Mathematical Toolkit **Fall 2024**

Lecture 9: October 30, 2024

Lecturer: Avrim Blum (notes based in part on notes from Madhur Tulsiani)

Probability theory is a mathematical framework used to model uncertainty and variability in nature. It is by no means the only contender for this role, but has weathered many trials through time. A good deal of probability theory was developed long before being formalized in the way that we're familiar with now, which is due to Kolmogorov. One could cite the works of Laplace, Poisson, Gauss, to name a few. So in some sense the formalization we present here is not strictly necessary, at least for most simple problems. But it does place the whole field on a very stable foundation, which is also helpful whenever something challenges our grasp of this otherwise intuitive discipline.

1 Basics of probability: the finite case

We recall very briefly the basics of probability and random variables. For a more detailed introduction, please see the lecture notes by Terry Tao, linked from the course homepage.

1.1 Finite probability spaces

Let Ω be a finite set. Let $\nu : \Omega \to [0,1]$ be a function such that

$$\sum_{\omega \in \Omega} \nu(\omega) \;=\; 1.$$

We often refer to Ω as a *sample space* or *outcome space* and the function ν as a *probability distribution* on this space. An *event* can be thought of as a subset of outcomes i.e., any $A \subseteq \Omega$ defines an event, and we define its probability as

$$\mathbb{P}\left[A\right] \;=\; \sum_{\omega \in A} \nu(\omega)\,.$$

The elements $\omega \in \Omega$ are often called "elementary events" (and associated with their singleton sets).

1.2 Random Variables and Expectation

In a finite probability space, a *real-valued random variable* over Ω is any function $X : \Omega \to \mathbb{R}$. So a random variable is technically neither random (it's quite deterministic) nor a variable (it's a function), but it's a terminology that has stuck.

For example, if you roll two dice, we might define random variable X_1 to be the value of die 1, random variable X_2 to be the value of die 2, and $X = X_1 + X_2$ to be the sum of the two dice.

It will often be natural to go back and forth between random variables and events. For instance, given a random variable X and a value b we can define the event "$X = b$" as $\{\omega \in \Omega : X(\omega) = b\}$. In the other direction, given an event A, it is often convenient to define an *indicator random variable* X_A as $X_A(\omega) = 1$ if $\omega \in A$ and $X_A(\omega) = 0$ otherwise.

In a finite probability space, we define the expectation of a random variable X as:

$$\mathbb{E}[X] \ := \ \sum_{\omega \in \Omega} \nu(\omega) \cdot X(\omega) \, .$$

In other words, the expectation of a random variable X is just its average value over Ω, where each elementary event ω is weighted according to its probability. For instance, if we roll a single die and look at the outcome, the expected value is 3.5, because all six elementary events have equal probability. Often one groups together the elementary events according to the different values of the random variable and rewrites the definition like this:

$$\mathbb{E}[X] \ = \ \sum_{a} \mathbb{P}(X = a) \cdot a.$$

An extremely useful fact about expectation is that is a linear transformation. In particular, if X and Y are random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proposition 1.1 (Linearity of Expectation) *For any two random variables X and Y, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

Proof: This follows directly from the definition.

$$\mathbb{E}[X + Y] = \sum_{\omega \in \Omega} \nu(\omega) \cdot (X(\omega) + Y(\omega)) = \sum_{\omega \in \Omega} \nu(\omega) \cdot X(\omega) + \sum_{\omega \in \Omega} \nu(\omega) \cdot Y(\omega) = \mathbb{E}[X] + \mathbb{E}[Y].$$

■

(Also, obviously, $\mathbb{E}[cX] = c\,\mathbb{E}[X]$).

Example: Card shuffling Suppose we unwrap a fresh deck of cards and shuffle it until the cards are completely random. How many cards do we expect to be in the same position as they were at the start? To solve this, let's think formally about what we are asking. We are looking for the expected value of a random variable X denoting the number of cards that end in the same position as they started. We can write X as a sum of indicator random variables X_i, one for each card, where $X_i = 1$ if the ith card ends in position i and $X_i = 0$ otherwise. These X_i are easy to analyze: $\mathbb{P}(X_i = 1) = 1/n$ where n is the number of cards. $\mathbb{P}(X_i = 1)$ is also $\mathbb{E}[X_i]$. Now we use linearity of expectation:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = 1.$$

So, this is interesting: no matter how large a deck we are considering, the expected number of cards that end in the same position as they started is 1.

