1 Probability spaces

Here are some basic probability notions we will be using throughout the course. We will only deal with finite probability spaces which makes the setup easier. A *probability space* (Ω, p) consists of a finite set Ω of *sample points* together with a *probability distribution*, which assigns to every $\omega \in \Omega$ a probability p_ω. The probabilities are non-negative and their sum is one, i.e. $\sum_{\omega\in\Omega} p_\omega = 1$. The *uniform distribution* assigns the same probability $1/|\Omega|$ to all the sample points.

Suppose you have a coin that comes up heads (H) 1/3 of the time and tails (T) 2/3 of the time. This can be modeled as a probability space with sample space $\{\mathtt{H}, \mathtt{T}\}$ and probability distribution $p_\mathtt{H} = 1/3, p_\mathtt{T} = 2/3$. If you toss the coin twice, you expect the pattern HH to occur 1/9 of the time, HT and TH to occur 2/9 of the time each, and TT to occur the remaining 4/9 of the time. This can be modeled as a probability space with sample space $\mathtt{HH}, \mathtt{HT}, \mathtt{TH}, \mathtt{TT}$ and probability distribution $p_\mathtt{HH} = 1/9, p_\mathtt{HT} = 2/9, p_\mathtt{TH} = 2/9, p_\mathtt{TT} = 4/9$.

This shows a general way of obtaining a large probability space from smaller ones: Given two probability spaces (Ω, p) and (Ω', p'), their *product space* is $(\Omega \times \Omega', p \cdot p')$, where $p \cdot p'$ assigns probability $p_\omega \cdot p'_{\omega'}$ to the sample point (ω, ω'). We can extend this operation to any number of probability spaces. The n-th power of a probability space is obtained by taking the product with itself n times.

Notice that a product of probability spaces under uniform distributions has a uniform distribution.

2 Events and random variables

An *event* is a subset of the sample space. The probability of event E is the sum of the probabilities of all the sample points inside E:

$$\Pr_{\Omega,p}[E] = \sum_{\omega\in E} p_\omega.$$

Usually we just write $\Pr[E]$ for $\Pr_{\Omega,p}[E]$.

For example consider the probability space $\{\mathtt{H}, \mathtt{T}\}^9$ under the uniform distribution modelling sequences of 9 tosses of a random coin. Then the event H_1 consisting of those sample points whose first entry is H has probability $\Pr[H_1] = 1/2$. This can be calculated from the above formula since there are 2^8 different sample points that start with H each with probability 2^{-9}. But there is an easier way: Since our probability space is a product of nine spaces and the event only depends on the first component (i.e., whether $\omega \in H_1$ or not is completely determined by the first coordinate of ω), then we can think of H_1 as an event over $\{\mathtt{H}, \mathtt{T}\}$ under the uniform distribution, and clearly $\Pr[H_1] = p_\mathtt{H} = 1/2$.

Now consider the event M consisting of those sample points that have more Hs than Ts. What is $\Pr[M]$? Now M depends on all nine components, so we have to resort to a different calculation.

Let $\overline{M}$ be the event of those sample points that have more Ts than Hs. To each sample point in M we can uniquely associate a sample point in $\overline{M}$ by turning each H into a T and vice versa: Thus HHHTTTHHH becomes TTTHHHTTT. This is a one-to-one, probability preserving map from M to $\overline{M}$, so $\Pr[M] = \Pr[\overline{M}]$. Since M and $\overline{M}$ partition the sample space (they are disjoint and cover the whole space), we have $\Pr[M] + \Pr[\overline{M}] = \Pr[\Omega] = 1$, so it must be that $\Pr[M] = \Pr[\overline{M}] = 1/2$.

A *random variable* over a probability space is a function from Ω to some set of values. We can view events as random variables that take values true (if ω is in the event) and false (if not). If X is a real-valued random variable, its expectation is given by

$$\mathrm{E}[X] = \sum_{\omega \in \Omega} p_\omega X(\omega).$$

Let N be the random variable over $\{\mathtt{H}, \mathtt{T}\}^9$ that counts the number of Hs in the sample. For example, $N(\mathtt{HHHTTTHHH}) = 6$. How can we calculate $\mathrm{E}[N]$? Doing a direct calculation is difficult. One easier way to do it is to notice that if two sample points ω and ω' are associated by the above map, then $N(\omega) + N(\omega') = 9$, so by pairing up the points we get $\mathrm{E}[N] = 9/2$.

