

Mathematical Preparation for Finance

A wild ride through mathematics

Kaisa Taipale

Contents

Preface	v
1 Introduction to probability	1
1.1 Brief introduction to sets	2
1.1.1 Sizes of sets via bijections	2
1.1.2 “Paradoxes” of set theory	4
1.1.3 Notation for set theory	6
1.2 Axioms of probability	8
1.2.1 Definitions	8
1.3 Counting	9
1.3.1 Triangular numbers	10
1.3.2 Factorials	11
1.3.3 Combinations and permutations	14
1.4 Binomial theorem	15
1.5 Geometric and arithmetic series	17
1.6 Binomial trees	20
1.7 Continuity property	21
1.8 Compound experiments	24
2 Geometry problems in probability	27
2.1 Function families	28
2.2 Derivatives and integrals	29
2.2.1 Volume and area	29
2.2.2 Differentiation as infinitesimal approximation	32

2.2.3	Average rate of change and Mean Value Theorem . . .	34
2.2.4	Integration and its disguises	36
2.3	Buffon's needle problem	38
2.4	Stick-breaking problems	43
3	Probability rules	49
3.1	Basic rules of probability	49
3.1.1	Complements and inclusion-exclusion	49
3.1.2	The hat check problem	51
3.1.3	The Secret Santa problem	52
3.2	Conditional probability	54
3.2.1	Law of Total Probability	55
3.2.2	Recursive games	56
3.3	Bayes' theorem	57
4	First applications in finance	61
4.1	Binomial stock pricing	61
4.1.1	Additive model	61
4.1.2	Multiplicative model	63
4.2	Arbitrage	64
4.3	Short- and long-term approximation	66
4.3.1	Long-term approximation	68
4.3.2	First differential equation	68
5	Discrete random variables and transformations of variables	73
5.1	Discrete random variables	73
5.1.1	Linearity of expectation – best thing ever	77
5.1.2	Multiplicativity of expectation?	80
5.2	Variance: average of squared deviations	82
5.3	Series	84
5.3.1	Convergence, divergence, well-defined	85
5.4	Moment generating functions	87
5.5	Linear and affine linear transformations	88
5.6	Specific discrete random distributions	91

5.6.1	Bernoulli	91
5.6.2	Binomial	92
5.6.3	Geometric	93
5.6.4	Hypergeometric	94
5.6.5	Poisson	95
5.6.6	Negative binomial	96
5.6.7	Transformations of discrete random variables	98
6	Continuous Random Variables	99
6.1	Geometry problems	100
6.1.1	Max or min in unit square	100
6.1.2	Absolute values	102
6.1.3	Circles: you mean I need to remember trig substitution?	103
6.2	Expected value and variance	104
6.3	Transformations of continuous random variables	104
6.4	Markov and Chebyshev	106
6.5	Important distributions	107
6.5.1	Exponential distribution and Poisson	107
6.5.2	Normal distribution and lognormal distribution	112
6.6	Central Limit Theorem	114
7	Random walks	117
7.1	Simple symmetric random walk	117
7.1.1	Expectation and variance	118
7.2	Asymmetric random walks	121
7.2.1	Expected value and variance	123
7.3	Scaling time or space	125
7.3.1	Special properties	128
7.4	Arithmetic Brownian motion	128
7.5	Geometric Brownian motion	130
7.6	Solving Brownian motion problems	131
8	Linear algebra	135
8.1	What is a vector? What is a matrix?	135

8.2	Linear combinations and matrix multiplication	138
8.2.1	Linear combinations	138
8.2.2	Matrix multiplication and dot products	139
8.3	Geometry and linear algebra	141
8.3.1	Planes and parameterizations	142
8.3.2	Projection	144
8.3.3	Determinants	146
8.3.4	Cross products	147
8.4	Useful inequalities for vectors	148
8.5	Some vectors in finance	149
8.6	Linear transformations	151
8.7	Bases	155
8.8	Applications to financial math	159
8.9	Invertible transformations	162
9	Spectral theorem and portfolio management	165
9.1	Orthogonal matrices and orthonormal bases	165
9.2	Gram-Schmidt orthogonalization	166
9.3	Rotation and scaling	168
9.4	Complex Numbers	170
9.4.1	Taylor series	172
9.5	Changes of basis and coordinates	172
9.5.1	Changing basis alone	172
9.5.2	Changing linear transformations into a new basis	173
9.6	Eigenvalues and eigenvectors	175
9.7	Quadratic forms and definiteness	180
9.8	Power series of matrices	181
9.9	Applications to financial math	182
9.10	Complex vectors	182
9.11	Complex matrices	183
9.12	The spectral theorem	185
9.13	Singular value decomposition	186
9.14	Applications of SVD	188

10	Joint distributions	189
10.1	Jointly distributed discrete random variables	189
10.1.1	Marginal probability mass function	190
10.1.2	Examples of discrete jointly distributed random variables	190
10.1.3	Conditional probability mass function	192
10.1.4	Independence	193
10.1.5	Multivariate versions	193
10.2	Jointly distributed continuous random variables	194
10.2.1	Marginal and conditional probability density functions	195
10.3	Covariance and correlation, again	197
10.4	Multivariate change of variables	199
10.5	Bivariate and multivariate normal	201
10.6	Visualizing the bivariate normal distribution	204
11	Optimization and Newton's method	207
11.1	Single-variable optimization	207
11.1.1	Single-variable unconstrained minima and maxima . .	208
11.1.2	Single-variable constrained optimization	210
11.2	Newton's method, single-variable	211
11.2.1	Newton's method for single-variable optimization . . .	213
11.3	Multivariate Taylor approximations	214
11.3.1	Zeroes of multivariate functions	216
11.4	Multivariate optimization	218
11.4.1	Unconstrained optimization	218
11.4.2	Multivariate Newton's method for optimization	219
11.4.3	Constrained optimization	220
12	Differential equations	221
12.1	Equilibrium: the concept	222
12.2	A single ordinary differential equation	222
12.3	So I can model a caribou population: what about money? . . .	225
12.4	Systems of Differential Equations	226
12.5	Equilibria	229
12.6	Straight-line solutions	231

12.7 Back to Black-Scholes for a minute 233

Preface

This book is being written for students in my Preparation for Financial Mathematics class at the University of Minnesota. It's an idiosyncratic tour through probability, calculus, and linear algebra, attempting to mix in financial applications throughout. Unlike many authors, I am making *no attempt whatsoever* to write a “self-contained” exposition: I'm assuming you know some calculus and linear algebra and are willing to look up a lot. Perhaps it is better to view this as a workbook than a reference book.

My students next take a course, FM 5011, in which probability is presented in terms of measure theory. You'll deal with the sigma-algebra that consists of sets of events and you'll put a probability measure \mathbb{P} on that σ -algebra. Then, in the rest of the first day, you'll proceed through probability distributions for random variables, noting the binomial, Poisson, uniform, and normal distributions among others. You'll review joint distributions and expectations. That's lecture one of 5011; by lecture two you'll cover stochastic processes.

In this book, we will take a more non-technical approach and get our hands dirty. The aim is that you will *understand* all those distributions, and have a beginning familiarity with Brownian motion, so that you can assimilate the technical notions presented in 5011.

This book has benefited from the feedback of my students through several years: in essentially random order, I'd like to thank them here. In 2017-2018, Yanqiu Tan, Tianzi Guo, Thara Ali Said, Jacob Gotto, David Rokhinson, Ameya Phadke, and Alec Hamer all made suggestions for improvement. In previous years, Joseph Ogega, Chong Wang, and Guining Zhang made suggestions – and since they're all graduated, you should hire them for great math finance jobs. In 2018-2019, I need to thank Boris Alyurov and Michele Knud-

x

PREFACE

sen.

Chapter 1

Introduction to probability

We're plunging right into probability. Several generations of students have now found that probability is subtle, confusing, and requires a different mindset than calculus and linear algebra. You want a full year to absorb these subtleties!

Broadly, our goal for this semester is to gain a firm foundation in basic probability so that we can proceed to understanding geometric Brownian motion, which is widely used as a basic model for stock prices. At the end of the semester (end of [chapter 7](#)) I want you to understand the transition from random walk to Brownian motion, and how these relate to the Black-Scholes differential equation and the binomial tree model for option pricing. The goal for this first chapter is to get you acquainted with the basics of probability, grounding you in the idea of looking at a *sample space* and finding the probability of an *event* occurring in this sample space when you conduct a *chance experiment*. You will get a lot of practice with such problems.

By the end of this chapter, I want you to understand the axioms of probability:

Definition 1.0.1. The probability measure on a sample space is denoted by P . It assigns to each set A in the sample space Ω a probability, satisfying the following axioms:

- $P(A) \geq 0$ for each subset A .
- $P(A) = 1$ when A is equal to the sample space.

$$\bullet P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i) \text{ for every collection of pairwise disjoint subsets } A_1, A_2, \dots$$

The set notation is explained in [Section 1.1](#) and the probability terminology is explained in [Section 1.2](#).

I also want you to be able to solve problems like the following:

Eight teams are in the semifinals of an international badminton tournament. The eight teams consist of two teams each from China, India, Denmark, and Korea. What is the probability that the two teams from each country end up playing against each other in each of the semifinal matches? That is, China 1 plays China 2, Denmark 1 plays Denmark 2, etc.

For these, you'll need the counting techniques outlined in [Sections 1.3](#) and [1.4](#).

1.1 Brief introduction to sets

1.1.1 Sizes of sets via bijections

A finite set is one that has a positive integer number of elements. These are the sets we got comfortable with in kindergarten and first grade: five apples, seven chairs, a million dollars. (Well...) Another way to say that a set A is finite is to say that its elements are in *bijection* with a set $\{1, \dots, n\}$ for n a positive integer. A “bijection” is a matching that pairs each element in A with one and exactly one element in $\{1, \dots, n\}$. Then we use the notation $|\cdot|$ for cardinality of a set to write that the size of A is $|A| = n$. (Note that this is the same notation as absolute value and magnitude, so you must understand the context of the statement to correctly interpret $|A|$.)

More generally, a bijection between two sets A and B is a matching (function, map) that is *injective* and *surjective*. A map $f : A \rightarrow B$ is injective, or “one-to-one,” if $f(a_1) = f(a_2)$ in B means that $a_1 = a_2$ in A . This means no two distinct elements in A can map to the same element in B . A map $f : A \rightarrow B$ is surjective, or “onto,” if for every $b \in B$ there's an $a \in A$ such

that $f(a) = b$. This means every element in B is “hit” by the map f from the set A .

Why make things so complicated? The concept of bijection is useful in talking about sizes of infinity. For instance, the set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ of natural numbers is a certain “size” of infinity. We call this size of infinity *countable*, because we can count through it. Any other set we can put in bijection with the counting numbers \mathbb{N} will also be called *countably infinite*.

Put the integers \mathbb{Z} in bijection with \mathbb{N} . You can do it! Can you reconcile your bijection with your intuition that there should be somehow twice as many integers as natural numbers? Infinity is mysterious.

Another “size” of infinity comes up when looking at *uncountable* sets. How many real number \mathbb{R} are there? Can you put the real numbers \mathbb{R} in bijection with the natural numbers \mathbb{N} ?

Prove to yourself that you *can't* put \mathbb{R} and \mathbb{N} in bijection.

The set of real numbers \mathbb{R} is an uncountable set, and so is the set of irrational numbers – let’s call that \mathbb{I} for now. (Irrational numbers are numbers that can’t be expressed as a ratio of two integers. Examples include π and e . The set of rational numbers is denoted by \mathbb{Q} .)

With uncountable sets like the real numbers we can construct very weird subsets that can’t be assigned a probability measure consistent with how probability should work. Basically, we can break the intuitive notion we have of the size or “measure” of a set: if we proceed naively, we could have a set that looks like it’s of size 1 and size 0 at the same time, in which case $1 = 0$ and all is lost. That’s why σ -algebras are introduced in later classes: using σ -algebras and more sophisticated ideas we can formalize the notion of measure and make sure probability works over uncountable sets. Our approach in this book: ignore this problem. Let’s stick with “nice” subsets of uncountable spaces like intervals and their complements. But before we stop considering complicated sets forever, I’ll throw out a few situations that seem paradoxical, to show you why you’ll need another year of math to truly understand measure theory.

1.1.2 “Paradoxes” of set theory

Above, I asked you to put \mathbb{Z} in bijection with \mathbb{N} . Hopefully you did that! Now, think about the rational numbers, \mathbb{Q} . This is a countable set, as you can see from the picture below:

\mathbb{P}	1	2	3	4	5	6	7	8	9	...
1	1	2	3	4	5	6	7	8	9	...
2	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	$\frac{7}{2}$	4	$\frac{9}{2}$...
3	$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{4}{3}$	$\frac{5}{3}$	$\frac{6}{3}$	$\frac{7}{3}$	$\frac{8}{3}$	$\frac{9}{3}$...
4	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	1	$\frac{5}{4}$	$\frac{6}{4}$	$\frac{7}{4}$	$\frac{8}{4}$	$\frac{9}{4}$...
...

In particular, that means that the set of rational numbers between zero and one is countable. Let's consider the cardinality of the set of *all* real numbers between zero and one. It's uncountable, and here's how to prove it. I use what's called Cantor's diagonalization argument. Assume you can list all the numbers between zero and one, writing them out as decimals. Put that list together:

\mathbb{N}	potential list for $[0,1]_{\mathbb{R}}$
1	0.72345...
2	0.63871...
3	0.08214...
4	0.15023...

New number:
0.8433...

A clever friend comes along and says, “I can give you a number between zero and one that is not on that list!” You say, “Impossible – I listed them all!” She says, “No: I'll give you a number whose i th digit after the decimal is one different than the i th digit of the i th number on your list, for every digit.” You are flabbergasted... but she is right! The number she gives you is between zero and one and is not on your list, no matter what was on your list.

This proves that $[0, 1]$ is an uncountable set of numbers. Moreover, it proves \mathbb{R} is uncountable. But we know that the length of the interval $[0, 1]$ is one.

Somehow we have a countable set $[0, 1]_{\mathbb{Q}}$ of rational numbers between zero and one and an uncountable set $[0, 1]_{\mathbb{I}}$ of irrational numbers between zero and one and their “sizes” add to one. It turns out that the *measure* of $[0, 1]_{\mathbb{I}}$ is one and the measure of $[0, 1]_{\mathbb{Q}}$ is zero.

From this example you might get the idea that countable sets are always of measure zero or that uncountable sets have positive measure. Nope. Here’s another puzzle, again involving the mathematician Georg Cantor:

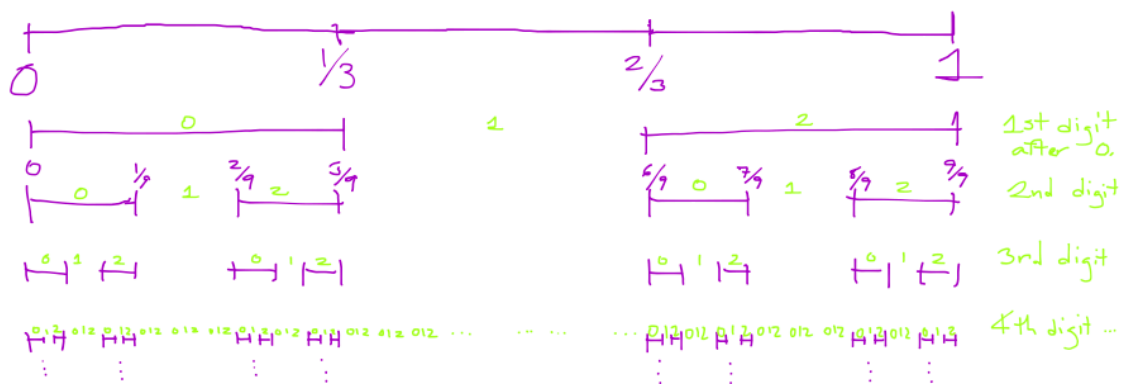
We can define a set C called the Cantor middle-thirds set. Take the interval $[0, 1]$ and remove the middle interval $(\frac{1}{3}, \frac{2}{3})$. Then from the two remaining intervals, remove the middle third. Keep removing the middle third of each remaining interval (forever). You’re left with a dusting of points. How big is it? Well, what do you mean by “how big”?

The “length” of the interval (its measure) can be found by using a geometric series. You start with an interval of length 1, take out $1/3$, take out $2/9$, take out $4/27$

$$\sum_{n=0}^{\infty} \frac{2^n}{3^{n+1}} = \frac{1}{3} \left(\frac{1}{1 - 2/3} \right) = 1.$$

So this is a set of measure zero (again, we have no rigorous definition here). But how many elements does it have?

Take a look at the picture below to see a way to label each point in the set:



This gives a ternary representation of each point $x \in [0, 1] \subset \mathbb{R}$ – ternary means “base three,” as opposed to binary or decimal. For each step of the middle-interval-subtraction, write 0 if the point is in the left interval, 1 if it is

in the middle, and 2 if the point is in the right interval. (These numbers are in green in the picture.) The points in the Cantor set C then consist of every number that has an expression only in 0s and 2s. For example, $1/3 = .0\bar{2}$ and $2/9 = .02\bar{0}$, while $1/2$ starts with .1 and so is not in the Cantor set. Now you can use Cantor's diagonalization argument to prove C is uncountable: try to make a list of every number in the Cantor set and have your clever friend come by and give you a number that is in the Cantor set but not on your list!

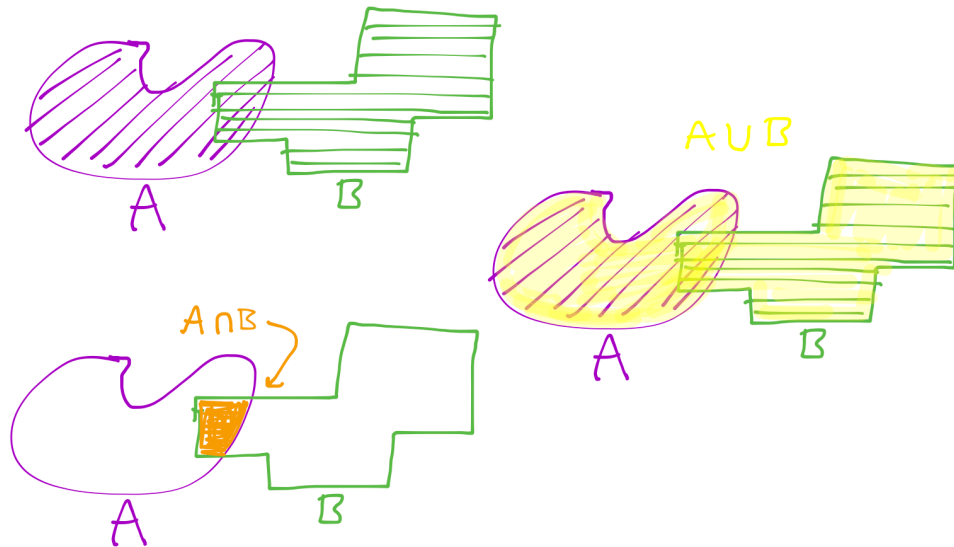
The Cantor middle-thirds set, then, is uncountable but of measure zero. What's the probability of choosing a point in the Cantor set if you choose a number randomly between zero and one? That's for next year!

The bottom line, for us: we will only consider sets that are discrete and of measure zero, or sets that are made from intervals in \mathbb{R} . An interval $[a, b]$ or (a, b) or $[a, b)$ for $b > a$ has measure $b - a$ and an uncountable number of points, and we don't have to deal with any paradoxes.

1.1.3 Notation for set theory

We express that a is an element of the set A by writing $a \in A$. As you may have noticed above, we often use curly brackets $\{ \ , \}$ to indicate we are discussing a set: $\{1, 2, \dots, n\}$ is the set of integers one through n , while $\{z | z/3 \in \mathbb{Z}\}$ would be the set z of numbers that satisfy the condition that $z/3$ is an integer. Read the $|$ sign (or, in some books, the $:$ sign) as "such that." There are often many ways to write the same set: $\{z | z/3 \in \mathbb{Z}\}$, $\{z | z \bmod 3 = 0\}$, or $\{3x\}$ for $x \in \mathbb{Z}$.

You will need to be comfortable with the interactions of sets as well as finding the cardinality of a given set or the measure of an "easy" set (one consisting of intervals). *Unions* and *intersections* of sets are the two big concepts. The union of two sets is the set of all elements in either set: $A \cup B = \{c | c \in A \text{ or } c \in B\}$. The intersection of sets is what you think it is: $A \cap B = \{c | c \in A \text{ and } c \in B\}$. Visually, this is



If you have a countably infinite number of sets, like $A_i = \{A_1, A_2, \dots\}$, you can write

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup A_3 \cup \dots$$

$$\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap A_3 \cap \dots$$

Two sets are disjoint if they don't overlap – if their intersection is empty. We can write $A \cap B = \emptyset$, using \emptyset to represent the empty set.

We also write A^c for the *complement* of a set A . This is the set of all elements that are not in A . For instance, if $A \subset \mathbb{Z}$ and $A = \{z \geq 0\}$, then A^c contains the elements in \mathbb{Z} that are less than zero: $A^c = \{z < 0\}$. Notice that $A \cap A^c = \emptyset$.

Last, we have notation for set “subtraction.” We can pick out all elements of a set B that are not also in A by writing $B \setminus A$.

Many of these concepts will be revisited in [section 3.1](#). This should be enough to get us started, though, and as you work with probability problems you will start to see why set theory is so useful.

1.2 Axioms of probability

1.2.1 Definitions

When you conduct a probability experiment or a chance experiment, you toss the die or flip the coin or throw the dart – whatever it may be. Then an event occurs: the outcome of your experiment. The set of all possible outcomes of your experiment is called the sample space, Ω , and an outcome in the sample space will be denoted $\omega \in \Omega$.

A sample space with a collection of events and an assignment of probabilities to the events is called a *probability space*. The probability measure on a sample space is denoted by P . It assigns to each set A in the sample space Ω a probability, satisfying the following axioms:

- $P(A) \geq 0$ for each subset A .
- $P(A) = 1$ when A is equal to the sample space.
- $P(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$ for every collection of pairwise disjoint subsets A_1, A_2, \dots

Notice that the set A can be a single event or a more. For instance, you can consider the event of rolling a six when you roll a die (set of size one) or the event of rolling an odd number (set of size three).

A *probability measure* P on Ω is defined by assigning a probability to every event ω in a finite or countably infinite Ω , so that $P(\omega) \geq 0$ and $1 = \sum_{\omega \in \Omega} P(\omega)$, and letting

$$P(A) = \sum_{\omega \in A} P(\omega).$$

Some specific examples and terminology: Two events are *mutually exclusive* if they can't both happen. For instance, you can't roll a two and a five on the same roll of a die. You can't flip both heads and tails when you flip a coin.

A lot of the problems you'll encounter in the next few pages will feature finitely many outcomes $\omega_1, \dots, \omega_N$ which are all equally likely. Then by the

axioms of probability you can deduce that $P(\omega_i) = 1/N$ for $i = 1, \dots, N$ and each event A has probability

$$P(A) = \frac{N(A)}{N},$$

where $N(A)$ is the number of outcomes in set A .

1.3 Counting

Let's go back to the questions asked at the beginning of the section:

Eight teams are in the semifinals of an international badminton tournament. The eight teams consist of two teams each from China, India, Denmark, and Korea. What is the probability that the two teams from each country end up playing against each other in each of the semifinal matches? That is, China 1 plays China 2, Denmark 1 plays Denmark 2, etc.

In one throw, you roll two dice. You win if the sum of the numbers on the two dice is 8 or more. What's the probability that you win?

How can you solve these problems? You probably have an intuitive approach, but I am going to push you to try a systematic and algorithmic approach. Learning this systematic approach will help you solve many probability problems, I promise. Here it is:

If each individual event ω in a finite sample space Ω is equally likely,

1. Identify the sample space Ω precisely and mathematically.
2. Write out the event or set of events $A \subset \Omega$ whose probability you want. Use the same notation and setup as you used when identifying the sample space Ω .
3. Calculate the size of Ω and the size of A .
4. Divide: $P(A) = |A|/|\Omega|$.

To do this, you need to know how to count.

1.3.1 Triangular numbers

We'll start with the dice problem and work through the problem-solving algorithm I suggested.

Step one: the sample space, the set of all possible outcomes of the experiment, is the set of combinations of numbers on the two dice. To be careful, I'll make one a red die and one a blue die – I want to keep track of them. Why? If the outcomes are written (number on red die, number on blue die), then (2, 1) and (1, 2) are both possible outcomes. If you don't keep track of the dice, you may think that the combination of 1 and 2 is as likely as the combination of 3 and 3, instead of being twice as likely.

So, $\Omega = (r, b)$, where r is the number on the red die and b is the number on the blue die.

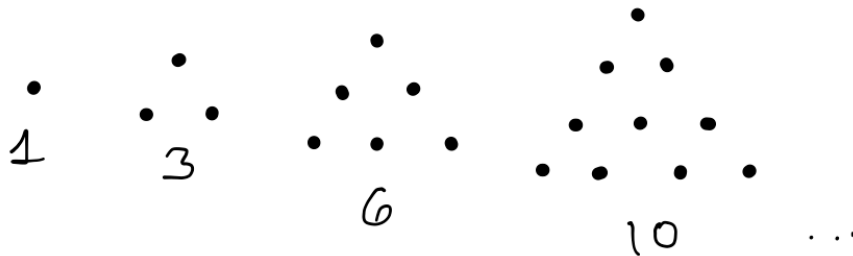
Step two: identify the events that give us a “win.” We can write this as $A = \{(r, b) \in \Omega \mid r + b \geq 8\}$, and draw both Ω and A :

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

The elements of the set A are highlighted with yellow.

Step three: How big is Ω ? See our previous drawing. Ω has 36 elements, all equally likely, so $|\Omega| = 36$.

How do we find $|A|$? You see the triangle of size $\sum_{i=1}^5 i$ highlighted with yellow in the sketch of Ω . You can count by hand, or you can use the wonder of triangular numbers. The triangular numbers are



I want a formula for these numbers so that I don't need to count by hand. Look at $\sum_{i=1}^n i$ and do a demonstration that can be made into a proof:

This picture proof is easily made rigorous to give

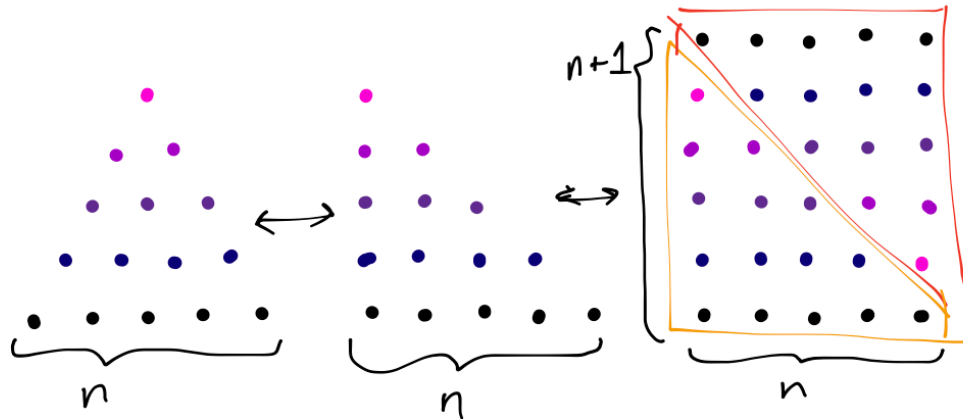
$$\sum_{i=1}^n i = \frac{n(n+1)}{2}.$$

Step four: $P(A) = |A|/|\Omega| = 15/36$.

1.3.2 Factorials

For many counting problems we need to know the number of possible orderings of a set of objects. Factorials are the building blocks of ordered counting. Say you need to know how many distinct five-letter strings can be made from the set of letters $\{a, b, c, d, e\}$, using each letter once. (In computer science a string is a sequence of characters.) It's clear that you have five choices for the first letter of the string, four choices remaining for the second letter, three for the third, and so on. Counting in this way you could make

$$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5!$$



different strings from $\{a, b, c, d, e\}$ without repeating any letters. We read this as “five factorial.” For a positive integer n ,

$$n \cdot (n - 1) \cdot \dots \cdot 2 \cdot 1 = n!$$

and by convention we say $0! = 1$. On the vocabulary side, we have just computed the number of *permutations* of the set $\{a, b, c, d, e\}$. Notice I didn’t say “ordered permutations” – permutations are by definition ordered.

Now let’s try the badminton problem:

Eight teams are in the semifinals of an international badminton tournament. The eight teams consist of two teams each from China, India, Denmark, and Korea. What is the probability that the two teams from each country end up playing against each other in each of the semifinal matches? That is, China 1 plays China 2, Denmark 1 plays Denmark 2, etc.

Step 1: Identify the sample space Ω precisely and mathematically. First, come up with your own sample space.

There are several possibilities for the sample space, but I will use all 8-tuples of teams – that is, all strings that can be created from

$$T = \{C1, C2, I1, I2, D1, D2, K1, K2\}.$$

Order matters because in this set-up I am assuming that the first team named plays the second, the third plays the fourth, and so on. I will write this as an 8-tuple, $(X1, X2, X3, \dots, X8)$ where the X_i are all distinct teams.

$$\Omega = \{(X1, X2, X3, \dots, X8) | X_i \in T, X_i \neq X_j.\}$$

(Compare this to the choice you made. What do you think?)

Step 2: Write out the event or set of events $A \subset \Omega$ whose probability you want. Use the same notation and setup as you used when identifying the sample space Ω . We're looking for the probability that each country plays itself in the semifinals, and we are looking for 8-tuples that satisfy this outcome condition. Examples of this include $(I1, I2, K2, K1, D2, D1, C1, C2)$ and $(K2, K1, I1, I2, D2, D1, C1, C2)$. Thus A is the set of 8-tuples in Ω with $X1$ and $X2$ from the same country, $X3$ and $X4$ from the same country, $X5$ and $X6$ from the same country, and $X7$ and $X8$ from the same country.

Step 3: Calculate the size of Ω and the size of A . The size of Ω is easy:

$$|\Omega| = 8!$$

The size of A is more subtle: you have 8 choices for the first team and 1 choice for the second (if the first team is $K1$, the second must be $K2$ and vice versa). Then 6 choices for the third team and 1 for the fourth. Continue on to get

$$|A| = 8 \cdot 1 \cdot 6 \cdot 1 \cdot 4 \cdot 1 \cdot 2 \cdot 1 = 2^4 \cdot 4!.$$

Another way to think about the size of A : there are $4!$ orders for the countries within the 8-tuple and given every country order ($KIDC$ or $CDIK$) there are 2^4 ways of ordering the teams ($K1$ then $K2$ versus $K2$ then $K1$). This gives $2^4 \cdot 4!$ immediately.

Step 4: Divide: $P(A) = |A|/|\Omega|$. Thus

$$P(A) = \frac{2^4 \cdot 4!}{8!} = \frac{2^4}{8 \cdot 7 \cdot 6 \cdot 5} = \frac{1}{105}.$$

A problem-solving note: you might argue that there are just four games, so counting $(C1, K1, C2, K2, I1, I2, D1, D2)$ as a different event than $(C2, K2, C1, K1, I1, I2, D1, D2)$ is redundant. On a physical level this is true, but it turns out that “overcounting” in our sample space helps us organize events nicely and we cancel out the “overcount” in our characterization of A . Sometimes it’s much easier to use some seeming “overcount” rather than the most parsimonious presentation of the events.

