16 Independence

16.1 Definitions

Suppose that we flip two fair coins simultaneously on opposite sides of a room. Intuitively, the way one coin lands does not affect the way the other coin lands. The mathematical concept that captures this intuition is called *independence*:

Definition 16.1.1. Events A and B are independent if $\Pr[B] = 0$ or if

$$\Pr\left[A \mid B\right] = \Pr[A]. \tag{16.1}$$

In other words, A and B are independent if knowing that B happens does not alter the probability that A happens, as is the case with flipping two coins on opposite sides of a room.

16.1.1 Potential Pitfall

Students sometimes get the idea that disjoint events are independent. The *opposite* is true: if $A \cap B = \emptyset$, then knowing that A happens means you know that B does not happen. So disjoint events are *never* independent—unless one of them has probability zero.

16.1.2 Alternative Formulation

Sometimes it is useful to express independence in an alternate form:

Theorem 16.1.2. *A and B are independent if and only if*

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]. \tag{16.2}$$

Proof. There are two cases to consider depending on whether or not $\Pr[B] = 0$.

Case 1 ($\Pr[B] = 0$)**:** If $\Pr[B] = 0$, A and B are independent by Definition 16.1.1. In addition, Equation 16.2 holds since both sides are 0. Hence, the theorem is true in this case.

Case 2 ($\Pr[B] > 0$)**:** By Definition 15.1.1,

$$\Pr[A \cap B] = \Pr\left[A \mid B\right] \Pr[B].$$

So Equation 16.2 holds if

$$\Pr\left[A \mid B\right] = \Pr[A],$$

which, by Definition 16.1.1, is true iff A and B are independent. Hence, the theorem is true in this case as well. ■

16.2 Independence Is an Assumption

Generally, independence is something that you *assume* in modeling a phenomenon. For example, consider the experiment of flipping two fair coins. Let A be the event that the first coin comes up heads, and let B be the event that the second coin is heads. If we assume that A and B are independent, then the probability that both coins come up heads is:

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

In this example, the assumption of independence is reasonable. The result of one coin toss should have negligible impact on the outcome of the other coin toss. And if we were to repeat the experiment many times, we would be likely to have $A \cap B$ about 1/4 of the time.

There are, of course, many examples of events where assuming independence is *not* justified, For example, let C be the event that tomorrow is cloudy and R be the event that tomorrow is rainy. Perhaps $\Pr[C] = 1/5$ and $\Pr[R] = 1/10$ in Boston. If these events were independent, then we could conclude that the probability of a rainy, cloudy day was quite small:

$$\Pr[R \cap C] = \Pr[R] \cdot \Pr[C] = \frac{1}{5} \cdot \frac{1}{10} = \frac{1}{50}.$$

Unfortunately, these events are definitely not independent; in particular, every rainy day is cloudy. Thus, the probability of a rainy, cloudy day is actually $1/10$.

Deciding when to *assume* that events are independent is a tricky business. In practice, there are strong motivations to assume independence since many useful formulas (such as Equation 16.2) only hold if the events are independent. But you need to be careful lest you end up deriving false conclusions. We'll see several famous examples where (false) assumptions of independence led to trouble over the next several chapters. This problem gets even trickier when there are more than two events in play.

16.3 Mutual Independence

16.3.1 Definition

We have defined what it means for two events to be independent. What if there are more than two events? For example, how can we say that the flips of n coins are all independent of one another?

Events $E_1, \ldots, E_n$ are said to be *mutually independent* if and only if the probability of any event E_i is unaffected by knowledge of the other events. More formally:

Definition 16.3.1. A set of events $E_1, E_2, \ldots, E_n$, is *mutually independent* if $\forall i \in [1, n]$ and $\forall S \subseteq [1, n] - \{i\}$, either

$$\Pr\left[\bigcap_{j\in S} E_j\right] = 0 \quad \text{or} \quad \Pr[E_i] = \Pr\left[E_i \;\middle|\; \bigcap_{j\in S} E_j\right].$$

In other words, no matter which other events are known to occur, the probability that E_i occurs is unchanged for any i.

