

# Absolute basics of probability for COMS 4236 \*

All of the probability we use will be over finite sample spaces - we will never have to worry about subtle issues like measurability or what is/isn't a legal event. The basic notions discussed in this brief note are

- sample spaces;
- probability distributions;
- events, compound events, and independence;
- random variables, expectation, and linearity of expectation.

This note is highly informal and contains only the absolute basics of terminology about these topics.

## 1 Sample spaces

As mentioned above we will only ever consider finite sample spaces. A *sample space*  $S$  is the set of all possible outcomes of some “probabilistic experiment.” The notion of a probabilistic experiment and its corresponding sample space may be best illustrated through some concrete examples:

1. **Example 1:** One probabilistic experiment would be “Pick a uniform random person from the entire Earth’s population as of midnight EST on Jan 1 2023.” For this probabilistic experiment, the sample space  $S$  would simply be the set of all living people on Earth at midnight EST on Jan 1 2023.
2. **Example 2:** A different probabilistic experiment (closer to our concerns in COMS 4236) would be “Choose a random  $n$ -bit string.” In this case the sample space  $S$  would be the set  $S = \{0, 1\}^n$ .
3. **Example 3:** Finally, a third probabilistic experiment would be “Choose a random number between 1 and  $n$  where each number is  $i$  chosen with probability proportional to  $i^2$ .” In this case the sample space  $S$  would be the set  $[n] = \{1, \dots, n\}$ .

## 2 Probability distributions

A *probability distribution*  $\mathcal{D}$  over a sample space  $S$  is defined by a probability weight  $\mathcal{D}(s)$  associated with each outcome  $s \in S$ . These probability weights must be nonnegative (they can be zero) and they must sum to 1; thus we have

$$\mathcal{D}(s) \geq 0 \text{ for all } s \in S \quad \text{and} \quad \sum_{s \in S} \mathcal{D}(s) = 1.$$

---

\*We gratefully acknowledge Rocco Servedio who prepared these notes for COMS 4236.

A probabilistic experiment naturally corresponds to a probability distribution over the relevant sample space. Returning to Example (1.) from above, for each person  $s$  in the world we would have  $\mathcal{D}(s) = 1/N$  where  $N$  is the total number of people in the world. For Example (2.), we would have  $\mathcal{D}(x) = \frac{1}{2^n}$  for each  $x \in \{0, 1\}^n$ . For Example (3.), since  $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$ , we would have  $\mathcal{D}(i) = \frac{i^2}{n(n+1)(2n+1)/6}$  for each  $i \in [n]$ . The intuition to have in mind is that we pick a random element of the sample space  $S$  according to  $\mathcal{D}$ .

This last example illustrates that (of course) not all probability distributions need to put equal weights on all possible outcomes; when a distribution puts equal weight on all points in the sample space  $S$  we say it is the *uniform distribution* over  $S$ . Examples (1.) and (2.) above correspond to uniform distributions, Example (3.) does not.

Instead of writing  $\mathcal{D}(x)$  to denote the probability of outcome  $x \in S$ , we sometimes write  $\Pr_{\mathcal{D}}[x]$  or just  $\Pr[x]$  if the distribution  $\mathcal{D}$  is clear from the context.

### 3 Events, compound events and independence

An *event* is simply a subset  $A \subseteq S$  of the sample space. The probability of an event  $A$  under distribution  $\mathcal{D}$  over  $S$  is simply  $\Pr_{\mathcal{D}}[A] = \sum_{s \in A} \Pr_{\mathcal{D}}[s]$ . (Think of an event as “something that either does or doesn’t happen” when the probabilistic experiment takes place, i.e. when a random  $s$  is drawn according to  $\mathcal{D}$ .)

For Example (1.) above, one event would be the subset of human beings who are complexity theorists; the probability of this event would be the probability that a randomly selected person is a complexity theorist. For Example (2.), one event would be the set of all  $n$ -bit strings with exactly  $n/2$  many ones (say that  $n$  is even); since there are  $\binom{n}{n/2}$  elements in this set and the distribution is uniform, the probability of this event (i.e. the probability that a uniform random  $n$ -bit string has exactly half of its coordinates 1 and exactly half 0) would be  $\binom{n}{n/2}/2^n$ , which is  $\Theta(1/\sqrt{n})$  by **Stirling’s approximation**. For Example (3.), one event would be the set  $\{1, 2\}$ ; if  $n = 6$ , then for this event  $A$  under the distribution  $\mathcal{D}$  described above we would have  $\Pr[A] = \frac{1^2 + 2^2}{1^2 + \dots + 6^2} = \frac{5}{91}$ .

Given two events  $A, B \subseteq S$ , the *compound event* corresponding to  $A$  and  $B$  is written  $A \wedge B$  or  $A \cap B$ ; its probability is

$$\Pr[A \wedge B] = \sum_{s \in A \cap B} \Pr[s].$$

(As mentioned above, sometimes it’s more natural to think of an event  $A$  as “a condition that may or may not be satisfied when a random  $s$  is drawn from  $\mathcal{D}$ .” The notation  $A \wedge B$  captures this; its intuitive meaning is that when  $s$  is drawn, it satisfies both condition  $A$  and condition  $B$ . In the context of Example (1.),  $A$  might be that a randomly selected person has brown hair, and  $B$  might be that a randomly selected person is a complexity theorist;  $A \wedge B$  would be the condition that a randomly selected person is a brown-haired complexity theorist. A more formal take on the situation is that  $A$  is a subset of people, namely the set of brown-haired people, and  $B$  is another subset of people, namely the set of complexity theorists, and consequently  $A \cap B$  is the intersection of these two subsets, namely the set of all brown-haired complexity theorists.)