But there is an even easier way. One useful property of any pair of random variables is linearity of expectation:

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

This formula is especially useful when we can break up a random variable into a sum of simpler ones. For each position $1 \le i \le 9$, we can define a random variable N_i that takes value 1 if the ith coordinate of the sample point is H and 0 if not. Then

$$N = N_1 + N_2 + \cdots + N_9$$

and therefore

$$\mathrm{E}[N] = \mathrm{E}[N_1] + \mathrm{E}[N_2] + \cdots + \mathrm{E}[N_9]$$

But now random variable N_1 only depends on the first component of the sample space, we have

$$\mathrm{E}[N_1] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = 1/2$$

and similarly for $\mathrm{E}[N_2] = 1/2, \ldots, \mathrm{E}[N_9] = 1/2$. Putting these together we get $\mathrm{E}[N] = 9 \cdot (1/2) = 9/2$.

Notice one thing we did in this calculation: We took an event "the ith coordinate is H" and associated a random variable N_i that takes value 1 if the event is true, and 0 if not. This is a useful trick; such random variables are called *indicator random variables* of the corresponding event.

3 Independence

Two events E and E' are *independent* if $\Pr[E \cap E'] = \Pr[E] \cdot \Pr[E']$. Independent events arise naturally from product spaces: If E is an event over (Ω, p) and E' is an event over (Ω', p'), they both naturally extend to events over the product space $(\Omega \times \Omega', p \cdot p')$, and you can check they satisfy the independence requirement.

Consider $\{\mathtt{H}, \mathtt{T}\}^9$ under the uniform distribution and the following events:

- H_i of those ω such that the ith entry of ω is `H`
- E such that ω contains an even number of `H`
- M such that ω contains more `H`s than `T`s.

Then H_1 and H_2 are independent since they can be viewed as events over distinct components of a product probability space. H_1 and E are also independent, although they both depend on the first entry of ω. Seeing this requires a little bit of work. H_1 and M are not independent, and neither are E and M, but I don't know an easy way to see this.

Events $E_1, \ldots, E_k$ are *independent* if for every subset $S \subseteq \{1, \ldots, k\}$,

$$\Pr\Big[\bigcap_{i \in S} E_i\Big] = \prod_{i \in S} \Pr[E_i].$$

They are *pairwise independent* if $\Pr[E_i \cap E_j] = \Pr[E_i]\Pr[E_j]$ for every $i \neq j$. Events $H_1, \ldots, H_9$ above are independent, and so are $H_1, \ldots, H_8, E$, but $H_1, \ldots, H_9, E$ are not independent: $\Pr[H_1 \cap \cdots \cap H_9 \cap E] = 0$ while $\Pr[H_1] \cdots \Pr[H_9]\Pr[E] = 2^{-10}$. However $H_1, \ldots, H_9, E$ are pairwise independent.

Random variables $X_1, \ldots, X_k$ are *independent* if for every collection of real-valued functions $f_1, \ldots, f_k$:

$$\mathrm{E}[f_1(X_1) \cdots f_k(X_k)] = \mathrm{E}[f_1(X_1)] \cdots \mathrm{E}[f_k(X_k)].$$

You should check that a collection of events is independent if and only if their indicator random variables are independent.

4 More on random variables and conditioning

We defined a random variable X as a function from a sample space Ω (with probability distribution p) to some set of values Ξ. The random variable X induces a probability space over Ξ as follows: The probability q_ξ of $\xi \in \Xi$ is given by $q_\xi = \Pr_{\Omega,p}[X = \chi]$. For example, let N be a random variable denoting the number of `H`s with respect to the probability space $\{\mathtt{H}, \mathtt{T}\}^9$ under the uniform distribution. Then N induces a probability distribution over the space $\{0, 1, \ldots, 9\}$ that assigns probability 2^{-9} to 0, $9 \cdot 2^{-9}$ to 1, and in general $\binom{9}{k}2^{-9}$ to k.

From this perspective random variables themselves can be viewed as a "basic object", while probability spaces and events inside them as "derived quantities". Let's go back to the the example of $\{\mathtt{H}, \mathtt{T}\}^9$ under the uniform distribution and see how this perspective plays out.