For example, if we toss 100 fair coins at different times, we might reasonably assume that the tosses are mutually independent since the probability that the ith coin is heads should be $1/2$, no matter which other coin tosses came out heads.

16.3.2 Alternative Formulation

Just as Theorem 16.1.2 provided an alternative definition of independence for two events, there is an alternative definition for mutual independence.

Theorem 16.3.2. *A set of events $E_1, E_2, \ldots, E_n$ is mutually independent iff $\forall S \subseteq [1, n]$,*

$$\Pr\left[\bigcap_{j\in S} E_j\right] = \prod_{j\in S} \Pr[E_j].$$

The proof of Theorem 16.3.2 uses induction and reasoning similar to the proof of Theorem 16.1.2. We will not include the details here.

Theorem 16.3.2 says that $E_1, E_2, \ldots, E_n$ are mutually independent if and only

if all of the following equations hold for all distinct i, j, k, and l:

$$\begin{aligned}
\Pr[E_i \cap E_j] &= \Pr[E_i] \cdot \Pr[E_j] \\
\Pr[E_i \cap E_j \cap E_k] &= \Pr[E_i] \cdot \Pr[E_j] \cdot \Pr[E_k] \\
\Pr[E_i \cap E_j \cap E_k \cap E_l] &= \Pr[E_i] \cdot \Pr[E_j] \cdot \Pr[E_k] \cdot \Pr[E_l] \\
&\vdots \\
\Pr[E_1 \cap \cdots \cap E_n] &= \Pr[E_1] \cdots \Pr[E_n].
\end{aligned}$$

For example, if we toss n fair coins, the tosses are mutually independent iff for all $m \in [1, n]$ and every subset of m coins, the probability that every coin in the subset comes up heads is 2^{-m}.

16.3.3 DNA Testing

Assumptions about independence are routinely made in practice. Frequently, such assumptions are quite reasonable. Sometimes, however, the reasonableness of an independence assumption is not so clear, and the consequences of a faulty assumption can be severe.

For example, consider the following testimony from the O. J. Simpson murder trial on May 15, 1995:

Mr. Clarke: When you make these estimations of frequency—and I believe you touched a little bit on a concept called independence?

Dr. Cotton: Yes, I did.

Mr. Clarke: And what is that again?

Dr. Cotton: It means whether or not you inherit one allele that you have is not—does not affect the second allele that you might get. That is, if you inherit a band at 5,000 base pairs, that doesn't mean you'll automatically or with some probability inherit one at 6,000. What you inherit from one parent is what you inherit from the other.

Mr. Clarke: Why is that important?

Dr. Cotton: Mathematically that's important because if that were not the case, it would be improper to multiply the frequencies between the different genetic locations.

Mr. Clarke: How do you—well, first of all, are these markers independent that you've described in your testing in this case?

Presumably, this dialogue was as confusing to you as it was for the jury. Essentially, the jury was told that genetic markers in blood found at the crime scene matched Simpson's. Furthermore, they were told that the probability that the markers would be found in a randomly-selected person was at most 1 in 170 million. This astronomical figure was derived from statistics such as:

- 1 person in 100 has marker A.
- 1 person in 50 marker B.
- 1 person in 40 has marker C.
- 1 person in 5 has marker D.
- 1 person in 170 has marker E.

Then these numbers were multiplied to give the probability that a randomly-selected person would have all five markers:

$$\begin{aligned}\Pr[A \cap B \cap C \cap D \cap E] &= \Pr[A] \cdot \Pr[B] \cdot \Pr[C] \cdot \Pr[D] \cdot \Pr[E] \\ &= \frac{1}{100} \cdot \frac{1}{50} \cdot \frac{1}{40} \cdot \frac{1}{5} \cdot \frac{1}{170} \\ &= \frac{1}{170{,}000{,}000}.\end{aligned}$$

The defense pointed out that this assumes that the markers appear mutually independently. Furthermore, all the statistics were based on just a few hundred blood samples.

