Learning Objectives:

- Review the basic notion of probability theory that we will use.
- Sample spaces, and in particular the space $\{0,1\}^n$
- Events, probabilities of unions and intersections.
- Random variables and their expectation, variance, and standard deviation.
- Independence and correlation for both events and random variables.
- Markov, Chebyshev and Chernoff tail bounds (bounding the probability that a random variable will deviate from its expectation).

17
Probability Theory 101

> *"God doesn't play dice with the universe"*, Albert Einstein

> *"Einstein was doubly wrong ... not only does God definitely play dice, but He sometimes confuses us by throwing them where they can't be seen."*, Stephen Hawking

> *" 'The probability of winning a battle' has no place in our theory because it does not belong to any [random experiment]. Probability cannot be applied to this problem any more than the physical concept of work can be applied to the 'work' done by an actor reciting his part."*, Richard Von Mises, 1928 (paraphrased)

> *"I am unable to see why 'objectivity' requires us to interpret every probability as a frequency in some random experiment; particularly when in most problems probabilities are frequencies only in an imaginary universe invented just for the purpose of allowing a frequency interpretation."*, E.T. Jaynes, 1976

Before we show how to use randomness in algorithms, let us do a quick review of some basic notions in probability theory. This is not meant to replace a course on probability theory, and if you have not seen this material before, I highly recommend you look at additional resources to get up to speed.[1] Fortunately, we will not need many of the advanced notions of probability theory, but, as we will see, even the so-called "simple" setting of tossing n coins can lead to very subtle and interesting issues.

[1] Harvard's STAT 110 class (whose lectures are available on youtube) is a highly recommended introduction to probability. See also these lecture notes from MIT's "Mathematics for Computer Science" course.

17.1 RANDOM COINS

The nature of randomness and probability is a topic of great philosophical, scientific and mathematical depth. Is there actual random-

ness in the world, or does it proceed in a deterministic clockwork fashion from some initial conditions set at the beginning of time? Does probability refer to our uncertainty of beliefs, or to the frequency of occurrences in repeated experiments? How can we define probability over infinite sets?

These are all important questions that have been studied and debated by scientists, mathematicians, statisticians and philosophers. Fortunately, we will not need to deal directly with these questions here. We will be mostly interested in the setting of tossing n random, unbiased and independent coins. Below we define the basic probabilistic objects of *events* and *random variables* when restricted to this setting. These can be defined for much more general probabilistic experiments or *sample spaces*, and later on we will briefly discuss how this can be done. However, the n-coin case is sufficient for almost everything we'll need in this course.

If instead of "heads" and "tails" we encode the sides of each coin by "zero" and "one", we can encode the result of tossing n coins as a string in $\{0,1\}^n$. Each particular outcome $x \in \{0,1\}^n$ is obtained with probability 2^{-n}. For example, if we toss three coins, then we obtain each of the 8 outcomes $000, 001, 010, 011, 100, 101, 110, 111$ with probability $2^{-3} = 1/8$ (see also Fig. 17.1). We can describe the experiment of tossing n coins as choosing a string x uniformly at random from $\{0,1\}^n$, and hence we'll use the shorthand $x \sim \{0,1\}^n$ for x that is chosen according to this experiment.

An *event* is simply a subset A of $\{0,1\}^n$. The *probability of* A, denoted by $\Pr_{x\sim\{0,1\}^n}[A]$ (or $\Pr[A]$ for short, when the sample space is understood from the context), is the probability that an x chosen uniformly at random will be contained in A. Note that this is the same as $|A|/2^n$ (where $|A|$ as usual denotes the number of elements in the set A). For example, the probability that x has an even number of ones is $\Pr[A]$ where $A = \{x : \sum_{i=0}^{n-1} x_i = 0 \mod 2\}$. In the case $n = 3$, $A = \{000, 011, 101, 110\}$, and hence $\Pr[A] = \frac{4}{8} = \frac{1}{2}$. It turns out this is true for every n:

Lemma 17.1

$$\Pr_{x\sim\{0,1\}^n}[\sum_{i=0}^{n-1} x_i \text{ is even }] = 1/2 \tag{17.1}$$

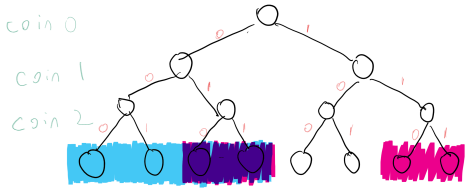

Figure 17.1: The probabilistic experiment of tossing three coins corresponds to making $2 \times 2 \times 2 = 8$ choices, each with equal probability. In this example, the blue set corresponds to the event $A = \{x \in \{0,1\}^3 \mid x_0 = 0\}$ where the first coin toss is equal to 0, and the pink set corresponds to the event $B = \{x \in \{0,1\}^3 \mid x_1 = 1\}$ where the second coin toss is equal to 1 (with their intersection having a purplish color). As we can see, each of these events contains 4 elements (out of 8 total) and so has probability $1/2$. The intersection of A and B contains two elements, and so the probability that both of these events occur is $2/8 = 1/4$.

P To test your intuition on probability, try to stop here and prove the lemma on your own.

Proof of Lemma 17.1. Let $A = \{x \in \{0,1\}^n : \sum_{i=0}^{n-1} x_i = 0 \mod 2\}$. Since every x is obtained with probability 2^{-n}, to show this we need to

show that $|A| = 2^n/2 = 2^{n-1}$. For every $x_0, \ldots, x_{n-2}$, if $\sum_{i=0}^{n-2} x_i$ is even then $(x_0, \ldots, x_{n-1}, 0) \in A$ and $(x_0, \ldots, x_{n-1}, 1) \notin A$. Similarly, if $\sum_{i=0}^{n-2} x_i$ is odd then $(x_0, \ldots, x_{n-1}, 1) \in A$ and $(x_0, \ldots, x_{n-1}, 0) \notin A$. Hence, for every one of the 2^{n-1} prefixes $(x_0, \ldots, x_{n-2})$, there is exactly a single continuation of $(x_0, \ldots, x_{n-2})$ that places it in A.

