Probability theory 101

Boaz Barak

**Disclaimer:** The MS-Word version of this text has significant formatting issues compared to the PDF and HTML versions.

# Probability Theory 101

### 

* Review the basic notion of probability theory that we will use.
* Sample spaces, and in particular the space
* 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).

*“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

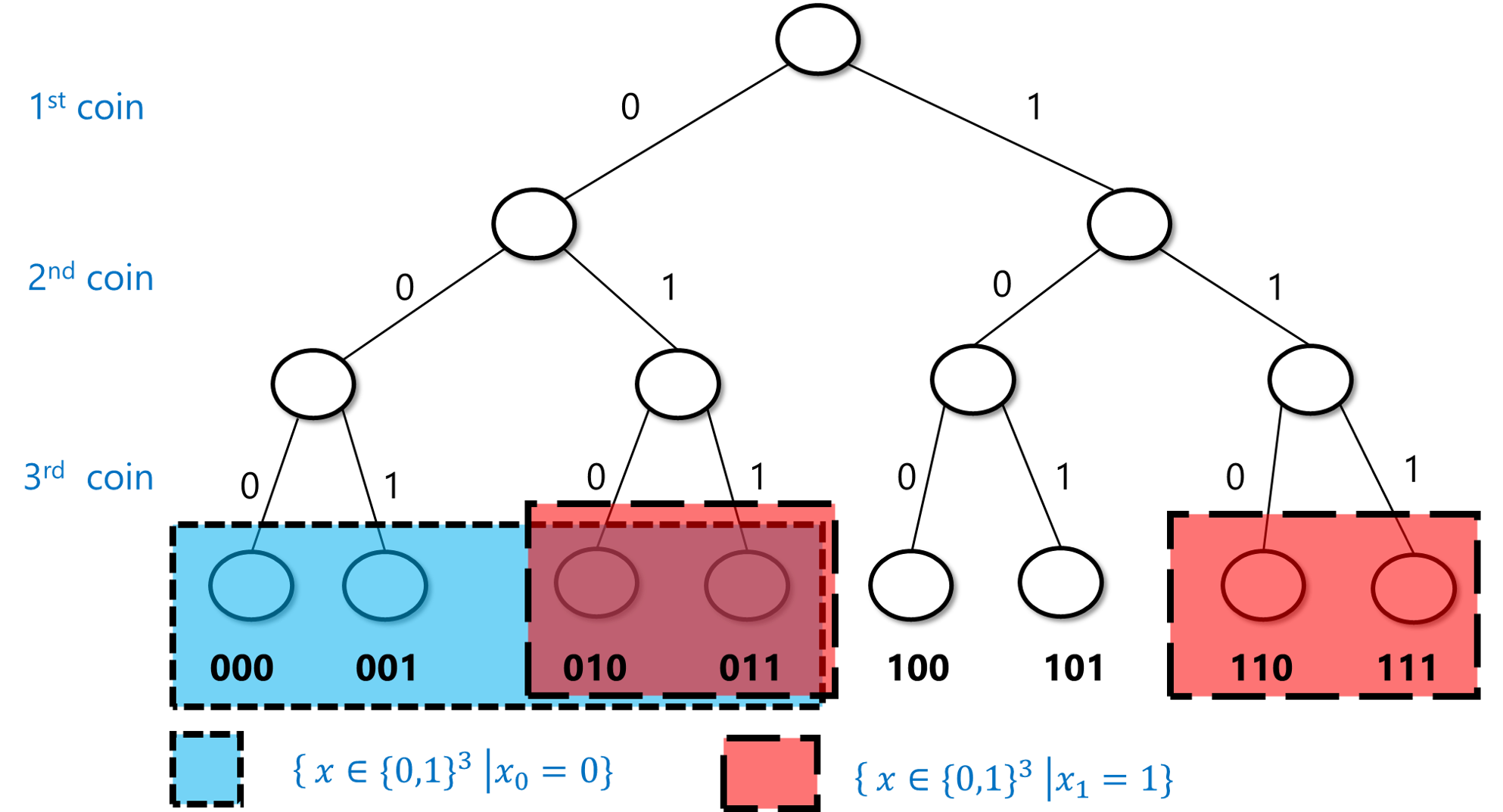
Before we show how to use randomness in algorithms, let us do a quick review of some basic notions in probability theory. This is not meant to replace a course on probability theory, and if you have not seen this material before, I highly recommend you look at additional resources to get up to speed. 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 coins can lead to very subtle and interesting issues.