After the trial, the jury was widely mocked for failing to "understand" the DNA evidence. If you were a juror, would *you* accept the 1 in 170 million calculation?

16.4 Pairwise Independence

The definition of mutual independence seems awfully complicated—there are so many subsets of events to consider! Here's an example that illustrates the subtlety of independence when more than two events are involved. Suppose that we flip three fair, mutually-independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.
- A_2 is the event that coin 2 matches coin 3.

- A_3 is the event that coin 3 matches coin 1.

Are A_1, A_2, A_3 mutually independent?

The sample space for this experiment is:

$$\{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}.$$

Every outcome has probability $(1/2)^3 = 1/8$ by our assumption that the coins are mutually independent.

To see if events A_1, A_2, and A_3 are mutually independent, we must check a sequence of equalities. It will be helpful first to compute the probability of each event A_i:

$$\begin{aligned}\Pr[A_1] &= \Pr[HHH] + \Pr[HHT] + \Pr[TTH] + \Pr[TTT] \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{2}.\end{aligned}$$

By symmetry, $\Pr[A_2] = \Pr[A_3] = 1/2$ as well. Now we can begin checking all the equalities required for mutual independence in Theorem 16.3.2:

$$\begin{aligned}\Pr[A_1 \cap A_2] &= \Pr[HHH] + \Pr[TTT] \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \Pr[A_1]\Pr[A_2].\end{aligned}$$

By symmetry, $\Pr[A_1 \cap A_3] = \Pr[A_1] \cdot \Pr[A_3]$ and $\Pr[A_2 \cap A_3] = \Pr[A_2] \cdot \Pr[A_3]$ must hold also. Finally, we must check one last condition:

$$\begin{aligned}\Pr[A_1 \cap A_2 \cap A_3] &= \Pr[HHH] + \Pr[TTT] \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4} \\ &\neq \Pr[A_1]\Pr[A_2]\Pr[A_3] = \frac{1}{8}.\end{aligned}$$

The three events A_1, A_2, and A_3 are not mutually independent even though any two of them are independent! This not-quite mutual independence seems weird at first, but it happens. It even generalizes:

Definition 16.4.1. A set $A_1, A_2, \ldots,$ of events is *k-way independent* iff every set of k of these events is mutually independent. The set is *pairwise independent* iff it is 2-way independent.

So the sets A_1, A_2, A_3 above are pairwise independent, but not mutually independent. Pairwise independence is a much weaker property than mutual independence.

For example, suppose that the prosecutors in the O. J. Simpson trial were wrong and markers A, B, C, D, and E appear only *pairwise* independently. Then the probability that a randomly-selected person has all five markers is no more than:

$$\begin{aligned}\Pr[A \cap B \cap C \cap D \cap E] &\leq \Pr[A \cap E] \\ &= \Pr[A] \cdot \Pr[E] \\ &= \frac{1}{100} \cdot \frac{1}{170} \\ &= \frac{1}{17{,}000}.\end{aligned}$$

The first line uses the fact that $A \cap B \cap C \cap D \cap E$ is a subset of $A \cap E$. (We picked out the A and E markers because they're the rarest.) We use pairwise independence on the second line. Now the probability of a random match is 1 in 17,000—a far cry from 1 in 170 million! And this is the strongest conclusion we can reach assuming only pairwise independence.

On the other hand, the 1 in 17,000 bound that we get by assuming pairwise independence is a lot better than the bound that we would have if there were no independence at all. For example, if the markers are dependent, then it is possible that

everyone with marker E has marker A,

everyone with marker A has marker B,

everyone with marker B has marker C, and

everyone with marker C has marker D.