■

We can also use the *intersection* ($\cap$) and *union* ($\cup$) operators to talk about the probability of both event A *and* event B happening, or the probability of event A *or* event B happening. For example, the probability p that x has an *even* number of ones *and* $x_0 = 1$ is the same as $\Pr[A \cap B]$ where $A = \{x \in \{0,1\}^n : \sum_{i=0}^{n-1} x_i = 0 \mod 2\}$ and $B = \{x \in \{0,1\}^n : x_0 = 1\}$. This probability is equal to $1/4$. (It is a great exercise for you to pause here and verify that you understand why this is the case.)

Because intersection corresponds to considering the logical AND of the conditions that two events happen, while union corresponds to considering the logical OR, we will sometimes use the $\wedge$ and $\vee$ operators instead of $\cap$ and $\cup$, and so write this probability $p = \Pr[A \cap B]$ defined above also as

$$\Pr_{x \sim \{0,1\}^n}\left[\sum_i x_i = 0 \mod 2 \wedge x_0 = 1\right]. \tag{17.2}$$

If $A \subseteq \{0,1\}^n$ is an event, then $\overline{A} = \{0,1\}^n \setminus A$ corresponds to the event that A does *not* happen. Since $|\overline{A}| = 2^n - |A|$, we get that

$$\Pr[\overline{A}] = \frac{|\overline{A}|}{2^n} = \frac{2^n - |A|}{2^n} = 1 - \frac{|A|}{2^n} = 1 - \Pr[A] \tag{17.3}$$

This makes sense: since A happens if and only if $\overline{A}$ does *not* happen, the probability of $\overline{A}$ should be one minus the probability of A.

R

Remark 17.2 — Remember the sample space. While the above definition might seem very simple and almost trivial, the human mind seems not to have evolved for probabilistic reasoning, and it is surprising how often people can get even the simplest settings of probability wrong. One way to make sure you don't get confused when trying to calculate probability statements is to always ask yourself the following two questions: **(1)** Do I understand what is the **sample space** that this probability is taken over?, and **(2)** Do I understand what is the definition of the **event** that we are analyzing?.

For example, suppose that I were to randomize seating in my course, and then it turned out that students sitting in row 7 performed better on the final: how surprising should we find this? If we started out with

the hypothesis that there is something special about the number 7 and chose it ahead of time, then the event that we are discussing is the event A that students sitting in number 7 had better performance on the final, and we might find it surprising. However, if we first looked at the results and then chose the row whose average performance is best, then the event we are discussing is the event B that there exists *some* row where the performance is higher than the overall average. B is a superset of A, and its probability (even if there is no correlation between sitting and performance) can be quite significant.

17.1.1 Random variables

Events correspond to Yes/No questions, but often we want to analyze finer questions. For example, if we make a bet at the roulette wheel, we don't want to just analyze whether we won or lost, but also *how much* we've gained. A (real valued) *random variable* is simply a way to associate a number with the result of a probabilistic experiment. Formally, a random variable is simply a function $X : \{0,1\}^n \to \mathbb{R}$ that maps every outcome $x \in \{0,1\}^n$ to a real number $X(x)$.[2] For example, the function $sum : \{0,1\}^n \to \mathbb{R}$ that maps x to the sum of its coordinates (i.e., to $\sum_{i=0}^{n-1} x_i$) is a random variable.

[2] In many probability texts a random variable is always defined to have values in the set $\mathbb{R}$ of real numbers, and this will be our default option as well. However, in some contexts in theoretical computer science we can consider random variables mapping to other sets such as $\{0,1\}^*$.

The *expectation* of a random variable X, denoted by $\mathbb{E}[X]$, is the average value that that this number takes, taken over all draws from the probabilistic experiment. In other words, the expectation of X is defined as follows:

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}^n} 2^{-n} X(x) \; . \tag{17.4}$$

If X and Y are random variables, then we can define $X + Y$ as simply the random variable that maps a point $x \in \{0,1\}^n$ to $X(x) + Y(x)$. One basic and very useful property of the expectation is that it is *linear*:

Lemma 17.3 — Linearity of expectation.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{17.5}$$

Proof.

$$\begin{aligned} \mathbb{E}[X+Y] = \sum_{x \in \{0,1\}^n} 2^{-n} \left(X(x) + Y(x)\right) = \\ \sum_{x \in \{0,1\}^b} 2^{-n} X(x) + \sum_{x \in \{0,1\}^b} 2^{-n} Y(x) = \\ \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned} \tag{17.6}$$

■

Similarly, $\mathbb{E}[kX] = k\,\mathbb{E}[X]$ for every $k \in \mathbb{R}$. For example, using the linearity of expectation, it is very easy to show that the expectation of the sum of the x_i's for $x \sim \{0,1\}^n$ is equal to $n/2$. Indeed, if we write $X = \sum_{i=0}^{n-1} x_i$ then $X = X_0 + \cdots + X_{n-1}$ where X_i is the random variable x_i. Since for every i, $\Pr[X_i = 0] = 1/2$ and $\Pr[X_i = 1] = 1/2$, we get that $\mathbb{E}[X_i] = (1/2)\cdot 0 + (1/2)\cdot 1 = 1/2$ and hence $\mathbb{E}[X] = \sum_{i=0}^{n-1} \mathbb{E}[X_i] = n \cdot (1/2) = n/2$.

P

If you have not seen discrete probability before, please go over this argument again until you are sure you follow it; it is a prototypical simple example of the type of reasoning we will employ again and again in this course.

