1.2 Lecture 2 (5/Oct) Some basics

1.2.1 Measure

Frequently one can "get by" with a naïve treatment of probability theory: you can treat random variables quite intuitively so long as you maintain Bayes' law for conditional probabilities of events:

$$\Pr(A_1|A_2) = \frac{\Pr(A_1 \cap A_2)}{\Pr(A_2)}$$

However, that's not good enough for all situations, so we're going to be more careful, and methodically answer the question, "what is a random variable?" (For a philosophical and historical discussion of this question see Mumford in [56].)

First we need measure spaces. Let's start with some standard examples.

1. $\mathbb{Z}$ with the counting measure.
2. $\mathbb{R}$ with the Lebesgue measure, i.e., the measure (general definition momentarily) in which intervals have measure proportional to their length: $\mu([a,b]) = b - a$ for $b \geq a$.
3. $[0,1]$ with the Lebesgue measure.
4. A finite set with the uniform probability measure.

As we see, a measure μ assigns a real number to (some) subsets of a universe; if, as in the last two examples, we also have

$$\mu(\text{universe}) = 1 \tag{1.1}$$

then we say the measure space is a probability space or a sample space.

Let's see what are the formal properties we want from these examples.

As we just hinted, we don't necessarily assign a measure to all subsets of the universe; only to the *measurable sets*. In order to make sense of this, we need to define the notion of a σ-algebra (also known as a σ-field).

A σ-algebra $(M, \tilde{M})$ is a set M along with a collection $\tilde{M}$ of subsets of M (called the *measurable sets*) which satisfy: (1) $\emptyset \in \tilde{M}$, and (2) $\tilde{M}$ is closed under complement and countable intersection.

It follows also that $M \in \tilde{M}$ and $\tilde{M}$ is closed under countable union (de Morgan). By induction this gives a *stability* property: we can take any finite sequence of the form, a countable union of countable intersections of ... of countable unions of measurable sets, and the result will be a measurable set.

A measure space is a σ-algebra $(M, \tilde{M})$ together with a *measure* μ, which is a function

$$\mu : \tilde{M} \to [0, \infty] \tag{1.2}$$

that is *countably additive*, that is, for any pairwise disjoint $S_1, S_2, \ldots \in \tilde{M}$,

$$\mu(\bigcup S_i) = \sum \mu(S_i). \tag{1.3}$$

So, (1.2) and (1.3) give us a measure space, and if we also assume (1.1) then we have a probability space.

Let us see some properties of measure spaces:

I. $\mu(\emptyset) = 0$ since $\mu(\emptyset) + \mu(\emptyset) = \mu(\emptyset \cup \emptyset) = \mu(\emptyset)$.

II. The *modular identity* $\mu(S) + \mu(T) = \mu(S \cap T) + \mu(S \cup T)$ holds because necessarily $S - T, T - S$ and $S \cap T$ are measurable, and both sides of the equation may be decomposed into the same linear combination of the measures of these sets. (The set $S - T$ is $S \cap (\neg T)$.) This identity is sometimes also called the lattice or valuation property.

III. From the modular identity and nonnegativity, $S \subseteq T \Rightarrow \mu(S) \leq \mu(T)$.

1.2.2 Measurable functions, random variables and events

A *measurable function* is a mapping X from one measure space, say $(M_1, \tilde{M}_1, \mu_1)$, into another, say $(M_2, \tilde{M}_2, \mu_2)$, such that pre-images of measurable sets are measurable, that is to say, if $T \in \tilde{M}_2$, then $X^{-1}(T) \in \tilde{M}_1$.

If M_1 is a probability space we call X a random variable.

The *range* of the random variable, M_2, can be many things, for example:

- $M_2 = \mathbb{R}$, with the σ-field consisting of rays (a, ∞), rays $[a, \infty)$, and any set formed out of these by closing under the operations of complement and countable union. (In CS language, any other measurable set is formed by a finite-depth formula whose leaves are rays of the aforementioned type, and each internal node is either a complementation or a countable union.)

  Sometimes it is convenient to use the "extended real line," the real line with ∞ and $-\infty$ adjoined, as the base set.