## Random coins

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

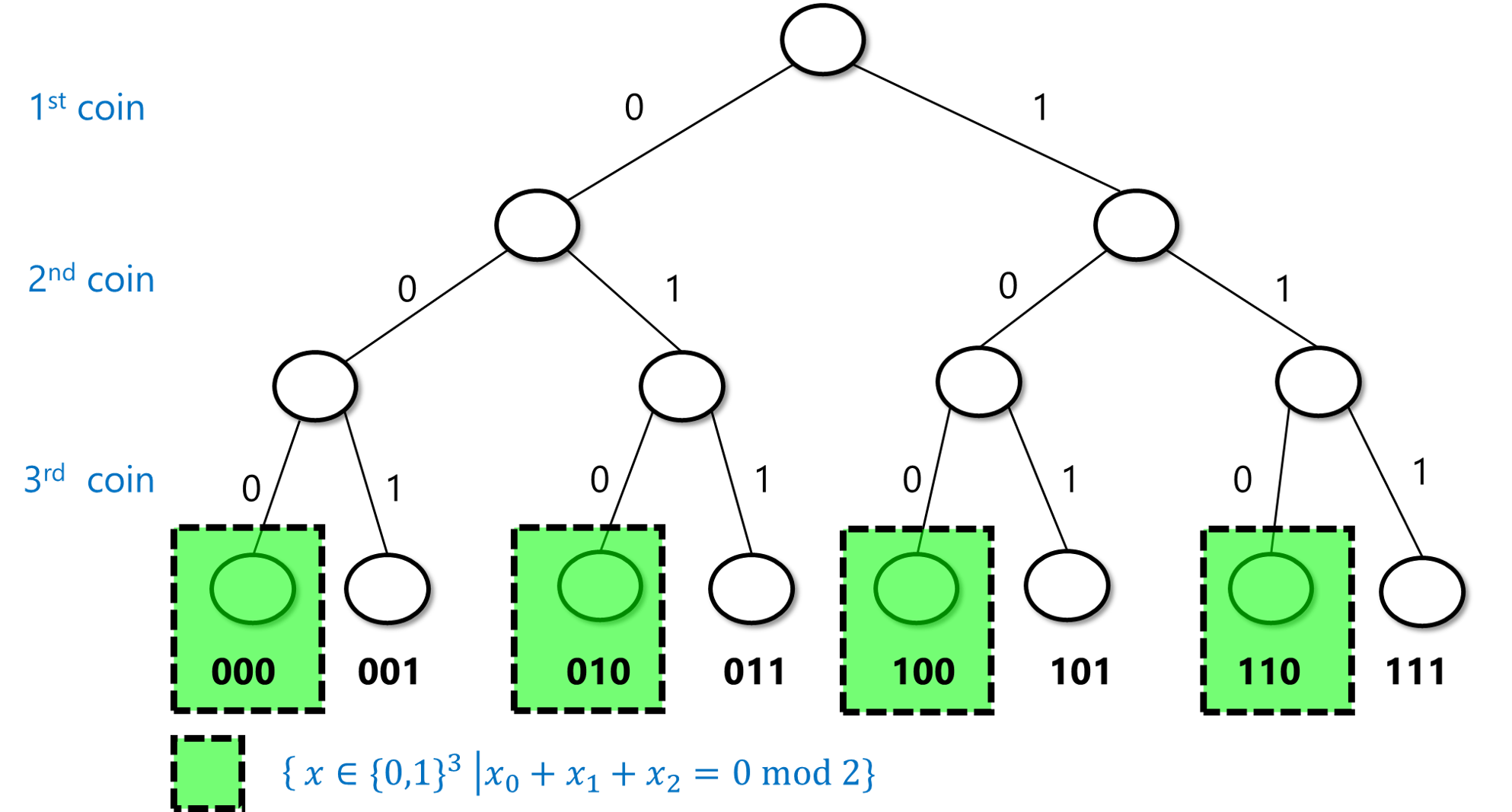
These are all important questions that have been studied and debated by scientists, mathematicians, statisticians and philosophers. Fortunately, we will not need to deal directly with these questions here. We will be mostly interested in the setting of tossing random, unbiased and independent coins. 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 -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 coins as a string in . Each particular outcome is obtained with probability . For example, if we toss three coins, then we obtain each of the 8 outcomes with probability (see also coinexperimentfig). We can describe the experiment of tossing coins as choosing a string uniformly at random from , and hence we’ll use the shorthand for that is chosen according to this experiment.



The probabilistic experiment of tossing three coins corresponds to making choices, each with equal probability. In this example, the blue set corresponds to the event where the first coin toss is equal to , and the pink set corresponds to the event where the second coin toss is equal to (with their intersection having a purplish color). As we can see, each of these events contains elements (out of total) and so has probability . The intersection of and contains two elements, and so the probability that both of these events occur is .

An *event* is simply a subset of . The *probability of* , denoted by (or for short, when the sample space is understood from the context), is the probability that an chosen uniformly at random will be contained in . Note that this is the same as (where as usual denotes the number of elements in the set ). For example, the probability that has an even number of ones is where . In the case , , and hence (see eventhreecoinsfig). It turns out this is true for every :



The event that if we toss three coins then the sum of the ’s is even has probability since it corresponds to exactly out of the possible strings of length .

### 

For every ,

### 

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

We prove the lemma by induction on . For the case it is clear since is even and is odd, and hence the probability that is even is . Let . We assume by induction that the lemma is true for and we will prove it for . We split the set into four disjoint sets , where for , is defined as the set of such that has even number of ones and and similarly is the set of such that has odd number of ones and . Since is obtained by simply extending -length string with even number of ones by the digit , the size of is simply the number of such -length strings which by the induction hypothesis is . The same reasoning applies for , , and . Hence each one of the four sets is of size . Since has an even number of ones if and only if (i.e., either the first coordinates sum up to an even number and the final coordinate is or the first coordinates sum up to an odd number and the final coordinate is ), we get that the probability that satisfies this property is

using the fact that and are disjoint and hence .