1.3.3 Combinations and permutations

What if we wanted instead ordered subsets of n items? For instance, all the possible 4-tuples $(Y1, Y2, Y3, Y4)$ of badminton teams that would go to the quarterfinals, or all the distinct two-letter strings we could make by picking two elements from $\{a, b, c, d, e\}$ without repetition? This second example involves picking two elements in order: for the first character in the string we have five choices, for the second we have four choices. We write this number

$$5 \cdot 4 = {}_5 P_2 = {}^5 P_2 = P(5, 2),$$

of which I prefer $P(5, 2)$, or in China we call it an *arrangement* instead and write

$$5 \cdot 4 = A_2^5.$$

Generalize this to creating a k -letter string from n distinct elements: count down from n choices for the first letter to $n - k + 1$ choices for the last of the k letters. We get

$$\frac{n!}{(n - k)!} = {}_n P_k = {}^n P_k = P(n, k) = A_k^n.$$

Combinations instead refer to selections of k items from n for which order does not matter. Imagine that you grab two socks from a drawer full of n socks and you want to see if they are a pair. Or you could choose three letters from $\{a, b, c, d, e\}$ and not care about their order. We know how to count the distinct strings we get by choosing three letters in order (it’s $P(5, 3)$) and now just need

to divide by orderings of the three letters (a factorial, $3!$). Thus the number of ways to grab three letters from $\{a, b, c, d, e\}$ is

$$\frac{P(5, 3)}{3!} = \frac{5!}{2!3!} = \binom{5}{3} = C(5, 3).$$

Here I point out a warning: I've also seen C_5^3 , C_3^5 , $C_{5,3}$, ${}_5C_3$, and 5C_3 . This is horrible, as it seems that any combination of C , 5 , and 3 is used somewhere. DON'T USE THESE! Simply use the binomial notation $\binom{5}{3}$. It is unambiguous and best. We set notation for the number of ways to pick k items (unordered) from n items to be

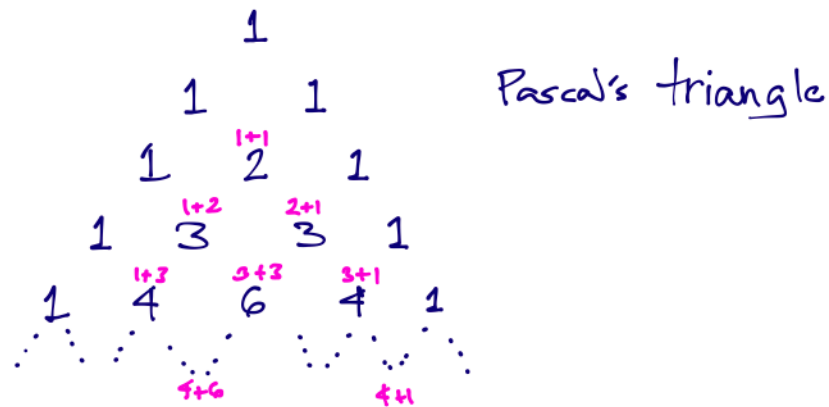
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Take this as your definition, and check out [Section 1.4](#) to see why I called this binomial notation.

1.4 Binomial theorem

The coefficients of the expansion of the innocent little expression $(x + y)^n$, for n an integer, will appear so many times in the rest of your life that you will be astounded. These numbers are truly essential in doing probability or financial mathematics.

Expand a few instances of $(x + y)^n$ where n is a positive integer: you'll start to notice a pattern. In school you may have learned that you can cleverly get these coefficients via Pascal's triangle. First, Pascal's triangle alone:



You may remember that $(x + y)^n$ expands nicely as

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k$$

as long as n is a positive integer. Since we defined $0! = 1$, we have $\binom{n}{0} = \binom{n}{n} = 1$. Here is how this compares with Pascal's triangle:

$$\begin{aligned}
 (x+y)^0 &= 1 \\
 (x+y)^1 &= 1x + 1y && \text{Binomial expansion} \\
 (x+y)^2 &= 1x^2 + 2xy + 1y^2 \\
 (x+y)^3 &= 1x^3 + 3x^2y + 3xy^2 + 1y^3 \\
 (x+y)^4 &= 1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4
 \end{aligned}$$

Pause here to ponder more deeply, and experiment with these problems:

Specialize $(x + y)^n$ to the situation $x = 1, y = 1$: what does this imply about the relationship between binomial coefficients and 2^n ?

How is $\binom{n}{k}$ related to $\binom{n}{n-k}$?

Consider flipping a fair coin n times. What is the probability that you get k heads and $n - k$ tails?

If you evaluate $(x + y)^n$ at $x = 1, y = 1$ (that's what I mean when I say "specialize" – it means "look at the special case"), you get 2^n . Looking at that with the binomial expansion, you see that

$$(1 + 1)^n = \sum_{k=0}^n \binom{n}{k} 1^{n-k} 1^k = \sum_{k=0}^n \binom{n}{k}.$$

Knowing that the sum of binomial coefficients $\binom{n}{k}$ from $k = 0$ to $k = n$ is 2^n actually comes in quite handy when you're looking at random walks and coin-flipping problems.

Grabbing k items out of n items is the same as leaving $n - k$ items behind, so $\binom{n}{k} = \binom{n}{n-k}$. Thinking about what you want and what you *don't* want is a great problem-solving tool – it sometimes lets you flip a problem into a more tractable problem.

It's easy to compute how many ways you can get k heads out of n tosses – that's our binomial coefficient! So there are $\binom{n}{k}$ ways to get k heads and $n - k$ tails from n total tosses. That's not a probability, though. How many total patterns of heads and tails are there? The sum over all j of $\binom{n}{j}$ is 2^n as we just saw above, so the size of the sample space of all coin flip sequences is 2^n and thus the probability of exactly k heads is

$$\frac{\binom{n}{k}}{2^n}.$$

1.5 Geometric and arithmetic series

You may remember arithmetic and geometric series from algebra class sometime in the past. First we'll review their definitions and some terminology and then I'll justify why we are talking about them in a probability and finance context.

An *arithmetic sequence* is a sequence a_0, \dots, a_n, \dots with a constant *difference* d between each two consecutive terms (so $a_i - a_{i-1} = d$ for all $i > 0$).

An arithmetic series then is the sum of such terms,

$$\sum_{i=0}^n a_i = \sum_{i=0}^n (a_0 + id).$$

The triangular numbers discussed earlier are one of the most basic arithmetic sequences, since you're looking at sums of $1, 1 + 2, 1 + 2 + 3, \dots$ (so $d = 1$).

Arithmetic sequences and series mainly come up as a tool in probability problems – you find you need to add up a bunch of numbers all separated by 37, and then you do it. In finance, true arithmetic sequences come up mainly when considering constant payouts (if you're not considering the time value of money). Many financial instruments give a constant payout or demand a recurring constant fee, and if you're ignoring the time value of money you might consider using an arithmetic sequence to model this. In this era of near-zero interest rates, using an arithmetic sequence to model my monthly Crossfit gym payments is reasonable in the short term.

A *geometric sequence* is a sequence a_0, \dots, a_n, \dots with a constant *ratio* r between each two consecutive terms (so $a_i/a_{i-1} = r$ for all $i > 0$). The geometric sum is then

$$\sum_{i=0}^n a_i = \sum_{i=0}^n a_0 r^i.$$

You've probably encountered this when looking at compound interest – classic problems you might remember doing would involve compounding monthly, daily, etc. In each of those situations, you are multiplying your current principal by $(1 + r)$ for the appropriate interest rate r . If you need an infinite number of terms in your sum, you get

$$\sum_{i=0}^{\infty} a_0 r^i = \frac{a_0}{1 - r}.$$

Geometric sequences also come up a lot in probability problems, and in fact there's a probability distribution called a "geometric distribution". What's the probability that you need exactly k rolls of a die until you roll the number 6 for

the first time? It's

$$\left(\frac{5}{6}\right)^k \frac{1}{6}.$$

What's the probability that it will take less than five rolls to get that 6 for the first time? You'll sum up these probabilities (since they correspond to mutually exclusive events) and that'll be a geometric series.

There are nice formulas for sums of these sequences. Derive them:

You know that

$$\sum_{i=0}^n a_i = \sum_{i=0}^n (a_0 + id).$$

Now use what you know about triangular numbers and sums of constants to come up with a closed formula for this sum, depending on a_0 , n , and d .

You know that

$$S_n = \sum_{i=1}^n a_i = \sum_{i=1}^n a_1 r^{i-1}.$$

I'm calling this sum S_n for a reason. Do something clever here: compare S_n and rS_n . In fact, look at $S_n - rS_n$. Write out the terms of $S_n - rS_n = (1-r)S_n$, then solve for S_n . This will give you a nice closed form presentation for S_n .

Terminology break: "closed form" and "closed formula" mean a function where you can just plug in a few numbers and directly compute an answer, as opposed to something recursive, or a formula with a limit in it, or where you have to do an infinite number of operations. Example: you could get an expression like

$$f(x) = \sum_{i=0}^{\infty} \frac{x}{2^i}$$

which as written implies you've got to sum up an infinite number of terms. On the other hand, you could use geometric series to work out that

$$f(x) = \sum_{i=0}^{\infty} \frac{x}{2^i} = x \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i = x \cdot \frac{1}{1 - 1/2} = x \cdot 2.$$

The expression $f(x) = 2x$ is something you can do in one "move".

1.6 Binomial trees

Now let's put into financial action a bit of what you've learned. We'll talk about stock movements here, foreshadowing random walks and Brownian motion. Make sure you understand *every detail* here, as these foundational ideas do need to be foundational if you want to do financial math!

Consider a stock that either goes up in price or down in price with equal probability. What is the probability that it goes up k times and down $n - k$ times?

A few ingredients to consider: the probability the stock goes up is $1/2$; the probability the stock goes down is $1/2$; the number of paths with k up-steps and $n - k$ down-steps is $\binom{n}{k} = \binom{n}{n-k}$. So the probability of going up exactly k times (or down exactly k times) is

$$\binom{n}{k} \frac{1}{2^n}.$$

Consider a stock whose price goes up by ten dollars or goes down by ten dollars, with equal probability, on any given day. Its initial price is S_0 . If in n days it has gone up k times and gone down $n - k$ times, what is its final price? Since a stock price can't be negative, what is the probability that the stock price is zero?

Write S_n for the price in dollars on day n :

$$S_n = S_0 + 10k - 10(n - k) = S_0 + 20k - 10n.$$

What's the probability of this price? If we don't care about negative prices, the probability of price S_n on day n is just

$$\binom{n}{k} \frac{1}{2^n}$$

from the previous problem. However, stock prices can't go negative (although you can short stocks).

The real-life probability of a stock price of zero is basically zero, as you can't buy and sell stocks with share price zero. Even if a company goes

bankrupt and stock is still bought and sold, it's not at a price of zero – Blockbuster Video's stock traded at prices from 4 to 23 cents a share for a while after it went into bankruptcy. When stock prices fall to near zero they're commonly delisted from stock markets, and then move to over-the-counter pink sheets (a very weird market).

This is a situation in which our model and reality conflict. An additive model of stock prices gives plenty of situations in which stock prices go to zero or negative numbers. If you use such a model, you'll need to decide if you want to take $\max(0, S_n)$ for your price on day n , and whether the probability of the stock price going to zero really reflects the probability that the company will go bankrupt.

Consider a stock whose price goes up by ten percent or goes down by ten percent, with equal probability, on any given day. Its initial price is S_0 . If in n days it has gone up k times and gone down $n - k$ times, what is its final price? What is the probability that the stock's price is zero after n days?

In this multiplicative model, the price on day n is

$$S_n = S_0(1 + 0.1)^k(1 - 0.1)^{n-k} = S_0 \cdot 1.1^k \cdot 0.9^{n-k}.$$

Since we're multiplying a positive number S_0 by more positive numbers (percent increase/decrease), S_n will never be zero in this model!

1.7 Continuity property

“Probability is a continuous set function.” What does that mean? Why do we care?

Our short discussion here will be a foreshadowing of measure theory, a topic not covered in this book! On a first reading, you might choose to skip this section – but you may find it useful to ponder the complexities of this topic.

Let

$$\{E_n\} = E_1 \subset E_2 \subset E_3 \subset \dots$$

be a nondecreasing sequence of sets. For the union $E = \cup_{i=1}^{\infty} E_i$, use the

notation $E = \lim_{n \rightarrow \infty} E_n$. Then the continuity property is

$$\lim_{n \rightarrow \infty} P(E_n) = P(\lim_{n \rightarrow \infty} E_n).$$

Essentially, the theorem is that you can interchange the probability function and the limit. Limits characterize continuity, which is why this is called the continuity property of probability. You can also translate this result to a non-increasing sequence, in which case $\lim_{n \rightarrow \infty} E = \bigcap_{i=1}^{\infty} E_i$.

What is an example of a set of sets like this? Here's an example made up for this situation – you probably won't encounter it in real life. Say you've got a game of darts that keeps getting tougher, and you're terrible at darts. The dart board is a disk of radius one foot; we'll assume you can hit the board with a dart every time but your aim is random on the dart board (every point is equally likely). Here are my made-up rules: On the first round, you need to hit the center of the disk, a region with radius 1/2 foot, to score. To score on the second round, you need to hit the disk in the middle with radius 1/4. To score on the third round, you need to hit the disk with radius 1/8. To score on the n th round, you need to hit the disk with radius $1/2^{n-1}$. Since you are so bad at darts, the probability of hitting the circular region A with radius r feet, $r < 1$, is

$$P(A) = \frac{\pi r^2}{\pi 1^2} = \frac{r^2}{1}.$$

What's the probability you score as the radius goes to zero? This probability is zero: the probability of scoring on round n , $P(E_n)$, is $1/4^{n-1}$. Since

$$\lim_{n \rightarrow \infty} P(E_n) = 0,$$

we know

$$P(\lim_{n \rightarrow \infty} E_n) = 0.$$

A related question is this: what's the probability that you score infinitely many times? That's also zero. The Borel-Cantelli Lemma gives this extension:

Let A_1, A_2, \dots be an infinite sequence of subsets of the sample space Ω . Define

the set C as the set of events ω that occur in infinitely many A_k . Then if

$$\sum_{n=1}^{\infty} P(A_n) < \infty,$$

we have

$$P(C) = 0.$$

So what's the probability that you score infinitely many times in my made-up dart game? Zero.

You can also get from this a corollary about things that *always* happen instead of *never* happening.

Let A_1, A_2, \dots be an infinite sequence of subsets of the sample space Ω , all of which are *independent*. Define the set D as the set of events ω that occur in all A_k for $k \geq m$, for some m , and such that ω doesn't occur in only finitely many A_i . Then if

$$\sum_{n=1}^{\infty} P(A_n) = \infty,$$

we have

$$P(D) = 1.$$

For an example of this type of situation, consider some poor student doomed to toss a coin forever. Let's say the event A_n occurs if in the n th block of 100 coin tosses, you get 100 heads in a row. Each block of 100 coin tosses is independent of every other block of 100 coin tosses, and $P(A_n) = 1/2^{100}$. We can see that

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{2^{100}} = \infty,$$

so by the lemma $P(D) = 1$. That means that with probability one you'll get a run of 100 heads in an infinite series of coin tosses.

Don't ignore this topic even though this section seems very small. Turns out the Borel-Cantelli lemma comes up again when we look at Brownian motion.

1.8 Compound experiments

A compound experiment consists of several elementary experiments that are independent of each other. For instance, if you toss three coins that are physically independent from each other, you're conducting a compound probability experiment.

We'll revisit the concept of independence when we discuss random variables. For now, use your intuitive understanding of the idea: two events are independent if they don't influence each other.

The probability of A and B is

$$P(A \cap B) = P(AB) = P(A)P(B)$$

as long as the events in A are physically independent of the events in B . Notice that some sources use the notation AB to mean the intersection of events A and B , and we will use this notation too.

Notice that the formula itself has no reference to a sample space here. Why not? We may mix and match sample spaces when discussing independent experiments. For instance, if we want to know the probability of rolling a three on a six-sided die *and* getting heads on a coin toss, we could look at the sample space $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{H, T\}$ and do a counting-based argument, or we could use the rule above to find

$$P(3 \text{ on die} \cap H) = P(3 \text{ on die})P(H) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12},$$

using independent calculations from the sample space of possible die rolls and coin tosses.

Another version of a compound experiment involves repeated independent probability experiments. Here are three different examples:

(A) Toss a coin seven times in a row. What is the probability that you get the sequence $HHHTTTH$?

(B) You take turns tossing a coin with a friend. The first person to toss tails owes the other one dollar and when tails appears, the game ends. What is the probability that you toss tails first and owe your friend a dollar?

(C) You hire an undergraduate to toss a fair coin for you an infinite number of times. If the undergraduate gets heads p times in a row, or tails p times in a row, we call it a *run of length p* . Prove that with probability one the undergraduate will get a run of length p at some point in this experiment.

(D) You are gambling on coin tosses: on each turn, you flip a single fair coin. Every time you gets heads, you pay her a dollar, and every time you get tails, she gives you a dollar. You have \$500. What is the probability that you have \$1000 before you run out of money?

Problem (A) is well within your reach. The probability that you get the sequence $HHHTTTH$ can be found using counting techniques or by multiplying the probabilities of independent outcomes. Make sure you can do this both ways.

Problem (B) has an infinite sample space

$$\Omega = \{T, HHT, HHHHT, HHHHHHT, \dots\} \cup \{HHHHHHHHHHHHHHHHH\dots\}.$$

Questions for you: Is this countable or uncountable? What is the probability that tails is never tossed at all? Hints: notice that all the events in the first subset of Ω are mutually exclusive, so you can sum up their probabilities. See if you can calculate the probability of each event and then sum them.

Problem (C) can be done using the Borel-Cantelli theorem discussed earlier – see if you can use this for a proof!

Problem (D) is an example of a stopping time problem. These types of problems were incredibly important in the beginnings of probability: in some sense, modern probability theory was born out of gambling, and when you are out of money you can't gamble any more. Figuring out the probability of reaching net worth W before net worth zero is pretty important, even in modern capitalism. The start-up concept of “runway,” for instance, is how much time you have before your start-up goes bankrupt, and depends on how much money you have and how fast you are spending it (your “burn rate”). Figuring out your runway is in essence a stopping time calculation.

Chapter 2

Geometry problems in probability

All the questions we addressed in the last chapter happened to be discrete probability problems. However, a lot of probability problems are geometric problems, with calculations taking place on a space with an uncountable number of points: throwing darts at a dartboard or dropping needles on a lined page, for instance. The probability of hitting a particular point is zero, for reasons we'll discuss, but the tools of geometry and calculus can help us.

Why are these problems of interest to someone who cares about finance? First, many problems in finance can be fruitfully considered by looking at continuous analogs. While stocks are priced in fractions of a cent on the New York Stock Exchange, the Black-Scholes equation for option pricing takes a continuous analogue and works rather well. Brownian motion relies on a continuous limit of discrete processes. Being able to use both continuous and discrete techniques as appropriate is an important tool in mathematical finance.

To solve geometric probability problems, you need to dredge up your knowledge of area and volumes. Calculating the size of a set with an uncountably infinite number of points is the calculation of an area or volume. Sometimes it's enough to know that the area of a disk of radius r is πr^2 , and sometimes you will need to integrate. By the end of this chapter I want you to be practiced in solving probability problems that require geometric techniques, like this:

Consider the square with corners at $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. Pick a point (x, y) in this square at random. What is the probability that the product of the coordinates, xy , is greater than one half?

This will require some review of calculus which will be done in the sections on function families ([Section 2.1](#)) and derivatives and integrals ([Section 2.2](#)). We'll also do a tiny bit of linear algebra in talking about affine linear transformations ([Section 5.5](#)). You may find that you need extra resources to shore up your memory of these topics or you may find them almost easy enough to skip – do the right thing. At the end of the chapter we'll examine some famous problems that use geometric techniques.

2.1 Function families

Here is a quick run-through of functions you'll encounter in this book:

- Linear functions and affine linear functions
- Quadratic functions, also known as conics
- Polynomials of all degrees
- Rational functions
- Power functions
- Exponential functions
- Logarithmic functions
- Trigonometric functions and their inverses
- Compositions of all the above

If you need a refresher on any of these, check out Paul's Online Calculus (and pre-calculus) Notes. Another extraordinarily effective way to refresh your memory is to volunteer for a few shifts of homework help for middle-school

and high-school students: you can often find a local library or Boys & Girls club that offers homework help to kids. Trying to explain algebra and pre-calculus concepts to a twelve- or seventeen-year-old will force you to think about these ideas as you never have before.

2.2 Derivatives and integrals

In this book, I'm assuming you took a calculus class in the past. You may need a reference, though. Any calculus book should have differentiation and integration covered, so find your favorite calculus book (or the cheapest).

Instead of rewriting a calculus textbook, I will concentrate on pointing out connections – building a higher-level appreciation of calculus and how it relates to probability and discrete mathematics. Calculus as we are taught it fundamentally takes place in the continuous world. We can't take the derivative at a discontinuous point of a function. This continuous point of view is always an approximation for finance, but it's very powerful!

2.2.1 Volume and area

Mathematicians don't tend to differentiate (haha!) between area, volume, hypervolume, and length: to a mathematician, they are all some sort of generalized "area." Think about why you see the $|\cdot|$ notation for absolute value of a number, length of an interval, size of a set, and determinant of a matrix. How are these quantities related? How are they different? Realizing the underlying unity of these ideas you were taught in different times and places will give you a lot of mathematical power. In this section, I want to point out how calculus relates to simple ideas of circumference or perimeter, area, and volume. Where will you put this to work, you ask? In transformations of random variables, in integration by parts, in deducing probability density functions from the structure of a probability problem, and in short-term and long-term approximation of financial and other quantities.

Circles and spheres. You remember that the area of a circle with radius r

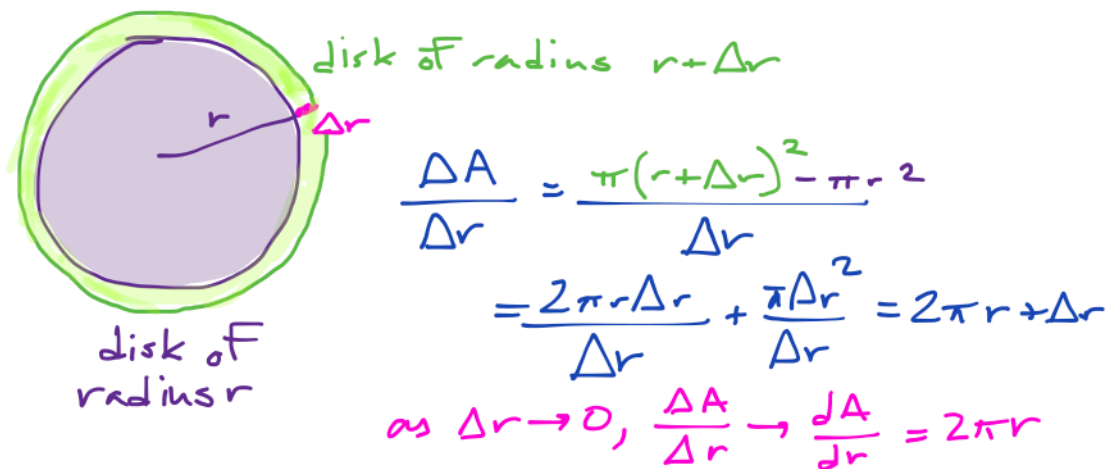
is $A(r) = \pi r^2$ and the circumference is $C(r) = 2\pi r$. Notice that

$$\frac{dA}{dr} = C(r).$$

The derivative of area is circumference. For spheres, the volume of a sphere with radius r is $V(r) = \frac{4}{3}\pi r^3$, while surface area is $A(r) = 4\pi r^2$. Again, differentiating the higher-dimensional volume gives the volume in the dimension one lower, if we expand what “volume” means to us:

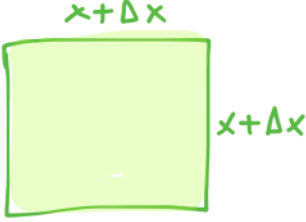
$$\frac{dV}{dr} = A(r).$$

Why does this work? A picture of the two-dimensional disk might help:



Notice that we use the notation Δr for a very small change in the r variable.

Squares. Try the naive analogue for a square with side length x : area of this square is x^2 and perimeter is $4x$ – the relationship doesn’t seem to hold! Can you see why?



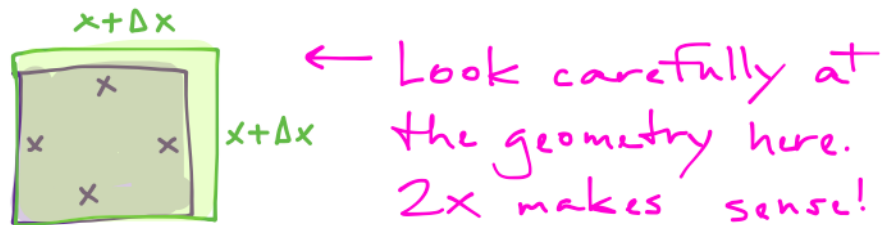
$$\frac{\Delta A}{\Delta x} = \frac{(x + \Delta x)^2 - x^2}{\Delta x}$$

$$= \frac{2x\Delta x + \Delta x^2}{\Delta x}$$

$$= 2x + \Delta x$$

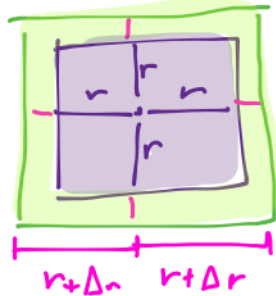
As $\Delta x \rightarrow 0$, $\frac{\Delta A}{\Delta x} \rightarrow \frac{dA}{dx} = 2x \neq \text{perimeter}$. Why?

Now look carefully at what you are actually doing here:



It turns out the procedure is working just fine – we are simply solving a different problem. The change in side length x gives us only half the perimeter of the square.

A more accurate analogue comes when we realize that our thinking about the square needs to be adjusted to match the reasoning for the circle. Let's use a new variable, $r = x/2$, and draw a new picture:



Let $x=2r$ and add Δr to each side. Now the geometry will give us perimeter.

$$\begin{aligned}\frac{\Delta A}{\Delta r} &= \frac{(2r+2\Delta r)^2 - (2r)^2}{\Delta r} \\ &= \frac{8r\Delta r + 4\Delta r^2}{\Delta r} = 8r + 4\Delta r\end{aligned}$$

As $\Delta r \rightarrow 0$, $\frac{\Delta A}{\Delta r} \rightarrow \frac{dA}{dr} = \text{Perimeter}$.

Here, notice that the area of the square is $A(r) = (2r)^2 = 4r^2$ and the perimeter is $P(r) = 8r$. Our change of viewpoint shows that the relationship we'd conjectured earlier holds.

2.2.2 Differentiation as infinitesimal approximation

The previous discussion of area and volume and their relationship might prompt you to look at how quickly a function $f(x)$ changes as you increase or decrease x . That's a derivative. Translating the sentence into mathematics, you considered the change

$$f(x + \Delta x) - f(x)$$

as the quantity

$$\Delta x$$

changed. The ratio is

$$\frac{f(x + \Delta x) - f(x)}{\Delta x},$$

which might start ringing alarm bells in your brain. That's right: a common definition for the derivative is

$$\frac{d}{dx}f(x) := \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

(Here the notation “:=” can be read as “is defined as”.) Our drawings in the previous discussions can be made into mathematically precise arguments using this definition!

This can be turned around. We can predict the value of $f(x)$ a “few moments” later, when the input is $x + \Delta x$. We want to predict $f(x + \Delta x)$. Turn our definition of the derivative inside-out to get the following:

If $f(x)$ is our output, then we can write the approximation (not equation!)

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x \quad (2.1)$$

Since we're working with a single input, x , and often graph this on the x-y plane, we can add $y = f(x)$ and write

$$y + \Delta y = f(x + \Delta x) \approx f(x) + f'(x)\Delta x \quad (2.2)$$

This allows us to figure out the change Δy in output and relate it to the change Δx in x :

$$\Delta y \approx f'(x)\Delta x. \quad (2.3)$$

Relate this to our discussion of volume and area to cement your understanding.

To make this all more precise, we can characterize how “good” or “bad” the approximation is. We introduce little-oh notation, $o(\cdot)$. Taking our definition from the National Institute of Standards and Technology, we say

- In words, $f(n) = o(g(n))$ if $f(n)$ becomes insignificant relative to $g(n)$ as n approaches infinity.
- With more symbols, “for all $c > 0$ there exists some $k > 0$ such that $0 \leq f(n) < cg(n)$ for all $n \geq k$. The value of k must not depend on n , but may depend on c .”

- With limits, this means $\lim_{n \rightarrow \infty} \frac{o(g(n))}{g(n)} = 0$.

This means we can write the equation

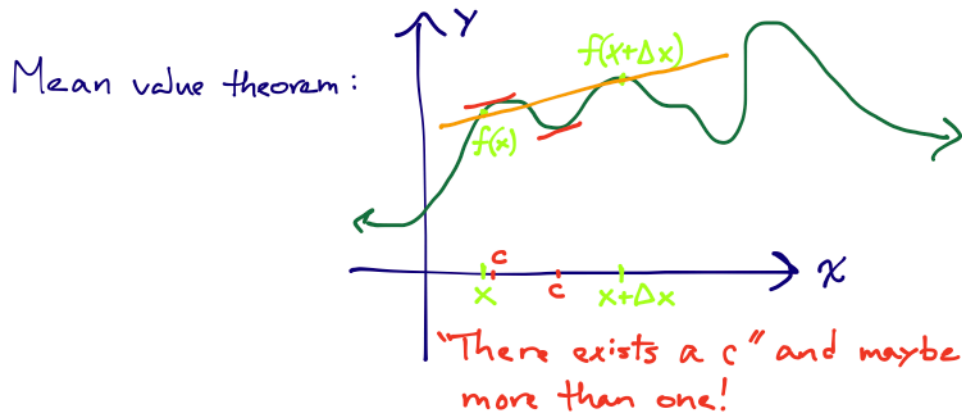
$$f(x + \Delta x) = f(x) + f'(x)\Delta x + o(\Delta x).$$

2.2.3 Average rate of change and Mean Value Theorem

For $f(x)$ a differentiable function,

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(c)$$

for some value of c between x and $x + \Delta x$. This means that at some point c , the derivative of f is equal to the average rate of change. A simple picture illustrates this nicely:



Remember, this requires that $f(x)$ is a differentiable function – in particular, $f(x)$ has to be continuous (at least near x). This Mean Value Theorem is very convenient for proofs and calculations in the continuous world, but it won't help you in the discrete world of stock prices. Just because the price of a share of Microsoft (MSFT) was \$50.05 at 12:55 pm on June 17, 2016 and was \$50.03 at 1:00 pm on the same day doesn't mean that it had a derivative of 2 cents per five minutes at some point in the middle!