1.3 Conditioning

Conditioning on an event E is equivalent to restricting the probability space to the set E. We then consider the conditional probability measure ν_E defined as

$$\nu_E(\omega) \;=\; \begin{cases} \frac{\nu(\omega)}{\mathbb{P}[E]} & \text{if } \omega \in E \\ 0 & \text{otherwise} \end{cases} .$$

Thus, one can define the conditional probability of an event F as

$$\mathbb{P}\left[F \mid E\right] \;=\; \sum_{\omega \in F} \nu_E(\omega) \;=\; \sum_{\omega \in E \cap F} \frac{\nu(\omega)}{\mathbb{P}\left[E\right]} \;=\; \frac{\mathbb{P}\left[E \wedge F\right]}{\mathbb{P}\left[E\right]} .$$

For a random variable X and an event E, we similarly define the *conditional expectation* of X given E as

$$\mathbb{E}\left[X \mid E\right] \;=\; \sum_{\omega} \nu_E(\omega) \cdot X(\omega) ,$$

with ν_E as above. Verify the following identities.

Proposition 1.2 (Total Probability and Total Expectation) *Let Ω be a finite sample space with probabiliy measure ν. Let $E, F \subseteq \Omega$ be events, and $X : \Omega \to \mathbb{R}$ be a random variable. Then*

1. $\mathbb{P}\left[F\right] \;=\; \mathbb{P}\left[E\right] \cdot \mathbb{P}\left[F \mid E\right] + \mathbb{P}\left[E^c\right] \cdot \mathbb{P}\left[F \mid E^c\right]$,
2. $\mathbb{E}\left[X\right] \;=\; \mathbb{P}\left[E\right] \cdot \mathbb{E}\left[X \mid E\right] + \mathbb{P}\left[E^c\right] \cdot \mathbb{E}\left[X \mid E^c\right]$,

where $E^c = \Omega \setminus E$.

Example: a random walk stock market Suppose there is a stock with the property that each day, it has a 50:50 chance of going either up or down by $1, unless the stock is at 0 in which case it stays there. You start with $m. Each day you can buy or sell as much as you want, until at the end of the year all your money is converted back into cash. What is the best strategy for maximizing your expected gain?

The answer is that no matter what strategy you choose, your expected gain by the end of the year is 0 (i.e., you expect to end with the same amount of money as you started). Let's prove that this is the case.

Define random variable X_t to be the gain of our algorithm on day t. Let X be the overall gain at the end of the year. Then,

$$X \quad = \quad X_1 + \ldots + X_{365}.$$

Notice that the X_t's can be highly dependent, based on our strategy. For instance, if our strategy is to pull all our money out of the stock market the moment that our wealth exceeds \$m, then X_2 depends strongly on the outcome of X_1. Nonetheless, by linearity of expectation,

$$\mathbb{E}[X] \quad = \quad \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_{365}].$$

Finally, no matter how many shares s of stock we hold at time t, $\mathbb{E}[X_t|s] = 0$. So, using Proposition 1.2, whatever probability distribution over s is induced by our strategy, $\mathbb{E}[X_t] = 0$. Since this holds for every t, we have $\mathbb{E}[X] = 0$.

1.4 Independence

Two events A and B are *independent* if $\mathbb{P}(A \wedge B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. For events of nonzero probability, we can more intuitively write this as A and B are independent if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. One can verify that this is equivalent to $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. In other words, restricting to one event does not change the probability of the other event. Independence is a joint property of events and the probability measure: one cannot make judgment about independence without knowing the probability measure.

Two random variables X and Y defined on the same finite probability space are defined to be independent if for all values x and y, the events "$X = x$" and "$Y = y$" are independent. Equivalently, they are independent if $\mathbb{P}\{X = x \mid Y = y\} = \mathbb{P}\{X = x\}$ for all non-zero probability events $\{X = x\} := \{\omega : X(\omega) = x\}$ and $\{Y = y\} := \{\omega : Y(\omega) = y\}$.

So far, we have considered just two events or two random variables. We say n events $A_1, ..., A_n$ are *mutually independent* (sometimes we will just say "independent") if for all subsets $S \subseteq \{1, ..., n\}$ we have:

$$\mathbb{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbb{P}(A_i).$$

We say n random variables $X_1, ..., X_n$ are mutually independent if for all values $x_1, ..., x_n$, the events "$X_1 = x_1$", ..., "$X_n = x_n$" are mutually independent. There is also a weaker notion that is often useful called *pairwise independence*. We say n events are pairwise independent if all pairs are independent, and likewise for random variables. Can you think of three events that are pairwise independent but not mutually independent?