Let H_i be a random variable that indicates the ith position of a sample point ω, i.e. $H_i = \mathtt{H}$ if the ith entry is an `H` and `T` otherwise. Then the variables $H_1, \ldots, H_9$ are independent and their values determine the value of any other random variable over this space. So a random variable is now a *function* of $H_1, \ldots, H_9$. Since events are also random variables that take true-false values, they can be viewed as *boolean-valued functions* with inputs $H_1, \ldots, H_9$. When we want to emphasize this dependency we write

$$\Pr_{H_1,\ldots,H_9}[E] = \Pr_{H_1,\ldots,H_9}[E(H_1, \ldots, H_9)]$$

and similarly for a real-valued random variable X

$$\mathrm{E}_{H_1,\ldots,H_9}[X] = \mathrm{E}_{H_1,\ldots,H_9}[X(H_1, \ldots, H_9)].$$

Now think of the following scenario. Suppose we have already tossed the first two coins, but we do not know what their outcomes are. We want to know what is the expected number N of Hs with respect to the remaining seven coin tosses. This expectation is not a single number, but a *function* of H_1 and H_2. In this specific case, we can say that

$$\mathrm{E}_{H_3,\ldots,H_9}[N](H_1,H_2) = \begin{cases} 11/2, & \text{if } H_1H_2 = \mathtt{HH} \\ 9/2, & \text{if } H_1H_2 = \mathtt{HT} \text{ or } \mathtt{TH} \\ 7/2, & \text{if } H_1H_2 = \mathtt{TT}. \end{cases}$$

But since H_1 and H_2 are random variables, $\mathrm{E}_{H_3,\ldots,H_9}[N]$ itself becomes a real-valued random variable on the probability space $\{\mathtt{H},\mathtt{T}\}^2$ (determined by H_1 and H_2) under the uniform distribution. It induces the following probability distribution:

$$\mathrm{E}_{H_3,\ldots,H_9}[N] = \begin{cases} 11/2, & \text{with probability } 1/4 \\ 9/2, & \text{with probability } 1/2 \\ 7/2, & \text{with probability } 1/4. \end{cases}$$

Taking the expectation of $\mathrm{E}_{H_3,\ldots,H_9}[N]$ in this space we get

$$\mathrm{E}_{H_1,H_2}\big[\mathrm{E}_{H_3,\ldots,H_9}[N]\big] = \frac{11}{2}\cdot\frac{1}{4} + \frac{9}{2}\cdot\frac{1}{2} + \frac{7}{2}\cdot\frac{1}{4} = \frac{9}{2}$$

so $\mathrm{E}_{H_1,H_2}[\mathrm{E}_{H_3,\ldots,H_9}[N]] = \mathrm{E}_{H_1\ldots,H_9}[N]$.

In general, if we have a pair of independent random variables X, Y and a function f from the range of (X,Y) to the real numbers we can write

$$\mathrm{E}_{X,Y}[f(X,Y)] = \mathrm{E}_X[\mathrm{E}_Y[f(X,Y)]].$$

As a special case, when f is the indicator of an event F, we get

$$\Pr_{X,Y}[F(X,Y)] = \mathrm{E}_X[\Pr_Y[F(X,Y)]].$$

Conditioning What happens when X and Y are not independent? The notation becomes more complicated but we can do something similar. For any pair of random variables X, Y, we let $Y \mid X$ be the following function. The inputs of $Y \mid X$ are those x in the range of X that occur with nonzero probability. The outputs of $Y \mid X$ are random variables that take values in the range of Y and are described by the following probabilities

$$(Y \mid X)(x) \text{ takes value } y \text{ with probability } \frac{\Pr_{X,Y}[X = x \text{ and } Y = y]}{\Pr_X[X = x]}.$$

We usually write $Y \mid X = x$ for $(Y \mid X)(x)$. Then for any function f we can write

$$\mathrm{E}_{X,Y}[f(X,Y)] = \mathrm{E}_X\big[\mathrm{E}_{Y|X}[f(X,Y) \mid X]\big].$$

When the explicit dependences of f on X and Y are not relevant we can represent $f(X,Y)$ by a random variable R and write $\mathrm{E}[R] = \mathrm{E}_X[\mathrm{E}[R \mid X]]$, where it is understood that the inner expectation is conditional. Similarly by identifying events with their indicator functions

$$\Pr[F] = \mathrm{E}_X\big[\Pr[F \mid X]\big].$$

5 Concentration inequalities

Concentration inequalities tell us how likely a real-valued random variable is to deviate from its expectation. The simplest and most universal one is Markov's inequality, which tells us that a nonnegative random variable is unlikely to exceed its mean value by too much:

Theorem 1 (Markov's inequality)**.** *Let X be a random variable taking nonnegative real values. Then for every $t \geq 1$*

$$\Pr_X\big[X \geq t\,\mathrm{E}[X]\big] \leq 1/t.$$

This has a simple explanation: If a random variable takes large values with large enough probability, then its average must also be large.