You can derive many of the rules of differentiation using the short-term approximation formula, and doing so is actually very good practice to prepare for the world of stochastic calculus.

Example: to derive the power rule using this idea of short-term approximation, we need the binomial theorem:

$$(x + y)^p = \sum_{k=0}^p \binom{p}{k} x^{p-k} y^k$$

Using $f(x) = x^p$, look at

$$f(x + \Delta x) = (x + \Delta x)^p = \sum_{k=0}^p \binom{p}{k} x^{p-k} (\Delta x)^k.$$

Then short-term approximation tells us that

$$\sum_{k=0}^p \binom{p}{k} x^{p-k} (\Delta x)^k \approx x^p + f'(x) \Delta x.$$

Since we'd like to solve for $f'(x)$, rearrange:

$$f'(x) \approx \frac{\sum_{k=0}^p \binom{p}{k} x^{p-k} (\Delta x)^k - x^p}{\Delta x},$$

which simplifies nicely to

$$f'(x) \approx \frac{\sum_{k=1}^p \binom{p}{k} x^{p-k} (\Delta x)^k}{\Delta x}.$$

I'll rewrite it even more pointedly:

$$f'(x) \approx \frac{px^{p-1}\Delta x + \binom{p}{2}x^{p-2}(\Delta x)^2 + \dots + (\Delta x)^p}{\Delta x} = px^{p-1} + \Delta x \left(\sum_{k=2}^p \binom{p}{k} x^{p-k} (\Delta x)^{k-1} \right).$$

As $\Delta x \rightarrow 0$, this gives us $f'(x) = px^{p-1}$, the power rule we learned in calculus. You can make this more precise by using limits throughout and using the $o(x)$ notation to keep track of the error!

2.2.4 Integration and its disguises

Integration takes on many roles in the mathematical foundations of finance:

- Integration is a useful tool for calculation in solving probability problems.
- Riemann sums and their limits occur in problems and in proofs in probability.
- Probabilistic methods (Monte Carlo techniques) can be used for numerical integration, including value-at-risk and expected shortfall calculations.
- Understanding the relationships between integrals ($\int f(x)dx$), Riemann sums ($\sum_{i=1}^n f^*(i)$), and probabilistic or statistical ideas like expected values and moments is crucial to your ability to progress in mathematical finance.

Let's begin with the first comment: integration is necessary to solve geometric probability problems.

Consider the square with corners at $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. Pick a point (x, y) in this square at random. What is the probability that the product of the coordinates, xy , is greater than one half?

In the last chapter, we dealt with finite and countably infinite sample spaces. There, we could assign a probability to each individual possible outcome $\omega \in \Omega$ and sum over these outcomes. Here, we could try this but we encounter a problem. What's the probability that I pick the point $(0.5, 0.6)$ exactly? A mathematical point has area zero, and the chance that I hit exactly that point is zero. However, the chance that I pick a point in a continuous region A of the unit square is the area of A (because the area of the unit square is one, and when we say "pick a point at random" without specifying further, every point is equally likely). How can this be? How can the sum of infinitely many zeroes be a positive number?

You actually encountered this in calculus. Recall that

$$\int_a^b 3dx = 3(b - a),$$

for instance, and

$$\int_a^a f(x)dx = 0$$

for any continuous function $f(x)$. The integral over a point is always zero even though the integral over an interval may not be. This is because \mathbb{R} is *uncountable* as you proved in Chapter 1, and uncountable sets present dramatically more subtle situations than finite or countable sets.

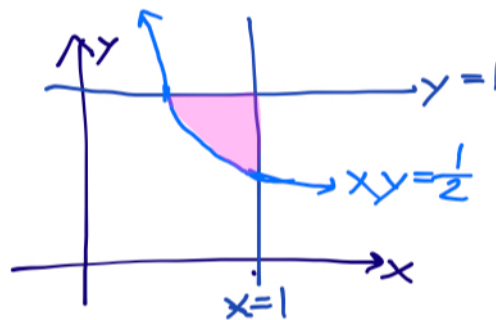
Back to our problem: Pick a point (x, y) in the unit square at random. What is the probability that the product of the coordinates, xy , is greater than one half?

Step 1: identify our sample space Ω . It's all the points in the unit square.

Step 2: identify the set A of outcomes we're interested in. Here

$$A = \{(x, y) \in \Omega \mid xy \geq \frac{1}{2}\}.$$

Step 3: Calculate the sizes of Ω and $A \subset \Omega$. We know Ω has area 1. What is the area of A ? Draw a picture:



Find the shaded area using an integral. Notice that it's the integral of $1 - \frac{1}{2x}$:

$$\int_{1/2}^1 dx - \int_{1/2}^1 \frac{dx}{2x} = \left(1 - \frac{1}{2}\right) - \frac{\ln x}{2} \Big|_{1/2}^1 = \frac{1}{2} \left(1 - (\ln 1 - \ln \frac{1}{2})\right) = \frac{1}{2} - \frac{\ln 2}{2}.$$

We're done! yay! WRONG. While trivial in this case, remember to divide by the size of the original sample space.

Step 4: Calculate $P(A) = \frac{|A|}{|\Omega|}$. Since $|\Omega| = 1$, we have

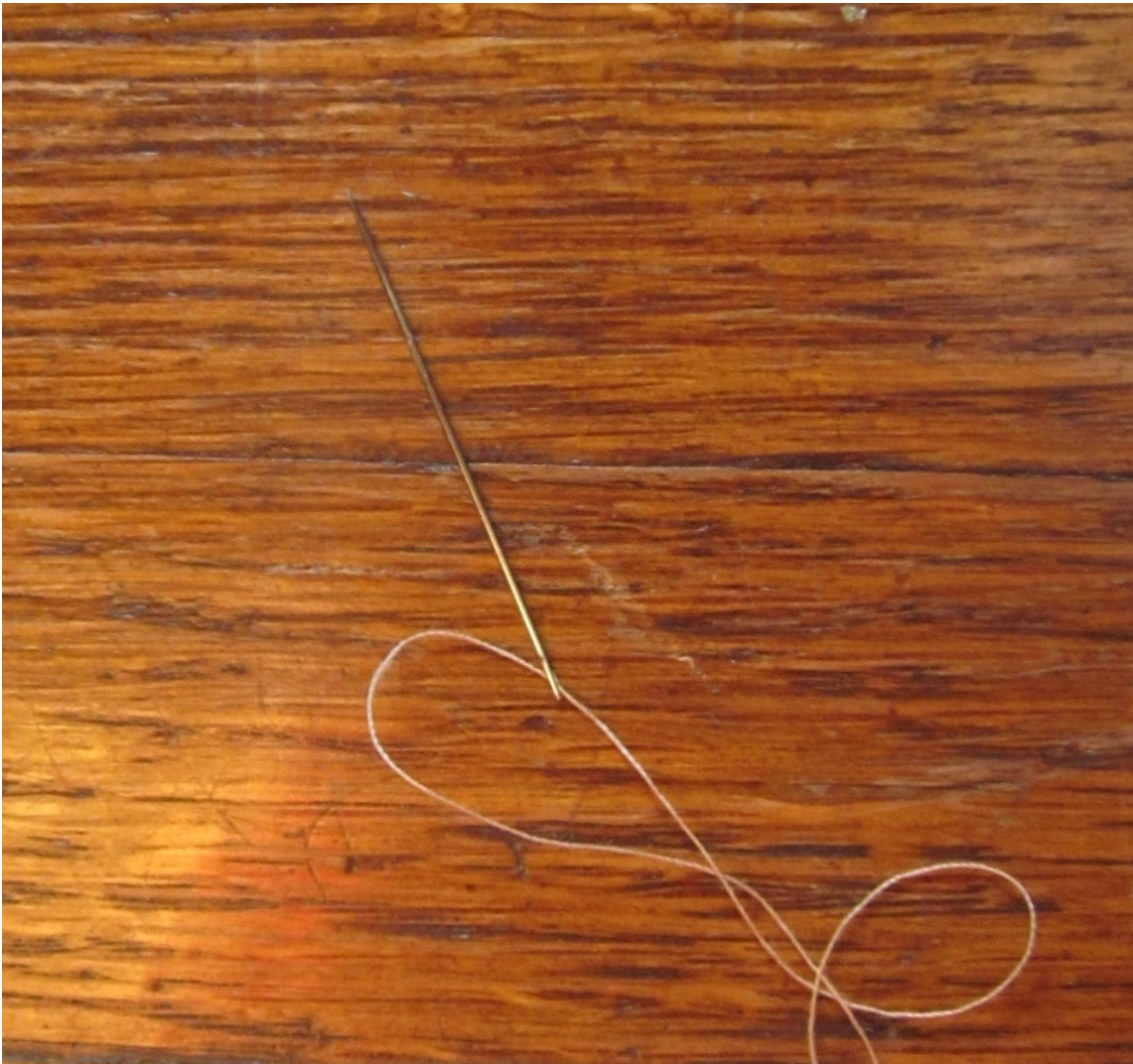
$$P(A) = \frac{1}{2} - \frac{\ln 2}{2}.$$

More examples of probability problems that use integration will be given in the last two sections of this chapter.

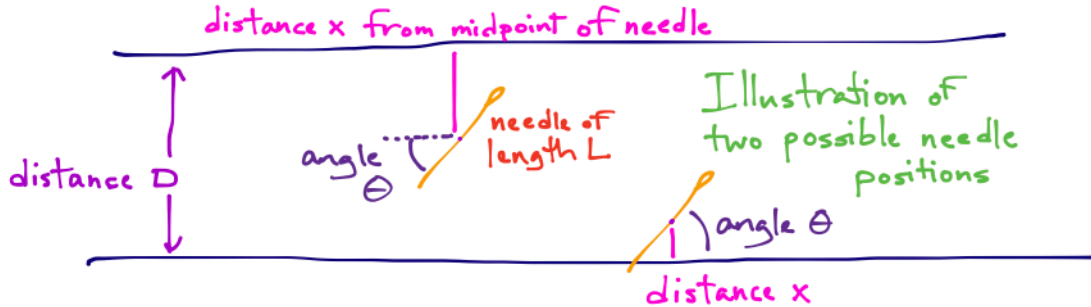
2.3 Buffon's needle problem

Buffon's needle problem is a classic probabilistic problem. It was first written down by Georges-Louis Leclerc, Comte de Buffon, a French intellectual who lived in the 1700s. He asked the following question in 1733: Consider a floor ruled with parallel lines. Drop a needle onto the floor. What is the probability that the needle will cross one of the lines?

There are two situations to consider, one with a needle shorter than the distance between the lines and one with a needle longer than the distance between the lines. We'll start by considering a short needle.



An illustration to set our notation is useful:



The length of the needle is L , the distance between the lines on the floor is D , and the angle that the needle makes with the lines is θ . We'll measure θ counterclockwise from the horizontal. In a given probability experiment (dropping the needle once) we'll label the distance from the middle of the needle to the nearest line on the floor by x .

There are several ways to solve Buffon's problem. For simplicity, I will present a method that follows our format of specifying Ω , finding the event set A that we desire, and comparing the sizes of these sets.

Using the notation of the illustration, we write

$$\Omega = \{(\theta, x) \mid 0 \leq \theta \leq \pi, 0 \leq x \leq \frac{D}{2}\}.$$

For yourself, check the following: Why don't we specify how far "over" the needle falls (horizontal position)? Why don't we use $0 \leq \theta < 2\pi$?¹

The event set A , then, is when

1. the needle is close enough to the line to possibly cross the line, and
2. the angle of the needle is such that it actually does cross the line.

¹We don't specify horizontal position because it's not necessary. If we specify horizontal position in Σ and A they'll just cancel out, because horizontal translation doesn't affect whether or not the needle crosses the line. Similarly, we restrict θ so that each position for the needle is "counted" only once.

Mathematically, write

$$A = \{(\theta, x) \in \Omega \mid \frac{L}{2} \sin \theta > x\}.$$

The area of Ω is $\frac{D\pi}{2}$, as it's a rectangle in the $\theta - x$ plane with side lengths $\frac{D}{2}$ and π . To find the area of A in Ω you must integrate either $\sin \theta$ with respect to θ or $\sin^{-1} x$ with respect to x . One choice is clearly easier than the other :

$$|A| = \int_0^\pi \frac{L}{2} \sin \theta d\theta = L.$$

Then

$$P(A) = \frac{|A|}{|\Omega|} = \frac{2L}{D\pi}.$$

The long-needle version of the problem (with $L > D$) can be solved in a similar way, but since the needle can cross more than one line a little more analysis is needed.



For now, we will ignore the long needle version of Buffon's problem. We'll come back to it in the section about joint probability density functions, a great tool which will allow the solution to be given in just a line or two.

2.4 Stick-breaking problems

Stick-breaking problems give students of probability all kinds of trouble. They're great! Deceptively simple, these stick-breaking problems illustrate how small changes to a problem can dramatically change the techniques necessary to solve it. They also show that these simple mathematical puzzles are actually structures that show up in all sorts of natural situations, like Dirichlet processes. Bizarrely, the last version is apparently an interview question at Goldman-Sachs. (See <http://www.wallstreetoasis.com/blog/interview-questions-ib-analyst-goldman-sachs> for reference.)

Version one: A stick of length L is broken in two places. The break points are independent of each other, and chosen at random (uniformly) on the stick. What is the probability that a triangle can be formed using the three pieces of the stick?

Version two: You and a friend are breaking the stick now: you each grasp one end of the stick and, together, break it at one point. Then you take the piece of the stick left in *your* hand and break it again. What is the probability that a triangle can be formed using the three pieces of the stick?

Version three: You and a friend are breaking the stick now: you each grasp one end of the stick and, together, break it at one point. Then you take the *longer* of the two pieces and break it again. What is the probability that a triangle can be formed using the three pieces of the stick?

So.... solutions? Try these problems first and then look at the solutions one by one, as you might be able to clear up misunderstandings about the structure of the problem by looking at solutions individually.

Before discussing solutions, I want to bring up a subtle but important point: you can use variables in this problem to represent either *lengths* or *locations*. That is, x might represent the length of a piece of the stick, or it might represent a location along the stick. This makes a big difference in that locations are given with respect to a single origin point.

Solution to version one: First, let $L = 1$. We can rescale and just say that L inches (or meters) is 1 unit of our own measurement system. Call the distance from one end of the stick to the nearest breakpoint x , and the distance from the

other end of the stick to its nearest breakpoint y . The two breaks happen at the exact same point of the stick with probability zero, so we don't have to worry about the contribution from breaking the stick into only two pieces. We'll have three pieces of the stick after breaking, of lengths x , y , and $1 - x - y$. The most important thing to notice here is the domain for x and y : x and y can be any values between 0 and 1 such that $x + y \leq 1$, and on this triangular set, every point (x, y) is as likely as any other!

To form a triangle, these lengths must satisfy the triangle inequality (the sum of any two sides of a triangle must be larger than or equal to the third side):



$$(1) \quad x + y \geq 1 - x - y,$$

$$(2) \quad x + 1 - x - y \geq y,$$

$$(3) \quad y + 1 - x - y \geq x.$$

Simplify to get

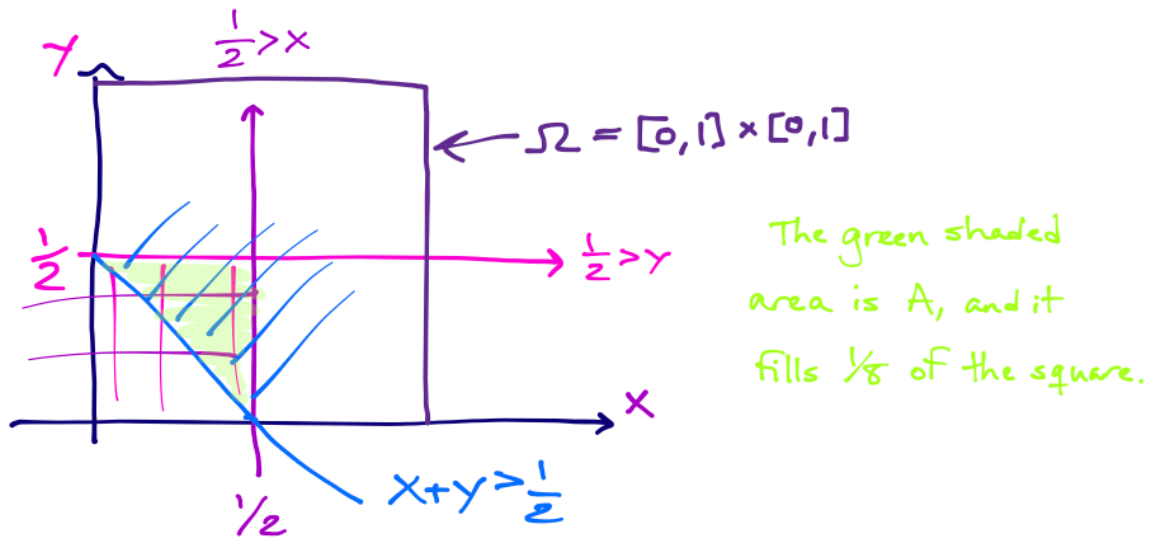
$$(1') \quad x + y \geq 1/2,$$

$$(2') \quad 1/2 \geq y,$$

$$(3') \quad 1/2 \geq x.$$

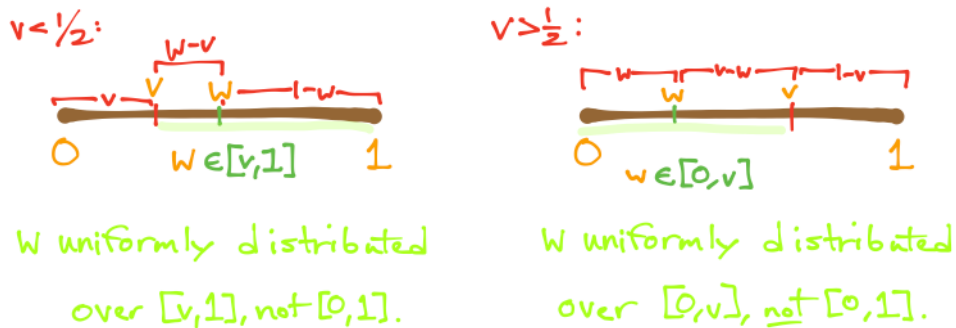
Since we know $0 \leq x \leq 1$ and $0 \leq y \leq 1$, and also $x + y \leq 1$, graph the sample space $\Omega = \{(x, y) | 0 \leq x, y \leq 1, x + y \leq 1\}$ and the event space $A = \{(x, y) | 0 \leq x, y \leq 1/2, x + y \geq 1/2\}$. You can use symmetry to see that

$$P(A) = \frac{1/8}{1/2} = \frac{1}{4}.$$



Solution to version three: You and a friend each grasp one end of the stick and, together, break it at one point. Then you take the longer piece and break it.

Here we will use locations rather than lengths. Again rescale the stick so it's length one, and label one end of the stick zero and the other end one. The first breakpoint $v \in [0, 1]$ occurs anywhere along the stick. If $v < \frac{1}{2}$, then the second breakpoint w will be chosen at random somewhere in the interval $[v, 1]$, while if $v > \frac{1}{2}$ the second breakpoint w is chosen at random on the interval $[0, v]$. Now, the problem is that the possible values of w depend on v – these are not independent quantities anymore, and w is not uniformly distributed on $[0, 1]$! We only have the tools for uniform distributions at this point.



To solve the problem today, we need to turn it into a problem with uniform distributions (every point as likely as every other point). The alternative is

to learn about continuous random variables and probability density functions, which won't happen immediately. So.... Let's use this rescaling trick again.

- If $v < \frac{1}{2}$, then let $v = x$ and $w = y \cdot (1 - v)$, for $y \in [0, 1]$. In essence, we're letting $v = x$ be whatever it is and then treating the stick of length $1 - v$ as its own new stick to rescale to length 1, and y tells us where along this the next breakpoint is.
- if $v > \frac{1}{2}$, let $v = x$ and $w = y \cdot v$, y again in $[0, 1]$. Again, we now have y representing the fraction of the longer stick at which the next break happens.
- Putting these situations together, we have x chosen randomly (uniformly) in the interval $[0, 1]$, and we have y chosen uniformly in the interval $[0, 1]$, and we've transformed the variables so that each point (x, y) in the unit square is as likely as any other.

Now we have to do some algebra and calculus, because we need to figure out which points (x, y) in the unit square correspond to (v, w) which give a triangle. Use the triangle inequalities for (v, w) and convert:

For $v < \frac{1}{2}$, we have three lengths: v , $w - v$, and $1 - w$. The triangle inequalities are

$$v + w - v > 1 - w,$$

$$v + 1 - w > w - v,$$

and

$$1 - w + w - v > v.$$

Simplifying,

$$w > 1/2,$$

$$2v + 1 > 2w,$$

and

$$1/2 > v.$$

Convert to x and y :

$$y(1 - x) > 1/2$$

and

$$1 + 2x > 2y(1 - x).$$

This gives us two inequalities,

$$y > \frac{1}{2(1 - x)}$$

and

$$y < \frac{1 + 2x}{2(1 - x)}.$$

We must integrate between these two curves:

$$\int_0^{1/2} \frac{1 + 2x}{2(1 - x)} - \frac{1}{2(1 - x)} dx = \ln 2 - 0.5.$$

Now carry out the same process for $v > 1/2$ and add the answers to get the total probability, which is

$$2[\ln 2 - 0.5].$$

The solution to the second stick problem is just $\ln 2 - 0.5$ – adapt the argument above to get this answer!

Chapter 3

Probability rules

3.1 Basic rules of probability

3.1.1 Complements and inclusion-exclusion

Essentially every concept in basic set theory gives rise to an analogous rule in probability.

If two sets of events A and B are not disjoint, there is an overlap $A \cap B = AB \neq \emptyset$. This intersection of A and B is overcounted when we look at the union of A and B , so we must remove it to find the size of $A \cup B$. In symbols,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Probabilistically, this implies

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

We encountered disjoint sets earlier: A and B are disjoint if $A \cap B = \emptyset$. In this case, the sizes of the sets add:

$$|A \cup B| = |A| + |B|.$$

This is a special case of the previous observation because $|\emptyset| = 0$. When A and B are disjoint sets, then we call the events in A and B *mutually exclusive*

events. The probability of either of two mutually exclusive events is the sum of the probabilities:

$$P(A \cup B) = P(A) + P(B),$$

which is again a special case because the probability of nothing happening ($P(\emptyset)$) is zero.

Another set-theoretic idea is that of the *complement* A^c of a set A . We consider A as a subset of some sample space Ω , so $A \subset \Omega$. Then the complement of A is everything in Ω except the elements of A :

$$A^c = \{a \in \Omega | a \notin A\}.$$

From the axioms of probability, we know that $P(\Omega) = 1$ for Ω the entire sample space. From the definition of A^c , we know that A and A^c are disjoint sets that partition Ω . Since $A \cup A^c = \Omega$, we also have $P(A) + P(A^c) = P(\Omega) = 1$. This implies

$$P(A^c) = 1 - P(A).$$

Take the preceding rules a little further to deduce the *inclusion-exclusion* rule. Instead of two sets A and B , what if you've got multiple subsets A_1, \dots, A_n of Ω , and you want to compute $P(A_1 \cup \dots \cup A_n)$? If all the A_i are pairwise disjoint you can simply sum the probabilities of the individual sets A_i . However, if there are overlaps, you need to consider intersections, triple intersections, quadruple intersections...

$$|\cup_{i=1}^n A_i| = \sum_{i=1}^n |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \dots + (-1)^{n-1} |A_1 \cap \dots \cap A_n|.$$

Probabilistically,

$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P(\cap_{i=1}^n A_i).$$

All useful tools in your toolbox!!

3.1.2 The hat check problem

A classic type of inclusion-exclusion problem has a form like the following:

- At a theater men leave their hats at the hat check counter. The irresponsible clerk loses everyone's ticket and so after the performance, he randomly returns hats to men who come back to retrieve them. What is the probability that no men get the correct hat returned? (Back when men wore hats and left them at a hat check before a movie, women also wore hats but kept them on as the hat was considered part of the woman's outfit.)
- You're sending thank-you notes after your wedding. After writing them all, you have a pile of notes and a pile of addressed envelopes. A "helpful" friend stuffs all the envelopes and sends the notes but doesn't realize the two piles are not in the same order – so random letters are put in random envelopes. What is the probability that at least one person gets the correct thank-you note?
- Rabbits are playing in a field outside their burrows. They are suddenly surprised by an eagle, so each rabbit runs to a separate hole and escapes into a burrow. What is the probability that no rabbit escapes down its own hole?¹

These are variations on what might be called the "derangement" problem. Derangement does not refer to being a deranged person, despite what your probability homework might lead you to believe or feel. Instead, a derangement is a permutation with no fixed points – a rearrangement of objects so that no object stays in the same place.

Let's outline a general way of solving these problems. Say there are n elements in the set of objects (hats, letters, rabbits). Let A_i be the event that object i ends up in the correct place (hat with correct man, letter with correct envelope, rabbit in own burrow). This is a key point: rather than looking at all possible arrangements of hats, look instead at hat i only.

¹Asked by Jarrad Smith on MathForum.org in 1999

Set-theoretically, then, the event that any hat returns to its correct owner is $\cup_{i=1}^n A_i$. We can use inclusion-exclusion to find $P(\cup_{i=1}^n A_i)$. Here are the pieces: $P(A_i) = 1/n$, $P(A_i A_j) = \frac{1}{n} \frac{1}{n-1}$, $P(A_i A_j A_k) = \frac{1}{n} \frac{1}{n-1} \frac{1}{n-2}$, etc. Combine them to get:

$$P(\cup_{i=1}^n A_i) = \binom{n}{1} \frac{1}{n} - \binom{n}{2} \frac{1}{n} \frac{1}{n-1} + \binom{n}{3} \frac{1}{n} \frac{1}{n-1} \frac{1}{n-2} - \dots + (-1)^n P(\cap_{i=1}^n A_i).$$

Notice this simplifies:

$$\binom{n}{k} \frac{(n-k)!}{n!} = \frac{n!}{(n-k)!k!} \frac{(n-k)!}{n!} = \frac{1}{k!},$$

so

$$P(\cup_{i=1}^n A_i) = \sum_{k=1}^n (-1)^{k-1} \frac{1}{k!}.$$

This is the probability that at least one object goes to the right place. To find the probability that no objects go to the right place, subtract the above from 1.

3.1.3 The Secret Santa problem

Here's a slightly different problem. What makes it different?

There are ten people in a family that wants to exchange gifts, and they have agreed to have a "Secret Santa" gift exchange. This means that each person randomly draws the name of a person not themselves and gives a gift only to that person, rather than giving all nine others a gift. What is the probability that (at least) two people are assigned to each other? That is, among the ten people, what is the probability that Anna is to give a gift to Boyuan and Boyuan is to give a gift to Anna – and possibly, Chimamanda and Danica are assigned to each other as well?

One way to try to solve this is to number the people from one through ten and look at all permutations of one through ten that do *not* contain two-cycles. This is certainly doable. You can use the Online Encyclopedia of Integer sequences to look up this number if you can't compute it yourself (check it out at <http://oeis.org/A000266>).

Another way to look at it, though, is to look at the events where two people *are* assigned each other. Again, rather than looking at ten-tuples, instead consider events A_k that correspond to two people being assigned to each other. The probability we desire, then, is $P(\cup_{k=1}^{45} A_k)$. (Why 45? There are ten people, from whom we draw a pair – that’s ten choose two pairs.)

The probability of any pair A_k being assigned to each other can be calculated easily. It’s $\frac{1}{9 \cdot 9}$. Then we need to compute the probability of $A_i \cap A_k$. These events may be mutually exclusive, in which case the probability is zero, but if the events are not mutually exclusive, then $P(A_i \cap A_k) = \frac{1}{9^4}$. Any particular triple intersection is either impossible (probability zero) or has probability $P(A_i \cap A_j \cap A_k) = \frac{1}{9^6}$, and you can follow the trend to compute the quadruple and quintuple intersections. Then to complete the the calculation we need to do some counting.

How many A_i ? We know there are $\binom{10}{2}$ events A_i . How many non-mutually-exclusive A_i and A_k ? To make sure they’re not mutually exclusive events, we pick 2 from 10 and then 2 from the remaining 8, so $\binom{10}{2} \binom{8}{2}$. Continue with this line of reasoning:

$$P(\cup_{i=1}^{45} A_i) = \binom{10}{2} \frac{1}{9^2} \tag{3.1}$$

$$- \binom{10}{2} \binom{8}{2} \frac{1}{9^4} \tag{3.2}$$

$$+ \binom{10}{2} \binom{8}{2} \binom{6}{2} \frac{1}{9^6} \tag{3.3}$$

$$- \binom{10}{2} \binom{8}{2} \binom{6}{2} \binom{4}{2} \frac{1}{9^8} \tag{3.4}$$

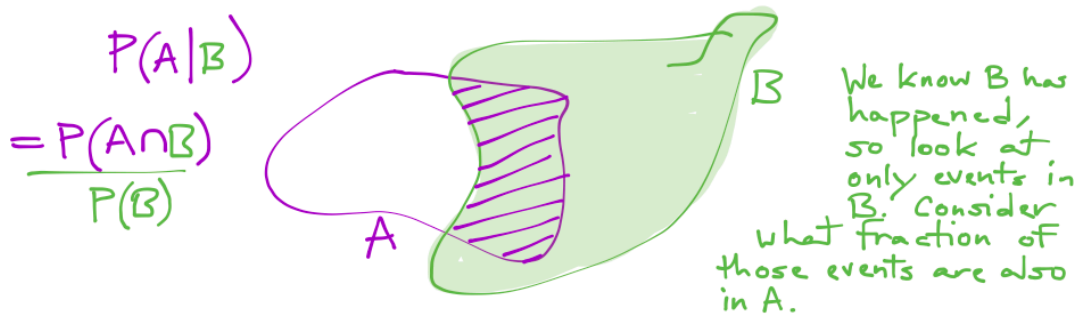
$$+ \binom{10}{2} \binom{8}{2} \binom{6}{2} \binom{4}{2} \binom{2}{2} \frac{1}{9^{10}} \tag{3.5}$$

3.2 Conditional probability

Define the probability of the event A given that event B happens as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Think about how to visualize this: I see it as



Notice that re-writing this gives

$$P(A|B)P(B) = P(A \cap B) = P(AB).$$

At times this is a useful way to compute the probability of two events happening simultaneously.

As an example, let's revisit a problem you did via counting in [Chapter 1](#).

Eight teams are in the semifinals of an international badminton tournament. The eight teams consist of two teams each from China, India, Denmark, and Korea. What is the probability that the two teams from each country end up playing against each other in each of the semifinal matches? That is, China 1 plays China 2, Denmark 1 plays Denmark 2, etc.

You can nicely do this with conditional probabilities instead of counting techniques:

Some team is picked to play in match 1 with probability 1. This team is from country A. Given that country A is picked to play in the first match, the probability that their opponent will be from the same country is $\frac{1}{7}$. Given

these events, any remaining team can be picked to play in the second semifinal match – say it's from country B. Given all these, the probability that country B's opponent is also from country B is $\frac{1}{5}$. Given...