If A is an event, then 1_A is the random variable such that $1_A(x)$ equals 1 if $x \in A$, and $1_A(x) = 0$ otherwise. Note that $\Pr[A] = \mathbb{E}[1_A]$ (can you see why?). Using this and the linearity of expectation, we can show one of the most useful bounds in probability theory:

Lemma 17.4 — Union bound. For every two events A, B, $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$

P

Before looking at the proof, try to see why the union bound makes intuitive sense. We can also prove it directly from the definition of probabilities and the cardinality of sets, together with the equation $|A \cup B| \leq |A| + |B|$. Can you see why the latter equation is true? (See also Fig. 17.2.)

Proof of Lemma 17.4. For every x, the variable $1_{A\cup B}(x) \leq 1_A(x) + 1_B(x)$. Hence, $\Pr[A \cup B] = \mathbb{E}[1_{A\cup B}] \leq \mathbb{E}[1_A + 1_B] = \mathbb{E}[1_A] + \mathbb{E}[1_B] = \Pr[A] + \Pr[B]$. ■

The way we often use this in theoretical computer science is to argue that, for example, if there is a list of 100 bad events that can happen, and each one of them happens with probability at most $1/10000$, then with probability at least $1 - 100/10000 = 0.99$, no bad event happens.

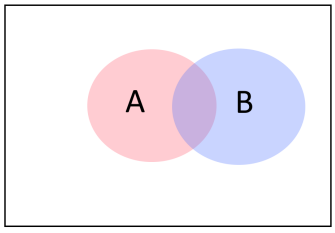

Figure 17.2: The *union bound* tells us that the probability of A or B happening is at most the sum of the individual probabilities. We can see it by noting that for every two sets $|A \cup B| \leq |A| + |B|$ (with equality only if A and B have no intersection).

17.1.2 Distributions over strings

While most of the time we think of random variables as having as output a *real number*, we sometimes consider random variables whose output is a *string*. That is, we can think of a map $Y : \{0,1\}^n \to \{0,1\}^*$ and consider the "random variable" Y such

that for every $y \in \{0,1\}^*$, the probability that Y outputs y is equal to $\frac{1}{2^n}|\{x \in \{0,1\}^n \mid Y(x) = y\}|$. To avoid confusion, we will typically refer to such string-valued random variables as *distributions* over strings. So, a *distribution* Y over strings $\{0,1\}^*$ can be thought of as a finite collection of strings $y_0, \ldots, y_{M-1} \in \{0,1\}^*$ and probabilities $p_0, \ldots, p_{M-1}$ (which are non-negative numbers summing up to one), so that $\Pr[Y = y_i] = p_i$.

Two distributions Y and Y' are *identical* if they assign the same probability to every string. For example, consider the following two functions $Y, Y' : \{0,1\}^2 \rightarrow \{0,1\}^2$. For every $x \in \{0,1\}^2$, we define $Y(x) = x$ and $Y'(x) = x_0(x_0 \oplus x_1)$ where $\oplus$ is the XOR operations. Although these are two different functions, they induce the same distribution over $\{0,1\}^2$ when invoked on a uniform input. The distribution $Y(x)$ for $x \sim \{0,1\}^2$ is of course the uniform distribution over $\{0,1\}^2$. On the other hand Y' is simply the map $00 \mapsto 00$, $01 \mapsto 01$, $10 \mapsto 11$, $11 \mapsto 10$ which is a permutation over the map $F : \{0,1\}^2 \rightarrow \{0,1\}^2$ defined as $F(x_0x_1) = x_0x_1$ and the map $G : \{0,1\}^2 \rightarrow \{0,1\}^2$ defined as $G(x_0x_1) = x_0(x_0 \oplus x_1)$

17.1.3 More general sample spaces.

While in this chapter we assume that the underlying probabilistic experiment corresponds to tossing n independent coins, everything we say easily generalizes to sampling x from a more general finite or countable set S (and not-so-easily generalizes to uncountable sets S as well). A *probability distribution* over a finite set S is simply a function $\mu : S \rightarrow [0,1]$ such that $\sum_{x \in S} \mu(s) = 1$. We think of this as the experiment where we obtain every $x \in S$ with probability $\mu(s)$, and sometimes denote this as $x \sim \mu$. An *event* A is a subset of S, and the probability of A, which we denote by $\Pr_\mu[A]$, is $\sum_{x \in A} \mu(x)$. A *random variable* is a function $X : S \rightarrow \mathbb{R}$, where the probability that $X = y$ is equal to $\sum_{x \in S \text{ s.t. } X(x)=y} \mu(x)$.

[3] TODO: add exercise on simulating die tosses and choosing a random number in $[m]$ by coin tosses

17.2 CORRELATIONS AND INDEPENDENCE

One of the most delicate but important concepts in probability is the notion of *independence* (and the opposing notion of *correlations*). Subtle correlations are often behind surprises and errors in probability and statistical analysis, and several mistaken predictions have been blamed on miscalculating the correlations between, say, housing prices in Florida and Arizona, or voter preferences in Ohio and Michigan. See also Joe Blitzstein's aptly named talk "Conditioning is the Soul of Statistics".[4]

Two events A and B are *independent* if the fact that A happens makes B neither more nor less likely to happen. For example, if we

[4] Another thorny issue is of course the difference between *correlation* and *causation*. Luckily, this is another point we don't need to worry about in our clean setting of tossing n coins.

think of the experiment of tossing 3 random coins $x \in \{0,1\}^3$, and we let A be the event that $x_0 = 1$ and B the event that $x_0 + x_1 + x_2 \geq 2$, then if A happens it is more likely that B happens, and hence these events are *not* independent. On the other hand, if we let C be the event that $x_1 = 1$, then because the second coin toss is not affected by the result of the first one, the events A and C are independent.