In such a scenario, the probability of a match is

$$\Pr[E] = 1/170.$$

So a stronger independence assumption leads to a smaller bound on the probability of a match. The trick is to figure out what independence assumption is reasonable. Assuming that the markers are *mutually* independent may well *not* be reasonable unless you have examined hundreds of millions of blood samples. Otherwise, how would you know that marker D does not show up more frequently whenever the other four markers are simultaneously present?

We will conclude our discussion of independence with a useful, and somewhat famous, example known as the Birthday Paradox.

16.5 The Birthday Paradox

Suppose that there are 100 students in a class. What is the probability that some birthday is shared by two people? Comparing 100 students to the 365 possible birthdays, you might guess the probability lies somewhere around $1/3$—but you'd be wrong: the probability that there will be two people in the class with matching birthdays is actually $0.999999692\ldots$. In other words, the probability that all 100 birthdays are different is less than 1 in 3,000,000.

Why is this probability so small? The answer involves a phenomenon known as the *Birthday Paradox* (or the *Birthday Principle*), which is surprisingly important in computer science, as we'll see later.

Before delving into the analysis, we'll need to make some modeling assumptions:

- For each student, all possible birthdays are equally likely. The idea underlying this assumption is that each student's birthday is determined by a random process involving parents, fate, and, um, some issues that we discussed earlier in the context of graph theory. The assumption is not completely accurate, however; a disproportionate number of babies are born in August and September, for example.

- Birthdays are mutually independent. This isn't perfectly accurate either. For example, if there are twins in the class, then their birthdays are surely not independent.

We'll stick with these assumptions, despite their limitations. Part of the reason is to simplify the analysis. But the bigger reason is that our conclusions will apply to many situations in computer science where twins, leap days, and romantic holidays are not considerations. After all, whether or not two items collide in a hash table really has nothing to do with human reproductive preferences. Also, in pursuit of

generality, let's switch from specific numbers to variables. Let m be the number of people in the room, and let N be the number of days in a year.

We can solve this problem using the standard four-step method. However, a tree diagram will be of little value because the sample space is so enormous. This time we'll have to proceed without the visual aid!

Step 1: Find the Sample Space

Let's number the people in the room from 1 to m. An outcome of the experiment is a sequence $(b_1, \ldots, b_m)$ where b_i is the birthday of the ith person. The sample space is the set of all such sequences:

$$\mathcal{S} = \{\, (b_1, \ldots, b_m) \mid b_i \in \{1, \ldots N\}\,\}.$$

Step 2: Define Events of Interest

Our goal is to determine the probability of the event A in which some pair of people have the same birthday. This event is a little awkward to study directly, however. So we'll use a common trick, which is to analyze the *complementary* event $\overline{A}$, in which all m people have different birthdays:

$$\overline{A} = \{\, (b_1, \ldots, b_m) \in \mathcal{S} \mid \text{all } b_i \text{ are distinct}\,\}.$$

If we can compute $\Pr[\overline{A}]$, then we can compute what really want, $\Pr[A]$, using the identity

$$\Pr[A] + \Pr[\overline{A}] = 1.$$

Step 3: Assign Outcome Probabilities

We need to compute the probability that m people have a particular combination of birthdays $(b_1, \ldots, b_m)$. There are N possible birthdays and all of them are equally likely for each student. Therefore, the probability that the ith person was born on day b_i is $1/N$. Since we're assuming that birthdays are mutually independent, we can multiply probabilities. Therefore, the probability that the first person was born on day b_1, the second on b_2, and so forth is $(1/N)^m$. This is the probability of every outcome in the sample space, which means that the sample space is uniform. That's good news, because, as we have seen, it means that the analysis will be simpler.