We saw that for any two random variables X and Y we have $\mathbb{E}[X] + \mathbb{E}[Y] = \mathbb{E}[X+Y]$. However, it is not in general the case that $\mathbb{E}[X] \cdot \mathbb{E}[Y] = \mathbb{E}[X \cdot Y]$ (for example, suppose X and Y are indicator random variables for the same event of probability p; then the LHS is p^2 but the RHS is p). Nonetheless, we *do* get this property when X and Y are independent.

Proposition 1.3 *Let $X, Y : \Omega \to \mathbb{R}$ be two independent random variables. Then*

$$\mathbb{E}[X \cdot Y] \;=\; \mathbb{E}[X] \cdot \mathbb{E}[Y]\,.$$

Proof:

$$\begin{aligned}
\mathbb{E}[X] \cdot \mathbb{E}[Y] &= \left(\sum_a \mathbb{P}(X=a)\cdot a\right) \cdot \left(\sum_b \mathbb{P}(Y=b)\cdot b\right) \\
&= \sum_{a,b} a \cdot b \cdot \mathbb{P}(X=a) \cdot \mathbb{P}(Y=b) \\
&= \sum_{a,b} a \cdot b \cdot \mathbb{P}(X=a \wedge Y=b) \quad \text{(by independence)} \\
&= \sum_c \sum_{(a,b):ab=c} a \cdot b \cdot \mathbb{P}(X=a \wedge Y=b) \quad \text{(grouping)} \\
&= \sum_c c \cdot \mathbb{P}(X \cdot Y = c) = \mathbb{E}[X \cdot Y]\,.
\end{aligned}$$

■

Example: Universal hashing. A *hash function* is a function $h : U \to \{0, .., M-1\}$ where U is a universe of inputs, and typically $|U| \gg M$. For example, we might hash strings into the range $\{1, ..., 10000\}$ to use as a lookup table. A desirable property of a hash function is that for the subset S of U that you actually care about (for instance, strings corresponging to English words) that you do not get too many collisions (distinct elements hashing to the same location), especially when $|S| \approx M$. One convenient way to construct such a hash function is to use randomization.

A challenge here is that we can't just have h pick a random number when given an input $s \in S$ because we need to be able to find it again (h has to actually be a function). Also, requiring the R.V.'s $X_s = h(s)$ be mutually independent for all $s \in S$ would require too large a hash function. However, it turns out that *pairwise independence* will be sufficient and allow h to be compact and efficient enough to be useful.

Definition 1.4 *A randomized algorithm H to construct hash functions $h : U \to \{0, \ldots, M-1\}$ is* **universal** *if for all $s \neq s'$ in U, we have*

$$\mathop{\mathbb{P}}_{h \leftarrow H}[h(s) = h(s')] \leq 1/M.$$

Note that if the R.V.'s $X_s = h(s)$ are uniformly distributed in $\{0, ..., M-1\}$ and pairwise independent, then H will be universal.

Proposition 1.5 *If H is universal, then for any set $S \subseteq U$, for any $s \in U$ (e.g., that we might want to lookup), if we construct h at random according to H, the* **expected** *number of collisions between s and other elements in S is at most $|S|/M$.*

Proof: Each $s' \in S$ ($s' \neq s$) has at most a $1/M$ chance of colliding with s by definition of "universal". Define indicator R.V. $C_{s,s'}$ for the event that s and s' collide, and $C_s = \sum_{s' \in S, s' \neq s} C_{s,s'}$ as the total number of collisions. By linearity of expectation, $\mathbb{E}[C_s] \leq |S|/M$. ■

Can we actually construct universal hash families? Here is one approach: Let's say inputs are u-bits long. Say the table size M is power of 2, so an index is b-bits long with $M = 2^b$. What we will do is pick h to be a random linear transformation from $\mathbb{F}_2^u$ to $\mathbb{F}_2^b$ (i.e., a random b-by-u matrix over $\mathbb{F}_2$).

Claim 1.6 *For any $s \neq s'$, $\mathbb{P}_h[h(s) = h(s')] = 1/M = 1/2^b$.*

Proof: If $s \neq s'$ there must exist some index i such that $s_i \neq s'_i$, and for concreteness say $s_i = 0$ and $s'_i = 1$. Imagine we first choose all of h but the ith column. Over the remaining choices of ith column, $h(s)$ is fixed. However, each of the 2^b different settings of the ith column gives a different value of $h(s')$ (in particular, every time we flip a bit in that column, we flip the corresponding bit in $h(s')$). So there is exactly a $1/2^b$ chance that $h(s) = h(s')$. ■

1.5 The countable case

Everything defined above can also be extended to countable spaces but we need to be careful about the convergence of the above summations.