The formal definition is that events A and B are *independent* if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$. If $\Pr[A \cap B] > \Pr[A] \cdot \Pr[B]$ then we say that A and B are *positively correlated*, while if $\Pr[A \cap B] < \Pr[A] \cdot \Pr[B]$ then we say that A and B are *negatively correlated* (see Fig. 17.1).

If we consider the above examples on the experiment of choosing $x \in \{0,1\}^3$ then we can see that

$$\begin{aligned} \Pr[x_0 = 1] &= \tfrac{1}{2} \\ \Pr[x_0 + x_1 + x_2 \geq 2] = \Pr[\{011, 101, 110, 111\}] &= \tfrac{4}{8} = \tfrac{1}{2} \end{aligned} \tag{17.7}$$

but

$$\Pr[x_0 = 1 \ \wedge \ x_0 + x_1 + x_2 \geq 2] = \Pr[\{101, 110, 111\}] = \tfrac{3}{8} > \tfrac{1}{2} \cdot \tfrac{1}{2} \tag{17.8}$$

and hence, as we already observed, the events $\{x_0 = 1\}$ and $\{x_0 + x_1 + x_2 \geq 2\}$ are not independent and in fact are positively correlated. On the other hand, $\Pr[x_0 = 1 \wedge x_1 = 1] = \Pr[\{110, 111\}] = \frac{2}{8} = \frac{1}{2} \cdot \frac{1}{2}$ and hence the events $\{x_0 = 1\}$ and $\{x_1 = 1\}$ are indeed independent.

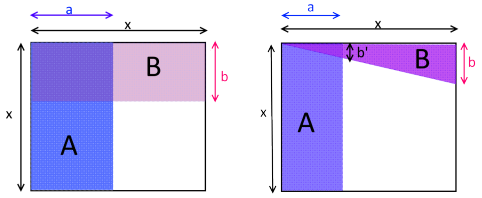

Figure 17.3: Two events A and B are *independent* if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$. In the two figures above, the empty $x \times x$ square is the sample space, and A and B are two events in this sample space. In the left figure, A and B are independent, while in the right figure they are negatively correlated, since B is less likely to occur if we condition on A (and vice versa). Mathematically, one can see this by noticing that in the left figure the areas of A and B respectively are $a \cdot x$ and $b \cdot x$, and so their probabilities are $\frac{a \cdot x}{x^2} = \frac{a}{x}$ and $\frac{b \cdot x}{x^2} = \frac{b}{x}$ respectively, while the area of $A \cap B$ is $a \cdot b$ which corresponds to the probability $\frac{a \cdot b}{x^2}$. In the right figure, the area of the triangle B is $\frac{b \cdot x}{2}$ which corresponds to a probability of $\frac{b}{2x}$, but the area of $A \cap B$ is $\frac{b' \cdot a}{2}$ for some $b' < b$. This means that the probability of $A \cap B$ is $\frac{b' \cdot a}{2x^2} < \frac{b}{2x} \cdot \frac{a}{x}$, or in other words $\Pr[A \cap B] < \Pr[A] \cdot \Pr[B]$.

R

Remark 17.5 — Disjointness vs independence. People sometimes confuse the notion of *disjointness* and *independence*, but these are actually quite different. Two events A and B are *disjoint* if $A \cap B = \emptyset$, which means that if A happens then B definitely does not happen. They are *independent* if $\Pr[A \cap B] = \Pr[A]\Pr[B]$ which means that knowing that A happens gives us no information about whether B happened or not. If A and B have nonzero probability, then being disjoint implies that they are *not* independent, since in particular it means that they are negatively correlated.

Conditional probability: If A and B are events, and A happens with nonzero probability then we define the probability that B happens *conditioned on* A to be $\Pr[B|A] = \Pr[A \cap B]/\Pr[A]$. This corresponds to calculating the probability that B happens if we already know that A happened. Note that A and B are independent if and only if $\Pr[B|A] = \Pr[B]$.

More than two events: We can generalize this definition to more than two events. We say that events $A_1, \ldots, A_k$ are *mutually independent* if knowing that any set of them occurred or didn't occur does not change the probability that an event outside the set occurs. Formally, the condition is that for every subset $I \subseteq [k]$,

$$\Pr[\wedge_{i\in I} A_i] = \prod_{i\in I} \Pr[A_i]. \tag{17.9}$$

For example, if $x \sim \{0,1\}^3$, then the events $\{x_0 = 1\}$, $\{x_1 = 1\}$ and $\{x_2 = 1\}$ are mutually independent. On the other hand, the events $\{x_0 = 1\}$, $\{x_1 = 1\}$ and $\{x_0 + x_1 = 0 \mod 2\}$ are *not* mutually independent, even though every pair of these events is independent (can you see why? see also Fig. 17.4).

17.2.1 Independent random variables

We say that two random variables $X : \{0,1\}^n \to \mathbb{R}$ and $Y : \{0,1\}^n \to \mathbb{R}$ are independent if for every $u, v \in \mathbb{R}$, the events $\{X = u\}$ and $\{Y = v\}$ are independent.[5] In other words, X and Y are independent if $\Pr[X = u \wedge Y = v] = \Pr[X = u]\Pr[Y = v]$ for every $u, v \in \mathbb{R}$. For example, if two random variables depend on the result of tossing different coins then they are independent:

Lemma 17.6 Suppose that $S = \{s_0, \ldots, s_{k-1}\}$ and $T = \{t_0, \ldots, t_{m-1}\}$ are disjoint subsets of $\{0, \ldots, n-1\}$ and let $X, Y : \{0,1\}^n \to \mathbb{R}$ be random variables such that $X = F(x_{s_0}, \ldots, x_{s_{k-1}})$ and $Y = G(x_{t_0}, \ldots, x_{t_{m-1}})$ for some functions $F : \{0,1\}^k \to \mathbb{R}$ and $G : \{0,1\}^m \to \mathbb{R}$. Then X and Y are independent.