- $M_2 =$ names of people eligible for a draft which is going to be implemented by a lottery. The σ-field here is 2^{M_2}, namely the power set of M_2.

- $M_2 =$ deterministic algorithms for a certain computational problem, with the counting measure. On a countably infinite set M_2, just as on a finite set, we can use the power set as the σ-field. The counting measure assigns to $S \subseteq M_2$ its cardinality $|S|$.

Events
With any measurable subset T of M_2 we associate the *event* $X \in T$; if X is understood, we simply call this the event T. This event has the probability $\Pr(X \in T)$ (or if X is understood, $\Pr(T)$) dictated by

$$\Pr(X \in T) = \mu_1(X^{-1}(T)). \tag{1.4}$$

The *indicator* of this event is the function $[T]$ or $\mathbb{I}_T$,

$$\mathbb{I}_T : M_1 \to \{0, 1\} \subseteq \mathbb{R}$$
$$\mathbb{I}_T(y) = \begin{cases} 1 & y \in X^{-1}(T) \\ 0 & \text{otherwise} \end{cases}$$

The basic but key property is that

$$\Pr(X \in T) = \int \mathbb{I}_T \, d\mu = E(\mathbb{I}_T). \tag{1.5}$$

It follows that probabilities of events satisfy:

1. $\Pr(\emptyset) = 0$ ("something happens")
2. $\Pr(M_2) = 1$ ("only one thing happens")
3. $\Pr(A) \geq 0$
4. $\Pr(A) + \Pr(B) = \Pr(A \cap B) + \Pr(A \cup B)$

Note that events can themselves be thought of as random variables taking values in $\{0, 1\}$; indeed we will sometimes define an event directly, rather than creating out of some other random variable X and subset T of the image of X.

For the most part we will sidestep measure theory—one needs it to cure pathologies but we will be studying healthy patients. However I recommend Adams and Guillemin [3] or Billingsley [13].

Often when studying probability one may suppress any mention of the sample space in favor of abstract axioms of probability. For us the situation will be quite different. While starting out as a formality, explicit sample spaces will soon play a significant role.

Joint distributions
Given two random variables $X_1 : M \to M_1$, $X_2 : M \to M_2$ (where each M_i has associated with it a σ-field $(M_i, \tilde{M}_i)$), we can form the "product" random variable $(X_1, X_2) : M \to M_1 \times M_2$. The same goes for any countable collection of rvs on M, and it is important that we can do this for countable collections; for instance we want to be able to discuss unbounded sequences of coin tosses. Given a product rv

$$(X_1, X_2, \ldots) : M \to M_1 \times M_2 \times \ldots,$$

its *marginals* are probability distributions on each of the measure spaces M_i. These distributions are defined by, for $A \in \tilde{M}_i$,

$$\Pr(X_i \in A) = \Pr((X_1, X_2, \ldots) \in M_1 \times M_2 \times \ldots \times M_{i-1} \times A \times M_{i+1} \times \ldots$$

That is, you simply ignore what happens to the other rvs, and assign to set $A \in \tilde{M}_i$ the probability $\mu(X_i^{-1}(A))$.

$X_1, X_2, \ldots$ are *independent* if for any finite $S = \{s_1, \ldots, s_n\}$ and any $A_{s_1} \in \tilde{M}_{s_1}, \ldots, A_{s_n} \in \tilde{M}_{s_n}$, we have $\Pr((X_{s_1}, \ldots, X_{s_n}) \in A_{s_1} \times \cdots \times A_{s_n}) = \Pr(X_{s_1} \in A_{s_1}) \cdots \Pr(X_{s_n} \in A_{s_n})$.

(Note that $\Pr((X_1, X_2) \in A_1 \times A_2)$ is just another way of writing $\Pr((X_1 \in A_1) \wedge (X_2 \in A_2))$.)

Example: a pair of fair dice.
Let M be the set of 36 ways in which two dice can roll, each outcome having probability $1/36$. On this sample space we can define various useful functions: e.g., X_i = the value of die i ($i = 1, 2$); $Y = X_1 + X_2$. X_1 and X_2 are independent; X_1 and Y are not independent.