We can also use the *intersection* () and *union* () operators to talk about the probability of both event *and* event happening, or the probability of event *or* event happening. For example, the probability that has an *even* number of ones *and* is the same as where and . This probability is equal to for . (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 and operators instead of and , and so write this probability defined above also as

If is an event, then corresponds to the event that does *not* happen. Since , we get that

This makes sense: since happens if and only if does *not* happen, the probability of should be one minus the probability of .

### 

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 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 that there exists *some* row where the performance is higher than the overall average. is a superset of , and its probability (even if there is no correlation between sitting and performance) can be quite significant.

### 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 a function that maps every outcome to an element . For example, the function that maps to the sum of its coordinates (i.e., to ) is a random variable.

The *expectation* of a random variable , denoted by , is the average value that that this number takes, taken over all draws from the probabilistic experiment. In other words, the expectation of is defined as follows:

If and are random variables, then we can define as simply the random variable that maps a point to . One basic and very useful property of the expectation is that it is *linear*:

### 

### 

Similarly, for every .

Let be the random variable that maps to . Prove that .

We can solve this using the linearity of expectation. We can define random variables such that . Since each equals with probability and with probability , . Since , by the linearity of expectation

### 

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 is an event, then is the random variable such that equals if , and otherwise. Note that (can you see why?). Using this and the linearity of expectation, we can show one of the most useful bounds in probability theory:

### 

For every two events ,

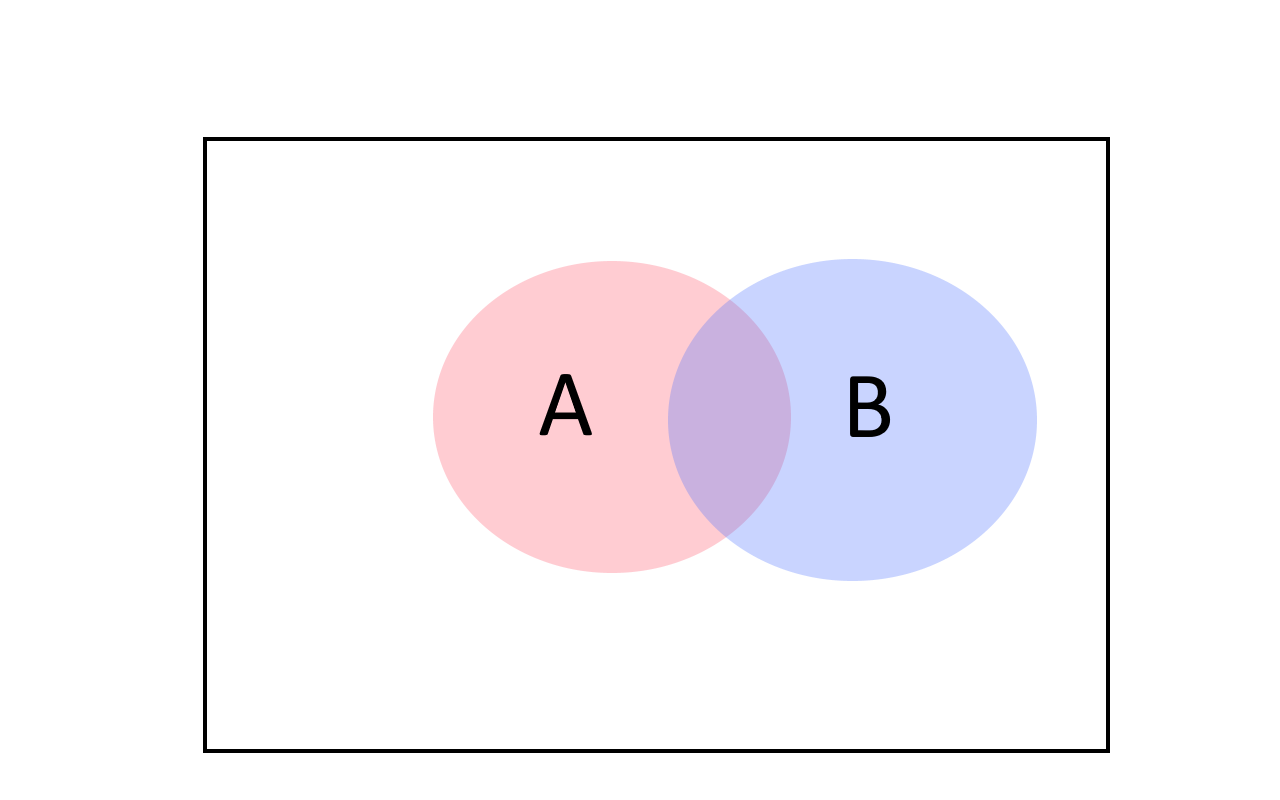
### 

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 . Can you see why the latter equation is true? (See also unionboundfig.)

### 

For every , the variable . Hence, .

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 , then with probability at least , no bad event happens.



The *union bound* tells us that the probability of or happening is at most the sum of the individual probabilities. We can see it by noting that for every two sets (with equality only if and have no intersection).

### 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 and consider the “random variable” such that for every , the probability that outputs is equal to . To avoid confusion, we will typically refer to such string-valued random variables as *distributions* over strings. So, a *distribution* over strings can be thought of as a finite collection of strings and probabilities (which are non-negative numbers summing up to one), so that .