P

The notation in the lemma's statement is a bit cumbersome, but at the end of the day, it simply says that if X and Y are random variables that depend on two disjoint sets S and T of coins (for example, X might be the sum of the first $n/2$ coins, and Y might be the largest consecutive stretch of zeroes in the second $n/2$ coins), then they are independent.

Proof of Lemma 17.6. Let $a, b \in \mathbb{R}$, and let $A = \{x \in \{0,1\}^k : F(x) = a\}$ and $B = \{x \in \{0,1\}^m : F(x) = b\}$. Since S and T are disjoint, we can reorder the indices so that $S = \{0, \ldots, k-1\}$ and $T = \{k, \ldots, k+m-1\}$ without affecting any of the probabilities. Hence we can write $\Pr[X = a \wedge X = b] = |C|/2^n$ where $C = \{x_0, \ldots, x_{n-1} : (x_0, \ldots, x_{k-1}) \in A \wedge (x_k, \ldots, x_{k+m-1}) \in B\}$. Another way to write this using string concatenation is that $C = \{xyz : x \in A, y \in B, z \in \{0,1\}^{n-k-m}\}$, and hence $|C| = |A||B|2^{n-k-m}$, which means that

$$\frac{|C|}{2^n} = \frac{|A|}{2^k}\frac{|B|}{2^m}\frac{2^{n-k-m}}{2^{n-k-m}} = \Pr[X = a]\Pr[Y = b]. \tag{17.10}$$

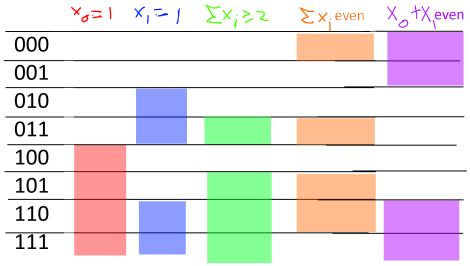

Figure 17.4: Consider the sample space $\{0,1\}^n$ and the events A, B, C, D, E corresponding to A: $x_0 = 1$, B: $x_1 = 1$, C: $x_0 + x_1 + x_2 \geq 2$, D: $x_0 + x_1 + x_2 = 0 mod 2$ and D: $x_0 + x_1 = 0 mod 2$. We can see that A and B are independent, C is positively correlated with A and positively correlated with B, the three events A, B, D are mutually independent, and while every pair out of A, B, E is independent, the three events A, B, E are not mutually independent since their intersection has probability $\frac{2}{8} = \frac{1}{4}$ instead of $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$.

[5] We use $\{X = u\}$ as shorthand for $\{x \mid X(x) = u\}$.

■

Note that if X and Y are independent random variables then (if we let S_X, S_Y denote all the numbers that have positive probability of being the output of X and Y, respectively) it holds that:

$$\mathbb{E}[XY] = \sum_{a\in S_X, b\in S_Y} \Pr[X=a \wedge Y=b]\cdot ab \stackrel{(1)}{=} \sum_{a\in S_X, b\in S_Y} \Pr[X=a]\Pr[Y=b]\cdot ab \stackrel{(2)}{=} \left(\sum_{a\in S_X} \Pr[X=a]\cdot a\right)\left(\sum_{b\in S_Y}\Pr[Y=b]b\right) \stackrel{(3)}{=} \mathbb{E}[X]\,\mathbb{E}[Y] \tag{17.11}$$

where the first equality ($=^{(1)}$) follows from the independence of X and Y, the second equality ($=^{(2)}$) follows by "opening the parentheses" of the righthand side, and the third inequality ($=^{(3)}$) follows from the definition of expectation. (This is not an "if and only if"; see Exercise 17.3.)

Another useful fact is that if X and Y are independent random variables, then so are $F(X)$ and $G(Y)$ for all functions $F, G : \mathbb{R} \to R$. This is intuitively true since learning $F(X)$ can only provide us with less information than does learning X itself. Hence, if learning X does not teach us anything about Y (and so also about $F(Y)$) then neither will learning $F(X)$. Indeed, to prove this we can write for every $a, b \in \mathbb{R}$:

$$\begin{aligned}\Pr[F(X)=a \wedge G(Y)=b] &= \sum_{x \text{ s.t.} F(x)=a, y \text{ s.t. } G(y)=b} \Pr[X=x\wedge Y=y] = \\ &\sum_{x \text{ s.t.} F(x)=a, y \text{ s.t. } G(y)=b} \Pr[X=x]\Pr[Y=y] = \\ &\left(\sum_{x \text{ s.t.} F(x)=a} \Pr[X=x]\right)\cdot\left(\sum_{y \text{ s.t.} G(y)=b}\Pr[Y=y]\right) = \\ &\Pr[F(X)=a]\Pr[G(Y)=b].\end{aligned} \tag{17.12}$$

17.2.2 Collections of independent random variables.

We can extend the notions of independence to more than two random variables: we say that the random variables $X_0, \ldots, X_{n-1}$ are *mutually independent* if for every $a_0, \ldots, a_{n-1} \in \mathbb{R}$,

$$\Pr\left[X_0 = a_0 \wedge \cdots \wedge X_{n-1} = a_{n-1}\right] = \Pr[X_0=a_0]\cdots\Pr[X_{n-1}=a_{n-1}]. \tag{17.13}$$

And similarly, we have that

Lemma 17.7 — Expectation of product of independent random variables. If $X_0, \ldots, X_{n-1}$ are mutually independent then

$$\mathbb{E}[\prod_{i=0}^{n-1} X_i] = \prod_{i=0}^{n-1} \mathbb{E}[X_i]. \tag{17.14}$$

Lemma 17.8 — Functions preserve independence. If $X_0, \ldots, X_{n-1}$ are mutually independent, and $Y_0, \ldots, Y_{n-1}$ are defined as $Y_i = F_i(X_i)$ for some functions $F_0, \ldots, F_{n-1} : \mathbb{R} \to \mathbb{R}$, then $Y_0, \ldots, Y_{n-1}$ are mutually independent as well.