Proof. Fix x and let I be the event $X \geq x$. We write

$$\mathrm{E}[X] = \mathrm{E}_I[\mathrm{E}[X \mid I]] = \mathrm{E}[X \mid I = \texttt{true}]\Pr[I = \texttt{true}] + \mathrm{E}[X \mid I = \texttt{false}]\Pr[I = \texttt{false}].$$

Now $\mathrm{E}_X[X \mid I = \texttt{false}] \geq 0$ and $\mathrm{E}_X[X \mid I = \texttt{true}] \geq x$, so we get

$$\Pr[I = \texttt{true}] \leq \frac{\mathrm{E}[X]}{x}.$$

Setting $x = t\,\mathrm{E}[X]$ proves the theorem. □

Here is one useful consequence of Markov's inequality:

Lemma 2. *Let F be an event and X be a random variable. If $\Pr[F] \geq p + q - pq$, then*

$$\Pr_X[\Pr[F \mid X] > p] \geq q.$$

Proof. We will work with $\overline{F}$, the complement of the event F. By assumption $\Pr[\overline{F}] \leq 1-p-q+pq = (1-p)(1-q)$. Then $\mathrm{E}_X[\Pr[\overline{F} \mid X]] \leq (1-p)(1-q)$. By Markov's inequality,

$$\Pr_X[\Pr[\overline{F} \mid X] \geq t(1-p)(1-q)] \leq 1/t.$$

Setting $t = 1/(1-q)$ we get

$$\Pr_X[\Pr[\overline{F} \mid X] \geq 1-p] \leq 1-q.$$

Taking the complement of the inner event we get

$$\Pr_X[\Pr[\overline{F} \mid X] < 1-p] \geq q$$

and finally, since $\Pr[\overline{F} \mid X] = 1 - \Pr[F \mid X]$ we get

$$\Pr_X[\Pr[F \mid X] > p] \geq q.$$

□

When we apply this inequality the $-pq$ term is often unimportant. For example suppose you know 20% of Hong Kong students do not get enough exercise. Let X be a random school. Setting $p = q = 10\%$ this claim tells you that in at least 10% of the schools, more than 10% of the students there do not get enough exercise.

More to come...

6 Comparing probability distributions

In cryptography we often want to compare two random variables X and Y taking values in some sample space Ξ. We are usually given some "sample" $\xi \in \Xi$ and we want to know if it "came from" X or from Y. We can rarely say this with certainty but we can make predictions based on the probability distributions induced by X and Y.

We say that X and Y are *identically distributed* (or *perfectly indistinguishable*) if for every event T over Ξ, $\Pr[T(X)] = \Pr[T(Y)]$. Intuitively, when two random variables are identically distributed, no experiment T can tell if we are looking at one or the other. You should convince yourself that X and Y are identically distributed *if and only if* they take every value with the same probability. For example, H_1 and H_2 in Section 4 are identically distributed.

Sometimes we have to deal with cases when X and Y are not identically distributed, but they are "very close" to one another. Let's consider random variables taking values in the set $\{0,1\}^n$. Suppose X is uniformly distributed and Y is uniformly distributed conditioned on never taking the value 0^n. That is, Y takes every value except 0^n with probability $1/(2^n - 1)$. Then X and Y are not identically distributed, because the event $Z(\xi)$ which is true when $\xi = 0^n$ and false when not distinguishes the two: $\Pr[Z(X)] = 2^{-n}$ while $\Pr[Z(Y)] = 0$.

However when n is large the difference between these two probabilities becomes tiny. And there is nothing special about Z here because it turns out that for *any* event T, $|\Pr[T(X)] - \Pr[T(Y)]| \leq 2^{-n}$. Intuitively, this says no matter which experiment we do, the chances that it tells X apart from Y is at most 2^{-n}. This motivates the following definition.

Definition 3. Two random variables X and Y taking values in Ξ are *ε-statistically indistinguishable* if for every event T over Ξ, $|\Pr[T(X)] - \Pr[T(Y)]| \leq \varepsilon$.

By playing with the above example you should be able to convince yourself that the quantity $|\Pr[T(X)] - \Pr[T(Y)]|$ is maximized by the event

$$T = \{\xi \colon \Pr[X = \xi] > \Pr[Y = \xi]\}$$

and this is true in general.