Two distributions and are *identical* if they assign the same probability to every string. For example, consider the following two functions . For every , we define and where is the XOR operations. Although these are two different functions, they induce the same distribution over when invoked on a uniform input. The distribution for is of course the uniform distribution over . On the other hand is simply the map , , , which is a permutation over the map defined as and the map defined as

### More general sample spaces.

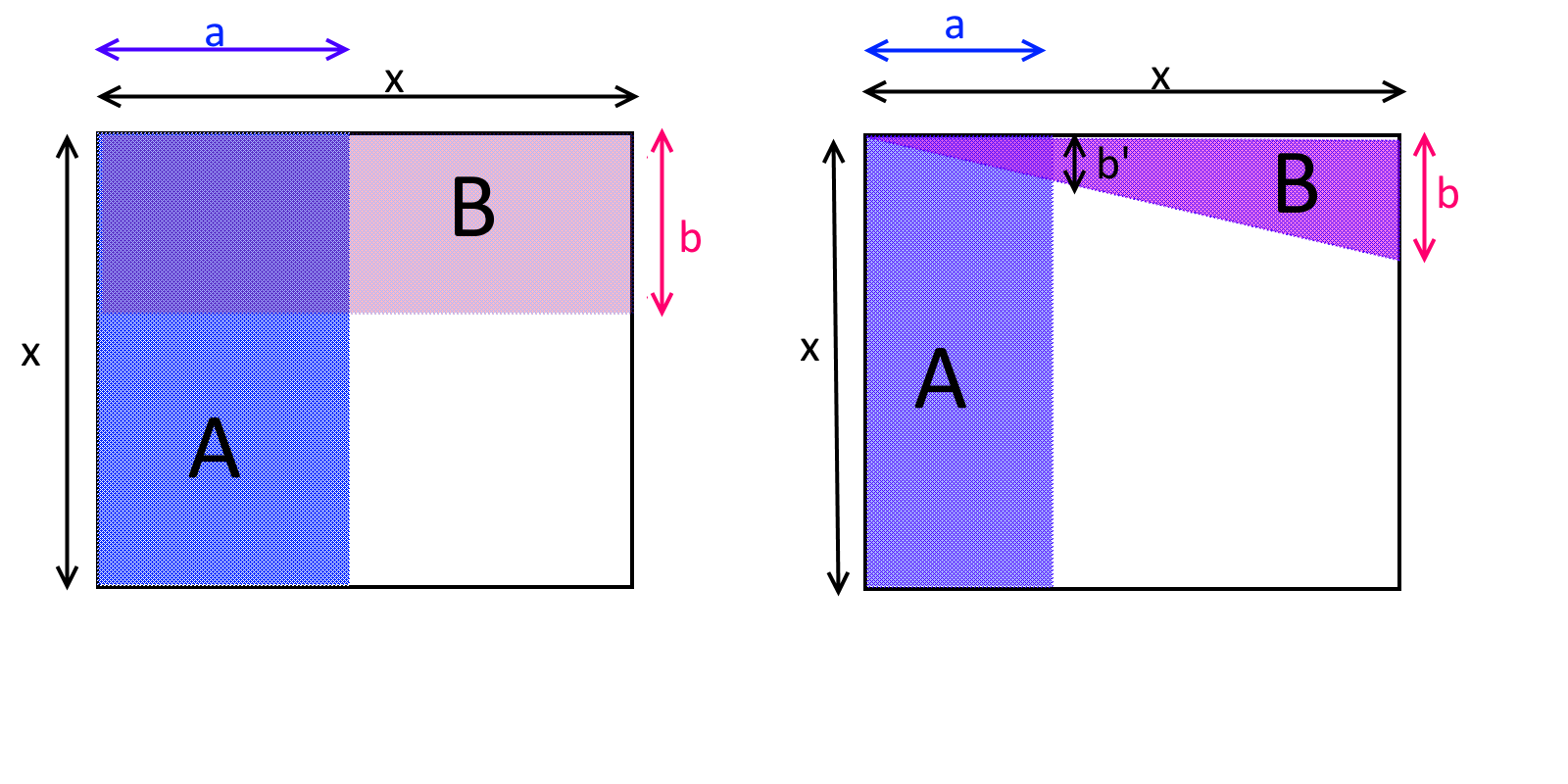
While in this chapter we assume that the underlying probabilistic experiment corresponds to tossing independent coins, everything we say easily generalizes to sampling from a more general finite or countable set (and not-so-easily generalizes to uncountable sets as well). A *probability distribution* over a finite set is simply a function such that . We think of this as the experiment where we obtain every with probability , and sometimes denote this as . An *event* is a subset of , and the probability of , which we denote by , is . A *random variable* is a function , where the probability that is equal to .

## 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”](https://youtu.be/dzFf3r1yph8). (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 coins.)

Two events and are *independent* if the fact that happens makes neither more nor less likely to happen. For example, if we think of the experiment of tossing random coins , and we let be the event that and the event that , then if happens it is more likely that happens, and hence these events are *not* independent. On the other hand, if we let be the event that , then because the second coin toss is not affected by the result of the first one, the events and are independent.