Maybe you see a pattern here! In the end, we get that the desired probability is

$$1 \cdot \frac{1}{7} \cdot 1 \cdot \frac{1}{5} \cdot 1 \cdot \frac{1}{3} = \frac{1}{105}.$$

Much less counting needed!

3.2.1 Law of Total Probability

One of the most useful ways to use conditional probability – even if you don't think you need it – is to partition a sample space or event space over mutually exclusive events. Say you're interested in the probability of A . Then partition Ω into mutually exclusive events B_i for $i = 1, \dots, n$, so that $\Omega = \cup_{i=1}^n B_i$. In this situation,

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i).$$

This is called the law of total probability.

This has some very straightforward applications and can also be useful in more subtle ways. First, a straightforward application. You are playing a game where a friend you trust not to cheat will give you a die or a coin, chosen randomly from a bag, and if you either roll a 6 or flip heads you'll win. What's the probability you'll win?

Partition the sample space into B_1 , the event that you're given the die, and B_2 , the event that you're given the coin. One of the two must happen, and they encompass the whole sample space! Each is equally likely (probability $1/2$). Then if A is the event that you win, $P(A|B_1) = 1/6$ and $P(A|B_2) = 1/2$. So

the final calculation is

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \quad (3.6)$$

$$= \frac{1}{6} \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{1}{2} \quad (3.7)$$

$$= \frac{1}{6}. \quad (3.8)$$

3.2.2 Recursive games

A more subtle application of this law of total probability is when one can condition over previous events in a sequence. Let's take a look at an example.

Roll a standard die until you "win" or "lose." You win if you roll a 2; you lose if you roll two odd numbers (they need not be consecutive). So if you roll 1 and then 3 you've lost. If you roll 4,4,6,6,4,4,4... you need to keep playing as you've neither won nor lost. If you roll 1 then 2, you won. What's the probability of winning, $P(A)$?

While there are several ways of doing this, let's look at conditioning on the result of the first roll.

- If you roll a 2 right away, you won. This happens with probability $1/6$.
- If you roll a 4 or a 6, you don't care about the roll. It does not affect the outcome.
- If you roll an odd number (1, 3, or 5) you have one strike against you. Start at the top of this list and go through the decision tree again, and if you end up here, you lose.

Alright. Let B_1 be the event that you roll a 2 on the first roll. Let B_2 be the event that you roll a 4 or 6. Let B_3 be the event that you roll a 1, 3, or 5 on the first roll. These partition the set of first rolls into mutually exclusive sets. Write $P(A) = p$. The law of total probability says

$$P(A) = p = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \quad (3.9)$$

$$= 1 \cdot \frac{1}{6} + p \cdot \frac{2}{6} + \frac{1}{6} \cdot \frac{3}{6}. \quad (3.10)$$

$P(A|B_3) = \frac{1}{4}$ because we want the probability we roll a 2 before a single odd number (3 odds, 1 two).

Solve for p . You get $p = \frac{7}{16}$.

This is just one example – you can get much more complex recurrence relations, for which more advanced mathematical techniques are needed. For instance, use these methods to consider the probability of getting two consecutive heads in n coin flips.

3.3 Bayes' theorem

Bayes' theorem is an extraordinarily useful tool, allowing its user to adjust estimates of probabilities when new evidence is made available. The theorem uses conditional probability to find the probability of a hypothesis H given evidence E . Start with the definition of conditional probability, which tells us:

$$P(E|H) = \frac{P(EH)}{P(H)}, \quad P(H|E) = \frac{P(EH)}{P(E)}.$$

Do a bit of algebra to notice the following:

$$P(H|E)P(E) = P(EH) = P(E|H)P(H).$$

From here, since we want to find $P(H|E)$, we can proceed in two ways: we can find the “odds form” of Bayes' theorem or we can find a formula for $P(H|E)$ directly. In either case, we will need to consider the probability that our hypothesis event H did not occur: we will write \overline{H} , for the complement of H in a set-theoretic sense or the negation of H in the context of logic.

The direct formula for Bayes' theorem is often used: use algebra to get from

$$P(H|E)P(E) = P(E|H)P(H)$$

to

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

For the final step, use the law of total probability to write $P(E)$ as a sum by splitting the event E over mutually exclusive H and \bar{H} :

$$P(E) = P(E|H)P(H) + P(E|\bar{H})P(\bar{H}).$$

This gives

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})},$$

a formula memorized by probability and statistics students around the world. The method can also be generalized to multiple hypotheses H_1, \dots, H_n as long as they're mutually exclusive.

The odds form is also easily derived: simply take the ratio of

$$P(H|E)P(E) = P(E|H)P(H)$$

and

$$P(\bar{H}|E)P(E) = P(E|\bar{H})P(\bar{H}).$$

The final result is

$$\frac{P(H|E)}{P(\bar{H}|E)} = \frac{P(H)P(E|H)}{P(\bar{H})P(E|\bar{H})}.$$

Why bother with this slightly less direct formulation? Because it presents the information with a nice structure.

- First, it states Bayes' result as "prior odds" times a "likelihood ratio", giving the "posterior odds" in which the evidence has been taken into account.
- Second, the odds format often allows you to use natural frequencies and avoid calculating probabilities. Natural frequencies provide information like, "It rains on average once every ten days," or, "An estimated 1 out of 133 Americans has celiac disease."

The only danger is that odds really are ratios of occurrences of dissimilar events – days with rain versus days without rain, for instance – rather than the fractions we're most familiar with, like days with rain as a proportion of all

days. Check what sets you're dealing with if you get confused about whether you have an odds ratio or fraction: odds compare the sizes of two sets that partition a larger set, while fractions and probabilities deal with the size of a subset in the numerator compared to the size of the whole set in the denominator.

Back to the statement of the theorem: The *prior odds* of the hypothesis event occurring are

$$\frac{P(H)}{P(\bar{H})}.$$

In the examples above, the prior odds of rain are 1 : 9 or $\frac{1}{9}$, and the a priori odds of an American having celiac disease are 1 : 132 or $\frac{1}{132}$. Taking evidence into account involves multiplying by the *likelihood ratio*, which gives information about how likely the evidence is if the hypothesis does or doesn't hold true. If $P(E|H) > P(E|\bar{H})$, the *posterior odds* describing the likelihood of H in the face of E are increased – the evidence has added to the probability that the hypothesis is true. Many of the most surprising results from Bayes' theorem, though, come because the likelihood ratio is not what people expect.

First let's consider a somewhat famous example that illustrates both the difficulties many people have with probabilistic reasoning and a good use of Bayes' theorem.

(Gigerenzer, "Calculated Risks," 2004) Doctors were told the following: you've got a patient with a positive mammogram (E). She is between 40 and 50 years old, with no symptoms or family history of breast cancer. The a priori probability that such a woman has breast cancer (H) is 0.8 percent. If she's got breast cancer, the mammogram will be positive 90% of the time, and if she does not have breast cancer, there's still a 7% chance that she has a positive mammogram. Given all this information, what's the probability that your patient actually has breast cancer?

What do you estimate, reader? Gerd Gigerenzer, a cognitive psychologist, asked doctors to do this experiment. Their estimates ranges from 1 percent to 90 percent. As a patient, how would you feel if you were told that there is 90% chance you have this cancer – especially when the true answer is around 9%?!

Using the probabilities as stated, we know that the prior odds that this patient has breast cancer are $\frac{0.8}{99.2}$. (Small odds!) The likelihood ratio is $\frac{90}{7}$. This is

big, but only about a factor of 13. The posterior odds, then, are $\frac{72}{694.4}$. When we solve for the probability $P(H|E)$, we find

$$P(H|E) = \frac{72}{72 + 694.4} \approx 0.09.$$

See Girgerenzer's book for more examples of how bad we are at estimating probabilities. These cognitive psychology experiments are very useful in showing us where we fall into cognitive traps, overestimating the probability of rare events, underestimating the probability of common events, and showing "anchoring" tendencies, in which we estimate that quantities are close to numbers coming from preconceived notions. While much of finance involves careful calculations in spreadsheets using approved models, we in numerical fields must still make quick estimates to check the validity of our work. Avoiding these cognitive traps will make you more effective in many ways!

Chapter 4

First applications in finance

It is amazing to me that a simple model – the binomial tree model for stock prices – is so effective in pricing various types of financial derivatives. In this chapter, we'll consider financial ideas (arbitrage and stock pricing) and additional tools useful for finance (approximation, including Taylor series). We'll also see why we need to step up our sophistication and introduce concepts like random variables and their expected values in the next chapter.

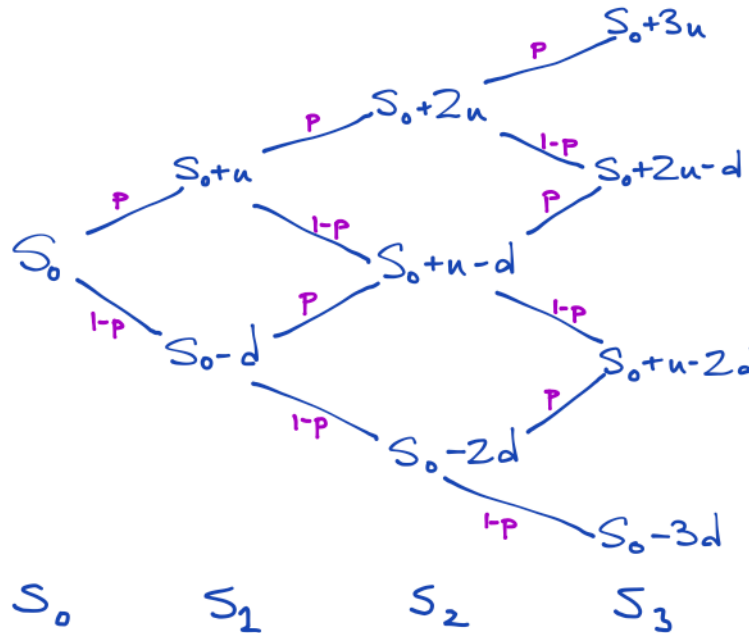
4.1 Binomial stock pricing

At 2:05 pm on a Tuesday, a stock has a given price listed on the New York Stock Exchange. As orders come in, that price will go up, go down, or stay the same. A first approximation – the binomial tree model for stock prices – is to pick a unit of time and to model the stock price at each unit step.

4.1.1 Additive model

The first idea you might have would be to start with an initial stock price, S_0 , and then add u dollars or subtract d dollars on each step. At time one, then, the price would be $S_1 = S_0 + u$ or $S_1 = S_0 - d$. Here, let $u, d > 0$. In constructing the model, you'd want to find the appropriate probability p of the stock price

increasing and the probability $1 - p$ of the price decreasing. Continue this over subsequent time steps.



Notice that the price at any time step t is the sum of two arithmetic series.

As always, you should evaluate the benefits and drawbacks of such a model. A benefit is that the model is quite simple, and easy to extend to any number of time steps desired. A huge drawback is that the model allows negative stock prices. This is unrealistic: while we can short a stock, the price of the stock itself can never be negative. In addition, consider the sizes of u and d : if S_0 is \$400, a change of \$1.00 in the stock price is not that significant, while if S_0 is \$4, a change of \$1.00 is an enormous change in value.

A last factor to consider is that this is a discrete model of a market that trades in continuous time. Whether this is a benefit or drawback is to be argued – what do you think? We will keep discussing the relationship between continuous and discrete models: remember that a theme of our initial probability discussion was the back-and-forth between combinatorial and geometric probability problems.

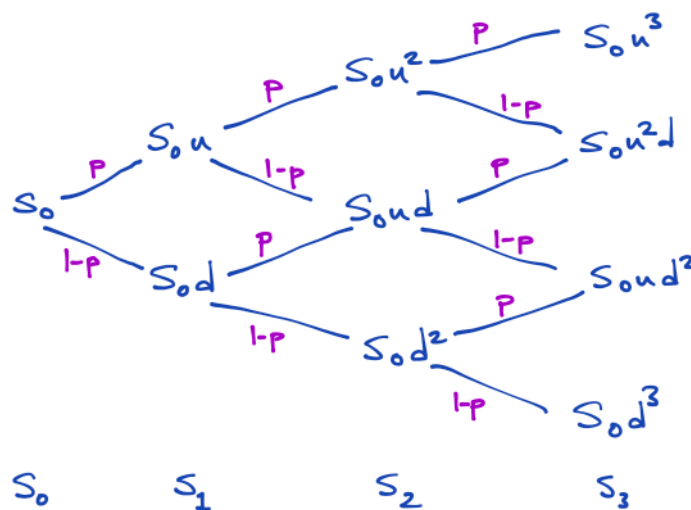
This additive tree model is really not very good for stock prices, but we shouldn't dismiss it entirely. It will lead us on a random walk through the cen-

tral limit theorem, normal distributions as a limit of repeated Bernoulli trials, and finally arithmetic Brownian motion.

4.1.2 Multiplicative model

To fix at least one problem with the additive binomial tree model, we can change to a multiplicative pricing model. The initial price S_0 changes by a percentage: with probability p the stock price increases by a factor of u and with probability $1 - p$ the stock price decreases by a factor of d . In general, we set $0 < d < 1 < u$. You will see why when we discuss arbitrage.

Construct the binomial tree as follows, then:



The stock price will never fall below zero in this model, and changes in price are proportional to the original price. This is still a discrete-time model, but financial mathematicians have found it quite useful. Here, notice that the maximum price at each time makes up a geometric sequence with common ratio u and the minimum price at each time makes up a geometric sequence with common ratio d .

From this model, you can do two things given u , d , and p : you can find the probability that the stock has a certain price at a given time step, and you can find all the possible prices at a given time step.

What are all the possible prices for this stock at time step n ?

Use the binomial theorem to get all prices at time n . S_n must take on a value that looks like $S_0 u^k d^{n-k}$ for some k between zero and n . Ask yourself: does k refer to steps “up” or steps “down” in price?

Is this model realistic? Can we simply say any numbers for u , d , and p ? Are these quantities related? What are the pros and cons?

4.2 Arbitrage

The no-arbitrage principle in finance and economics answers one of these questions. The no-arbitrage principle claims that “there is no free lunch” – you can’t make guaranteed money. If the no-arbitrage principle holds, then p is uniquely determined by u and d . We can show this rigorously later.

To make this discussion easier, we need to introduce the concepts of *random variable* and then *expected value*. Notice that in each of the binomial tree illustrations above there are notations along the bottom: S_0, S_1, S_2, S_3 . The notation S_i means essentially “the outcome of the probability experiment at step i .” It is a random variable, which is a confusing name for the following reasons: in much of math, a variable is a letter that we use in expressing a mathematical condition like $x + 3 = 7$ or $x^2 + 1 = y$, and the variables can take on any values that satisfy the condition, but in statistics, the random variable can only take on values in Ω , and the random variable comes with a certain probability distribution. In the next chapter we will talk about such distributions and in later study you’ll encounter the definition of a random variable X as a measurable function $X : \Omega \rightarrow E$ for E a measure space. For now, it’s enough to know that S_3 is a random variable that can take on each outcome that appears in step 3 of a binomial tree. A random variable has to take on a numerical value, not a value like H or T .

The next big idea is *expected value*. The expected value of a random variable is the weighted average of all the possible outcomes. Let’s consider only discrete random variables S_n with outcomes s_0, \dots, s_n . Each outcome has a particular probability, $P(S_n = s_i)$, and the sum of these probabilities is 1. The

expected value of S_n , then, is

$$E(S_n) = \sum_{i=0}^n s_i P(S_n = s_i),$$

which weights each outcome according to its probability. For example, let's say you flip a fair coin and you let the random variable X be 0 if the coin comes up heads, 1 if it comes up tails. The expected value of the outcome of X is $1/2$, from $0 \cdot 1/2 + 1 \cdot 1/2$. Again, a funny name, as you'd never expect to get the value $1/2$ from flipping a coin! But $1/2$ is the weighted average of the outcomes we allowed for X .

Back to no arbitrage: the “no free lunch” principle. One of the crazy paradoxes of finance is that people participate in modern finance because they want to make money, but the no-arbitrage principle states that everything is fairly priced by the market and so there is no guaranteed risk-free money to be made. One way to think about this is as follows: on a very small time scale, there may be inaccuracies in pricing in the market, but these inaccuracies will be found and exploited by so-called arbitrageurs, and thus supply and demand will push the inaccuracies to be corrected. Thus, in the long term, financial instruments are priced fairly. This philosophy has had enormous success because it seems to work pretty well over larger timescales, and because it has allowed us to get a unique “correct” price for many common financial instruments. A price that everyone agrees on makes trade much easier!

More concretely, the no-arbitrage principle would tell us that if the stock under consideration is fairly priced initially (at time zero) then the expected value of the stock at time one would be the same as its initial price. Why is this? Say the expected value of the stock at time one was definitely bigger than at time zero. You'd buy a lot of stock at time zero! (So would other people...) Similarly, if the expected value of the stock at time one was less than the value at time zero, you wouldn't buy it at all: you'd say it was overpriced and you'd put your money elsewhere.

Mathematically, this implies that

$$S_0 = E[S_1].$$

Since this expected value is just a weighted average, let's see what this means for one step of the multiplicative binomial tree:

$$S_0 = pS_0u + (1 - p)S_0d,$$

which simplifies to

$$1 = p(u - d) + d,$$

finally giving

$$p = \frac{1 - d}{u - d}.$$

This gives us the promised relationship between p , u , and d – and astoundingly requires no price information at all.

In this era of historically low interest rates, what we've done is even almost right. For most of financial history, though, we have needed to consider whether we should just put our money in the bank rather than investing in stocks. The interest you can earn on cash is referred to as the “time value of money.”

4.3 Short- and long-term approximation

We discussed the idea of short-term approximation in [Chapter 2](#), in [Section 2.2](#). As we alluded to, the idea of predicting value of $f(x)$ a “few moments” later, when the input is $x + \Delta x$, is quite powerful. How does this apply to finance? Take a moment to ponder this question and brainstorm some answers.

If $y = f(x)$ is the output of a function, then we can write the approximation (not equation!)

$$y + \Delta y = f(x + \Delta x) \approx f(x) + f'(x)\Delta x \quad (4.1)$$

We can rewrite this in many ways, including:

$$f(x + \Delta x) - f(x) \approx f'(x)\Delta x \quad (4.2)$$

and

$$\frac{f(x + \Delta x) - f(x)}{\Delta x} \approx f'(x). \quad (4.3)$$

What are our assumptions in making these short-term predictions?

This approximation is really only valid if $f(x)$ is a continuous and differentiable function. If f is not continuous in the area of our approximation, the approximation is total nonsense – $f(x)$ and $f(x + \Delta x)$ may be nowhere near each other. If f is not differentiable at a point where we'd like to look at this approximation, then $f'(x)$ won't exist or make sense.

Financially, the idea of short-term approximation comes about because we'd like to know the price of a stock tomorrow, for instance. At the beginning of this chapter we discussed binomial tree models for stock prices, which are necessarily discrete. Even though it's not "true", we can invent a continuous function $C(t)$ for the price of a stock depending on time, and we can approximate the discrete situation by using calculus and the continuous model. The Black-Scholes model, for instance, is a continuous (or stochastic) model of the price of an option on a stock. The partial differential equation

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS \frac{\partial C}{\partial S} - rC = 0$$

is called the Black-Scholes equation, implicitly describing the price C of a European call option in terms of S , the price of the underlying stock; time t ; standard deviation of stock returns σ ; and a risk-free interest rate r (that time value of money!). All of these derivatives can be discretized using the idea of short-term approximation.

The short-term approximation is a linear approximation. Let's interpret the equation

$$C(t + \Delta t) \approx C(t) + C'(t)\Delta t$$

as a linear approximation for the price of a stock at time $t + \Delta t$ given the starting price $C(t)$ at time t . Say t is 11 am on a Monday, and we want to guess where the price will be after lunch (at 1 pm) *if the stock behaves in a predictable way*. Then we could consider approximating $C'(t)$ from the morning's data, or from last week's data, or some other data set. This number would give us how fast the stock price is sinking or rising. The linear approximation is easy then.

4.3.1 Long-term approximation

We want to combine short-term predictions to make long-term predictions. How do we do this? What do we mean?

Our most important tool is the fundamental theorem of calculus. The fundamental theorem of calculus relates integrals and derivatives. One version of the FTC in words is, “net change is the sum of rates of change over the interval.” We can write this in symbols below.

Example 4.3.1. Explore

$$f(b) - f(a) = \int_a^b f'(x)dx.$$

An example: you want to look at total change in a stock price from time a to time b . That’s $C(b) - C(a)$. One way to look at the change in price is to look at how it stepped up and stepped down at every moment between time a and time b . If you sum up all the changes, $\Delta C \approx C'(t)\Delta t$ for each moment t , you get a version of the fundamental theorem of calculus.

Can you convince yourself that the fundamental theorem of calculus is a reinterpretation of the short-term approximation discussed above?

How do you relate the discrete and continuous versions of each of these concepts? What is each useful for?

4.3.2 First differential equation

A differential equation is an equation that relates quantities and their derivatives – for instance, the quantity $f(x)$ and its derivative with respect to x , $f'(x)$.

I want you to learn how to *read* these differential equations, and identify the assumptions implicit in the mathematics.

For example, what does this differential equation say?

$$f'(x) = af(x).$$

It says, “the rate of change of f with respect to x is proportional to the value of f at every value of x .” Think through the implications. Let’s say a is positive, for simplicity. Then if $f(x)$ is negative, $f'(x)$ must also be negative. If $f(x)$ is positive, $f'(x)$ must also be positive. If $f(x) = 0$, then the rate of change of $f(x)$ is also zero.

What is the solution to a differential equation? It’s a function that makes the differential equation true. For instance, $f(x) = 0$ makes the equation $f'(x) = af(x)$ true: $0 = 0$ is always true. On the other hand, $f(x) = 3$ is not a solution to this differential equation if $a \neq 0$, as $f'(x) = 0$ but $a \cdot 3 \neq 0$.

How can we solve this differential equation? I’ll group the methods as numerical methods, analytic methods, and graphical methods. Numerical methods find approximate numerical solutions and are great if you’ve got some computational power at your disposal. Analytic methods find you exactly-accurate formulas. In this class, many differential equations will have analytic solutions (closed formulas) but in real life many differential equations don’t have analytics solutions at all. Last, graphical methods like slope fields can help you understand the behavior of solutions and might give guidance as to what further techniques you’d like to pursue.

Let’s look at our first differential equation in a financial context. The most basic finance calculation many people do is the calculation of accrual of interest on principal. Using P for principal and t for time in days, we can write

$$\frac{dP}{dt} = rP(t),$$

for r the daily interest rate. Note that P could be positive (money in the bank) or negative (you owe the bank money).

Say $P(0) = 100$ (you start with \$100) and the daily interest rate is $r = 0.01$. Then if interest is accrued daily, you end up with $100 \cdot (1 + 0.01) = 101$ at the end of the day. But if interest is compounding continuously, this is only an approximation – a linear, short-term approximation. What’s the exact amount of interest you end up with?

Here’s where an analytic solution comes in handy. We’ll have to integrate

here. Do a bit of manipulation to convert

$$\frac{dP}{dt} = 0.01P(t)$$

to

$$\int \frac{dP}{P} = \int 0.01 dt.$$

Use your calculus knowledge to solve:

$$\ln |P| = 0.01t + c,$$

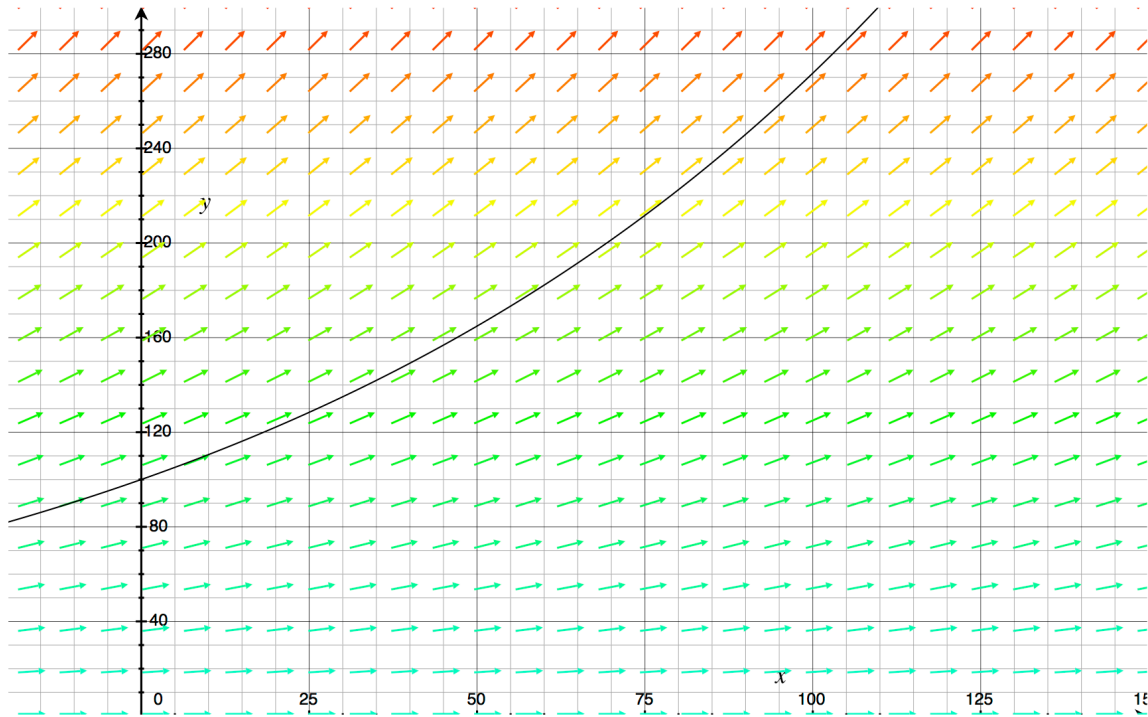
where c is a constant. Exponentiate to get

$$|P(t)| = e^{0.01t} e^c.$$

If we know that $P(0) = 100$, then we should have $e^c = 100$ – ah, e^c is the initial amount of principal. (For the general solution of the differential equation we can drop the absolute value on the left-hand side as if we’re starting with positive amounts of money, all future values will be positive, and if we start with and compound debt, we’ll just get more debt.) To find the exact value of $P(1)$, then, we just evaluate:

$$P(1) = 100e^{0.01} \approx 101.005.$$

Last, let’s look at a graphical exploration of the differential equation. This is a slope field: at points $(t, P(t))$, we draw little lines of slope $P'(t)$.



This allows us to quickly understand the qualitative behavior of solutions to

$$\frac{dP}{dt} = 0.01P(t).$$

These solutions must always “follow the arrows,” so it’s easy to sketch solutions to the differential equation as well!

Later we’ll consider numerical methods for solving differential equations, and we’ll see that Euler’s method, for instance, is really just a version of long-term approximation applied to differential equations.

Chapter 5

Discrete random variables and transformations of variables

This chapter will combine a number of concepts that aren't usually discussed in conjunction. First, we'll talk about discrete random variables, expected values, and variance. Second, we'll notice that series keep coming up. We'll talk about series in general, power series, and Taylor series. I'll throw in some simple but important finance applications. Last, we'll talk about linear and affine linear transformations of variables. This will be our first step toward incorporating linear algebra and multivariable calculus.

5.1 Discrete random variables

It's finally time to look seriously at *random variables*. A random variable is a function from Ω to \mathbb{R} : it always takes on numerical values. Remember that Ω is the set of possible outcomes of a probability experiment, so writing out a random variable as a function $X : \Omega \rightarrow \mathbb{R}$ is a way of assigning a numerical value to each outcome of the probability experiment. In general, we don't know what the outcome of a probability experiment is until the experiment is carried

out – that’s why we use X or another variable name to represent this outcome before knowing what it is!

Here are some examples of random variables. When you flip a coin, you get heads H or tails T , but the random variable associated to your coin flip might be X , with

$$\begin{aligned} X(H) &= 0 \\ X(T) &= 1, \end{aligned}$$

or

$$\begin{aligned} Y(H) &= 1 \\ Y(T) &= -1, \end{aligned}$$

or Z taking its value in any other two-element set in \mathbb{R} . The numerical values depend on the problem you’re trying to solve. For instance, we could play a game where you flip a coin and if it’s heads I give you five dollars and if it’s tails you give me eight dollars. You could decide then to represent the outcomes of the probability experiment (heads and tails) as taking values in the profit-loss space represented by $\{-8, 5\}$, and using the tools we’ll build you can easily and rigorously decide whether this is a game you want to play or not. You might write W for the random variable representing your winnings, with

$$W : \{H, T\} \rightarrow \{5, -8\}.$$

A *discrete* random variable is one that takes values in a finite or countably infinite subset of \mathbb{R} . The values that the random variable can take make up the *range* of the random variable, often denoted I . (Convince yourself that any random variable taking values on a continuous interval of \mathbb{R} can’t be a discrete random variable, using this definition.) The best way to get a feel for discrete random variables is to do examples.

(Aside: As you do problems, I *highly* recommend writing out what random variable you want to consider, very explicitly. A major cause of mistakes is mixing up the random variable you really want to consider and some other related quantity – for instance, doing a calculation with the outcomes $\{0, 1\}$ rather than the outcomes $\{5, -8\}$ in our coin-tossing game above, or mixing up “number of winning coin flips it takes to win a game” with “total number of coin flips it takes to win a game”.)

Example 5.1.1. Flip a coin three times. Let X be the random variable representing how many times heads appears in your three flips. Then $X : \Omega \rightarrow \{0, 1, 2, 3\} \subset \mathbb{R}$, and we call $I = \{0, 1, 2, 3\}$ the range of X .

Example 5.1.2. Roll two dice simultaneously. Consider the random variable given by taking the maximum roll.

Here, I would say X is the value on the first die and Y is the outcome of the second die, and I'd say $Z = \max(X, Y)$ is the random variable we'd like to consider. The range of Z is $I = \{1, 2, 3, 4, 5, 6\}$.

Alone, the concept of the discrete random variable doesn't seem very subtle or important. However, it's going to facilitate the discussion of

- the *probability mass function*,
- the *expected value* of a random variable, and
- the *variance* of a random variable.

These are basic probabilistic concepts which are very important in finance: once we consider portfolio optimization, you'll look to maximize your expected return (the expected value of the portfolio in the future) and minimize the volatility of the portfolio (the standard deviation of the value of the portfolio).

Let's define these concepts.

First, for a discrete random variable we can find the probability mass function, $f(k) = P(X = k)$ for $k \in I$.

Second, we can define the expected value (expectation, mean) using the probability mass function:

$$\mu = E(X) = \sum_{k \in I} kP(X = k).$$

Each outcome is weighted by its probability. The expected value function is written using the $E[\cdot]$ or $E(\cdot)$ notation, and when everyone reading knows which random variable we're talking about, then we can use the notation μ

(mu) for mean. Notice that $E[c] = c$ for c a constant, and convince yourself that $E[cx] = cE[X]$ as well. Figure out what $E[aX + b]$ is for a, b constants.