P

We leave proving Lemma 17.7 and Lemma 17.8 as Exercise 17.4 Exercise 17.5. It is good idea for you stop now and do these exercises to make sure you are comfortable with the notion of independence, as we will use it heavily later on in this course.

17.3 CONCENTRATION

The name "expectation" is somewhat misleading. For example, suppose that you and I place a bet on the outcome of 10 coin tosses, where if they all come out to be 1's then I pay you 100,000 dollars and otherwise you pay me 10 dollars. If we let $X : \{0,1\}^{10} \to \mathbb{R}$ be the random variable denoting your gain, then we see that

$$\mathbb{E}[X] = 2^{-10} \cdot 100000 - (1 - 2^{-10})10 \sim 90. \tag{17.15}$$

But we don't really "expect" the result of this experiment to be for you to gain 90 dollars. Rather, 99.9% of the time you will pay me 10 dollars, and you will hit the jackpot 0.01% of the times.

However, if we repeat this experiment again and again (with fresh and hence *independent* coins), then in the long run we do expect your average earning to be 90 dollars, which is the reason why casinos can make money in a predictable way even though every individual bet is random. For example, if we toss n coins, then as n grows, the number of coins that come up ones will be more and more *concentrated* around $n/2$ according to the famous "bell curve" (see Fig. 17.5).

Much of probability theory is concerned with so called *concentration* or *tail* bounds, which are upper bounds on the probability that a random variable X deviates too much from its expectation. The first and simplest one of them is Markov's inequality:

Theorem 17.9 — Markov's inequality. If X is a non-negative random variable then $\Pr[X \geq k\,\mathbb{E}[X]] \leq 1/k$.

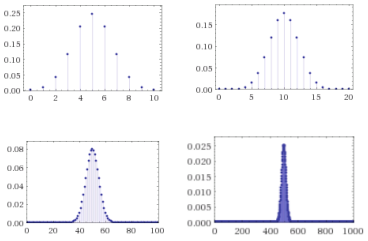

Figure 17.5: The probabilities that we obtain a particular sum when we toss $n = 10, 20, 100, 1000$ coins converge quickly to the Gaussian/normal distribution.

P

Markov's Inequality is actually a very natural statement (see also Fig. 17.6). For example, if you know that the average (not the median!) household income in the US is 70,000 dollars, then in particular you can deduce that at most 25 percent of households make more than 280,000 dollars, since otherwise, even if the remaining 75 percent had zero income, the top 25 percent alone would cause the average income to be larger than 70,000. From this example you can already see that in many situations, Markov's inequality will not be *tight* and the probability of deviating from expectation will be much smaller: see the Chebyshev and Chernoff inequalities below.

Proof of Theorem 17.9. Let $\mu = \mathbb{E}[X]$ and define $Y = 1_{X \geq k\mu}$. That is, $Y(x) = 1$ if $X(x) \geq k\mu$ and $Y(x) = 0$ otherwise. Note that by definition, for every x, $Y(x) \leq X/(k\mu)$. We need to show $\mathbb{E}[Y] \leq 1/k$. But this follows since $\mathbb{E}[Y] \leq \mathbb{E}[X/k(\mu)] = \mathbb{E}[X]/(k\mu) = \mu/(k\mu) = 1/k$. ■

Going beyond Markov's Inequality: Markov's inequality says that a (non-negative) random variable X can't go too crazy and be, say, a million times its expectation, with significant probability. But ideally we would like to say that with high probability, X should be very close to its expectation, e.g., in the range $[0.99\mu, 1.01\mu]$ where $\mu = \mathbb{E}[X]$. This is not generally true, but does turn out to hold when X is obtained by combining (e.g., adding) many independent random variables. This phenomenon, variants of which are known as "law of large numbers", "central limit theorem", "invariance principles" and "Chernoff bounds", is one of the most fundamental in probability and statistics, and is one that we heavily use in computer science as well.

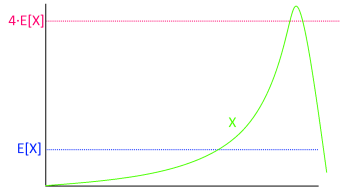

Figure 17.6: Markov's Inequality tells us that a non-negative random variable X cannot be much larger than its expectation, with high probability. For example, if the expectation of X is μ, then the probability that $X > 4\mu$ must be at most $1/4$, as otherwise just the contribution from this part of the sample space will be too large.

17.3.1 Chebyshev's Inequality

A standard way to measure the deviation of a random variable from its expectation is by using its *standard deviation*. For a random variable X, we define the *variance* of X as $\mathrm{Var}[X] = \mathbb{E}[(X - \mu)^2]$ where $\mu = \mathbb{E}[X]$; i.e., the variance is the average squared distance of X from its expectation. The *standard deviation* of X is defined as $\sigma[X] = \sqrt{\mathrm{Var}[X]}$. (This is well-defined since the variance, being an average of a square, is always a non-negative number.)

Using Chebyshev's inequality, we can control the probability that a random variable is too many standard deviations away from its expectation.

Theorem 17.10 — Chebyshev's inequality. Suppose that $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$. Then for every $k > 0$, $\Pr[|X - \mu| \geq k\sigma] \leq 1/k^2$.