Step 4: Compute Event Probabilities

We're interested in the probability of the event $\overline{A}$ in which everyone has a different birthday:

$$\overline{A} = \{\, (b_1, \ldots, b_n) \mid \text{all } b_i \text{ are distinct}\,\}.$$

This is a gigantic set. In fact, there are N choices for b_i, $N-1$ choices for b_2, and so forth. Therefore, by the Generalized Product Rule,

$$|\overline{A}| = \frac{N!}{(N-m)!} = N(N-1)(N-2)\cdots(N-m+1).$$

Since the sample space is uniform, we can conclude that

$$\Pr[\overline{A}] = \frac{|\overline{A}|}{N^m} = \frac{N!}{N^m(N-m)!}. \tag{16.3}$$

We're done!

Or are we? While correct, it would certainly be nicer to have a closed-form expression for Equation 16.3. That means finding an approximation for $N!$ and $(N-m)!$. But this is what we learned how to do in Section 9.6. In fact, since N and $N-m$ are each at least 100, we know from Corollary 9.6.2 that

$$\sqrt{2\pi N}\left(\frac{N}{e}\right)^N \quad \text{and} \quad \sqrt{2\pi(N-m)}\left(\frac{N-m}{e}\right)^{N-m}$$

are excellent approximations (accurate to within .09%) of $N!$ and $(N-m)!$, respectively. Plugging these values into Equation 16.3 means that (to within .2%)[1]

$$\begin{aligned}
\Pr[\overline{A}] &= \frac{\sqrt{2\pi N}\left(\frac{N}{e}\right)^N}{N^m\sqrt{2\pi(N-m)}\left(\frac{N-m}{e}\right)^{N-m}} \\
&= \sqrt{\frac{N}{N-m}}\,\frac{e^{N\ln(N)-N}}{e^{m\ln(N)}e^{(N-m)\ln(N-m)-(N-m)}} \\
&= \sqrt{\frac{N}{N-m}}\,e^{(N-m)\ln(N)-(N-m)\ln(N-m)-m} \\
&= \sqrt{\frac{N}{N-m}}\,e^{(N-m)\ln\left(\frac{N}{N-m}\right)-m} \\
&= e^{\left(N-m+\frac{1}{2}\right)\ln\left(\frac{N}{N-m}\right)-m}.
\end{aligned} \tag{16.4}$$

[1] If there are two terms that can be off by .09%, then the ratio can be off by at most a factor of $(1.0009)^2 < 1.002$.

We can now evaluate Equation 16.4 for $m = 100$ and $N = 365$ to find that the probability that all 100 birthdays are different is[2]

$$3.07\ldots \cdot 10^{-7}.$$

We can also plug in other values of m to find the number of people so that the probability of a matching birthday will be about $1/2$. In particular, for $m = 23$ and $N = 365$, Equation 16.4 reveals that the probability that all the birthdays differ is $0.49\ldots$. So if you are in a room with 23 other people, the probability that some pair of people share a birthday will be a little better than $1/2$. It is because 23 seems like such a small number of people for a match that the phenomenon is called the *Birthday Paradox*.

16.5.1 Applications to Hashing

Hashing is frequently used in computer science to map large strings of data into short strings of data. In a typical scenario, you have a set of m items and you would like to assign each item to a number from 1 to N where no pair of items is assigned to the same number and N is as small as possible. For example, the items might be messages, addresses, or variables. The numbers might represent storage locations, devices, indices, or digital signatures.

If two items are assigned to the same number, then a *collision* is said to occur. Collisions are generally bad. For example, collisions can correspond to two variables being stored in the same place or two messages being assigned the same digital signature. Just imagine if you were doing electronic banking and your digital signature for a \$10 check were the same as your signature for a \$10 million dollar check. In fact, finding collisions is a common technique in breaking cryptographic codes.[3]

In practice, the assignment of a number to an item is done using a hash function

$$h : S \to [1, N],$$

where S is the set of items and $m = |S|$. Typically, the values of $h(S)$ are assigned randomly and are assumed to be equally likely in $[1, N]$ and mutually independent.