Third, we can define variance:

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2].$$

Meditate on this definition for a while: it's the expected value of the square of the difference of X from the mean μ . Hmmmm.... how to understand this? Notice that again we have the $\text{var}(\cdot)$ function which we can apply to any random variable, and we have the notation σ^2 when it is clear which random variable is under discussion. Also, we can define the standard deviation σ here:

$$\text{stdev}(X) = \sigma = \sqrt{\text{var}(X)}.$$

Example 5.1.3. Again toss a coin three times and let X be the number of heads appearing.

Following the definitions above, we first compute the probability mass function (pmf).

$$P(X = 0) = \binom{3}{0} \left(\frac{1}{2}\right)^3 \quad (5.1)$$

$$P(X = 1) = \binom{3}{1} \left(\frac{1}{2}\right)^3 \quad (5.2)$$

$$P(X = 2) = \binom{3}{2} \left(\frac{1}{2}\right)^3 \quad (5.3)$$

$$P(X = 3) = \binom{3}{3} \left(\frac{1}{2}\right)^3 \quad (5.4)$$

$$(5.5)$$

We're counting the number of ways k heads can appear in three tosses, then multiplying by the probability of any combination. It's better to just write $P(X = k) = \binom{3}{k} \left(\frac{1}{2}\right)^3$, you'll see.

Second comes expected value: it is

$$E[X] = \sum_{k=0}^3 k \binom{3}{k} \left(\frac{1}{2}\right)^3,$$

which evaluates to

$$E[X] = \frac{1}{8}[0 + 3 + 6 + 3] = \frac{3}{2}.$$

Third, variance. If we use our formula above for variance, we need to compute the expected value of a new random variable, $(X - \frac{3}{2})^2$. The variance is

$$\text{var}(X) = \sum_{k=0}^3 (k - \frac{3}{2})^2 P(X = k) \quad (5.6)$$

$$= \frac{1}{8} \left[\frac{9}{4} + 3 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + \frac{9}{4} \right] \quad (5.7)$$

$$= \frac{3}{4}. \quad (5.8)$$

Frankly, that last calculation was mildly annoying. There's got to be a better way. Many advances in mathematics have been made through rigorous laziness – can we simplify this calculation?

5.1.1 Linearity of expectation – best thing ever

Fortunately, we can simplify our calculations of variance. The expected value function $E(\cdot)$ is linear! That is,

$$E[X + Y] = E[X] + E[Y]$$

as long as $E[X]$, $E[Y]$ are finite numbers.

Sketch a proof of this fact: consider two discrete random variables X and Y and call their sum $Z = X + Y$. You'll have to look at the probability of the joint event " $X = k$ and $Y = j$," which you can write as $P(X = k, Y = j)$. Figure out the pmf for $P(Z = \ell)$ and go from there.

In particular, this has the following lovely consequence: if we let $\mu =$

$E(X)$,

$$\text{var}(X) = E[(X - \mu)^2] \quad (5.9)$$

$$= E[X^2 - 2\mu X + \mu^2] \quad (5.10)$$

$$= E[X^2] - E[2\mu X] + E[\mu^2] \quad (5.11)$$

$$= E[X^2] - 2\mu E[X] + \mu^2 \quad (5.12)$$

$$= E[X^2] - 2\mu^2 + \mu^2 \quad (5.13)$$

$$= E[X^2] - \mu^2 \quad (5.14)$$

This is lovely: let's do another example calculation using this new formula for variance.

Example 5.1.4. Consider those two dice that you rolled earlier. Remember one die had a value X and the other a value of Y , and we want to examine $Z = \max(X, Y)$. The range of Z is $I = \{1, 2, 3, 4, 5, 6\}$.

The pmf for Z is $P(Z = k) = \frac{2k-1}{36}$. I got this by writing down the first few probabilities, noticing a pattern, and then mathematically formalizing it.

The expected value of Z is easier to compute using the formula for the pmf than it is to compute by force.

$$E[Z] = \sum_{k=1}^6 k \frac{2k-1}{36} \quad (5.15)$$

$$= \frac{1}{36} \sum_{k=1}^6 (2k^2 - k) \quad (5.16)$$

$$= \frac{1}{36} \left[2 \sum_{k=1}^6 k^2 - \sum_{k=1}^6 k \right] \quad (5.17)$$

$$= \frac{1}{36} \left[2 \cdot \frac{6 \cdot 7 \cdot 13}{6} - \frac{6 \cdot 7}{2} \right] \quad (5.18)$$

$$= \frac{161}{36} \quad (5.19)$$

Some magic happened in there: I used my knowledge of sums of consecutive integers and sums of consecutive squares. You'll remember the triangular

numbers $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ from a previous chapter, but you may not know the sum of squares formula. Here it is:

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}.$$

This is something I re-memorize when working through this section, and then forget when it's not important. Do what you like, but it is helpful to know by rote.

Variance is an even worse calculation, but at least we can simplify it by using our nice new formula.

$$\text{var}(X) = E[Z^2] - \mu^2 \tag{5.20}$$

$$= \sum_{k=1}^6 k^2 \frac{2k-1}{36} - \left(\frac{161}{36}\right)^2 \tag{5.21}$$

Here you'll need the sum of cubes. You know what's really cool?

$$\sum_{i=1}^n i^3 = \left(\sum_{i=1}^n i\right)^2.$$

The sum of the first n cubes is the square of the n th triangular number. So

$$\sum_{i=1}^n i^3 = \frac{n^2(n+1)^2}{4}.$$

This is called Nicomachus' theorem sometimes – Nicomachus lived in what's now Jordan, in the Middle East, in the first century BCE. The theorem was also discovered by Al-Karaji, an Arab mathematician, and Aryabhata, an Indian mathematician, all at or before the year 1000. (See <https://www.math.nmsu.edu/davidp/brid> for more information.) Then it was discovered in France in the 1300s. Thinking about the history of mathematical achievement and financial thought is humbling and instructive. What drove these three mathematicians in different areas to figure out this number-theoretic theorem that also pops up in probability?

How did Thales of Miletus have the idea to become the first options or futures guy? Why did the prophet Mohammed think so much about the ethics of futures contracts? While synthetic CDOs are pretty new, the basics of hedging and portfolio management have been thought about for thousands of years.

Back to variance: Using the identity, we can simplify the variance calculation to

$$\text{var}(X) = \frac{1}{36} \left[2 \left(\sum_{k=1}^6 k \right)^2 - \sum_{k=1}^6 k^2 \right] - \left(\frac{161}{36} \right)^2,$$

or

$$\text{var}(X) = \frac{1}{36} \left[\left(\frac{6^2 \cdot 7^2}{2} \right)^2 - \frac{6 \cdot 7 \cdot 13}{6} \right] - \left(\frac{161}{36} \right)^2 = \frac{791}{36} - \frac{161^2}{36^2} = \frac{2555}{36^2}.$$

This is about 1.97. Does this make sense? Always do a quick “gut check” to catch errors – the variance is not bigger than all possible outcome values, the variance is not negative... looks good!

Notice all this use of series. Calculations of variance and expected value for discrete random variables often require a lot of series, and in addition, we can package expectation and variance as the first two *moments* in a *moment generating function*. If you’d like to start reviewing series, skip to [Section 5.3](#). We’ll first discuss properties of expectation and variance just a bit more.

5.1.2 Multiplicativity of expectation?

It is so cool that expectation is linear. Really. If we were very lucky, expectation would also factor nicely: $E(XY)$ and $E(X)E(Y)$ would be equal. Is this true?

To see whether this would be true in general or not, go back to probability itself: does probability factor? That is, do we have $P(X = x, Y = y) = P(X = x)P(Y = y)$ for discrete random variables X and Y taking values in the appropriate ranges?

Remember the discussion in [Section 1.8](#) of compound experiments. Before having the language of random variables, we argued that if two events A and B are independent, then $P(AB) = P(A)P(B)$. Now we can translate this into

the language of random variables and *define* independence of discrete random variables as follows:

Definition 5.1.1. Two discrete random variables X and Y are independent if for all x in the range of X and all y in the range of Y , we have

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

If X and Y are not independent, this won't be true for all combinations of x and y – there will be some event $(X = x, Y = y)$ for which this equation is false.

A very simple example: flip a coin once. Let X be the number of heads, and Y be the number of tails. The probability of getting heads *and* tails on the same flip is zero (the events are mutually exclusive), while $P(X = 1)P(Y = 1) = 0.25$. It's easy to see that X and Y are related just by thinking about it: $X = 1 - Y$ relates the two nicely.

Another way of looking at this is to say that the joint probability mass function of independent random variables X and Y factors into the probability mass functions for X and Y .

Look at the definition of expected value and see if you can prove the following fact for discrete random variables: If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

as long as both $E(X)$ and $E(Y)$ exist and are finite.

This fact is also true for continuous random variables, using the following definition of independence:

Definition 5.1.2. Two random variables X and Y are independent if for all x in the range of X and all y in the range of Y , we have

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

You'll see why we needed to change the equalities to inequalities when we discuss continuous random variables.

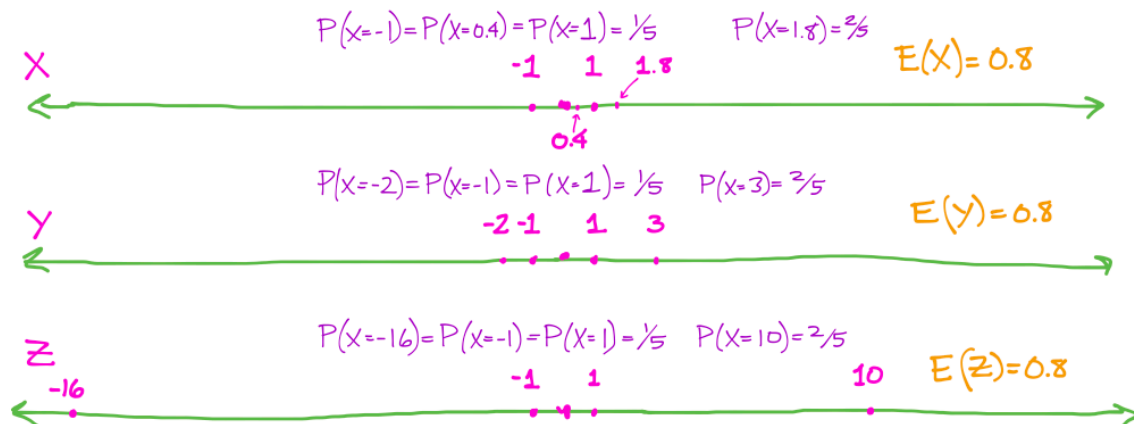
5.2 Variance: average of squared deviations

Variance is a subtle quantity, initially hard to grasp. It might help to contrast it with other measures of variability in a set of data. We want to look at how a data set is scattered, and there are many ways to do that:

- the range of the data is useful (max-min) but not very stable with respect to outliers;
- quartiles can give you a sense of the data but don't give one statistic by which to compare different data sets;
- the average of absolute values of differences from the mean gives a fine statistic of variability.

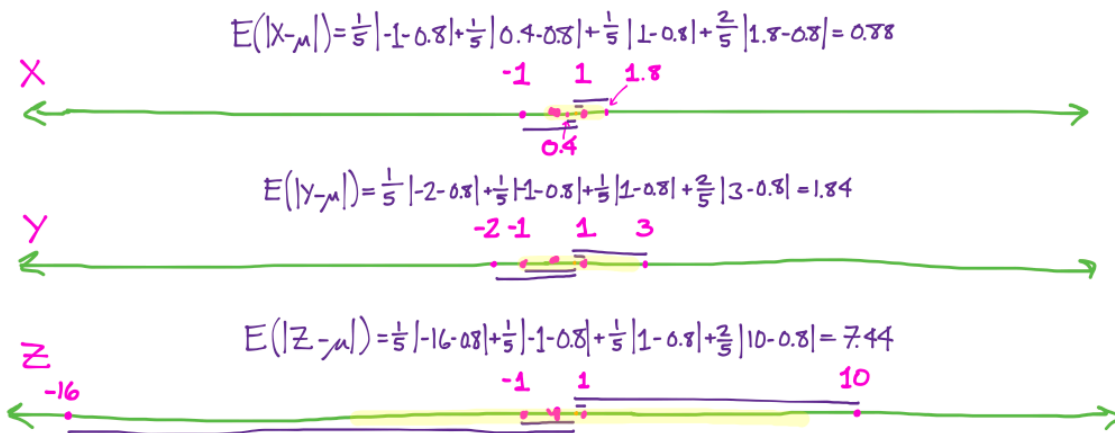
However, in the end one method has come to rule them all: average of squared deviation.

Let's explore measures of "scatter" through an example. We'll look at three random variables, X , Y , and Z , with probability mass functions given in the figure:



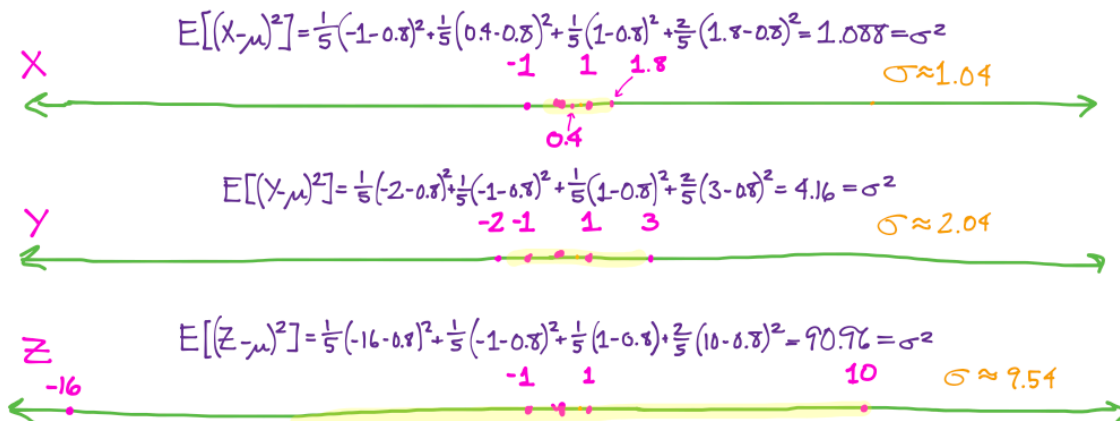
I've engineered each random variable so that they all have the same mean, $\mu = 0.8$. The deviation of each outcome $X = x$ from the mean would be $x - \mu$, and similarly for Y and Z . Try taking a weighted average of these deviations to give some measure of "scatter." Why is it zero each time? Because linearity

of expectation means that $E(X - \mu) = E(X) - \mu = 0$. Fine, use absolute deviation instead: take the expected value of $|X - \mu|$ to get the absolute mean deviation. Calculations for X , Y , and Z :



This seems fine, but for better or worse it doesn't give much weight to outliers. It's a statistic used in some time series calculations where you don't want to give too much weight to outliers, but maybe in some situations where you want to measure "scatter" weighting the outliers is part of the point!

Variance is instead the mean of *squared* deviations. Squaring the deviations results in positive numbers, so there's no cancellation of deviations to worry about. Take a look: the calculations of $E[(X - \mu)^2]$ etc. are below.



Outcomes further from the mean have a big effect on variance! Look at that $\sigma^2(Z) = 90.96$! Take the square root to get the standard deviation, which has the same units as the original random variable. This is the standard measure of “scatter” for data sets and probability distributions. The name “standard deviation” comes about because of this historical community decision, even though for some applications you might decide to use an alternative statistic.

5.3 Series

A *power series* is a function that is an endless sum of monomials – like a polynomial, but of infinite degree:

$$P(z) = \sum_{k=0}^{\infty} c_k z^k.$$

A power series can take in a real number or a complex number. Here we will consider only power series whose input is real; further on we’ll look at functions of complex numbers. However, throughout the book we will only consider real coefficients c_k .

Why consider these power series? Two reasons: to set a good framework for approximations (Taylor series), and to understand moment generating functions. We’ll concentrate on the first reason in this subsection.

Power series are a natural generalization of the “short-term approximation” or linear approximation of the last chapter. Often a more accurate short-term approximation can be gained by taking a higher-degree polynomial as the approximating function. We’ll talk about quadratic approximation, for instance, and some of you may have encountered cubic splines for numerical interpolation. These are polynomials, rather than power series, but you’ll find that it’s really powerful to have the power series framework in which to consider all these approximations.

Example 5.3.1. Let $f(x)$ be a function that can be differentiated as many times as we need. Determine a third degree polynomial $P(x) = ax^3 + bx^2 + cx + d$ so that $f(0) = P(0)$, $f'(0) = P'(0)$, $f''(0) = P''(0)$, and $f'''(0) = P'''(0)$.

The example above brings us directly to *Taylor series*, power series representations of a function at $x = a$. If we want to write a power series representation of a function $f(x)$ at $x = 0$ using a Taylor series, each coefficient c_k is given by the following formula:

$$c_k = \frac{f^{(k)}(0)}{k!}.$$

Here, $f^{(k)}(0)$ is the k^{th} derivative of $f(x)$ at $x = 0$, for $k \geq 1$, and $k! = k \cdot (k - 1) \cdots 2 \cdot 1$. Basically, we are trying to make the slope, convexity, and higher derivatives of our approximation match the original function as well as possible around the point $x = 0$. Then *if* all the derivatives exist and this infinite series converges, we have

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k.$$

In the following example, each series converges for all $x \in \mathbb{R}$.

Example 5.3.2. Find the Taylor series at $x = 0$ for e^x , $\cos x$, $\sin x$.

By contrast, the next two series converge only for $|x| < 1$.

Example 5.3.3. Find the Taylor series at $x = 0$ for $\frac{1}{1-x}$ (geometric series) and $\ln(1+x)$ (a harmonic series). What goes wrong at the edges $x = 1$ and $x = -1$ of the interval $|x| < 1$? Why am I not asking for the Taylor series for $\ln x$ centered at $x = 0$?

5.3.1 Convergence, divergence, well-defined

Three big concepts come up in working with series: convergence, divergence, and well-definedness. I just threw you into finding Taylor series and used the word “convergence” without defining it, because I expect the reader (you) to have some experience of these concepts. But let’s look carefully at them:

The Taylor series for $f(x) = \frac{1}{1-x}$ at $x = 0$ is $T(x) = \sum_{k=0}^{\infty} x^k$.

- For $x = 1/2$, for instance, this series works beautifully and sums to 2. (Use a geometric argument to justify if you like). The series *converges*, as it sums to a unique, finite, real number.
- What happens at $x = 1$, though? $T(1) = 1 + 1 + 1 + \dots$ does not sum to any finite number – it goes to infinity. We say that this series *diverges*. This reflects the behavior of the original function; $f(1)$ isn't defined because it prompts you to divide by zero.
- What about the series at $x = -1$? It's no problem to compute $f(-1) = \frac{1}{1-(-1)}$. However, $T(-1) = 1 - 1 + 1 - 1 + 1 - \dots$. Now, is this $1 + (-1 + 1) + (-1 + 1) + \dots$ or $(1 - 1) + (1 - 1) + (1 - 1) + \dots$? or something else? This sum is not *well-defined*: it doesn't have a unique, unambiguous answer. It is also divergent, then, because it's not convergent!

These concepts can be defined a bit more rigorously using limits. First I'll remind you of convergence and divergence of *sequences*:

Definition 5.3.1. A sequence of real numbers $\{a_j\}$ converges to a limit L if for any $\epsilon > 0$, there is some integer N so that $|a_j - L| < \epsilon$ for all $j > N$. If a sequence $\{a_j\}$ doesn't converge, it diverges.

Recall that there are several types of behavior for a divergent sequence: it could blow up or blow down to infinity or negative infinity (like $\{2^j\}$ or $\{-j\}$), or it could bounce around forever despite being bounded (like $\{\sin(\pi j)\}$ or $\{\sin j\}$), or it could do truly chaotic things. Sometimes it's easy to prove that a sequence converges or diverges. Sometimes it's really hard. Witness the Collatz conjecture: start with any positive number j . If j is even, divide it by 2 and make that the next term. If it's odd, make $3j + 1$ the next term. There's a conjecture that this sequence will always reach 1, no matter what you start with, but this is not proven yet and is remarkably hard. Try it!

Ok, back to series.

5.4 Moment generating functions

Remember that $f(x) = e^x$ is a pretty remarkable function. It is the only function that is its own derivative. You can phrase that as a differential equation if you like:

$$\frac{d}{dx}e^x = e^x,$$

or

$$\frac{dy}{dx} = y.$$

What does this differential equation mean for the Taylor series at $x = 0$? Well, since the equation gives

$$\sum_{k=0}^{\infty} a_k x^k = T(x) = T'(x) = \sum_{k=0}^{\infty} k a_k x^{k-1},$$

we can deduce $a_k = (k + 1)a_{k+1}$ for all k . Since $a_0 = 1$ as $e^0 = 1$, we know $a_1 = 1$, $a_2 = 1/2$, $a_3 = 1/(3 \cdot 2)$, and $a_k = 1/k!$. Hence the Taylor series for e^x is

$$T(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Why do we care about this? Because this series provides a great package for expectation, variance, etc. Let's define the moment generating function for the random variable X by

$$M_X(t) = E[e^{tX}].$$

Generating functions are used in combinatorics and all sorts of other fields to package numerical information in a series. The variable t is called a *formal variable*, and basically it is a placeholder that is there simply to give you its degree n (from t^n). This degree is part of the packaging that the generating function does. For the moment generating function, the coefficient of t^n is the n th *moment*. The n th moment of X is the expectation of X^n , $E[X^n]$.

Check it out:

$$e^{tX} = 1 + \frac{tX}{1!} + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots$$

So the moment generating function of X is defined to be

$$M_X(t) := E[e^{tX}] = E\left[\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}\right] = \sum_{k=0}^{\infty} E[X^k] \frac{t^k}{k!}.$$

Let's consider a discrete random variable X that takes only positive integer values. It has a probability mass function that takes the range $I \subset \mathbb{Z}_{\geq 0}$ and gives an output $P(X = j)$ for each $j \in I$. We can combine this with the generating function calculation above to calculate each $E[X^k]$, using $E[X^k] = \sum_{j \in I} j^k P(X = j)$, carefully regrouping terms:

$$M_X(t) = \sum_{j \in I} e^{tj} P(X = j).$$

You can see that the first moment, the coefficient of t when e^{tj} is expanded, gives you $E[X] = \mu$. The second moment gives you $E[X^2]$ which you can use to get $\text{var}(X) = E[X^2] - E[X]^2 = E[X^2] - \mu^2$. Higher moments won't be discussed now, but I will say that as we study different special discrete and continuous random variables, you'll see that we can classify their distributions using moment generating functions.

5.5 Linear and affine linear transformations

Above, we discussed the linearity of expectation. You can prove that

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

for a, b, c constants. This means that expectation is a linear operator: it satisfies

$$E[X + Y] = E[X] + E[Y]$$

and

$$E[aX] = aE[X],$$

the two conditions for linearity. You are familiar with several other linear operators, too:

$$\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$$

and

$$\frac{d}{dx}[af(x)] = a\frac{d}{dx}f(x),$$

so differentiation is a linear operator, and integration is as well (convince yourself!).

Quick quiz Is the function $\ln(x)$ linear? Is squaring linear? What about the functions sine and cosine? What about variance of a random variable?

You'll notice that none of the above mentioned functions or operators are linear! For instance, because squaring (taking x to x^2) is not linear, variance is not linear either. Check for yourself to see

$$\text{var}(aX + b) = a^2\text{var}(X).$$

What happened to the b ? Do the math to check it out.

A *linear* function is not exactly what you might think, though. Quick quiz: which of these is linear?

$$3x - 4y = 0 \quad 3x - 4y = 2.$$

Right: although both equations have a line as the graph, only one is linear in the technical sense. (We abuse this language all the time.) We say that $3x - 4y = 0$ is a linear equation and $f(x, y) = 3x + 4y$ is linear because $3x - 4y = 0$ describes a linear vector space in \mathbb{R}^2 .

Definition 5.5.1. A vector space, or linear space, is a set of vectors that is closed under addition and scalar multiplication.

In particular, the set $\{(x, y) \in \mathbb{R}^2 \mid 3x - 4y = 0\}$ is linear because if a point (a, b) and a point (c, d) both satisfy the equation, then you can check that

$(a + c, b + d)$ also satisfies the equation. Also, if (a, b) satisfies the equation, you can check that (ka, kb) also satisfies the equation for any $k \in \mathbb{R}$. Dead giveaway: the graph of $3x - 4y = 0$ is a line that goes through the origin.

By contrast, if you look at the equation $3x - 4y = 2$, you can see that these properties – closed under addition, closed under multiplication – are not true of the points that satisfy the equation. But the graph is a line! Yes: this equation is an *affine linear* equation. It's a shift of a linear equation, and its graph is a shift of a linear space. We call these shifts *affine transformations*.

Comfort with transformations of functions translates very well to probability applications. By transformations, we mean stretching, scaling, and translating functions. We can accomplish these tasks and more using vectors and matrices. Let's bravely step into the two-dimensional world and consider vectors in \mathbb{R}^2 .

Rewrite $(x - 2)^2 + (\frac{y}{3} - 1)^2 = 4$ as $u^2 + v^2 = 4$.

Notice that the above equation involves shifting x and shifting and scaling y . We can write this using two equations, one for u and one for v , or using matrices and vectors.

Set of equations: $u = x - 2$ and $v = \frac{y}{3} - 1$ gives the desired shift in the previous example. Using matrices and vectors, we can write this as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

This is scaling (matrix multiplication) and shift (affine transformation).

You need to get comfortable alternating between viewing $(a, b) \in \mathbb{R}^2$ as a vector and as a point in the plane.

Quiz: In general, if we scale in the x -direction by a and the y -direction by b , and shift by (b_1, b_2) , we can write: *what?*

To ponder: We could do these computations before without matrices; why bother? *To be discussed...* How can you “go backwards” and undo a scaling and translation?

5.6 Specific discrete random distributions

5.6.1 Bernoulli

The Bernoulli random variable and Bernoulli distribution are a building block for other discrete random variables and their distributions. The Bernoulli random variable X takes on the value of 1 with probability p or 0 with probability $1 - p$, with $0 \leq p \leq 1$. Here we identify Ω , the set of possible outcomes, with the range $\{0, 1\}$ of the random variable X . Notice this distribution obeys the axioms of probability: with our identification of Ω and $\{0, 1\}$, we have

$$P(\Omega) = 1,$$

we have

$$P(X = 0) = 1 - p \geq 0$$

and

$$P(X = 1) = p \geq 0,$$

and we have

$$P(X = 0 \text{ or } X = 1) = P(X = 0) + P(X = 1)$$

because the outcomes 0 and 1 are disjoint.

Bernoulli random variables are useful for the endless variety of coin-flipping problems, but you'll also find them useful as indicator random variables for solving more complicated problems. As you'll see you can use sums and products of Bernoulli random variables to come up with many of the following discrete random variables.

The expectation of X is

$$E(X) = p \cdot 1 + (1 - p) \cdot 0 = p$$

and the variance is

$$\text{var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p).$$

5.6.2 Binomial

The name comes from the binomial expansion, of course, which we've seen so many times:

$$(p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k}.$$

The binomial distribution asks, “In n coin flips, what is the probability of k heads?” Use the Bernoulli trials just discussed to rephrase this: in n Bernoulli trials, what is the probability of k successes?

Very explicitly, a random variable distributed binomially is a random variable X depending on the parameter n , taking on values in $\{0, \dots, n\}$. This can be interpreted as the number of successes k in n trials, where each trial is independent and success has probability p and failure has probability $q = 1 - p$. The probability of succeeding exactly k times in n trials is then

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

This expression is the probability mass function for the binomial random variable.

To figure out the expected value of X , let's use the idea of indicator random variables. Let X_i be the result of the i th trial, with $X_i = 1$ if you succeed and $X_i = 0$ if you fail. We already know $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Then $\sum_i X_i = X$ counts the number of successes you have. Using linearity of expectation, we know that

$$E(X) = E\left(\sum_{i=0}^n X_i\right) = \sum_{i=0}^n E(X_i).$$

This is an easy computation, since $E(X_i) = p \cdot 1 + (1 - p) \cdot 0 = p$. We have

$$E(X) = np.$$

Likewise, the variance computation for X is easy if we use the property of variance $\text{var}(aX + c) = a^2 \text{var}(X)$ and the fact that each trial is independent, so

$\text{var}(X_i + X_j) = \text{var}(X_i) + \text{var}(X_j)$ for $i \neq j$. (Independence is crucial here!!) We can just sum the variances of the indicator random variables to get

$$\text{var}(X) = \sum_{i=0}^n \text{var}(X_i) = np(1 - p).$$

At points in this math finance journey you may wonder, where is the finance? Here I refer you to the pioneering paper of Cox, Ross, and Rubinstein, in which they use a binomial distribution to provide a discrete model for derivatives prices that is easily implemented on a computer and in the continuous limit produces the Black-Scholes-Merton model for pricing European options. Go read it right now and be amazed that you can read some foundational literature!

Notice that one reason the Cox-Ross-Rubinstein and Black-Scholes-Merton basic models converge is that for large values of n , it's possible to approximate the (discrete) binomial distribution using the (continuous) normal distribution. We'll discuss this when we talk about the Central Limit Theorem.

5.6.3 Geometric

Say you're doing a series of Bernoulli trials again, but now you're considering how many trials X you must carry out until you reach your first success. Another way to think about this is how many failures $Y = X - 1$ occur before the first success. The probability mass function for either of these is built on ideas of conditional probability:

- you assume for $P(X = k)$ that you fail $k - 1$ times and then succeed on the k th trial, for k in $\{1, 2, \dots\}$, or
- you assume for $P(Y = k)$ that you fail k times and then succeed on the next trial, for k in $\{0, 1, 2, \dots\}$.

As before, let the probability of success on a given trial be p and the probability of failure be $q = 1 - p$. Writing out the first few probabilities in the probability mass function, verify that

$$P(X = 1) = p, \quad P(X = 2) = qp, \quad P(X = 3) = q^2p, \quad \dots$$

It's easy to package this into one function:

$$P(X = k) = q^{k-1}p, \quad k = 1, 2, 3, \dots$$

Follow the same process for Y :

$$P(Y = k) = q^k p, \quad k = 0, 1, 2, \dots$$

Instead of using indicator random variables to calculate expectation and variance, we'll use what we know about series. For $0 < p < 1$, we can do the following calculation (and if $p = 0$ or $p = 1$, we don't need to calculate anything!):

$$E(X) = \sum_{i=1}^{\infty} iP(X = i) = \sum_{i=1}^{\infty} i \cdot p \cdot (1-p)^{i-1} = p \sum_{i=1}^{\infty} i \cdot q^{i-1} = p(1-q)^{-2} = 1/p.$$

The identity we used here is that for $0 < p < 1$, $\sum_{i=1}^{\infty} i(1-p)^{i-1}$ is the derivative of $\frac{1}{(1-p)^2}$.

5.6.4 Hypergeometric

Now say you've got R red balls and W white balls in an urn ($N = R + W$ total items) and you're picking out n items all at once. What's the probability of k red balls in the n you pick out? This is the hypergeometric random variable X .

