much in the past now, although a few holdovers are still around, much like the Japanese soldiers who didn't realize that World War II was over. The second has survived up to these days, and it is of a very different nature (even if the adherents to the defeated "religion" in the first fight adhered to one of the "churches" in the second, and purported to connect the two arguments). Even if more substantial, this second issue ("subjective" vs. "frequentist" vs. "axiomatic" probability) has nothing at all to do with Probability Theory – it is an issue in Statistics. As a non-statistician, I will not pass a definitive judgement on the issue, but as a mathematician, I can't help believing that the active combatants are sorely misguided, since there is no point in trying to define a "best" or, worse, "correct" way to assign probabilities: this is a real-world problem, and its real-world solution depends on the real-world specifications.

1 The Axioms of Probability

1.1 The Additivity (or Continuity) Axiom

It is no problem to derive from basic logical arguments, the axiom

$$P\left[\bigcup_{j=1}^{n} A_j\right] \leq \sum_{j=1}^{n} P\left[A_j\right] \tag{1}$$

(countable additivity). This, by taking complements, becomes equivalent to

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots \supseteq A_n \Rightarrow$$

$$\Rightarrow P\left[\bigcap_{j=1}^{n} A_j\right] = P\left[A_n\right] \tag{2}$$

However, this is not nearly enough to tackle situations where we can legitimately talk about "infinitely many events" (we will be discussing this issue better later on). It is a fact that statements like (1), and (2) do *not* imply the corresponding statements for an *infinite* (even countable) collection of events. This has to be stated as a specific additional axiom. To do so is very much in the spirit of Lebesgue measure and integration theory, where facts about "infinite" sets are *always* meant as *limit statements* for *finite* sets.

Specifically, the axiom

$$P\left[\bigcup_{j=1}^{\infty} A_j\right] \leq \sum_{j=1}^{\infty} P\left[A_j\right] \tag{3}$$

or, equivalently,

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots \supseteq A_n \supseteq \ldots \Rightarrow$$
$$\Rightarrow P\left[\bigcap_{j=1}^{\infty} A_j\right] = P\left[\bigcap_{j=1}^{\infty} A_j\right]$$

has to be stated explicitly, and is *not* a consequence of (1). It is however a reasonable assumption, since our only hope of stating anything about the result of infinitely many operations has to be our ability of performing a finite approximation of infinity.

1.2 Some Consequences

1.2.1 Finite Probability Spaces

In probability theory, what really counts is not so much the definition of a sample space Ω, as the choice of the algebra of events, $\mathcal{F}$. Whatever the complexity of Ω, if $\mathcal{F}$ is a finite collection, we are dealing with a finite probability model, and no delicate problems can arise. This is, more or less, the realm that has staid outside the religious wars.

1.2.2 Countable Probability Spaces

Even if the sets in $\mathcal{F}$ are a countable collection, there is little in terms of delicate problems in the standard setup. However, already here, a pseudo-paradox has been concocted to challenge the least "intuitive" of the probability axioms, i.e. (3).

The "paradox" (which it is not) consists in the definition of "picking an integer at random". Here the problem is the very fuzzy definition of "at random". In the finite context, when you are picking a number at random from, say, a bag of balls for the purpose of determining a lottery winner, there is no big problem:

"at random" most often means that all possible outcomes, say, $1, 2, \ldots, N$, for some N, are *equally likely*. I.e.

$$P[\{j\}] = \frac{1}{N}; j = 1, 2 \ldots, N$$

$$P[\{j\}] = 0; j \neq 1, 2 \ldots, N$$

Now, if we would try to that for *all* integers, we would have trouble: since now what we called N would be, intuitively, ∞, we would be assigning probability "$\frac{1}{\infty}$" (i.e., 0) to each - but by countably additivity, this could not be, because, as one number will eventually be picked, we need to have $\sum_j 0 = 1$. Of course, if we let go of the countably additive requirement, this is no longer true – you could well assign probability 1 to the set of all integers, and probability 0 to any finite subset. The result is a finitely additive, but not countably additive probability.

This all good and well, until we sit down and try to figure out what the meaning of "picking an integer number at random" should be. To say that it makes sense to think of "all" integers as having the same probability, means that we can conceive a way to make, say, both 1, and $10^{10^{10^{\cdots^{10}}}}$, as well as *any other* integer, have the same probability of being chosen! In real life this seems kind of ludicrous, so it doesn't seem to be a real restriction to be asked to assign weights p_j, such that $\sum_j p_j = 1$ to each integer j. This implies, in particular, that, as $j \to \infty$, $p_j \to 0$, but, again, this seems a very reasonable "restriction" on whatever specific procedure we could cook up.

1.2.3 General Probability Spaces

The issues that arise in the general context are much more delicate technically, so the debate about complete additivity would seem more significant than in the countable case. However, as we will promptly see, the axiomatic system, based as it is on the Lebesgue integration theory, is extremely well suited to real life.