$X_1, \ldots : M \to T$ are *independent and identically distributed (iid)* if they are independent and all marginals are identical. If T is finite and the marginals are the uniform distribution, we say that the rv's are uniform iid.

Conditional Probabilities are defined by

$$\Pr(X \in A | X \in B) = \frac{\Pr(X \in A \cap B)}{\Pr(X \in B)} \tag{1.6}$$

provided the denominator is positive.

An old example. You meet Mr. Smith and find out that he has exactly two children, at least one of which is a girl. What is the probability that both are girls? Answer[1]: $1/3$.

Taking (1.6) and applying induction, we have that if $\Pr(A_1 \cap \ldots A_n) > 0$, then:

Chain rule for conditional probabilities

$$\begin{aligned} \Pr(A_1 \cap \ldots A_n) &= \Pr(A_n | A_1 \cap \ldots A_{n-1}) \cdot \Pr(A_{n-1} | A_1 \cap \ldots A_{n-2}) \\ &\quad \cdots \Pr(A_2 | A_1) \cdot \Pr(A_1). \end{aligned}$$

(If $\Pr(A_1 \cap \ldots A_n) = 0$ then because of the denominators, some of the conditional probabilities in the chain might not be well defined, but you can say that either $\Pr(A_1) = 0$ or there is some i s.t. $\Pr(A_i | A_1 \cap \ldots A_{i-1}) = 0$.)

[1] As usual in such examples we suppose that the sexes of the children are uniform iid. Some facts from general knowledge should be enough for you to doubt both uniformity and independence.

Real-valued random variables; expectations
If X is a real-valued rv on a sample space with measure μ, its expectation (aka average, mean or first moment) is given by the following integral

$$E(X) = \int X \, d\mu$$

which is defined in the Lebesgue manner by[2]

$$\int X \, d\mu = \lim_{h \to 0} \sum_{\substack{\text{integer} \\ -\infty < j < \infty}} jh \Pr(jh \leq X < (j+1)h) \tag{1.7}$$

provided this limit converges absolutely, which means that

$$\sup_{h>0} \sum_{\substack{\text{integer} \\ -\infty < j < \infty}} |jh| \Pr(jh \leq X < (j+1)h) < \infty. \tag{1.8}$$

It is not hard to innocently encounter cases where the integral is not defined. Stand a meter from an infinite wall, holding a laser pointer. Spin so you're pointing at a uniformly random orientation. If the laser pointer is not shining at a point on the wall (which happens with probability $1/2$), repeat until it does. The displacement of the point you're pointing at, relative to the point closest to you on the wall, is $\tan \alpha$ meters for α uniformly distributed in $(-\pi/2, \pi/2)$. You could be forgiven for thinking the average displacement "ought" to be 0, but the integral does not converge absolutely, because $\int_0^{\pi/2} \tan \alpha \, d\alpha = -\int_{\cos(0)}^{\cos(\pi/2)} \frac{1}{x} \, dx = -[\log x]|_{\cos(0)}^{\cos(\pi/2)} = +\infty$, using the substitution $x = \cos \alpha$.

To see the kind of problem that this can create, consider that for an integration definition to make sense, we ought to have the property that if $\lim a_i = a$ and $\lim b_i = b$, then $\lim \int_{a_i}^{b_i} f(\alpha) d\alpha = \int_a^b f(\alpha) d\alpha$. But in the present circumstance we can, for instance, take $a_i = -\arccos 1/i, b_i = \arccos 2/i$, and then $\int_{a_i}^{b_i} \tan \alpha \, d\alpha = -\log i + \log 2i = \log 2$ (rather than 0).

[2]One can be more scrupulous about the measure theory; see the suggested references. *But he knew little out of his way, and was not a pleasing companion; as, like most great mathematicians I have met with, he expected universal precision in everything said, or was for ever denying or distinguishing upon trifles, to the disturbance of all conversation. He soon left us.*
The Autobiography of Benjamin Franklin, chapter 5