While I haven't seen hypergeometric used a ton in financial mathematics, Henriksson and Merton used the distribution in their 1981 paper, "On Market Timing and Investment Performance. II." It also comes up in various exams you may need to pass on your career path. Very briefly, Henriksson and Merton look at what kind of results you'd get by choosing n investments at random from W losing investments and R winning investments. If forecasters do no better (or do worse!) than the hypergeometric distribution, then their methods are not useful. Of course there is much more to the paper; check it out to see for yourself.

What's the probability mass function for X , the number of red balls among n balls picked out of an urn of R red balls and W white balls? The probability

that $X = k$ is the ways to choose k red balls times the number of ways to choose $n - k$ white balls, over the total number of ways to pick n balls from N . That's

$$P(X = k) = \frac{\binom{R}{k} \binom{W}{n-k}}{\binom{R+W}{n}}.$$

To find the expected value of the number of red balls picked out of the urn of N total balls, use indicator variables again. Let X_i be 1 if the i th ball is red and 0 if the i th ball is white. Then the total number of red balls is $X = \sum_{i=1}^n X_i$. We can use linearity of expectation (best thing ever!) to break this into computing the $E(X_i)$ and summing. The probability that X_i is one is

$$P(X_i = 1) = \frac{R}{R + W}$$

and so

$$E(X_i) = 1 \cdot \frac{R}{R + W} + 0 \cdot \frac{W}{R + W}.$$

Thus

$$E(X) = \sum_{i=1}^n \frac{R}{R + W} = \frac{nR}{R + W}.$$

Use indicator variables to find the variance, too – you find that

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) + \sum_{i,j=1, j \neq i}^n \text{cov}(X_i, X_j).$$

Work through that and check your work:

$$\text{var}(X) = \frac{RWn(R + W - n)}{(R + W)^2(R + W - 1)}.$$

5.6.5 Poisson

The Poisson process is a bit different than the others discussed so far. It's not simply a combinatorial extension of Bernoulli trials. Instead, the Poisson

random variable is often used for events happening in time. How many earthquakes happen in a given time period? How many times does volatility spike over a certain time interval?

The Poisson distribution has been used to model market shocks via jump processes (adding jumps into a model of the market, with frequency described by a Poisson distribution).

Define a Poisson random variable X as one that takes on the values $k = 0, 1, 2, \dots$ with the probability mass function

$$P(X = k) = \begin{cases} \frac{\lambda^k e^{-\lambda}}{k!} & k = 0, 1, 2, \dots \\ 0 & \text{else} \end{cases}$$

Here λ is a parameter often called the event rate or rate parameter. It gives the average number of events per time interval.

5.6.6 Negative binomial

How many MBAs must a financial firm interview before finding exactly n good candidates? This is an example using the negative binomial distribution from *Analysis of Financial Time Series* by Tsay. If each candidate is independent and each applicant has probability p of being a good fit, and the total number of interviews necessary is the random variable Y , then in the literature both $X = Y - n$ and Y may be described as having the negative binomial distribution. I'll delve into this a bit because it's good practice with discrete probability, not because of any particular use for financial math.

First let's look at the probability mass function for Y , the total number of interviews necessary to get exactly n good candidates. Each interview is a Bernoulli trial with probability of success p and probability of failure $1 - p$. The number of good candidates n desired is a fixed positive integer! With the set-up given, where Y is the total number of interviews/Bernoulli trials necessary, the last interview or trial must be a success (if it were a failure, either it was not a necessary interview because n good candidates had already been identified, or not enough good candidates were yet identified and so the trials should continue). Call this last interview the k th interview. In the last

$k - 1$ interviews, there must have been $n - 1$ successes. How many ways can those be arranged? Use a binomial coefficient to count this: $\binom{k-1}{n-1}$ ways. There are n total successful trials, each with probability p , and $k - n$ unsuccessful trials, each with probability $1 - p$. Thus

$$P(Y = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$$

for $k = n, n + 1, n + 2, \dots$. (You should check for yourself that for k smaller than n the probability must be zero.)

We can split this random variable Y up as the sum of geometric random variables Y_i , where each Y_i is the number of interviews necessary to get from success $i - 1$ to success i . Use linearity of expectation and the fact that $E(Y_i) = 1/p$ to see that

$$E(Y) = E\left(\sum_{i=0}^n Y_i\right) = \sum_{i=0}^n E(Y_i) = \frac{n}{p}.$$

Note that these Y_i are independent, as well, so we can find variance this way too:

$$\text{var}(Y) = \text{var}\left(\sum_{i=0}^n Y_i\right) = \sum_{i=0}^n \text{var}(Y_i) = \frac{n(1-p)}{p^2}.$$

If you're reading a variety of sources on your probability journey, you might notice that Tijms, Tsay, and Grinstead and Snell define the negative binomial distribution as I have above. On the other hand, Wikipedia defines the negative binomial distribution as the number of successes you see until a predetermined number of failures has occurred – so if you flip “failure” and “success”, the definition in Wikipedia is looking at the random variable $X = Y - n$ instead of Y . (To repeat the set-up at the start of the section, X is the number of failures occurring before n successes, and Y is the total number of trials necessary for n successes.) This gives a slightly different presentation of the probability mass function because something else is being counted. I will not flip my successes and failures to match Wikipedia, but let's look at the probability mass function for X , the number of failures until n successes.

Let's think this through: if there are no failures, just n successes, then $X = 0$ and $P(X = 0) = p^n$. If there is one failure and n successes, then $X = 1$

and $P(X = 1) = \binom{n+1-1}{1} p^n (1-p)^1$. If there are k failures and n successes, $P(X = k) = \binom{n+k-1}{k} p^n (1-p)^{n-k}$.

5.6.7 Transformations of discrete random variables

One last thing that may help you grapple with discrete random variables is this theorem:

Theorem 5.6.1. If X is a discrete random variable and $Y = g(X)$, then Y has the probability mass function

$$P(Y = k) = \sum_{x \in \{g(x)=k\}} P(X = x)$$

and Y has expected value

$$E[Y] = \sum_{x \in I_X} g(x) P(X = x)$$

for I_X the range of values that X can take.

This is a very streamlined and useful theorem that applies to any type of transformation of X !

Chapter 6

Continuous Random Variables

In the previous chapter we considered Poisson random variables, for instance the number of earthquakes that occur in two years. While the number of earthquakes is necessarily discrete – an integer value – the time between two earthquakes can take values on a continuous domain. Times and distances are natural settings for continuous random variables. We often see continuous random variables coming up from geometry questions. In addition, continuous settings are often a nice idealized environment in which to approximate solutions to discrete financial problems. Stock prices technically don't take values in a continuous range, but using a continuous approximation for the distribution of stock prices allows for very fast computations.

Continuous random variables X are defined by the existence of a *probability density function*, or pdf, that characterizes the behavior of the random variable. This pdf $f(x)$ must satisfy the following properties:

$$F(x) = P(X < x) = \int_{-\infty}^x f(y)dy,$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1, \quad f(x) > 0 \forall x \in \mathbb{R}.$$

Two of these equations come from the axioms of probability: if $f(x)$ could take on negative values, we could get negative probabilities (from the definition of $F(x) = P(X < x)$), and if the integral of $f(x)$ over \mathbb{R} was not 1, we'd be violating the axiom $P(\Omega) = 1$.

The function $F(x) = P(X < x)$ is called the *cumulative distribution function* or cdf for the random variable X . Notice a few things about $F(x)$:

$$\lim_{x \rightarrow \infty} F(x) = 1$$

because of that axiom, and $F'(x) = f(x)$ by the Fundamental Theorem of Calculus. These will both come in handy!

A few more technical notes: since we here make the decision to define $F(x)$ as an integral of a pdf, $F(x)$ will be continuous for all x and differentiable at all but a finite number of points (differentiable almost everywhere). You can take a slightly different approach and define X to be a continuous random variable if it has a cdf $F(x)$ that is continuous for all x and differentiable at all but a *countable* number of points, but to think through the ramifications you need some measure-theoretic tools that I don't want to develop in this course.

6.1 Geometry problems

Geometry is a great place to start an examination of continuous random variables. For a little while, we can work in the sanitized and idealized world of pure geometric shapes – we'll delve into messy data soon enough. In this section, we'll just do example after example to illustrate how to work with pdfs and cdfs.

6.1.1 Max or min in unit square

Pick a point (x, y) at random in the unit square, $0 \leq x \leq 1$ and $0 \leq y \leq 1$ in \mathbb{R}^2 . What is the probability that the maximum of x and y is less than any given value z ?

To rephrase that, we can invent a new continuous random variable, $Z = \max(x, y)$. Now we ask, what is the cumulative distribution function $F(z) = P(Z < z)$?

This may be familiar from previous geometry problems, and there's a good chance you've already done this calculation. Draw a picture and notice that for any given z , the set of points (x, y) with both x and y less than z is a square with one vertex at the origin $(0, 0)$ and side lengths z . The cumulative distribution function is in fact

$$F(z) = P(Z < z) = \begin{cases} 0 & z < 0 \\ z^2 & 0 \leq z < 1 \\ 1 & 1 \leq z \end{cases} .$$

Check that when $z = 0$ we've got $F(0) = 0$, and that when $z > 1$ we've got $F(z) = 1$.

Working backward from this cdf, we can find the corresponding probability density function $f(z)$. It's just the derivative of $F(z)$, by the Fundamental Theorem of Calculus and the definition of $F(z)$. Thus

$$f(z) = \begin{cases} 0 & z < 0 \\ 2z & 0 \leq z < 1 \\ 0 & 1 \leq z \end{cases} .$$

Notice that this is not continuous, and we don't care. Moreover, if you're very sharp you'll say, hey, the function $F(z)$ was not differentiable at $z = 0$ or $z = 1$ – how can you assign a value to $f(z)$ there if you're just taking the derivative of $F(z)$? I would respond, you're right that $F(z)$ is not differentiable at zero or 1. To make $f(z)$ nicer, I'm just assigning values to $f(0)$ and $f(1)$ that are consistent with the demands that $F(z) = \int_{-\infty}^z f(x)dx$ and $f(z) \geq 0$. You can assign other values to $f(0)$ and $f(1)$ if you like! Yes, this leads to the troubling realization that “the” pdf of a random variable is not really unique, but measure theory will (someday) assure us that it won't change any probabilities. The values of the integrals are not affected.

Your turn: find the pdf and cdf for the continuous random variable Z , where $Z = \min(X, Y)$ for a point chosen at random in the unit square.

6.1.2 Absolute values

Another example problem is finding the cumulative distribution function and probability density function for $Z = |X - Y|$, where (X, Y) is a point chosen at random in the unit square $[0, 1] \times [0, 1] \subset \mathbb{R}^2$. The absolute value in this definition of Z is a complicating factor, which makes this a good example!

Again, we'll use basic geometry to figure out the cumulative distribution function (cdf) for Z . First, start with what you want to know and substitute:

$$P(Z \leq z) = P(|X - Y| \leq z) \quad (6.1)$$

$$= P(-z \leq X - Y \leq z). \quad (6.2)$$

Draw a picture of the unit square with the lines $-z = x - y$ and $x - y = z$, remembering that z is a constant of your choice, then figure out what set satisfies both desired inequalities. (A mental game you can play is this: imagine an outsider gives you a series of constants to try for z , ranging from negative to positive numbers. What scenarios occur?) In this case, we want all the points that are within distance z from the line $x = y$. Call this set of points A . For a value of z between 0 and 1, this A has area $1 - (1 - z)^2$. Thus

$$P(-z \leq X - Y \leq z) = 1 - (1 - z)^2,$$

giving us our cdf:

$$F(z) = \begin{cases} 0 & z < 0 \\ 1 - (1 - z)^2 & 0 \leq z \leq 1 \\ 1 & 1 < z \end{cases} .$$

Differentiating to get the pdf, we have

$$f(z) = \begin{cases} 0 & 0 < z \\ 2 - 2z & 0 \leq z \leq 1 \\ 0 & 1 < z \end{cases} .$$

Notice that this probability density function isn't continuous, and it takes values greater than one. Those are both fine! Probability density functions need not be continuous, and since they are not actually probabilities, they can also take values greater than one.

6.1.3 Circles: you mean I need to remember trig substitution?

Here's another example, inspired by a homework problem in Tijms' "Understanding Probability" book: Imagine you're on the High Roller Ferris wheel in Las Vegas. It is 550 feet high. The power goes out and you are stopped at a random point. What is the probability that you are stuck at a height greater than 412.5 feet?

Our strategy will go as follows:

1. Figure out what random variable X we want to know about.
2. Find a way to convert this X into a random variable Y about which we have more information.
3. Use the information about the cdf of Y to find the cdf of X .
4. Use the cdf of X to find the desired probability.

We want the cdf of the height X in feet of our car on the Ferris wheel. The difficulty is that X is not uniformly distributed – because of the shape of the Ferris wheel, we move faster through heights near 275 feet than heights near the top or bottom of the Ferris wheel. What related quantity is uniformly distributed? The angle of the spoke supporting our car on the Ferris wheel! The angle from the vertical of the spoke supporting our car moves smoothly from zero radians from the vertical, at the bottom of the wheel, to π radians from the bottom vertical, at the top of the wheel. We can use symmetry to consider the random variable Y uniformly distributed on the interval $[0, \pi]$, and then see height as $X = 275 - 275 \cos(Y)$.

Since Y is uniformly distributed on $[0, \pi]$, we have

$$P(Y \leq y) = \frac{y}{\pi}$$

for $y \in [0, \pi]$. Thus

$$P(X \leq 412.5) = P(275 - 275 \cos(Y) \leq 412.5) \quad (6.3)$$

$$= P(-275 \cos(Y) \leq 137.5) \quad (6.4)$$

$$= P(\cos(Y) \geq -0.5) \quad (6.5)$$

$$= P\left(Y \geq \frac{2\pi}{3}\right) \quad (6.6)$$

$$= \frac{2\pi}{3} \cdot \frac{1}{\pi}. \quad (6.7)$$

6.2 Expected value and variance

The expected value of a continuous random variable X is very analogous to the discrete case, just using an integral instead of the corresponding sum:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx.$$

Likewise, while the variance of a continuous random variable X is again defined as $\text{var}(X) = E[(X - \mu)^2]$, it can again by linearity be written as $\text{var}(X) = E(X^2) - \mu^2$:

$$\text{var}(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

6.3 Transformations of continuous random variables

Above you've used the fact that the probability density function is the derivative of the cumulative distribution function several times, and we can use this again to prove a general theorem for nice transformations of continuous random variables. Here, "nice" means monotonic. A function is monotonically increasing on an interval if its derivative is always positive on that interval, and it's monotonically decreasing if its derivative is always negative on that interval. You'll see where this is a useful condition.

6.3. TRANSFORMATIONS OF CONTINUOUS RANDOM VARIABLES 105

Let $g(X) = Y$ be your transformation of the continuous random variable X . Assume you know the probability density function $f_X(x)$ for X , and thus the cumulative distribution function. What's the probability density function of Y ? Find it via calculus: first, $P(Y < y) = P(g(X) < y)$. If only we could get X alone, as we understand $P(X < x)$!

Aha – if we can *invert* the function $g(x)$, this is possible. That's where “monotonic” comes in. If a function of x is monotonic on an interval, it can be inverted on that interval. There's a unique function $g^{-1}(y)$ such that $g^{-1}(g(x)) = x$ and $g(g^{-1}(y)) = y$. The graph of the function $g(x)$ passes the “horizontal line test” you may have learned at some point – a horizontal line intersects with the graph of $g(x)$ at only one point in the interval, which means that the graph of the reflection over the line $y = x$ is a function. That reflection is the graph of the inverse.

Back to the proof: if $g(x)$ is a strictly increasing function,

$$P(Y < y) = P(g(X) < y) \tag{6.8}$$

$$= P(X < g^{-1}(y)) \tag{6.9}$$

$$= \int_{-\infty}^{g^{-1}(y)} f_X(x) dx. \tag{6.10}$$

Thus by the fundamental theorem of calculus and the chain rule,

$$h_Y(y) = \frac{d}{dy} P(Y < y) \tag{6.11}$$

$$= \frac{d}{dy} \int_{-\infty}^{g^{-1}(y)} f_X(x) dx \tag{6.12}$$

$$= f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y). \tag{6.13}$$

$$\tag{6.14}$$

If $g(x)$ is strictly decreasing,

$$P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) \quad (6.15)$$

$$= \int_{g^{-1}(y)}^{\infty} f_X(x) dx \quad (6.16)$$

and by the fundamental theorem of calculus and the chain rule

$$h_Y(y) = \frac{d}{dy} P(Y < y) \quad (6.17)$$

$$= \frac{d}{dy} \int_{g^{-1}(y)}^{\infty} f_X(x) dx \quad (6.18)$$

$$= -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y). \quad (6.19)$$

We can make the formula more compact by writing

$$h_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

6.4 Markov and Chebyshev

Here I'll introduce two inequalities that do happen to hold for both discrete and continuous random variables.

Markov's inequality says that if X is a random variable taking only non-negative values, then for any $\alpha > 0$ we have

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

Notice the lack of conditions on the random variable X !

We can cleverly transform this into a related inequality, Chebyshev's inequality, by letting $X = Y - \mu$ and using an absolute value. Chebyshev's inequality says that for Y with mean μ and variance σ^2 ,

$$P(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}.$$

Why bother with these inequalities? First, they give us rough but useful bounds on probabilities. You can detect lies and mistakes with these bounds. Second, they are very useful in various proofs. Neither inequality has a lot of hypotheses that the random variable must satisfy, and so you can apply them in many situations.

6.5 Important distributions

Here, since we're focusing on applications to financial mathematics, the discussion will first relate the exponential distribution to the Poisson distribution you saw last chapter, and then discuss the normal and lognormal distributions and their applications in financial math. Last, we'll touch on a few other continuous distributions that come up in data analysis and finance.

6.5.1 Exponential distribution and Poisson

A continuous random variable X has the exponential distribution with parameter $\lambda > 0$ if it has the probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases}$$

One common use of the exponential distribution is in modeling the time until a rare event occurs, or the time between rare events. Conceptually, this is why it's related to the discrete Poisson distribution – we'll formalize this in a theorem below.

The expected value of X is $E(X) = 1/\lambda$, and the variance of X is $\text{var}(X) = 1/\lambda^2$. The cumulative distribution function is

$$P(X \leq x) = F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases}$$

Computing all of these is a good exercise (do it!).

Unfortunately, there is notational ambiguity when people write about the exponential distribution. Another way of expressing the probability density function is

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x \geq 0 \\ 0 & \text{else} \end{cases},$$

where $\beta = \frac{1}{\lambda} > 0$ is the mean and the standard deviation of the exponential random variable. When this notation is used, β is called a *survival parameter*. This comes from models in which the random variable X represents the time a system or organism manages to survive. How do you know which notation an author is using? Look at units: $E(X)$ and X must have the same units. Say X is the time before some event happens. Then β will have time units and λ will have units “events per unit time”.

Example 6.5.1. Imagine that the time to failure of a type of hard drive widely used by a large company is exponentially distributed, and on average the hard drives break every eight years. What is the probability that a hard drive will break this year?

First identify the parameter. We have “on average the hard drive breaks in eight years”, so $E(X) = 1/\lambda = 8$ years. That means $\lambda = \frac{1}{8}$. Then $P(X \leq 1) = 1 - e^{-1/8}$, which is an approximately 11.75 percent chance the hard drive will break in a year.

What’s the chance that a hard drive will break in eight years or less? Same process: $P(X \leq 8) = 1 - \frac{1}{e}$, which is approximately 63.2 percent. Why isn’t this fifty percent? Because the mean and the median are different quantities. The exponential distribution has a long right tail, in this case pulling the mean to 8 years while the median time to breakdown is about 5.5 years. (Can you find the exact median time to failure, using the definition of the median as t such that $P(X < t) = .5$?)

A very important property of the exponential distribution is *memorylessness*. This is an amazing property that roughly says that it doesn’t matter when you start counting: the time to your rare event is always (probabilistically) the same. More precisely,

$$P(X > t + s | X > s) = P(X > t)$$

for all $s, t > 0$. This is also easy to prove! Take a look:

$$P(X > t + s | X > s) = \frac{P(X > t + s)}{P(X > s)} \quad (6.20)$$

$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \quad (6.21)$$

$$= e^{-\lambda t} \quad (6.22)$$

$$= P(X > t). \quad (6.23)$$

The exponential distribution is the only continuous distribution with this property! (Which discrete distribution also has this property?)

Example 6.5.2. You're considering the hard drive that fails on average every eight years again. One of the hard drives has been working for three years. What is the probability that it will fail in the next year if failure is truly modeled exponentially?

Yep, memorylessness says that the previous three years of service don't matter. They don't make failure any more or less likely, so again you have the approximately 11.75 percent chance of failure in the next year.

In practice, failure rates are usually increasing or decreasing, rather than constant (as in a pure exponential distribution). Many components are more likely to fail as they age. Amazingly, spacecraft tend to be less likely to fail as they get older (see "Impact of the space environment on spacecraft lifetimes" by Baker and Baker). This happens when catastrophic events are likely to reveal themselves early, and if they don't occur the device is likely to work a long time. In these cases, the memorylessness property does *not* hold, and another distribution is more appropriate.

Last, let's make formal the relationship between the exponential distribution and the Poisson distribution. Let X_1, \dots, X_n be exponentially distributed random variables, all independent and identically distributed with parameter λ . Interpret each X_i as the time between event i and event $i + 1$ in a sequence of events. For any $t > 0$, define $N(t)$ as the number of events occurring in the time interval $[0, t)$. Then we have the following theorem:

Theorem 6.5.1. With X_i as above interpreted as waiting times between events, and $N(t)$ the number of events occurring in the interval $[0, t)$, we have

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

for $k = 0, 1, 2, \dots$. Notice that this means $N(t)$ is Poisson with $E(N(t)) = \lambda t$.

Likewise, you can prove that if a sequence of events is Poisson-distributed, then the waiting times between events are exponentially distributed.

It is very useful to be able to switch back and forth between consideration of the Poisson and exponential distributions. You saw some examples above, but I'll give three examples to be very explicit about which distribution to use when. Here let $N(\tau)$ be the number of events in time interval $[0, \tau)$ and let X_1 be the waiting time until the first event.

- What's the probability that no events occur in the time interval $[0, \tau)$? Use either: you're looking for $P(N(\tau) = 0)$ (probability of no events) or for $P(X_1 > \tau)$ (probability that first event happens after this interval).
- What's the probability that five or more events happen in the interval $[0, \tau)$? Poisson: it's easy to take $1 - \sum_{k=0}^5 P(N(\tau) = k)$. It is less easy to figure out the sum of X_i s, where X_i for $i > 1$ is the waiting time from the $i - 1$ th event to the i th event.
- What's the probability that you wait more than j minutes for the event? Exponential – it's tailor-made for this question. Here, you really care about the continuous random variable of time, rather than the number of events that may occur before or after j minutes.

In the second situation, you might think you could just look at the sum of five or more exponentially distributed random variables. The part I find annoying about this approach is that you need to figure out and discard the probabilities that $X_1 > \tau$, $X_1 + X_2 > \tau$, etc., and then look only at the probabilities that $\sum_{i=1}^k X_i > \tau$ for $k = 5, 6, 7, \dots$. It's just awfully complicated. However, it does bring up a question: what's the distribution of a sum of independent exponentially distributed random variables?

Write S_n for the sum of n exponentially distributed random variables. The sum S_n has the Erlang distribution. We can use convolution for continuous random variables to find the distribution for S_2 (although we'll call it Z for the next two paragraphs to cut down on the subscripts needed). Convoluting the probability density functions for X and Y gives the probability density function for $Z = X + Y$. Define the convolution of probability density functions $f_X(x)$ and $g_Y(y)$ to be

$$(f_X \star g_Y)(z) = \int_{-\infty}^{\infty} f_X(z - y)g_Y(y)dy \quad (6.24)$$

$$= \int_{-\infty}^{\infty} f_X(x)g_Y(z - x)dx. \quad (6.25)$$

The \star is the symbol for the convolution operation on two functions, just as $+$ denotes addition and \times and \cdot denote multiplication in \mathbb{R} . We'll put convolution to use right away by finding the distribution for the sum of two exponentially distributed random variables that have the same parameter λ . Just to keep our calculation cleaner, call the exponentially distributed random variables X and Y and their sum $Z = X + Y$.

Let $f_X(x) = \lambda e^{-\lambda x}$ and $g_Y(y) = \lambda e^{-\lambda y}$ be the probability density functions for the exponentially distributed random variables. Then:

$$h_Z(z) = (f_X \star g_Y)(z) = \int_{-\infty}^{\infty} f_X(z - y)g_Y(y)dy \quad (6.26)$$

$$= \int_0^z \lambda e^{-\lambda(z-y)} \lambda e^{-\lambda y} dy \quad (6.27)$$

$$= \int_0^z \lambda^2 e^{-\lambda z + \lambda y - \lambda y} dy \quad (6.28)$$

$$= \int_0^z \lambda^2 e^{-\lambda z} dy. \quad (6.29)$$

Wait, what? How do you get to change the limits of integration from $-\infty$

and ∞ to 0 and z ? The secret is that

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

and likewise for $g_Y(x)$. The part of the function that is zero is really important: $f_X(z - y) = 0$ whenever $z - y \leq 0$, so whenever $y \geq z$. Similarly, $g_Y(y) = 0$ whenever $y \leq 0$. Put those together: the integrand is zero if $y \leq 0$ or if $y \geq z$, so the integrand is only non-zero if $0 < y < z$. Having cleared that up, finish the calculation:

$$h_Z(z) = \begin{cases} \lambda^2 e^{-\lambda z} & z > 0 \\ 0 & z \leq 0. \end{cases}$$

You can extend this to the sum of n independent exponentially distributed random variables by induction.

6.5.2 Normal distribution and lognormal distribution

Here we'll talk about three things: the standard normal distribution, the normal distribution in general, and then the lognormal distribution. This discussion may motivate you to understand why transformations of random variables are important to understand more abstractly.

A continuous random variable Z has the standard normal distribution if it has the probability density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

for $z \in \mathbb{R}$. Here we are using Z for the random variable and $\phi(z)$ for the pdf very consciously – these are traditional notations that indicate we're dealing with the standard normal rather than a more general normal distribution. The standard normal distribution is the only normal distribution with mean 0 and variance 1. The cumulative distribution function for Z is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}y^2} dy.$$

There is no nicer formula for this cdf; in particular, there is no closed form. Today we often use technology to compute values of $\Phi(z)$, but we also make use of the traditional z-table for looking up values of $\Phi(z)$ given values of z .

A continuous random variable X that is normally distributed with mean μ and variance σ^2 can be obtained as an affine linear transformation of Z . Let's introduce a nice notation for indicating that a random variable X is normally distributed with mean μ and variance σ : write

$$X \sim N(\mu, \sigma^2).$$

It's a property of normal distributions that if $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. In particular, we can see that $X \sim N(\mu, \sigma^2)$ is a transformation $\sigma Z + \mu$ of a standard normal random variable Z , as by the above property $X \sim N(\sigma \cdot 0 + \mu, \sigma^2 \cdot 1)$. This also tells us how to *standardize* a normal random variable X :

$$\frac{X - \mu}{\sigma} = Z$$

brings us back to the standard normal.

Why do we care? The probability density function $f(x)$ for $X \sim N(\mu, \sigma^2)$ has a lot of symbols:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

for $-\infty < x < \infty$. (Note $\sigma > 0$.) The cumulative distribution function, again, has no closed form:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}(y-\mu)^2/\sigma^2} dy.$$

There aren't tables to look up these values except when $\mu = 0$ and $\sigma = 1$. Before the availability of computers, using the affine linear transformation to get back to a standard normal cumulative distribution function was the only way to calculate values for $F(x)$, and even now, it's often still the easiest way as algorithms for computing $\Phi(z)$ are optimized in computational software.

Another really nice property of normally distributed random variables is that if X_1 and X_2 are *independent* and normally distributed with $X_1 \sim N(\mu_1, \sigma_1^2)$

and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2$ is also normally distributed. Because X_1 and X_2 are independent, both their means and variances add, and $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Now let's look at the lognormal distribution. A continuous random variable Y is lognormally distributed if $\ln(Y) \sim N(\mu, \sigma^2)$. Another way to look at this is that $Y = e^X$ for some X that is normally distributed. This relationship tells us that the cumulative distribution function for a lognormally distributed Y is

$$F(y) = P(Y \leq y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right)$$

for $y > 0$. As a consequence, we get the probability density function

$$f(y) = \begin{cases} \frac{1}{\sigma y \sqrt{2\pi}} e^{-\frac{1}{2}(\ln y - \mu)^2 / \sigma^2} & y > 0 \\ 0 & \text{else} \end{cases}$$

From this you can deduce that $E(Y) = e^{\mu + \frac{\sigma^2}{2}}$ and $\text{var}(Y) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$. Self-quiz: since $Y = e^X$, why don't we have $E(Y)$ equal to e^μ ? This is actually quite important to understand in the context of finance.

6.6 Central Limit Theorem

The Central Limit Theorem has been called “the queen of theorems” in probability, and indeed it is amazing – and one of the primary reasons we care about the normal distribution. In essence, the Central Limit Theorem says that under certain conditions any large enough number of independent and identically distributed random variables sum to a random variable that has an approximately normal probability density function.

Let's be more precise. Let $X_1, X_2, \dots, X_i, \dots$ be a sequence of independent, identically distributed, real-valued random variables with mean μ and standard deviation $\sigma > 0$. Consider

$$S_n = \sum_{i=1}^n X_i$$

for any natural number n ; it's a partial sum of the X_i s.

You, personally, can prove using linearity of expectation and properties of variance that $E(S_n) = n\mu$ and $\text{var}(S_n) = n\sigma^2$. It's worth pointing out that as $n \rightarrow \infty$ these quantities also approach infinity (if $\mu \neq 0$). So it's not correct to talk about the limiting distribution for S_n , despite the temptation. Instead, consider the standardization Z_n of S_n :

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Theorem 6.6.1. The Central Limit Theorem: as $n \rightarrow \infty$, the distribution of the standardization Z_n of S_n converges to the standard normal distribution.

Another way to state this is to say:

Theorem 6.6.2. For X_1, \dots independent and identically distributed random variables with mean μ and variance $\sigma > 0$, and $S_n = \sum_{i=1}^n X_i$,

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x) \quad \forall x \in \mathbb{R}.$$

How big must n be for the normal approximation to be “good”? In many situations, an n of at least 30 will be big enough, but truly it depends on the original distribution of the X_i . In addition, there is another technical problem when the X_i have a discrete distribution. Approximating the resulting discrete distribution of S_n or Z_n by a continuous distribution means that the event $S_n = k$ should correspond to $\{k - \Delta \leq S_n \leq k + \Delta\}$ for $\Delta \in [0, 1)$ – but different values of Δ give different approximations. The tradition is to use $\Delta = 1/2$ and not worry too much.