Giving up on continuity/countable additivity for probabilities is most certainly a mathematically challenging and interesting program, but the motivation for (the very few) people who actually dare to take this plunge (as opposed to the somewhat larger number of people who say one *should* take this plunge, but have no technical ability to do so) is the "because it's there" syndrome, shared with extreme mountain climbers - it has little or no basis in applications at all.

So, what are we talking about here? Suppose we are given a general sample space Ω. We would like to build a probability model on such a space. To do so, we have to define what we mean by "event", and assign a probability set function on such events. This function should satisfy all axioms of probability, but it turns out that if we include "countable additivity" among the axioms, there is a caveat theorem about our possible choice of events:

Fact: Suppose Ω is uncountable, and we choose the collection of *all subsets* of Ω as the (σ-)algebra of events, i.e. $\mathcal{F} = 2^{\Omega}$. Then, there is no way of defining a non trivial countably additive probability on $(\Omega, 2^{\Omega})$

Here, a "trivial" probability is one that gives positive probability only to an at most countable collection of points: i.e., there exist $\omega_1, \omega_2, \ldots$ such that $P[\{\omega_j\}] = p_j \geq 0$, and $\sum p_j = 1$.

In plain words, this means that there is no effectively uncountable model that also allows the assignment of a probability to any *arbitrary* subset of Ω. This "limitation" offended the sensibility of some of the pioneers in this field, who argued (also in connection to their extremist view of "subjective probabilities", which we will discuss in the next section) that such a constrained on the "free" spirit was unwarranted. Those familiar with 19th and 20th idealism will immediately recognize the abstract foundation of such and similar assertions. Since idealism, as a strong philosophical school has been essentially destroyed by World War II, this objection should seem to us very quaint

In fact, the restriction we mentioned makes very much sense to the practically minded. The "unreachable" subsets of Ω mentioned above, turn out to be the "non measurable" sets in measure theory. These sets, in turn, can be defined only through a "construction" based on the "uncountable version of the Axiom of Choice" - essentially they are constructed by choosing an "arbitrary" element in Ω (under some conditions) uncountably many times. While this is not a problem from a mathematical point of view (at least, if you accept this strong version of the Axiom of Choice, like most mathematicians do, after it has been shown by Gödel that such an axiom is independent of the usual axioms, but it cannot introduce any new contradictions beyond what may - or, hopefully, may not - already be present in the mathematical system).

So, the subsets that can actually be assigned a probability arise in the following way, which is much more reasonable than any talk about "unfettered action by the spirit". The model for the procedure is provided by $\Omega = [0, 1]$, the unit interval (there is a theorem proving that any probability model can be mapped on the triple $\{[0, 1], \mathcal{B}, \lambda\}$, where $\mathcal{B}$ is the σ-algebra of "Borel sets" (which is what we will get promptly - a more complete explanation will have to wait for your *Real Analysis* class), and λ is "Lebesgue measure" (same considerations apply).

1. We pick a family `F` of subsets of Ω that is a) large enough so that there can be only one probability that takes assigned values on the elements, b) small and "concrete" enough so that it is easy (or, at least, feasible) to assign a probability to its elements. For $\Omega = [0, 1]$, we could (and usually do) choose the *intervals*: subsets of the form $\{x : a \leq x \leq b\}$.

2. We assign a probability P on elements of this family, in a way that is consistent with the axioms (e.g., if we can write an element A of `F`, as $A = \bigcup_j B_j$, where the B_j are disjoint subsets of Ω, all elements of `F`, it is true that $P[A] = \sum_j P[B_j]$). For our example case, we would define $\lambda([a, b]) = b - a$.

3. We now use countable additivity to extend this definition to all sets that are unions of intervals, intersections of unions of intervals, unions of intersections of unions of intervals,... In theory, this sequence goes on a

countable number of times - and then starts afresh on the resulting family for another countable number of times, and so on, for a countable number of countable number of times, and then, again, ... (etc., etc., ...)

The principle involved here is called "transfinite induction", and is a formalization for (very roughly speaking) performing a huge sequence of limits. Although you will rarely find this "constructive" path to the construction of a σ-algebra of events $\mathcal{F}$, by taking limits starting from a smaller set $\mathtt{F}$, the fact that this is one way of building the "σ-algebra generated by $\mathtt{F}$" (i.e., the smallest σ-algebra containing all the elements in $\mathtt{F}$) is mentioned with proof in older (mainly pre-World War II) books on measure theory and probability. Only sets that cannot be reached starting from intervals through this infinitely infinite sequence of limits cannot be assigned probabilities...

Summing up, the "axiomatic" approach gives us a very realistic way of addressing probabilities: we cannot expect to assign a "probability" to a situation which we cannot describe directly, nor even as the limit of limits, of limits, of limits,... of situations we can describe. The payoff of this "limitation" (which is not a limitation at all) is that we are able to determine probabilities by a limiting approach. In other words, we can safely work with finite possibilities (which are the only ones our minds can seriously handle), and more general schemes that can be reliably *approximated* by finite schemes. It is the basic philosophy of Lebesgue measure and integration theory, and it seems a very realistic foundation.