We have that

$$\Pr[A \wedge B] = \Pr[A|B] \cdot \Pr[B], \quad \text{where} \quad \Pr[A|B] = \frac{\sum_{s \in A \cap B} \Pr[s]}{\sum_{s \in B} \Pr[s]} = \frac{\Pr[A \wedge B]}{\Pr[B]}. \quad (1)$$

$\Pr[A|B]$  is the *conditional probability of  $A$  given  $B$* . In Example (2.), we might have that  $A$  is the set of all  $n$ -bit strings with an even number of 1’s and  $B$  is the set of all  $n$ -bit strings that have

their first three bits all being 1; then  $\Pr[A|B]$  would be the fraction of all  $n$ -bit strings of the form  $111*^{n-3}$  that have an even number of 1's.

An easy consequence of Equation (1) is the following:

$$\Pr[A] = \overbrace{\Pr[A \wedge B] + \Pr[A \wedge \overline{B}]}^{\text{"law of total probability"}} \leq \Pr[B] + \Pr[A|\overline{B}] \Pr[B] \leq \Pr[B] + \Pr[A|\overline{B}];$$

we will use this on several occasions.

Events  $A, B$  are said to be *independent* if  $\Pr[A \wedge B] = \Pr[A] \cdot \Pr[B]$ . Let's return to Example (2.), where again the probabilistic experiment is drawing a uniform  $n$ -bit string (call it  $x$ ). Are the events  $A$  and  $B$  described above independent? (Yes, assuming  $n > 3$ ; think about why...). Let  $C$  be the event "at least half of the bits in  $x$  are 1." Are events  $B$  and  $C$  independent? (No; again, think about why...)

Intuitively, independence between events is very powerful and useful because "independent repetitions of a random experiment drives probabilities down very fast" — if we perform a random experiment which has "success probability"  $p$  independently  $k$  times, the probability that all  $k$  occurrences result in success is only  $p^k$ . (Note that if the original sample space for a probabilistic experiment is  $S$ , then the sample space corresponding to  $k$  executions of the probabilistic experiment is  $S^k$ , the set of all  $k$ -tuples of elements of  $S$ .)

## 4 Random variables, expectation, and linearity of expectation.

Given a sample space  $S$  and a distribution  $\mathcal{D}$  over  $S$ , a *random variable* is a function  $X$  from  $S$  to  $\mathbb{R}$ . In the context of Example (1.), one possible random variable would be the function which, on input a person on Earth, outputs their height in centimeters. In Example (3.), one possible random variable would be the function  $X(s) = 3s + 4$ .

The *expectation* of a random variable  $X$  is "the average value it takes over a random draw from  $\mathcal{D}$ "; more precisely, it is

$$\mathbf{E}[X] := \sum_{s \in S} X(s) \mathcal{D}(s) = \sum_a a \cdot \Pr[X = a].$$

Returning to the random variable  $X$  described above for Example (1.),  $\mathbf{E}[x]$  would just be the average height in centimeters of a random person on earth. In Example (3.), with  $n = 3$  (so  $S = \{1, 2, 3\}$ ) we would have

$$\mathbf{E}[X] = (3 \cdot 1 + 4) \cdot \frac{1^2}{1^2 + 2^2 + 3^2} + (3 \cdot 2 + 4) \cdot \frac{2^2}{1^2 + 2^2 + 3^2} + (3 \cdot 3 + 4) \cdot \frac{3^2}{1^2 + 2^2 + 3^2}.$$

The most important thing to know about expectation is the principle of *linearity of expectation*. This says that given any collection of random variables  $X_1, \dots, X_t$  over a sample space  $S$ , *whether or not they are independent* we have

$$\mathbf{E}[X_1 + \dots + X_t] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_t]$$

(think about why this is true!) This principle can often simplify your life; for example, consider Example (2.) where the random variable  $X$  is the number of 1's in a uniformly random  $n$ -bit string. What is  $\mathbf{E}[X]$ ? A direct application of the definition of expectation gives

$$\mathbf{E}[X] = \sum_{i=0}^n i \cdot \frac{\binom{n}{i}}{2^n},$$

(since a uniform random  $n$ -bit string has  $i$  ones with probability  $\binom{n}{i}/2^n$ ). It can be shown that this evaluates to  $n/2$ , but linearity of expectation makes this very easy to see: we have  $X = X_1 + \dots + X_n$  where  $X_i$  is 1 if the  $i$ -th bit of the string is 1. Applying linearity of expectation we get that

$$\mathbf{E}[X] = \mathbf{E}[X_1 + \dots + X_n] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_n] = n \cdot (1/2)$$

since each of the  $n$  random variables  $X_i$  has  $\mathbf{E}[X_i] = 1/2$ .