Chapter 7

Random walks

Random walks bring us from discrete probability to continuous motion. We know how to look at the results of sequential coin flips. We now know about the normal distribution, as well. Last, we know about the Central Limit Theorem, which gives us some results about sums of independent identically distributed random variables. However, in finance we care about the final distributions of stock prices and returns and we often care about the path taken as well. And we know that this financial info inhabits a netherworld between discrete and continuous

7.1 Simple symmetric random walk

A simple symmetric walk has a name that makes sense. Consider a walk along a line of integers. You start at zero and move right or left one integer unit with equal probability. Mathematically, we can say that for the i th step X_i , we have pmf

$$P(X_i = 1) = 1/2 \tag{7.1}$$

$$P(X_i = -1) = 1/2. \tag{7.2}$$

The simple symmetric random walk $W^{(1)}(t)$ is an integer-time stochastic process $\{W^{(1)}(1), W^{(1)}(2), W^{(1)}(3), \dots\}$ where each $W^{(1)}(n) = S_n$ is the sum

of the steps X_i that occurred before time j :

$$S_n = \sum_{i=1}^n X_i.$$

We set $W^{(1)}(0) = S_0 = 0$. Then notice that each $W^{(1)}(n) = S_n$ gives the *position* at time n (or after n steps, since we're taking unit-time steps). If you want to know the distance from the origin at time n , you look at $|S_n|$, while the total number of steps taken is just n .

The simple symmetric walk is a great setting in which to ask a lot of questions:

- What's $E(W^{(1)}(n))$?
- What's $\text{var}(W^{(1)}(n))$?
- What's the probability that $W^{(1)}(n)$ falls below a certain value or hits a certain higher value?
- How long would it take on average to reach a certain value?

We won't answer all of these questions yet, but you can imagine why they're important and why they came up: if I bet \$20 on each round of blackjack on the Strip in Las Vegas, how long do I expect to play until I go broke? (Tip: you can find lower-priced tables if you go to old Las Vegas and get off the Strip). If I switch to \$5 tables, how long could I expect to play? I do need to throw in the caveat that these questions more properly belong in the asymmetric random walk section due to the house advantage.

7.1.1 Expectation and variance

The expectation and variance of the simple symmetric random walk $W^{(1)}(n)$ are easy calculations that show off the power of the linearity of expectation and the delight of independent random variables. Remember that each step is of size one. Since this walk is symmetric and starts at zero, notice that $E(X_i) = 0$ for all i . (Check!) Thus

$$E(W^{(1)}(n)) = E\left(\sum_{i=1}^n X_i\right) \quad (7.3)$$

$$= \sum_{i=1}^n E(X_i) \quad (7.4)$$

$$= \sum_{i=1}^n 0 \quad (7.5)$$

$$= 0. \quad (7.6)$$

Variance is another easy calculation: you should check that $\text{var}(X_i) = 1$ (check the signs!), and remember that the steps X_i are independent from each other.

$$\text{var}(W^{(1)}(n)) = \text{var}\left(\sum_{i=1}^n X_i\right) \quad (7.7)$$

$$= \sum_{i=1}^n \text{var}(X_i) \quad (7.8)$$

$$= \sum_{i=1}^n 1 \quad (7.9)$$

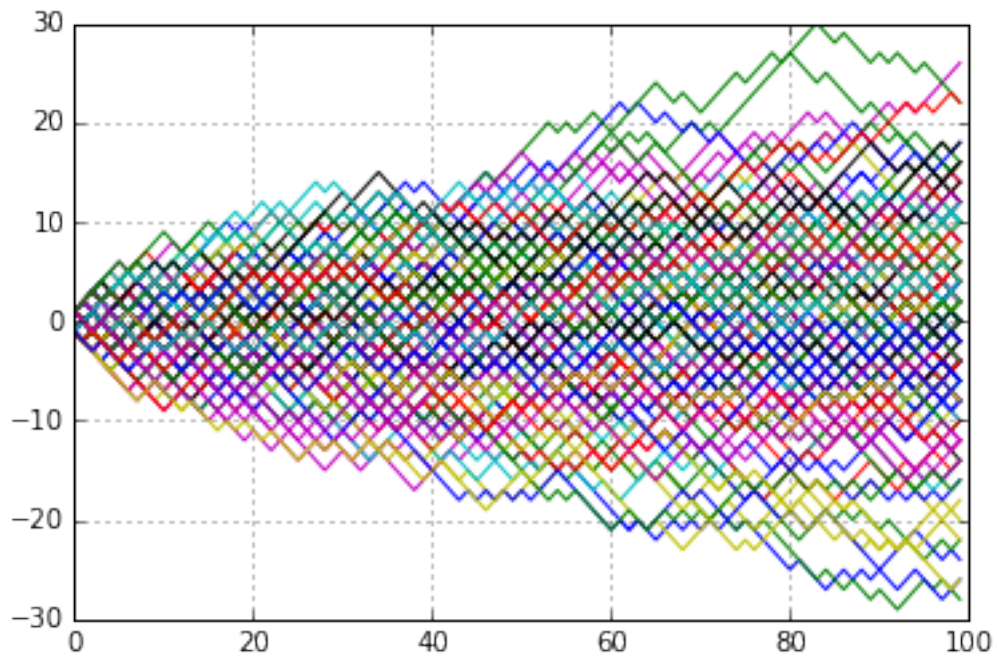
$$= n. \quad (7.10)$$

Thus the standard deviation of the position $W^{(1)}(n)$ after n independent steps is \sqrt{n} , which may or may not be surprising to you. While simple, this property of the simple symmetric random walk will play a large role below.

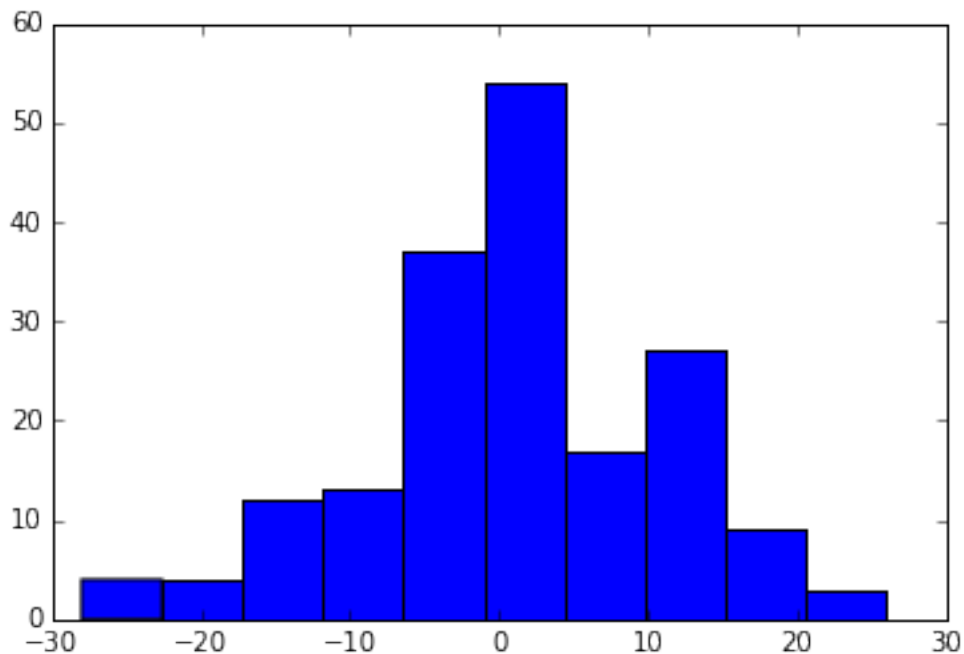
For finite n , we can approximate the distribution of $W^{(1)}(n)$ using the Central Limit Theorem. We can say that for large n ,

$$W^{(1)}(n) \approx \mathcal{N}(0, n).$$

It's useful to look at a picture of this set of random walks. Here I generate 180 simple symmetric random walks in Python:



Notice that there's a cone of possible paths, with the position S_n bounded above by n and below by $-n$ every step of the way. If we plot a histogram of the S_{100} values of these 180 paths, we see that the validity of approximating position S_n using the normal distribution begins to emerge:



As we consider asymmetrical random walks, we want to see how the cone of possible paths and the histogram of endpoints change, and as we transition to smaller and smaller steps in the walk, we want to maintain this cone of possible paths.

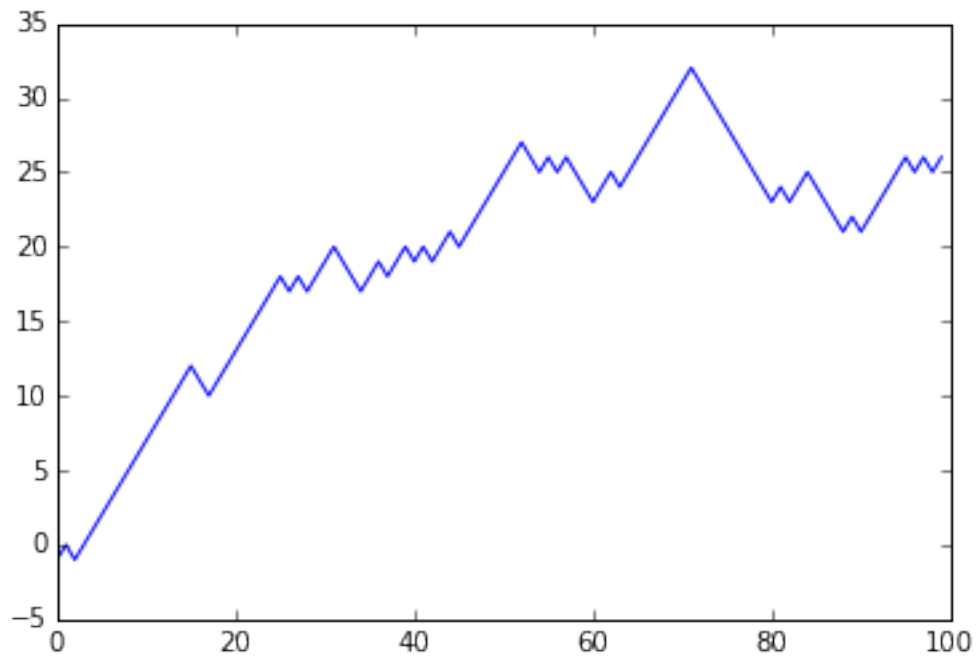
7.2 Asymmetric random walks

The first generalization of the simple symmetric random walk is the random walk with asymmetric probabilities. All we do is consider

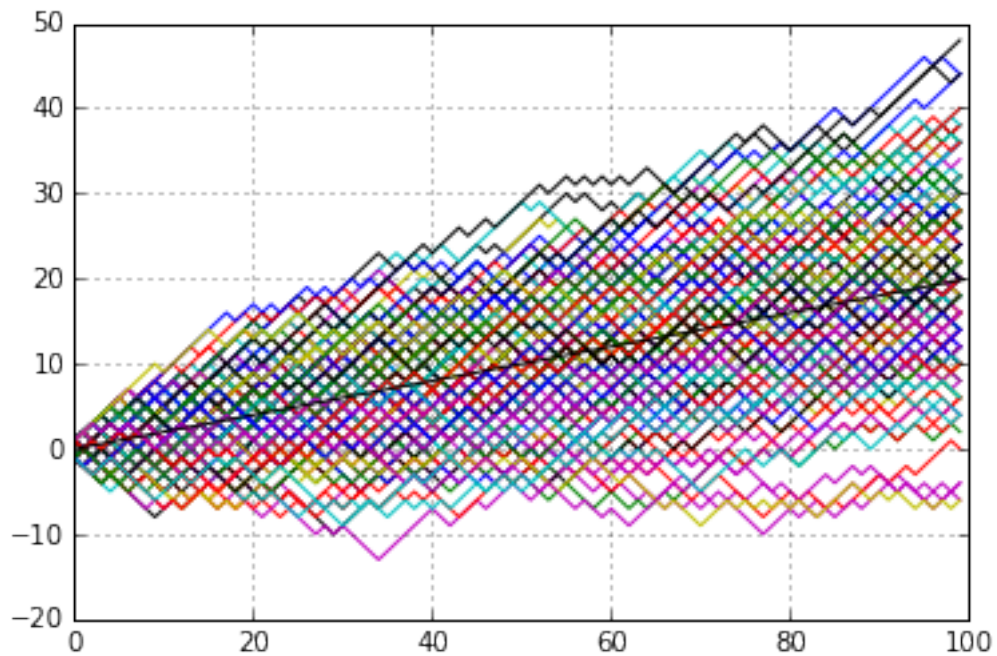
$$P(X_i = 1) = p, \quad P(X_1 = -1) = q,$$

where $p + q = 1$. How does this look?

Here is a picture of a random walk with one step of size one for every unit of time, with $p = .6$:



We can look at the following simulation of 180 such random walks. Note the black line across the top, corresponding to $(0.6 - 0.4)t$:



Notice that changing p and q doesn't change the cone of possible paths – these simple random walks are still bounded by n above and n below at step n – but it does radically change the general trajectory of the paths. Let's quantify this by looking at the expected value and variance of S_n .

7.2.1 Expected value and variance

Since $E(X_i) = p \cdot 1 + q \cdot (-1) = p - q = 2p - 1$, we have the following calculation of expectation S_n :

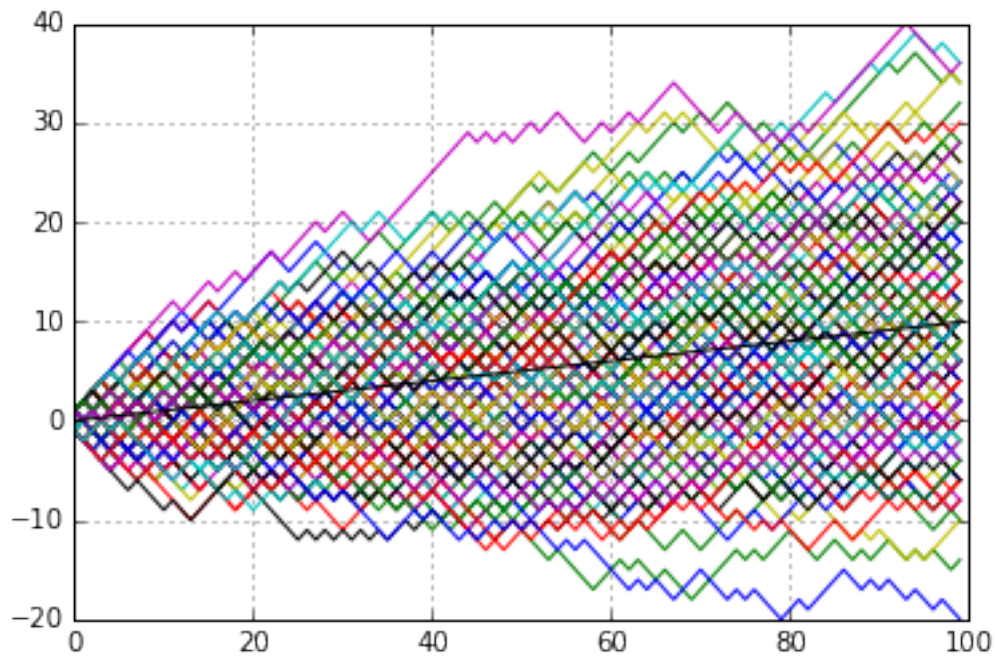
$$E(W^{(1)}(n)) = E\left(\sum_{i=1}^n X_i\right) \tag{7.11}$$

$$= \sum_{i=1}^n E(X_i) \tag{7.12}$$

$$= \sum_{i=1}^n 2p - 1 \tag{7.13}$$

$$= n(2p - 1). \tag{7.14}$$

We plotted this as a line on the figure consisting of simulated walks above. Here's another example, with $p = 0.55$.



The variance calculation is mildly more interesting:

$$\text{var}(X_i) = (p \cdot 1^2 + q \cdot (-1)^2) - (2p - 1)^2 = 4p(1 - p),$$

and with independence of the steps,

$$\text{var}(W^{(1)}(n)) = \text{var}\left(\sum_{i=1}^n X_i\right) \quad (7.15)$$

$$= \sum_{i=1}^n \text{var}(X_i) \quad (7.16)$$

$$= \sum_{i=1}^n 4pq \quad (7.17)$$

$$= 4npq. \quad (7.18)$$

For what value of p is this variance maximized? What does that mean in terms of the walks we can take?

You can calculate the value of p maximizing $\text{var}(W^{(1)}(n))$ using calculus or using symmetry of the parabola $4p(1 - p)$: no matter what method you use, variance is maximized for $p = 1/2$. Figure out how to justify this to yourself using the pictures above. Also confirm to yourself that variance is minimized for $p = 1$ or $q = 1$, and understand why.

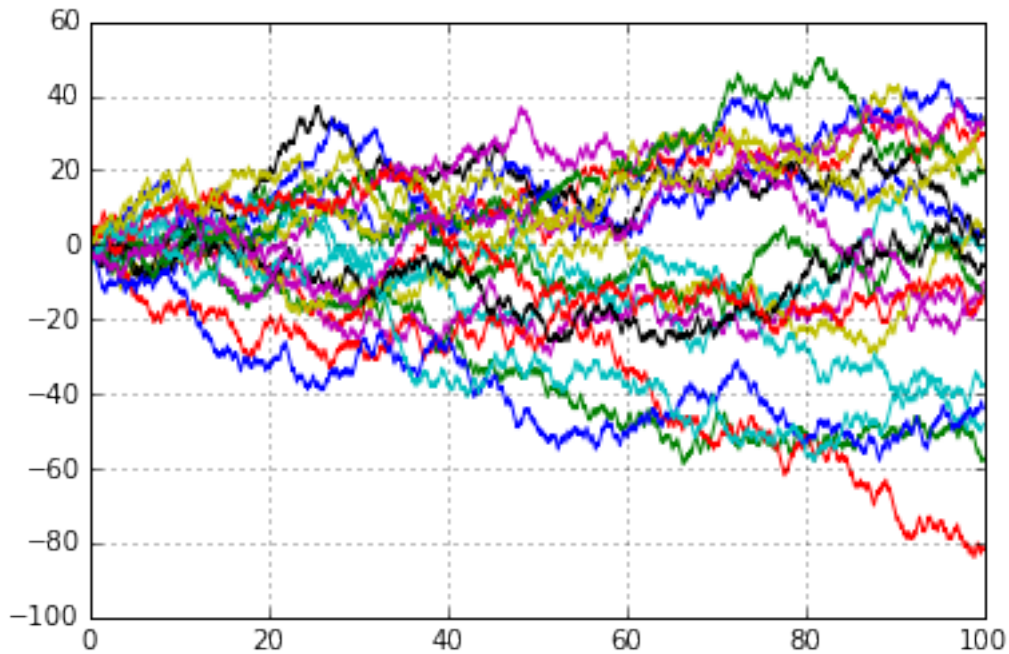
7.3 Scaling time or space

Instead of adjusting the probabilities, we could scale the frequency of the steps or the size of the steps. Let's return to the symmetric random walk so that we can mess with one idea at a time!

Return to the symmetric probability setting $P(X_i = 1) = 1/2$, but take two steps in every unit time – or three, or k , for k a positive integer (for simplicity). Before, each step took one time unit, and we conflated “number of steps” and “total time units”. Now we need to be more careful. Let's say that we take k steps in one unit of time, so in t units of time we take $n = kt$ steps in the random walk. Another way to think about this is that in one unit of time we take k steps, each taking time length $\Delta = 1/k$. We'd like to design a discrete-time stochastic process $W^{(k)}(t)$ that just gives a finer and finer random walk, without blowing up the position of the walker (as k changes we want variance of $W^{(k)}(t)$ for any particular time t to stay constant).

For example, here are twenty random walks that each have

- a step distance of 1 unit per step
- ten steps per unit time (so $k = 10$, while $\Delta = 0.1$.)
- all over 1,000 total steps or 100 units of time



Look how some paths get near 40 or -40 just 20 time units in. The variance of this random walk process is much larger than our previous random walks: for this particular set of 20 trials, we have a variance at time 100 of 1022.51. Variance is about ten times bigger than the time length of the random walk, and that's no coincidence. What if we let $k = 12345$? Then variance would be about $12345t$ at time t .

As k grows, we'll simply move away from the origin more rapidly as we walk – the variance of the position at time t will get larger as k gets larger. This does not preserve the properties of random walks that we wanted. We want a finer random walk (more steps k per unit time) that keeps variance at time t constant even as k changes. Scaling only the frequency of steps we take (time) while leaving the length of the step (space) at one results in a variance of kt at time t .

Solution: as we scale time, we must scale space as well. *If we move more often we've got to take smaller steps* in order to keep variance constant with respect to step frequency. We want the variance of our stochastic process $W^{(k)}(t)$ to be T at time $t = T$, not kT . This can be accomplished by scaling distance

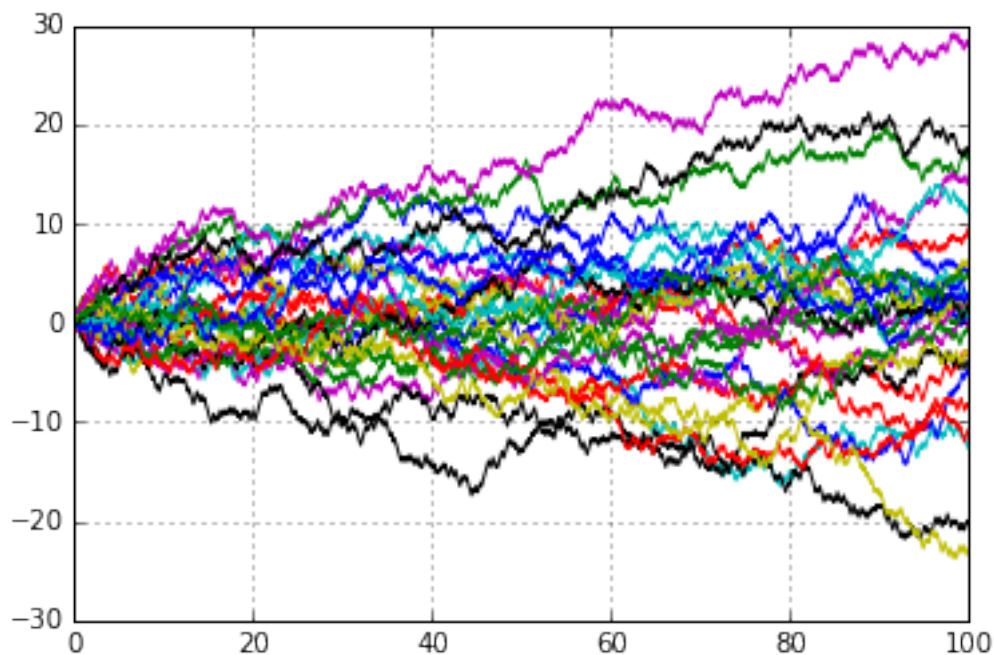
to take out that extra k :

$$W^{(k)}(t) = \sum_{i=1}^{n=kt} \frac{1}{\sqrt{k}} X_i = \sum_{i=1}^{n=kt} Y_i,$$

for $Y_i = \frac{1}{\sqrt{k}} X_i$.

Take a look. In this set of 30 random walks, we took

- ten steps per unit time (so $k = 10$, while $\Delta = 0.1$)
- each step of length $1/\sqrt{10}$, about 0.316
- all over 1,000 total steps or 100 units of time.



At time 100, the variance for the final endpoints of these random walks is 105.769 – much closer to 100, as desired.

The image above has a random selection of random walks, rather than all possible ones. With the scaling $Y_i = \frac{1}{\sqrt{k}} X_i$, what is the maximum distance one

could go in time t , starting from the starting point of $W^{(k)}(0) = 0$? With kt steps of length $\frac{1}{\sqrt{k}}$ at time t , the maximum distance traveled would be $\frac{kt}{\sqrt{k}} = \sqrt{kt}$. Note that this does grow with k .

7.3.1 Special properties

Three special properties stand out for these random walks that will also be very important in the discussion of Brownian motion:

- **Independent increments:** for $0 < s < t < l < k$, $W^{(k)}(t) - W^{(k)}(s)$ and $W^{(k)}(k) - W^{(k)}(l)$ are independent.
- **Expected value is zero for our symmetric set-up,** $E(W^{(k)}(t)) = 0$, and variance is proportional to interval: $\text{var}(W^{(k)}(t)) = t$.
- **By the Central Limit Theorem,** as $n = kt$ grows large, $W^{(k)}(t)$ is well-approximated by a normal distribution with the above mean and variance.

Once we transition to Brownian motion, we'll see that independence of increments and the normality of the distribution at a given time t are important carryovers. The difference between our random walks and Brownian motion is that Brownian motion is continuous.

7.4 Arithmetic Brownian motion

Here we will only talk about one-dimensional Brownian motions. (When we say one dimension, we mean one space dimension – we're not counting time.) It's possible to look at two-dimensional or multi-dimensional Brownian motions, which could be useful when modeling a portfolio of stocks, for instance.

Brownian motion is a *continuous-time process*. In the random walks above, we only took steps at times that were multiples of $1/k$ for k an integer. They were discrete-time stochastic processes. Brownian motion involves steps at all times $t > 0$, $t \in \mathbb{R}$.

Let's build up from the simple symmetric random walk, $W^{(1)}(t)$. Let $W^{(k)}(t)$ be the scaled symmetric random walk defined before. We can write a scaled random walk with drift μ and “diffusion coefficient” σ as

$$B^{(k)}(t) = \sigma W^{(k)}(t) + \mu t.$$

Recall this will have the properties that

- For $0 \leq s < t < u < v$, we have independent increments $B^{(k)}(t) - B^{(k)}(s)$ and $B^{(k)}(v) - B^{(k)}(u)$.
- For $0 \leq s < t$, the distribution of $B^{(k)}(t) - B^{(k)}(s)$ depends only on $t - s$.
- In particular, for large s and t , $0 \leq s < t$, $B^{(k)}(t) - B^{(k)}(s)$ is approximated by the distribution $\mathcal{N}(\mu(t - s), \sigma^2(t - s))$.

Take the limit as $k \rightarrow \infty$, and get a *stochastic process*. Taking the limit preserves the properties of stationary and independent increments which are normally distributed. So let B_t be the limiting process

$$B_t = \lim_{k \rightarrow \infty} \sigma W^{(k)}(t) + \mu t.$$

Now we have a real-valued process B_t , $t \geq 0$, with

- Independent increments: for $0 < s < t < u < v$, $B_t - B_s$ and $B_v - B_u$ are independent.
- For $0 \leq s < t$, the increment $B_t - B_s$ has distribution $N(\mu(t - s), \sigma^2(t - s))$.
- With probability one, B_t is continuous!

This is called a Brownian motion with drift μ and diffusion coefficient σ . The “standard” version, with $\mu = 0$ and $\sigma = 1$, is called standard Brownian motion or the Wiener process, and we often denote this by W_t .

The relationship between any two random walks $B^{(k)}$ and $B^{(\ell)}$ is just a change of coordinates, scaling space and time if k and ℓ are different and scaling the drift and diffusion if they differ. The Wiener process (Brownian motion) is the limit of a simple symmetric random walk as k goes to infinity (as step size goes to zero). Thus Brownian motion is the continuous-time limit of a random walk.

Note that if we're being very specific, we could call this an arithmetic Brownian motion. Here we are drawing the distinction between *arithmetic* and *geometric*, and this corresponds exactly to the distinction between an additive binomial tree and a multiplicative binomial tree.

7.5 Geometric Brownian motion

Geometric Brownian motion is the next logical step. If B_t is an (arithmetic) Brownian motion, we can make a geometric Brownian motion S_t by defining

$$S_t = S_0 e^{B_t}.$$

This is again a continuous stochastic process, and we can show it's got properties similar to the arithmetic Brownian motion B_t . The process S_t has stationary and independent increments, just like B_t . On the other hand, you can see that at time zero we start with S_0 , any number we like, although we generally use a positive number in financial modeling. (In particular we don't use $S_0 = 0$ because that would result in a constant function.) Last, since at a particular t_1 we have normally distributed B_{t_1} , we also have a lognormally distributed S_{t_1} . Everything you learned about the lognormal distribution applies at each moment in time to a geometric Brownian motion.

Previously, we saw that (arithmetic) Brownian motion comes about from scaling and taking the continuous limit of an (additive) random walk. Geometric Brownian motion comes instead as the limit of a multiplicative random walk: look at an initial stock price, S_0 , multiplied by factors L_i at each time step:

$$S_n = S_0 L_1 L_2 \dots L_n.$$

If we model these L_i as independent and identically distributed random variables, applying the logarithm allows us to write the equation additively and then the Central Limit Theorem applies.

$$\ln S_n = \ln S_0 + \sum_{i=1}^n \ln L_i$$

The random variables $\ln L_i$ are still independent and identically distributed. You can imagine taking more and more multiplicative factors L_i at shorter and shorter time steps, just as we did with arithmetic Brownian motion.

7.6 Solving Brownian motion problems

How do we solve these problems?

Example 7.6.1. With a standard Brownian motion W_t , solving problems is often easy! Since $W_t \sim N(0, t)$, we can standardize and normalize with ease. Let Z be the standard normal continuous random variable. Take a look:

$$P(W_4 < 0) = P(Z < 0) = 1/2.$$

Here we're definitely using symmetry of the normal distribution. A step up:

$$P(W_{100} < W_{80}) = P(W_{100} - W_{80} < 0) = P(Z < 0) = 1/2.$$

Now we're using stationary increments: we know $W_{100} - W_{80} \sim N(0, 100 - 80)$. Mildly more difficult:

$$P(W_{100} < W_{80} + 2) = P(W_{100} - W_{80} < 2) \tag{7.19}$$

$$= P(W_{20}/\sqrt{20} < 2/\sqrt{20}) \tag{7.20}$$

$$= P(Z < 2/\sqrt{20}) \tag{7.21}$$

$$= \Phi(2/\sqrt{20}). \tag{7.22}$$

Since $2/\sqrt{20} \approx 0.447$, this probability is between .6700 and .6736 using a z -table. And now let's combine some conditions:

$$P(W_3 < W_2 + 2 \text{ and } W_1 < 0) = P(W_3 - W_2 < 2 \text{ and } W_1 - W_0 < 0) \quad (7.23)$$

$$= P(W_3 - W_2 < 2)P(W_1 - W_0 < 0) \quad (7.24)$$

$$= P(Z < 2)P(Z < 0) \quad (7.25)$$

$$= \Phi(2) \cdot \frac{1}{2}. \quad (7.26)$$

Here I write $W_1 - W_0$ to really emphasize that I'm looking at an increment that does not overlap with $W_3 - W_2$, even though they have the same distribution. The *distribution* of what happens between times 2 and 3 doesn't depend on prior behavior, even though actual position at time 3 certainly depends on position at time 2! So I'll end with a question we can't solve with current tools:

$$P(W_3 < 2 \text{ and } W_4 > 4).$$

Do you see any way to transform this into a question about non-overlapping, and thus independent, increments? I don't! In fact, that position at time 4 *is* influenced by position at time 2. We'll need to figure out how, which will bring us to considering the covariance matrix of $\{W_t\}$ at integer times.

Questions about Brownian motion with drift are similar, although standardizing requires another step.