For efficiency purposes, it is generally desirable to make N as small as necessary to accommodate the hashing of m items without collisions. Ideally, N would be only a little larger than m. Unfortunately, this is not possible for random hash functions. To see why, let's take a closer look at Equation 16.4.

[2]The possible .2% error is so small that it is lost in the ... after 3.07.

[3]Such techniques are often referred to as *birthday attacks* because of the association of such attacks with the Birthday Paradox.

By Theorem 9.6.1 and the derivation of Equation 16.4, we know that the probability that there are no collisions for a random hash function is

$$\sim e^{\left(N-m+\frac{1}{2}\right)\ln\left(\frac{N}{N-m}\right)-m}. \tag{16.5}$$

For any m, we now need to find a value of N for which this expression is at least 1/2. That will tell us how big the hash table needs to be in order to have at least a 50% chance of avoiding collisions. This means that we need to find a value of N for which

$$\left(N-m+\frac{1}{2}\right)\ln\left(\frac{N}{N-m}\right)-m \sim \ln\left(\frac{1}{2}\right). \tag{16.6}$$

To simplify Equation 16.6, we need to get rid of the $\ln\left(\frac{N}{N-m}\right)$ term. We can do this by using the Taylor Series expansion for

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$$

to find that[4]

$$\begin{aligned}
\ln\left(\frac{N}{N-m}\right) &= -\ln\left(\frac{N-m}{N}\right) \\
&= -\ln\left(1-\frac{m}{N}\right) \\
&= -\left(-\frac{m}{N} - \frac{m^2}{2N^2} - \frac{m^3}{3N^3} - \cdots\right) \\
&= \frac{m}{N} + \frac{m^2}{2N^2} + \frac{m^3}{3N^3} + \cdots.
\end{aligned}$$

[4]This may not look like a simplification, but stick with us here.

Hence,

$$\begin{aligned}\left(N-m+\frac{1}{2}\right)\ln\left(\frac{N}{N-m}\right)-m &= \left(N-m+\frac{1}{2}\right)\left(\frac{m}{N}+\frac{m^2}{2N^2}+\frac{m^3}{3N^3}+\cdots\right)-m\\ &= \left(m+\frac{m^2}{2N}+\frac{m^3}{3N^2}+\cdots\right)\\ &\quad -\left(\frac{m^2}{N}+\frac{m^3}{2N^2}+\frac{m^4}{3N^3}+\cdots\right)\\ &\quad +\frac{1}{2}\left(\frac{m}{N}+\frac{m^2}{2N^2}+\frac{m^3}{3N^3}+\cdots\right)-m\\ &= -\left(\frac{m^2}{2N}+\frac{m^3}{6N^2}+\frac{m^4}{12N^3}+\cdots\right)\\ &\quad +\frac{1}{2}\left(\frac{m}{N}+\frac{m^2}{2N^2}+\frac{m^3}{3N^3}+\cdots\right).\end{aligned} \tag{16.7}$$

If N grows faster than m^2, then the value in Equation 16.7 tends to 0 and Equation 16.6 cannot be satisfied. If N grows more slowly than m^2, then the value in Equation 16.7 diverges to negative infinity, and, once again, Equation 16.6 cannot be satisfied. This suggests that we should focus on the case where $N = \Theta(m^2)$, when Equation 16.7 simplifies to

$$\sim \frac{-m^2}{2N}$$

and Equation 16.6 becomes

$$\frac{-m^2}{2N} \sim \ln\left(\frac{1}{2}\right). \tag{16.8}$$

Equation 16.8 is satisfied when

$$N \sim \frac{m^2}{2\ln(2)}. \tag{16.9}$$

In other words, N needs to grow quadratically with m in order to avoid collisions. This unfortunate fact is known as the *Birthday Principle* and it limits the efficiency of hashing in practice—either N is quadratic in the number of items being hashed or you need to be able to deal with collisions.

6.042J / 18.062J Mathematics for Computer Science
Fall 2010