The formal definition is that events and are *independent* if . If then we say that and are *positively correlated*, while if then we say that and are *negatively correlated* (see coinexperimentfig).



Two events and are *independent* if . In the two figures above, the empty square is the sample space, and and are two events in this sample space. In the left figure, and are independent, while in the right figure they are negatively correlated, since is less likely to occur if we condition on (and vice versa). Mathematically, one can see this by noticing that in the left figure the areas of and respectively are and , and so their probabilities are and respectively, while the area of is which corresponds to the probability . In the right figure, the area of the triangle is which corresponds to a probability of , but the area of is for some . This means that the probability of is , or in other words .

If we consider the above examples on the experiment of choosing then we can see that

but

and hence, as we already observed, the events and are not independent and in fact are positively correlated. On the other hand, and hence the events and are indeed independent.

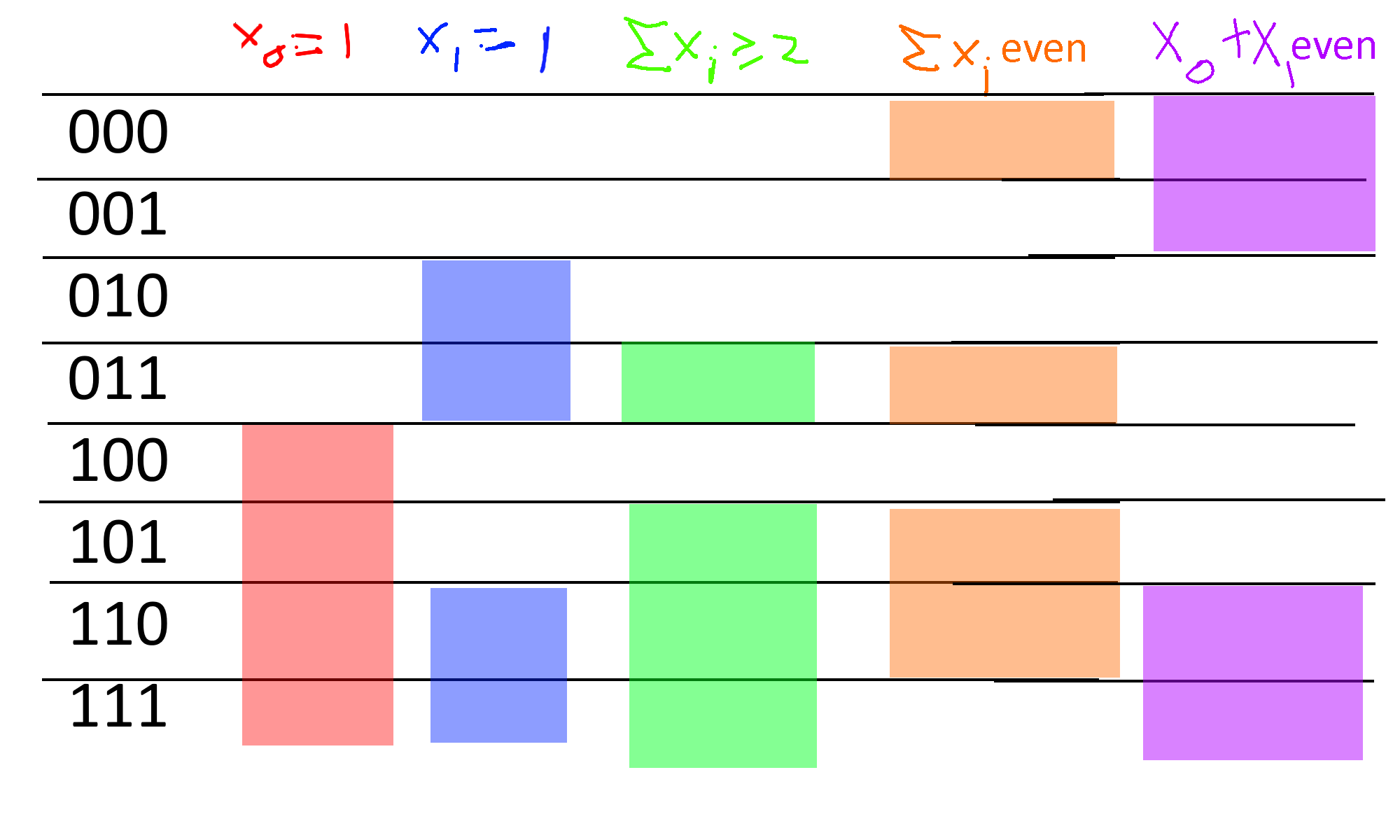
### 

People sometimes confuse the notion of *disjointness* and *independence*, but these are actually quite different. Two events and are *disjoint* if , which means that if happens then definitely does not happen. They are *independent* if which means that knowing that happens gives us no information about whether happened or not. If and 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 and are events, and happens with nonzero probability then we define the probability that happens *conditioned on*  to be . This corresponds to calculating the probability that happens if we already know that happened. Note that and are independent if and only if .

**More than two events:** We can generalize this definition to more than two events. We say that events 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 ,

For example, if , then the events , and are mutually independent. On the other hand, the events , and are *not* mutually independent, even though every pair of these events is independent (can you see why? see also independencecoinsfig).