2 How Do We Determine Probabilities?

Every Probability theorems seems to start with the words, "let a probability space $\{\Omega, \mathcal{F}, P\}$ be given...". The sample space and the σ-algebra might be given naturally by the problem, but how in the world are we going to determine P?

This is the point of most results in *probability* theory. If we are given the probabilities of a "large enough" collection of elements of $\mathcal{F}$, the theory helps us in determining the probability of any other event we might want to know.

The preceding paragraph simply begs the question: *how* are we going to determine these starting probabilities?

Clearly, the answer depends on the problem at hand. If we are working with a theoretical, abstract, system, we either don't care, or we have some extra-probabilistic information that guides us (e.g., in Statistical Physics, we rely on Classical or Quantum Mechanics for a setup).

In particular, this happens when, for whatever reason, we can squeak by with finite $\mathcal{F}$, and a "big enough" collection of elements in $\mathcal{F}$, that can be thought as "equiprobable". Specifically, this happens when it is reasonable to determine a "partition" of Ω (i.e. a collection of subsets A_i, such that $A_i \cap A_j = \emptyset$, when $i \neq j$, and $\bigcup_i A_i = \Omega$) where each of the "tiles" has the same probability: if there are n such tiles, $P[A_i] = \frac{1}{n}$. One consequence is that we are unable to

handle sets smaller than the elements A_j of the partition (they are called *atoms* of the resulting, finite, σ-algebra).

Since this is a reasonable goal in problems such as games of chance, and such games were the original motivation for probability investigations, this approach ("the classical model") seems very important. However it is not: while there are several cases where it is appropriate, most of them are indeed in the realm of games of chance. In more serious situations this is not so - either because $\mathcal{F}$ is not finite (and that closes the issue), or because it is simply impossible to find such a partition.

So much for the misconception of "equiprobable" is somewhat the natural or prevailing way to address the problem. There used to be a "classical" definition of probability, based on equiprobable events, but its extremely limited scope make it useless in all but very simple and niche fields.

2.1 The "Frequentist" Solution

A solution to the problem of determining probabilities was put forward by Von Mises in the first half of the 20th Century. To assign a probability to an event, we observe it a large number of times – i.e., we repeat the experiment that produces (or not) the event many, many times. After N repeats, we count the number of successes, say n, and decide that the probability of the event is approximately $p \simeq \frac{n}{N}$. The larger N, the more we can use $\frac{n}{N}$ as a reliable value for p. Unfortunately, this approach is logically faulty, because the assertion that there is a number p such that, as N grows, the frequency $\frac{n}{N}$ approaches p is itself dependent on the existence of a probability model *beforehand*. In other words, to assert that we can assign probability p to an event, we must already have assumed we have assigned it a probability. For this reason, the frequentist definition of probability is not used explicitly any more, although, informally, it provides a basis for empirical statistics.

2.2 The "Subjective" Solution

Faced with the circular argument of the "frequentists", a whole school of thought has come up with the assertion that probabilities are subjectively ascribed to events by modelers (hopefully, "experts" in the field), subject to minimal logical restrictions reflecting logical consistency. This approach is especially popular in the Social Sciences, Economy, and other areas where (as opposed, say, to physics) it is difficult, if not utterly impossible, to reproduce a given experiment precisely, and thus have no use for a "frequentist" theory, not only because of the circularity we have discussed, but because it is simply unworkable.

Of course, the limits of this approach are determined by its very nature: how do we get a communicable model if the basic parameters are "subjective", i.e., somewhat arbitrary, and subject to variation between "experts"? Also, "subjectivists" like to stress that this approach allows us to assign probabilities to one-time (non reproducible) events, but it is not altogether clear what the

purpose of such an assignment would be: like it or not, the usefulness of probability theory comes from its presumption to predict "relative frequencies" of occurrences, and is of very little practical use if no frequency of occurrences is going to be considered.

2.3 The Axiomatic Solution

Regardless of what your preferred philosophy would be, the axiomatic approach (which is now shared by the vast majority of professionals in this area) consists in the statement that Probability Theory is not really concerned with "what" probabilities are (just as Geometry is not really concerned with "what" a "point" or a "line" are supposed to "really be"). It concerns itself with statements that we can derive from a set of axioms, without worrying about how the numbers that we will call "probabilities" would be determined. We just assume that they are given.

In applications, we will chose the best method of assigning probabilities to specific concrete events as determined by circumstances. Hence, for games of chance, it might be preferable to resort to the "classic" approach, while for an economic model, we might want to refer to "expert opinions", and for a physical model, we might try to come up with probabilities from physical principles. When it is possible to repeat the experiment in a controllably equal environment, we can (and do) resort to statistical estimation of probabilities – based on the assumption that there is a probability model, and we simply don't know the appropriate values of the probabilities.

Regardless of how we set it up, after we have specified our probabilistic model, it will be the job of statistics to validate it against experiments.