Proof. The proof follows from Markov's inequality. We define the random variable $Y = (X - \mu)^2$. Then $\mathbb{E}[Y] = \mathrm{Var}[X] = \sigma^2$, and hence by Markov the probability that $Y > k^2\sigma^2$ is at most $1/k^2$. But clearly $(X - \mu)^2 \geq k^2\sigma^2$ if and only if $|X - \mu| \geq k\sigma$. ■

One example of how to use Chebyshev's inequality is the setting when $X = X_1 + \cdots + X_n$ where X_i's are *independent and identically distributed* (i.i.d for short) variables with values in $[0, 1]$ where each has expectation $1/2$. Since $\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] = n/2$, we would like to say that X is very likely to be in, say, the interval $[0.499n, 0.501n]$. Using Markov's inequality directly will not help us, since it will only tell us that X is very likely to be at most $100n$ (which we already knew, since it always lies between 0 and n). However, since $X_1, \ldots, X_n$ are independent,

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] \,. \tag{17.16}$$

(We leave showing this to the reader as Exercise 17.6.)

For every random variable X_i in $[0, 1]$, $\mathrm{Var}[X_i] \leq 1$ (if the variable is always in $[0, 1]$, it can't be more than 1 away from its expectation), and hence (17.16) implies that $\mathrm{Var}[X] \leq n$ and hence $\sigma[X] \leq \sqrt{n}$. For large n, $\sqrt{n} \ll 0.001n$, and in particular if $\sqrt{n} \leq 0.001n/k$, we can use Chebyshev's inequality to bound the probability that X is not in $[0.499n, 0.501n]$ by $1/k^2$.

17.3.2 The Chernoff bound

Chebyshev's inequality already shows a connection between independence and concentration, but in many cases we can hope for a quantitatively much stronger result. If, as in the example above, $X = X_1 + \ldots + X_n$ where the X_i's are bounded i.i.d random variables of mean $1/2$, then as n grows, the distribution of X would be roughly the *normal* or *Gaussian* distribution— that is, distributed according to the *bell curve* (see Fig. 17.5 and Fig. 17.7). This distribution has the property of being *very* concentrated in the sense that the probability of deviating k standard deviations from the mean is not merely $1/k^2$ as is guaranteed by Chebyshev, but rather is roughly e^{-k^2}.[6] That is, we have an *exponential decay* of the probability of deviation.

The following extremely useful theorem shows that such exponential decay occurs every time we have a sum of independent and bounded variables. This theorem is known under many names in different communities, though it is mostly called the Chernoff bound in the computer science literature:

[6] Specifically, for a normal random variable X of expectation μ and standard deviation σ, the probability that $|X - \mu| \geq k\sigma$ is at most $2e^{-k^2/2}$.

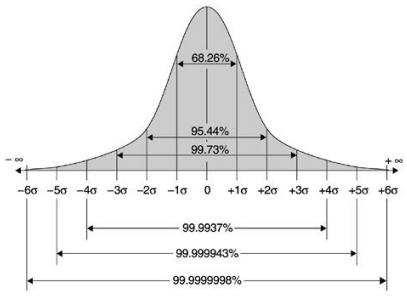

Figure 17.7: In the *normal distribution* or the Bell curve, the probability of deviating k standard deviations from the expectation shrinks *exponentially* in k^2, and specifically with probability at least $1 - 2e^{-k^2/2}$, a random variable X of expectation μ and standard deviation σ satisfies $\mu - k\sigma \leq X \leq \mu + k\sigma$. This figure gives more precise bounds for $k = 1, 2, 3, 4, 5, 6$. (Image credit:Imran Baghirov)

Theorem 17.11 — Chernoff/Hoeffding bound. If $X_1, \ldots, X_n$ are i.i.d random variables such that $X_i \in [0,1]$ and $\mathbb{E}[X_i] = p$ for every i, then for every $\epsilon > 0$

$$\Pr[\left|\sum_{i=0}^{n-1} X_i - pn\right| > \epsilon n] \leq 2 \cdot e^{-2\epsilon^2 n}. \tag{17.17}$$

We omit the proof, which appears in many texts, and uses Markov's inequality on i.i.d random variables $Y_0, \ldots, Y_n$ that are of the form $Y_i = e^{\lambda X_i}$ for some carefully chosen parameter λ. See Exercise 17.9 for a proof of the simple (but highly useful and representative) case where each X_i is $\{0,1\}$ valued and $p = 1/2$. (See also Exercise 17.10 for a generalization.)

[7]

[7] TODO: maybe add an example application of Chernoff. Perhaps a probabilistic method proof using Chernoff+Union bound.

17.4 LECTURE SUMMARY

- A basic probabilistic experiment corresponds to tossing n coins or choosing x uniformly at random from $\{0,1\}^n$.
- *Random variables* assign a real number to every result of a coin toss. The *expectation* of a random variable X is its average value, and there are several *concentration* results showing that under certain conditions, random variables deviate significantly from their expectation only with small probability.

17.5 EXERCISES

R

Remark 17.12 — Disclaimer. Most of the exercises have been written in the summer of 2018 and haven't yet been fully debugged. While I would prefer people do not post online solutions to the exercises, I would greatly appreciate if you let me know of any bugs. You can do so by posting a GitHub issue about the exercise, and optionally complement this with an email to me with more details about the attempted solution.

Exercise 17.1 Suppose that we toss three independent fair coins $a, b, c \in \{0,1\}$. What is the probability that the XOR of a,b, and c is equal to 1? What is the probability that the AND of these three values is equal to 1? Are these two events independent?

■

Exercise 17.2 Give an example of random variables $X, Y : \{0,1\}^3 \to \mathbb{R}$ such that $\mathbb{E}[XY] \neq \mathbb{E}[X]\,\mathbb{E}[Y]$.

■

Exercise 17.3 Give an example of random variables $X, Y : \{0,1\}^3 \to \mathbb{R}$ such that X and Y are *not* independent but $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

■

Exercise 17.4 — Product of expectations. Prove Lemma 17.7

■

Exercise 17.5 — Transformations preserve independence. Prove Lemma 17.8

■

Exercise 17.6 — Variance of independent random variables. Prove that if $X_0, \ldots, X_{n-1}$ are independent random variables then $\mathrm{Var}[X_0 + \cdots + X_{n-1}] = \sum_{i=0}^{n-1} \mathrm{Var}[X_i]$.