Consider the sample space and the events corresponding to : , : , : , : and : . We can see that and are independent, is positively correlated with and positively correlated with , the three events are mutually independent, and while every pair out of is independent, the three events are not mutually independent since their intersection has probability instead of .

### Independent random variables

We say that two random variables and are independent if for every , the events and are independent. (We use as shorthand for .) In other words, and are independent if for every . For example, if two random variables depend on the result of tossing different coins then they are independent:

### 

Suppose that and are disjoint subsets of and let be random variables such that and for some functions and . Then and are independent.

### 

The notation in the lemma’s statement is a bit cumbersome, but at the end of the day, it simply says that if and are random variables that depend on two disjoint sets and of coins (for example, might be the sum of the first coins, and might be the largest consecutive stretch of zeroes in the second coins), then they are independent.

### 

Let , and let and . Since and are disjoint, we can reorder the indices so that and without affecting any of the probabilities. Hence we can write where . Another way to write this using string concatenation is that , and hence , which means that

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

$$
\begin{gathered}
\mathbb{E}[ XY ] = \sum\_{a \in S\_X,b \in S\_Y} {\textstyle\Pr[X=a \wedge Y=b]}\cdot ab \; =^{(1)} \; \sum\_{a \in S\_X,b \in S\_Y} {\textstyle \Pr[X=a]\Pr[Y=b]}\cdot ab =^{(2)} \\
\left(\sum\_{a \in S\_X} {\textstyle \Pr[X=a]}\cdot a\right)\left(\sum\_{b \in S\_Y} {\textstyle \Pr[Y=b]}b\right) =^{(3)} \\
\mathbb{E}[X] \mathbb{E}[Y]
\end{gathered}
$$

where the first equality () follows from the independence of and , the second equality () follows by “opening the parentheses” of the righthand side, and the third inequality () follows from the definition of expectation. (This is not an “if and only if”; see noindnocorex.)

Another useful fact is that if and are independent random variables, then so are and for all functions . This is intuitively true since learning can only provide us with less information than does learning itself. Hence, if learning does not teach us anything about (and so also about ) then neither will learning . Indeed, to prove this we can write for every :

### Collections of independent random variables.

We can extend the notions of independence to more than two random variables: we say that the random variables are *mutually independent* if for every ,

And similarly, we have that

### 

If are mutually independent then

### 

If are mutually independent, and are defined as for some functions , then are mutually independent as well.

### 

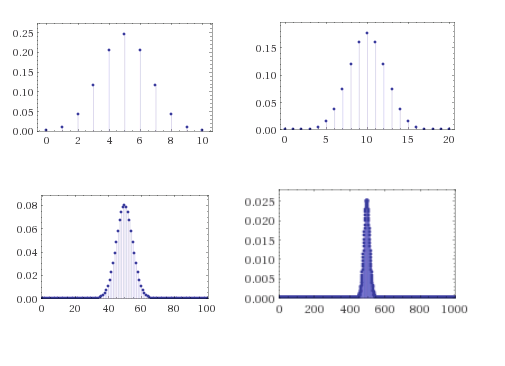
We leave proving expprod and indeplem as expprodex indeplemex. It is good idea for you stop now and do these exercises to make sure you are comfortable with the notion of independence, as we will use it heavily later on in this course.

## Concentration and tail bounds

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 ’s then I pay you 100,000 dollars and otherwise you pay me 10 dollars. If we let be the random variable denoting your gain, then we see that

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 close to 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 independent and unbiased coins, then as grows, the number of coins that come up ones will be more and more *concentrated* around according to the famous “bell curve” (see bellfig).



The probabilities that we obtain a particular sum when we toss coins converge quickly to the Gaussian/normal distribution.