2 Interesting random variables

Bernoulli random variables A *Bernoulli*(p) random variable X is defined as taking the value 1 with probability p and the value 0 with probability $1-p$. We can write this as

$\mathbb{P}[X = x] = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$. One may intuitively think of a Bernoulli random variable as the indicator function of "heads" in an outcome space $\Omega = \{\text{tails, heads}\}$ of a biased coin toss. Alternatively, we simply take the outcome space to be $\Omega = \{0,1\}$. More generally, indicator functions of events are Bernoulli random variables.

Finite Bernoulli i.i.d. sequence We can also think of a sequence of coin tosses, with

$$X_i = \begin{cases} 1 & \text{if toss i is heads} \\ 0 & \text{if toss i is tails} \end{cases} .$$

being n Bernoulli random variables in the probability space $\Omega_n = \{0,1\}^n$, i.e., $X_i(\omega) = \omega_i$. Define the product probability measure on this finite space using:

$$\mu_n(\omega) = \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i}.$$

Note that if $p = \frac{1}{2}$, we have $\mu_n(\omega) = \frac{1}{2^n}$, i.e., $\mathbb{P}_n$ is the uniform distribution over the outcome space, as all outcomes are equally likely.

Exercise 2.1 *For the outcome space defined above, verify that:*

- *For any fixed i, X_i is indeed a Bernoulli(p) random variable, and*
- *If $I \subset [n]$ and $J \subset [n]$ are disjoint, then any function of X_I and any function of X_j are independent random variables.*

As noted in the previous lecture, when the latter point holds, we simply say that $X_1, \cdots, X_n$ are independent. Furthermore since all the X_i have the same distribution, we call the sequence *i.i.d.*, meaning independent and identically distributed.

Binomial random variables Let Z_n be a random variable counting the number of heads associated with n independent biased coin tosses. We can model this in Ω_n above as $Z_n = \sum X_i$.

Let us calculate the expectation of Z. By linearity we have $\mathbb{E}[Z_n] = \sum \mathbb{E}[X_i]$. Since $Z_n = \sum X_i$, we have, $\mathbb{E}[Z_n] = \sum \mathbb{E}[X_i]$. Now,

$$\begin{aligned} \mathbb{E}[X_i] &= 1 \cdot \mathbb{P}[X_i = 1] + 0 \cdot \mathbb{P}[X_i = 0] \\ &= \mathbb{P}[X_i = 1] = p \end{aligned}$$

Hence $\mathbb{E}[Z_n] = np$. Note that we did not use independence in the above calculations. We just needed that for each i, $\mathbb{E}[X_i] = p$.

We do need independence, and namely the product probability measure, to calculate $\mathbb{P}(Z_n = k)$ for $k \in [n]$ (this is often called the *probability mass function*). First note that the shorthand $(Z_n = k)$ simply means $\{\omega \in \Omega : Z_n(\omega) = k\}$. Since all ω that have the same number (in this case k) of 1's have the same probability, we simply need to count how many such ω's there are, and multiply by this individual probability.

Exercise 2.2 *Verify that* $\mathbb{P}_n(Z_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Z is called a *Binomial*(n, p) random variable.

Infinite Bernoulli i.i.d. sequence and Geometric random variables We would like to generalize the Bernoulli sequence probability space to an infinite sequence. We would like to choose $\Omega = \{0, 1\}^{\mathbb{N}}$ as our outcome space, but this is not a countable set. We will come back to the issue of properly defining the probability space with this uncountable Ω.

For now, if we still consider the mental experiment of infinite i.i.d. Bernoulli(p) sequence of random variables $X_1, X_2, \cdots$, which we interpret once more as coin tosses. We define Y be the number of tosses till the first heads. If we are just interested in Y (the first heads rather than all outcomes of all tosses), we can take Ω to be $\mathbb{N}$.

Exercise 2.3 *Although we cannot define a countable probability space for the infinite i.i.d. Bernoulli sequence, show that if we just want define a space for* Y*, we can take* $\Omega = \mathbb{N}$ *and* $\mathbb{P}(i) = (1-p)^{i-1} \cdot p$ *for* $i \geq 1$.

Y is known as a *Geometric*(p) random variable.

Let us calculate $\mathbb{E}[Y]$, in a somewhat creative way. Let E be the event that the first toss is heads. Then by total expectation we have,

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[Y|E] \cdot \mathbb{P}[E] + \mathbb{E}[Y|E^c] \cdot \mathbb{P}[E^c] \\ &= 1 \cdot p + (1 + \mathbb{E}[Y]) \cdot (1-p) \end{aligned}$$

Thus we have, $\mathbb{E}[Y] = \frac{1}{p}$. The main observation that we used here is that, thanks to independence, when the first toss is *not* heads, then the problem resets (with the hindsight of one consumed toss).