■

Exercise 17.7 — Entropy (challenge). Recall the definition of a distribution μ over some finite set S. Shannon defined the *entropy* of a distribution μ, denoted by $H(\mu)$, to be $\sum_{x \in S} \mu(x) \log(1/\mu(x))$. The idea is that if μ is a distribution of entropy k, then encoding members of μ will require k bits, in an amortized sense. In this exercise we justify this definition. Let μ be such that $H(\mu) = k$.

1. Prove that for every one to one function $F : S \to \{0,1\}^*$, $\mathbb{E}_{x \sim \mu} |F(x)| \geq k$.
2. Prove that for every ϵ, there is some n and a one-to-one function $F : S^n \to \{0,1\}^*$, such that $\mathbb{E}_{x \sim \mu^n} |F(x)| \leq n(k+\epsilon)$, where $x \sim \mu$ denotes the experiments of choosing $x_0, \ldots, x_{n-1}$ each independently from S using the distribution μ.

■

Exercise 17.8 — Entropy approximation to binomial. Let $H(p) = p \log(1/p) + (1-p)\log(1/(1-p))$.[8] Prove that for every $p \in (0,1)$ and $\epsilon > 0$, if n is large enough then[9]

$$2^{(H(p)-\epsilon)n} \binom{n}{pn} \leq 2^{(H(p)+\epsilon)n} \tag{17.18}$$

where $\binom{n}{k}$ is the binomial coefficient $\frac{n!}{k!(n-k)!}$ which is equal to the number of k-size subsets of $\{0, \ldots, n-1\}$.

■

Exercise 17.9 — Chernoff using Stirling.

1. Prove that $\Pr_{x \sim \{0,1\}^n}[\sum x_i = k] = \binom{n}{k} 2^{-n}$.
2. Use this and Exercise 17.8 to prove the Chernoff bound for the case that $X_0, \ldots, X_n$ are i.i.d. random variables over $\{0,1\}$ each equaling 0 and 1 with probability $1/2$.

■

Exercise 17.10 — Poor man's Chernoff. Let $X_0, \ldots, X_n$ be i.i.d random variables with $\mathbb{E}\, X_i = p$ and $\Pr[0 \leq X_i \leq 1] = 1$. Define $Y_i = X_i - p$.

[8] While you don't need this to solve this exercise, this is the function that maps p to the entropy (as defined in Exercise 17.7) of the p-biased coin distribution over $\{0,1\}$, which is the function $\mu : \{0,1\} \to [0,1]$ s.y. $\mu(0) = 1-p$ and $\mu(1) = p$.

[9] **Hint:** Use Stirling's formula for approximating the factorial function.

1. Prove that for every $j_1, \ldots, j_n \in \mathbb{N}$, if there exists one i such that j_i is odd then $\mathbb{E}[\prod_{i=0}^{n-1} Y_i^{j_i}] = 0$.
2. Prove that for every k, $\mathbb{E}[(\sum_{i=0}^{n-1} Y_i)^k] \leq (10kn)^{k/2}$.[10]
3. Prove that for every $\epsilon > 0$, $\Pr[|\sum_i Y_i| \geq \epsilon n] \geq 2^{-\epsilon^2 n/(10000 \log 1/\epsilon)}$.[11]

■

Exercise 17.11 — Simulating distributions using coins. Our model for probability involves tossing n coins, but sometimes algorithms require sampling from other distributions, such as selecting a uniform number in $\{0, \ldots, M-1\}$ for some M. Fortunately, we can simulate this with an exponentially small probability of error: prove that for every M, if $n > k\lceil \log M \rceil$, then there is a function $F : \{0,1\}^n \to \{0, \ldots, M-1\} \cup \{\perp\}$ such that **(1)** The probability that $F(x) = \perp$ is at most 2^{-k} and **(2)** the distribution of $F(x)$ conditioned on $F(x) \neq \perp$ is equal to the uniform distribution over $\{0, \ldots, M-1\}$.[12]

■

Exercise 17.12 — Sampling. Suppose that a country has 300,000,000 citizens, 52 percent of which prefer the color "green" and 48 percent of which prefer the color "orange". Suppose we sample n random citizens and ask them their favorite color (assume they will answer truthfully). What is the smallest value n among the following choices so that the probability that the majority of the sample answers "green" is at most 0.05?

a. 1,000

b. 10,000

c. 100,000

d. 1,000,000

■

Exercise 17.13 Would the answer to Exercise 17.12 change if the country had 300,000,000,000 citizens?

■

Exercise 17.14 — Sampling (2). Under the same assumptions as Exercise 17.12, what is the smallest value n among the following choices so that the probability that the majority of the sample answers "green" is at most 2^{-100}?

a. 1,000

b. 10,000

c. 100,000

d. 1,000,000

[10] **Hint:** Bound the number of tuples $j_0, \ldots, j_{n-1}$ such that every j_i is even and $\sum j_i = k$.

[11] **Hint:** Set $k = 2\lceil \epsilon^2 n/1000 \rceil$ and then show that if the event $|\sum Y_i| \geq \epsilon n$ happens then the random variable $(\sum Y_i)^k$ is a factor of ϵ^{-k} larger than its expectation.

[12] **Hint:** Think of $x \in \{0,1\}^n$ as choosing k numbers $y_1, \ldots, y_k \in \{0, \ldots, 2^{\lceil \log M \rceil} - 1\}$. Output the first such number that is in $\{0, \ldots, M-1\}$.

e. It is impossible to get such low probability since there are fewer than 2^{100} citizens.

13

■

[13] TODO: add some exercise about the probabilistic method

17.6 BIBLIOGRAPHICAL NOTES