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

### 

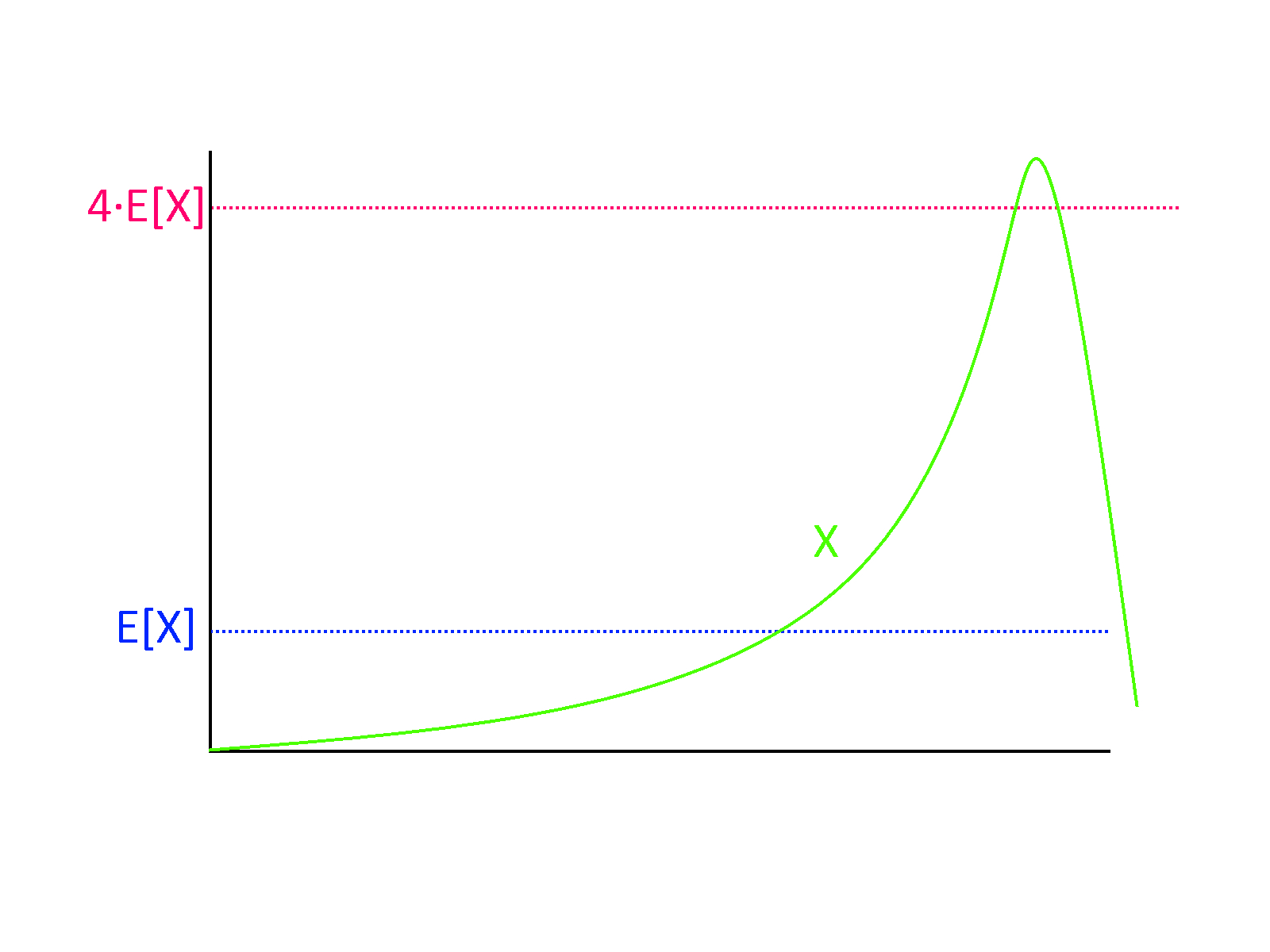
If is a non-negative random variable then for every , .

### 

Markov’s Inequality is actually a very natural statement (see also markovfig). 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 dollars. 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.

### 

Let and define . That is, if and otherwise. Note that by definition, for every , . We need to show . But this follows since .



Markov’s Inequality tells us that a non-negative random variable cannot be much larger than its expectation, with high probability. For example, if the expectation of is , then the probability that must be at most , as otherwise just the contribution from this part of the sample space will be too large.

**The averaging principle.** While the expectation of a random variable is hardly always the “typical value”, we can show that is guaranteed to achieve a value that is at least its expectation with positive probability. For example, if the average grade in an exam is points, at least one student got a grade or more on the exam. This is known as the *averaging principle*, and despite its simplicity it is surprisingly useful.

### 

Let be a random variable, then .

Suppose towards the sake of contradiction that . Then the random variable is always positive. By linearity of expectation . Yet by Markov, a non-negative random variable with must equal with probability , since the probability that is at most for every . Hence we get a contradiction to the assumption that is always positive.

### Chebyshev’s Inequality

Markov’s inequality says that a (non-negative) random variable 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, should be very close to its expectation, e.g., in the range where . In such a case we say that is *concentrated*, and hence its expectation (i.e., mean) will be close to its *median* and other ways of measuring ’s “typical value”. *Chebyshev’s inequality* can be thought of as saying that is concentrated if it has a small *standard deviation*.

A standard way to measure the deviation of a random variable from its expectation is by using its *standard deviation*. For a random variable , we define the *variance* of as where ; i.e., the variance is the average squared distance of from its expectation. The *standard deviation* of is defined as . (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.

### 

Suppose that and . Then for every , .

### 

The proof follows from Markov’s inequality. We define the random variable . Then , and hence by Markov the probability that is at most . But clearly if and only if .

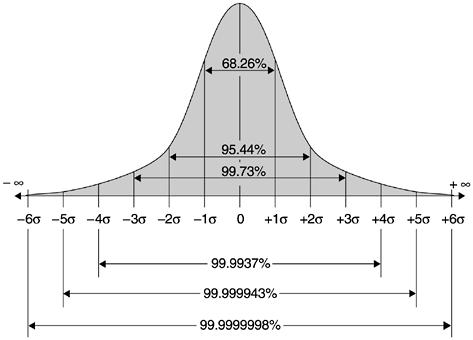
One example of how to use Chebyshev’s inequality is the setting when where ’s are *independent and identically distributed* (i.i.d for short) variables with values in where each has expectation . Since , we would like to say that is very likely to be in, say, the interval . Using Markov’s inequality directly will not help us, since it will only tell us that is very likely to be at most (which we already knew, since it always lies between and ). However, since are independent,

(We leave showing this to the reader as varianceex.)

For every random variable in , (if the variable is always in , it can’t be more than away from its expectation), and hence varianceeq implies that and hence . For large , , and in particular if , we can use Chebyshev’s inequality to bound the probability that is not in by .

### 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, where the ’s are bounded i.i.d random variables of mean , then as grows, the distribution of would be roughly the *normal* or *Gaussian* distribution that is, distributed according to the *bell curve* (see bellfig and empiricalbellfig). This distribution has the property of being *very* concentrated in the sense that the probability of deviating standard deviations from the mean is not merely as is guaranteed by Chebyshev, but rather is roughly . Specifically, for a normal random variable of expectation and standard deviation , the probability that is at most . That is, we have an *exponential decay* of the probability of deviation.



In the *normal distribution* or the Bell curve, the probability of deviating standard deviations from the expectation shrinks *exponentially* in , and specifically with probability at least , a random variable of expectation and standard deviation satisfies . This figure gives more precise bounds for . (Image credit:Imran Baghirov)

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](https://en.wikipedia.org/wiki/Chernoff_bound) in the computer science literature:

### 

If are i.i.d random variables such that and for every , then for every

We omit the proof, which appears in many texts, and uses Markov’s inequality on i.i.d random variables that are of the form for some carefully chosen parameter . See chernoffstirlingex for a proof of the simple (but highly useful and representative) case where each is valued and . (See also poorchernoff for a generalization.)

* A basic probabilistic experiment corresponds to tossing coins or choosing uniformly at random from .
* *Random variables* assign a real number to every result of a coin toss. The *expectation* of a random variable is its average value.
* There are several *concentration* results, also known as *tail bounds* showing that under certain conditions, random variables deviate significantly from their expectation only with small probability.

## Exercises

### 

Suppose that we toss three independent fair coins . What is the probability that the XOR of ,, and is equal to ? What is the probability that the AND of these three values is equal to ? Are these two events independent?

### 

Give an example of random variables such that .

### 

Give an example of random variables such that and are *not* independent but .

Let be an odd number, and let be the random variable defined as follows: for every , if and otherwise. Prove that .

### 

Prove expprod

### 

Prove indeplem

### 

Prove that if are independent random variables then .

### 

Recall the definition of a distribution over some finite set . Shannon defined the *entropy* of a distribution , denoted by , to be . The idea is that if is a distribution of entropy , then encoding members of will require bits, in an amortized sense. In this exercise we justify this definition. Let be such that .  
1. Prove that for every one to one function , .  
2. Prove that for every , there is some and a one-to-one function , such that , where denotes the experiments of choosing each independently from using the distribution .

### 

Let .[[1]](#footnote-85) Prove that for every and , if is large enough then[[2]](#footnote-86)

where is the binomial coefficient which is equal to the number of -size subsets of .

### 

1. Prove that .
2. Use this and entropybinomex to prove the Chernoff bound for the case that are i.i.d. random variables over each equaling and with probability .

### 

Let be i.i.d random variables with and . Define .  
1. Prove that for every , if there exists one such that is odd then .  
2. Prove that for every , .[[3]](#footnote-89)  
3. Prove that for every , .[[4]](#footnote-90)

### 

Our model for probability involves tossing coins, but sometimes algorithms require sampling from other distributions, such as selecting a uniform number in for some . Fortunately, we can simulate this with an exponentially small probability of error: prove that for every , if , then there is a function such that **(1)** The probability that is at most and **(2)** the distribution of conditioned on is equal to the uniform distribution over .[[5]](#footnote-92)

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 random citizens and ask them their favorite color (assume they will answer truthfully). What is the smallest value among the following choices so that the probability that the majority of the sample answers “green” is at most ?

1. 1,000
2. 10,000
3. 100,000
4. 1,000,000

### 

Would the answer to samplingex change if the country had 300,000,000,000 citizens?

Under the same assumptions as samplingex, what is the smallest value among the following choices so that the probability that the majority of the sample answers “green” is at most ?

1. 1,000
2. 10,000
3. 100,000
4. 1,000,000
5. It is impossible to get such low probability since there are fewer than citizens.

## Bibliographical notes

There are many sources for more information on discrete probability, including the texts referenced in notesmathchap. One particularly recommended source for probability is [Harvard’s [STAT 110](https://projects.iq.harvard.edu/stat110/home) class, whose lectures are available on [youtube](https://projects.iq.harvard.edu/stat110/youtube) and whose book is available [online](http://probabilitybook.net).

The version of the Chernoff bound that we stated in chernoffthm is sometimes known as [Hoeffding’s Inequality](https://en.wikipedia.org/wiki/Hoeffding%27s_inequality). Other variants of the Chernoff bound are known as well, but all of them are equally good for the applications of this book.

1. While you don’t need this to solve this exercise, this is the function that maps to the entropy (as defined in entropyex) of the -biased coin distribution over , which is the function s.y. and . [↑](#footnote-ref-85)
2. **Hint:** Use Stirling’s formula for approximating the factorial function. [↑](#footnote-ref-86)
3. **Hint:** Bound the number of tuples such that every is even and . [↑](#footnote-ref-89)
4. **Hint:** Set and then show that if the event happens then the random variable is a factor of larger than its expectation. [↑](#footnote-ref-90)
5. **Hint:** Think of as choosing numbers . Output the first such number that is in . [↑](#footnote-ref-92)