Example 7.6.2. Let X_t be a Brownian motion with drift $\mu = 2$ and diffusion $\sigma = 3$, so

$$X_t = 2t + 3W_t.$$

Find $P(X_4 < X_5 + 1)$.

$$P(X_4 < X_5 + 1) = P(X_4 - X_5 < 1) = P(X_5 - X_4 > -1).$$

At this point I always stop to check some things: $X_5 - X_4$ has what distribution again?

$$X_5 - X_4 = 2 \cdot 5 + 3W_5 - (2 \cdot 4 + 3W_4) = 2 + 3(W_5 - W_4).$$

Since $W_5 - W_4 \sim N(0, 1)$, we'll call it Z and write

$$X_5 - X_4 = 2 + 3Z.$$

Let's get back to the calculation:

$$P(X_4 < X_5 + 1) = P(Z > -1) \tag{7.27}$$

$$= 1 - P(Z < -1) \tag{7.28}$$

$$= 1 - \Phi(-1) \tag{7.29}$$

This is approximately 0.8413.

Likewise, questions about geometric Brownian motion usually just involve taking a logarithm and then following the steps shown above. In list form,

- take appropriate logarithms to go from geometric Brownian motion to arithmetic Brownian motion
- translate from arithmetic Brownian motion to normal random variables
- standardize those normal random variables
- use a z-table and/or symmetry of the normal distribution to get numbers.

As illustrated above, you might get stuck if you can't separate your computations into computations about independent increments – but otherwise the world is yours!

Chapter 8

Linear algebra

While I'm assuming you've encountered vectors and matrices in previous math classes, we'll start with a short review. Then in the next few chapters, we'll cover elements of linear algebra, multivariable calculus, and differential equations that provide a nice base for financial math. Financial information is almost always multivariate: as a portfolio manager, you manage multiple assets; as a risk analyst, you look at multiple risks. Combining these classical multivariate topics with statistics will give you access to powerful mathematical techniques.

8.1 What is a vector? What is a matrix?

Most generally, a *vector* is an element of a *vector space*. Often, we care about the vector spaces \mathbb{R}^n or \mathbb{C}^n , in which case

- the vector is an n -tuple of real or complex numbers (a list in which order matters), or
- (equivalently) a direction with a magnitude.

There are examples that are rather different, but we'll save those for discussion later.

We'll use the notation \vec{v} for a vector, and can write for instance

$$\vec{v} = [v_1, \dots, v_n] \in \mathbb{R}^n$$

for a vector whose components v_i for $i = 1, \dots, n$ are all real numbers.

A *vector space* or *linear space* is a set of vectors that is closed under scalar multiplication and vector addition. This word “scalar” refers to a number that scales things – a stretch factor – and so for a real vector space, a scalar is a real number, and for a complex vector space, a scalar is a complex number. Notice that any vector space *must* include the zero vector (which we write $\vec{0}$) because if scaling a vector by any scalar is an operation that keeps the output in the vector space, we've got to be able to scale by zero.

Here are some examples of vectors and vector spaces:

Example 8.1.1. Think of all the vectors in \mathbb{R}^3 that have zero for their z -coordinate. Call that vector space V , and write V as

$$V = \{\vec{v} = [x, y, z] \in \mathbb{R}^3 \mid z = 0\}.$$

Check that you can add or subtract two such vectors and still have $z = 0$. Check that you can multiply the vector by any number in \mathbb{R} and stay in V . Remember multiplication of a vector by a scalar is done element by element, so $c[x, y, 0] = [cx, cy, c \cdot 0] = [cx, cy, 0]$.

Example 8.1.2. Another vector space $W \subset \mathbb{R}^3$ is given by all the vectors $\vec{v} = [x, y, z]$ such that $x + y + z = 0$. You can graph this. Do you know what surface this gives in \mathbb{R}^3 ?¹ Vectors in this vector space include $[1, 2, -3]$ and $[-\pi, 38, -38 + \pi]$. Check for yourself that this vector space is closed under scalar multiplication and vector addition (that is, for all $\vec{v}, \vec{w} \in W$, $a\vec{v} + b\vec{w} \in W$ for scalars $a, b \in \mathbb{R}$).

If a vector space is contained in another vector space, we say it is a vector subspace of the larger vector space. In our examples immediately above, both V and W are vector subspaces of \mathbb{R}^3 , but since neither V nor W contains the

¹A plane that goes through the origin!

other one, neither is a vector subspace of the other. However, both V and W contain the intersection $V \cap W$, the set of vectors $\vec{v} = [x, y, z]$ in \mathbb{R}^3 with $z = 0$ and $x + y + z = 0$. That vector space, Q , is given by

$$Q = \{[x, y, z] \in \mathbb{R}^3 \mid x + y = 0, z = 0\}.$$

In fact, that's a line through the origin given by $x + y = 0$ and $z = 0$, and we could rewrite it as

$$Q = \{[x, -x, 0] \in \mathbb{R}^3\}.$$

If you like parameterizations (I do!) we could use a parameter $t \in \mathbb{R}$ and write this yet one more way:

$$Q = \{t[1, -1, 0] \in \mathbb{R}^3 \mid t \in \mathbb{R}\}.$$

What is a matrix? Mechanically, a matrix is made by stacking vectors as rows or columns to make a rectangular array of numbers. In this book we'll most frequently encounter matrices of numbers, but we also make matrices of symbols or expressions (you'll notice this in the section on rotation matrices, for instance).

An example of a matrix of real numbers would be

$$\begin{bmatrix} 1 & 2 & 3 \\ -2 & \pi & -1.5 \end{bmatrix}.$$

This is a two by three matrix. An example of a matrix of complex numbers would be

$$\begin{bmatrix} i & 2 & 3 - i \\ 0 & i\pi & -1.5i \end{bmatrix}.$$

This is also a 2×3 matrix. When we need a very general $m \times n$ (m by n) matrix called A , we can write

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}.$$

You can multiply matrices by scalars (just multiply each element by the scalar) and you can add matrices (also element by element). The observant among you might think, hmm, does that mean we can make a vector space out of matrices? Yes, you can! The vector space of all $m \times n$ matrices with real entries is often called $Mat_{m \times n}(\mathbb{R})$, or $M_{m \times n}$, or $\mathbb{R}^{m \times n}$. Somehow you have to indicate the dimensions of the matrix and what the entries are allowed to be.

8.2 Linear combinations and matrix multiplication

8.2.1 Linear combinations

First, let's look at the idea of linear combinations of vectors and relate it to matrices. Here's a very simple motivating example:

The equation $2x + y - 4z + w$ is a linear combination of the variables x, y, z and w . It can be expanded as a dot product,

$$[x, y, z, w] \cdot [2, 1, -4, 1],$$

or as a product of matrices,

$$\begin{bmatrix} 2 & 1 & -4 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}.$$

Here is another motivating example:

An example of a linear combination of the vectors $\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -7 \\ 2 \\ 4 \end{bmatrix}$ is given by

$$3 \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} -7 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ 7 \end{bmatrix}.$$

Using matrix multiplication, we could also write

$$\begin{bmatrix} 3 & -7 \\ 2 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ 7 \end{bmatrix}.$$

8.2.2 Matrix multiplication and dot products

Linear combinations of vectors can always be expressed via matrix multiplication, and matrix multiplication is built out of dot products. The *dot product* of two real vectors \vec{a} and \vec{b} in \mathbb{R}^n is

$$\vec{a} \cdot \vec{b} = [a_1, \dots, a_n] \cdot [b_1, \dots, b_n] = \sum_{i=1}^n a_i b_i.$$

We are *only* defining dot product for real vectors right now (complex vectors will show up in [Section 9.10](#)). For real vectors, $\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$, so dot product is commutative (not true for complex inner product!). Multiplication by scalars (real numbers) distributes over dot product, too: $\vec{a} \cdot (c\vec{b}) = c(\vec{a} \cdot \vec{b}) = (c\vec{a}) \cdot \vec{b}$. Often with commutativity we talk about associativity (for instance $(3+2)+1 = 3+(2+2)$), but for dot product this does not make sense: $\vec{a} \cdot (\vec{b} \cdot \vec{c})$ is not an operation that makes sense, as you can't dot a vector and a scalar.

Geometrically, the dot product $\vec{a} \cdot \vec{b}$ is related to the angle between the vectors \vec{a} and \vec{b} . Pick the smallest possible angle between the two vectors, θ between zero and π . Define

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos \theta.$$

If \vec{a} and \vec{b} are perpendicular to each other, then the angle between them is $\pi/2$ radians and $\vec{a} \cdot \vec{b} = 0$.

This uses the magnitude of \vec{a} and \vec{b} : for $\vec{a} \in \mathbb{R}^n$, we can define the magnitude $\|\vec{a}\|$ by

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2}.$$

This should look a lot like Euclidean distance to you (if you're not sure, just think about $\vec{a} \in \mathbb{R}^2$). It is exactly that, the distance from the tail of \vec{a} at the origin to the tip of the vector \vec{a} . Notice too then that $\vec{a} \cdot \vec{a} = \|\vec{a}\|^2$. This will be useful.

Matrix multiplication, then, is built from dot products as follows: Let C be an $m \times n$ matrix with rows given by \vec{c}_1 through \vec{c}_m . Let D be an $n \times p$ matrix with columns given by \vec{d}_1 through \vec{d}_p . We can write the matrix multiplication as

$$CD = \begin{bmatrix} \leftarrow & \vec{c}_1 & \rightarrow \\ \leftarrow & \vec{c}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \vec{c}_m & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{d}_1 & \vec{d}_2 & \dots & \vec{d}_p \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \vec{c}_1 \cdot \vec{d}_1 & \vec{c}_1 \cdot \vec{d}_2 & \dots & \vec{c}_1 \cdot \vec{d}_p \\ \vec{c}_2 \cdot \vec{d}_1 & \vec{c}_2 \cdot \vec{d}_2 & \dots & \vec{c}_2 \cdot \vec{d}_p \\ \vdots & & \ddots & \vdots \\ \vec{c}_m \cdot \vec{d}_1 & \vec{c}_m \cdot \vec{d}_2 & \dots & \vec{c}_m \cdot \vec{d}_p \end{bmatrix}$$

Make sure you know the difference between a row vector and column vector! A row vector looks like

$$\vec{r} = [r_1, r_2, \dots, r_m] \in \mathbb{R}^m,$$

for instance, while a column vector might be

$$\vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{R}^n.$$

Often people (like me) are rather sloppy and switch back and forth between writing a vector in \mathbb{R}^n as a row or column vector depending on how much paper they have. That's not terrible. But being clear in calculations about

whether you're using a row vector or column vector is important. For instance, what is the difference between multiplying two vectors as matrices and taking the dot product between two vectors with the same shape?

Example 8.2.1. Let $\vec{v} = [1, 2]$ and $\vec{w} = [-1, 1]$. Then compute the following, with the knowledge that the T stands for *transpose* (exchange rows and columns, or flip over the diagonal):

$$\vec{v} \cdot \vec{w}$$

$$\vec{v}\vec{w}$$

$$\vec{v}^T \vec{w}$$

$$\vec{v}\vec{w}^T$$

Now check your answers in the footnote. ²

8.3 Geometry and linear algebra

Linear algebra is, unsurprisingly, rather... linear. But it's got a lot of geometry going on with all those angles and lengths and volumes. In this section we'll discuss planes and parametrizations as a great first example, then tackle determinants, cross products, and projection. We'll focus on the geometric meaning as a means of beginning the integration³ of linear algebra with multivariable calculus. Some of these techniques are specific to low dimensions (\mathbb{R}^2 and \mathbb{R}^3) but they can give intuition to the generalizations to higher dimensions. As high-dimensional data analysis becomes an ever more important part of the landscape of mathematics, finance, and industry, this is useful!

²(Dot product is 1, second product makes no sense for dimension reasons, third is $\begin{bmatrix} -1 & 1 \\ -2 & 2 \end{bmatrix}$, and fourth is 1, the same as the dot product.)

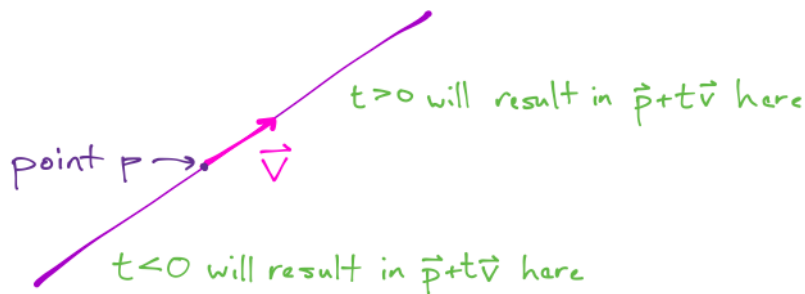
³haha

8.3.1 Planes and parameterizations

In this section we'll talk about equations of planes, but maybe it's good to start with the equation of a line as a test case. You can write the equation of a line L in \mathbb{R}^3 in several ways. One of my favorite ways to write a line in \mathbb{R}^3 is to parameterize it. I will use the parameter $t \in \mathbb{R}$. It's just a scalar. Then we need a direction for the line – call it \vec{v} – and a point that lies in the line – call that \vec{p} . A set-theoretic description of the points in the line, then, is

$$L = \{\vec{x} \in \mathbb{R}^3 \mid \vec{x} = t\vec{v} + \vec{p}\}.$$

When $t = 0$ we're just at the point \vec{p} , and as t ranges through the rest of \mathbb{R} , we just scale along the vector \vec{v} .



This concept of parameterization will be very useful in exploring planes and other multivariate curves, surfaces, solids, etc. Among other things, parameterization can help us understand the notion of dimension both in linear and nonlinear contexts.

We can use matrix multiplication or dot product to write the equation of a plane in \mathbb{R}^3 . For instance, the equation $3x - y + 2z = 5$ can be rewritten as $(3, -1, 2)(x, y, z)^T = 5$. This is a logical condition – a constraint – that picks out a certain set of points in \mathbb{R}^3 . The plane given by $3x - y + 2z = 5$ is an affine subspace of \mathbb{R}^3 , not a vector space in \mathbb{R}^3 . Why? First, check to see if the space $\{[x, y, z] \in \mathbb{R}^3 \mid 3x - y + 2z = 5\}$ is closed under scalar multiplication and vector addition. (Remember zero is a scalar!) What do you find? Second, ponder this question:

Example 8.3.1. What is the difference between a point in this plane and a vector in this plane? Find an example of a point in the plane and a vector lying in the plane.

You probably found your own examples, but $(0, -5, 0)$ and $(0, 0, 2.5)$ are two points in the plane – they satisfy the equation $3x - y + 2z = 5$. By contrast, $[0, -5, -2.5]$ is a vector *in* the plane. A few different ways to say that: it's a vector that lies parallel to the plane when based at the origin, so we can translate it to lie in the plane; it is a vector that goes between two points in the plane; it's a vector perpendicular to the normal vector $(3, -1, 2)$, which is enough to characterize the plane since we're in \mathbb{R}^3 .

A plane can be parametrized using two variables, for instance s and t , because it's a two-dimensional object. For example, let $x = s$, $y = t$, and $z = 2.5 - 1.5s + 0.5t$. Notice this satisfies our previous Cartesian equation, $3x - y + 2z = 5$! We can write the parametric vector-valued equation like this:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} s \\ t \\ 2.5 - 1.5s + 0.5t \end{pmatrix}.$$

If we wish to drop the x, y, z and write a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, we can write the below expressions as well as the previous:

$$f(s, t) = s \begin{pmatrix} 1 \\ 0 \\ -1.5 \end{pmatrix} + t \begin{pmatrix} 0 \\ 1 \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 2.5 \end{pmatrix}$$

or

$$f(s, t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1.5 & 0.5 \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 2.5 \end{pmatrix}.$$

Notice here that I've now written the expression as a linear shift of the linear combination of two vectors that lie in the plane, or as the linear shift of a matrix product. While simply a cosmetic rewrite, this points out a larger lesson. Explore this through examples:

Look at

$$(x_1 \ x_2 \ x_3) \begin{pmatrix} 3 & 2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix}$$

How can this product be expanded? Write the product in two ways, as a single row vector and as the linear combination of three row vectors in \mathbb{R}^2 .

Do this again for another product:

$$\begin{pmatrix} 3 & 2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Again, rewrite the product as a single column vector in \mathbb{R}^3 and as the linear combination of two column vectors in \mathbb{R}^3 .

What do you notice about the linear combinations and vector spaces that result?

This leads to the useful concepts of *row space* and *column space*, explored a little later.

8.3.2 Projection

Let's get back to lines for a bit. One common question that arises in physics, finance, and statistics is this: "I have a vector \vec{a} that I want to explore, and another vector \vec{b} as a reference of sorts. How much of vector \vec{a} points in the direction of \vec{b} ? What does this even mean?"

Examples you may have seen include the classic physics question about a block on a slippery slope. That block is going to slide down the slope, but how fast? What's the component in the direction of gravity, and what's the horizontal component? We can use similar ideas closer to finance, in ideas of decomposing a company's stock's movement into the component due to "the market" and the component due to the company itself. (Insert discussion of how this relates to alpha, beta here.***) Later, we'll employ Gram-Schmidt decomposition (Section ??) and singular value decomposition (Section 9.13)

to pursue these ideas.

For the moment, though, let's just look at projection. Let's define the projection of a vector \vec{a} onto another vector \vec{b} . Imagine that a projection is the shadow of \vec{a} on the line in direction \vec{b} if the sun is "directly overhead."



The algebraic formulation is

$$\text{proj}_{\vec{b}}\vec{a} = \frac{\vec{a} \cdot \vec{b}}{|\vec{b}|^2}\vec{b}.$$

Notice that this vector has a direction (it goes in the direction of \vec{b}) and a magnitude (the magnitude of the projection is $|\vec{a}| \cos(\theta)$, where θ is the angle between the vectors \vec{a} and \vec{b}). You could figure out this definition of projection yourself by looking at the natural geometric expression $|\vec{a}| \cos(\theta) \frac{\vec{b}}{|\vec{b}|}$ and using $\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos(\theta)$ to prove the formula in terms of the dot product.

In particular, this helps us write \vec{a} as the sum of a component in the direction of \vec{b} and a component perpendicular to \vec{b} . We can get the component perpendicular to \vec{b} by simply subtracting:

- check that

$$\vec{c} = \vec{a} - \frac{\vec{a} \cdot \vec{b}}{|\vec{b}|^2}\vec{b} = \vec{a} - \text{proj}_{\vec{b}}\vec{a}$$

is a vector orthogonal to \vec{b} , and check that

- check that $\text{proj}_{\vec{b}}\vec{a} + \vec{c} = \vec{a}$.

8.3.3 Determinants

One must reckon with determinants of square matrices at some point in linear algebra. It's time. We will in general automate computation of determinants, so I will include here only the 2×2 and a few comments on determinants and volume.

The determinant of the 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is $\det A = ad - bc$. This can be positive or negative or zero.

- Prove that if the determinant is zero, then the first column is a scalar multiple of the second column.
- Demonstrate to yourself that if the determinant is negative, then the linear transformation $T(\vec{x}) = A\vec{x}$ “switches the order of” the standard Euclidean basis vectors \vec{e}_1 and \vec{e}_2 (changing the orientation, as a reflection would)
- Draw some of examples of transformations of the unit square via A and demonstrate to yourself that the absolute value of the determinant, $|\det A|$, gives the area of the transformed unit square.

For larger $n \times n$ matrices, we often use the Laplace expansion. The Laplace expansion is a specific method of breaking an $n \times n$ determinant into n smaller $(n-1) \times (n-1)$ determinants. It stems from a larger idea, that you could actually compute determinants by taking $n!$ products of matrix entries and summing them with sign coming from number of inversions in the order of rows picked out: see picture *****

The Laplace expansion makes this systematic and hides much of the work:

- Set the following notation: $\hat{A}_{i,j}$ is the $(n-1) \times (n-1)$ submatrix of the $n \times n$ matrix A you get by dropping row i and column j .
- Pick the top row of A to expand along for this Laplace expansion. This is a choice; you could use any row or column.

- Expanding along the top row,

$$\det A = \sum_{j=1}^n (-1)^{j+1} a_{1,j} \det \hat{A}_{1,j}.$$

For large matrices, you'd not want to do this by hand, but as a recursive procedure it's a relatively easy algorithm to implement in a computer. Of course, determinants are already implemented in almost any programming language you'd care to use at work, often with a number of optimizations that we don't have time to cover in this class.

8.3.4 Cross products

The cross product of two vectors in \mathbb{R}^3 gives a vector that is perpendicular to both of the input vectors.

It's is a funny product, as it's only defined in \mathbb{R}^3 . Contrast this with dot product – $\vec{a} \cdot \vec{b}$ makes sense in any positive dimension – or in the next section, determinant, which makes sense for any square matrix. Why only \mathbb{R}^3 for the determinant?

Because of this odd constraint, the cross product is not that useful in finance, but it's common enough in linear algebra examples that we should not skip it here. Here's a definition:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ -(a_1 b_3 - a_3 b_1) \\ a_1 b_2 - a_2 b_1 \end{pmatrix}.$$

Check a few easy properties of this formula:

- $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$, so in particular this is not a commutative product! order matters!
- $\vec{a} \times c\vec{a} = \vec{0}$ for $c \in \mathbb{R}$, as a special case

I find this formula almost intolerable to remember as I am generally allergic to formulas, so instead I use one of the following methods: the “cover up each row” method or the “determinant of three-by-three/stripes” method. Both methods rely on knowing how to compute the determ

Cover up each row:

8.4 Useful inequalities for vectors

It’s important to mention a few special inequalities that are useful for linear algebra and for finance. First, the triangle inequality. Back in [Chapter 2](#), we encountered a version of the triangle inequality: two sides of a triangle have to sum to a number larger than or equal to the third side of the triangle. Now let’s use vectors to express the same idea. Look at a triangle with sides \vec{v} , \vec{w} , and $\vec{v} + \vec{w}$, with all vectors in \mathbb{R}^n . Then the triangle inequality is

$$\|\vec{v}\| + \|\vec{w}\| \geq \|\vec{v} + \vec{w}\|.$$

Why is this discussed right after all the material on dot products for real vectors? Because you can rewrite those magnitudes as dot products and use the Cauchy-Schwarz inequality to prove the triangle inequality in an elegant way.

The Cauchy-Schwarz inequality is really fundamental to any inner product space – the dot product is an example of a more general inner product. Using the notation we’ve been using in this chapter, the Cauchy-Schwarz inequality says

$$|\vec{v} \cdot \vec{w}| \leq \|\vec{v}\| \|\vec{w}\|.$$

You can prove this using the geometric characterization of the dot product as $\vec{v} \cdot \vec{w} = \|\vec{v}\| \|\vec{w}\| \cos \theta$, where θ is the angle between \vec{v} and \vec{w} .

Now I suggest you try using the Cauchy-Schwarz inequality to prove the triangle inequality. Use the fact that you can rewrite $\|\vec{v} + \vec{w}\|^2$ as $(\vec{v} + \vec{w}) \cdot (\vec{v} + \vec{w})$. Can you finish the proof?

8.5 Some vectors in finance

Linear algebra is used all over finance, and here I'll introduce four vectors that are useful in our further applications of linear algebra. First, we can represent a portfolio of stocks (or other assets) with the vector $\vec{x} = [x_1 \ \dots \ x_m]$. Interpreted, this means we have x_i shares of stock i , for m stocks $i = 1, \dots, m$. These numbers can be real numbers: a negative entry x_i would indicate holding a short position on the stock, and we can have non-integer entries via buying fractional shares or through investing in a mutual fund or exchange-traded fund.

Second, we could construct vectors that represent what happens to the price of these assets under a particular economic scenario. Say we're looking at a possible change in regulations, or in energy prices, or just have a forecast of what all the asset prices will be in one week. Then we could represent what happens under this hypothetical future scenario using a vector

$$\vec{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix}$$

where s_i represents the *change in price* of a share of asset i under the scenario. (We could equally well write a vector \vec{s}' in which s'_i represents the price of a share of asset i under the scenario, rather than the change in price.)

Notice that both our portfolio vector \vec{x} and our single-scenario change-in-price vector \vec{s} (or single-scenario price vector \vec{s}') both have m entries. This is because m assets are under consideration. Also notice that $\vec{x}\vec{s} = \vec{x} \cdot \vec{s}$ gives the expected change in the value in the portfolio \vec{x} given the occurrence of the scenario under consideration, while $\vec{x}\vec{s}' = \vec{x} \cdot \vec{s}'$ gives the overall value of the portfolio under the given scenario.

A third type of vector we could invent would look at the potential prices of a stock or asset A under various scenarios. For instance, what would happen to a pharmaceutical company's stock price if (I'll date myself) the Affordable Care Act is repealed? if President Trump decreases the Food and Drug Administration's regulatory responsibilities and allows "fast-tracking" of new drugs"? if the rules for H1B visa applicants are changed substantially? if changing

trade relations with India and China change the profile of the export market? Any of these changes could affect the stock price of a pharmaceutical company, and you might want to look at these scenarios using linear algebra as a first pass analysis. Say for stock A we consider n scenarios. Then a vector $\vec{a} = [a_1 \ a_2 \ \dots \ a_n]$ could represent the change in price of the asset A under each of the n scenarios, or a vector $\vec{a}' = [a'_1 \ a'_2 \ \dots \ a'_n]$ could represent the net price of the asset under each of the scenarios.

Fourth, an essential part of risk management and forecasting is working with the probabilities of these future scenarios. You may use “expert judgement” to come up with the probabilities of these future scenarios, or you may use the no-arbitrage principle to come up with probabilities based on the price changes forecast by your “expert judgement.” Either way, with n scenarios under consideration you’d want a vector

$$\vec{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}$$

with each element p_i giving the probability of the i th scenario. Note that if you’re considering this situation where these n scenarios cover all future events, you must have the probabilities adding up to one by the axioms of probability:

$$\sum_{i=1}^n p_i = 1.$$

This is the first time in our consideration of vectors in finance that we’ve had a condition like this, and it’s a little special – finally probability is starting to intersect with our multivariate work!

Again let’s look at a few matrix or dot products that are relevant. The product

$$\vec{a}\vec{p} = \vec{a} \cdot \vec{p} = \nu_A$$

would give the expected value of the change in price of the asset A under the n scenarios under consideration. If you’re confident in your scenario analysis

and the given probabilities and ν_A was positive, you'd have an expected profit on your hands! Buy that asset! It's not necessarily a guaranteed profit – under some scenarios you might still lose money – but it may be a “good bet” in investment terms. If you are looking at no-arbitrage pricing and probabilities, you'd expect to have

$$\vec{a}\vec{p} = \vec{a} \cdot \vec{p} = \nu_A = 0,$$

with no expected profit. (The product $\vec{a}'\vec{p}$ would just give the expected value, rather than change in value, of asset A over all the scenarios under consideration.)

Questions:

How would you represent a portfolio with 6 shares of stock A , 5 of stock B , and a short on 7 shares of stock C ?

Can you find a probability vector \vec{p} that will give you an expected profit of zero given a scenario vector for stock A of $\vec{a} = [1 \ 3 \ 0 \ 0.5]$? What about for $\vec{a} = [1 \ 3 \ 1 \ 0.5]$? What problems do you encounter for the second one, and why?⁴

8.6 Linear transformations

Consider a general $m \times n$ matrix A , with m rows and n columns. Figure out some way to remember that rows come first, columns second – easy for some to remember, but I'm a person who used to have trouble keeping left and right straight. My solution:

⁴Remember that probabilities must all be positive and less than or equal to 1, and for a probability vector \vec{p} elements must all add to one.



Royal Crown Cola reminds me that rows come first, columns second!
Quiz yourself:

- If we multiply the matrix A on the left by a row vector with ____⁵ elements, then we get a row vector with n elements.
- If we multiply A on the right by a column vector with n elements, then we get a column vector with ____⁶ elements.

In this way, we can think of multiplication by A as a function that transforms row vectors \vec{x} in \mathbb{R}^m to row vectors $\vec{x}A$ in \mathbb{R}^n , or as a function that transforms column vectors \vec{y} in \mathbb{R}^n to column vectors $A\vec{y}$ in \mathbb{R}^m .

Example 8.6.1. Is multiplication by a matrix A , on either the left or the right, a linear operation? (If necessary, remind yourself what *linear* means!)

We call these functions *linear transformations*, because they are nice general ways of transforming yourself from one linear space to another linear space.

Go back to our example of a row vector $(x \ y \ z)$ multiplied by a 3×2 matrix

$$A = \begin{pmatrix} 3 & 2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix}.$$

⁵ m

⁶ m

We say that A represents a linear transformation

$$R : \mathbb{R}^3 \rightarrow \mathbb{R}^2,$$

and we write the action as

$$R(\vec{x}) = \vec{x}A.$$

With this notation it's easy to see that multiplying by a matrix A is a linear operation: $(b\vec{x} + c\vec{y})A = b\vec{x}A + c\vec{y}A$ for any $b, c \in \mathbb{R}$ by properties of matrix multiplication, so $R(b\vec{x} + c\vec{y}) = bR(\vec{x}) + cR(\vec{y})$.

Example 8.6.2. What is the domain of the function $\vec{x}A$ with

$$A = \begin{pmatrix} 3 & 2 \\ -1 & 3 \\ 2 & 4 \end{pmatrix}?$$

What is the range, or *image*, of this function? The domain is \mathbb{R}^3 and range is... Well, in this case it's all of \mathbb{R}^2 but that takes some work to prove. We need to develop some more machinery.

The concept of range is a bit sophisticated for linear transformations. We call the possible outputs of R the image of R , and we notice it consists of all linear combinations of the rows of A . The notation for this is $\text{im}(R)$ or $\text{im}(A^T)$. Here, A^T is the transpose and we write this because historically mathematicians are prejudiced in favor of multiplying with the matrix on the left and the vector on the right, and converting $\vec{x}A$ to this format means taking $A^T\vec{x}^T$. Reconcile this with the notation for column space below. The term for “all possible linear combinations” is *span*. In sentences, we can say, *The image of R is the linear span of the rows of A* , or alternatively, *The image of R is the row span of the matrix A* . If the rows of A are written as $\vec{a}_1, \dots, \vec{a}_m$, then we can write $\text{span}(\vec{a}_1, \dots, \vec{a}_m)$ for this row span.

Example 8.6.3. Repeat this analysis instead multiplying by a column vector on the right: the linear transformation

$$L : \mathbb{R}^2 \rightarrow \mathbb{R}^3,$$

written as

$$L(\vec{x}) = A\vec{x}.$$

Here you get the *column span* of the matrix A for the image of L . If the columns of A are $\vec{c}_1, \dots, \vec{c}_n$, we can write $\text{span}(\vec{c}_1, \dots, \vec{c}_n)$ for the span of the columns, or we can write $\text{im}(A)$ for the image of the linear transformation $A\vec{y}$.

How do we know the dimension of the row span or the column span? We need to look at *linear independence* of rows and columns. This will give us the idea of the *rank* of a matrix. We'll define rank in [section 8.7](#), but first, we'll set up some financial concepts.

When we talked about the span of a set of vectors above, we were making a vector space by simply defining, for $\vec{v}_i \in \mathbb{R}^m$,

$$V = \text{span}(\vec{v}_1, \dots, \vec{v}_n) = \{\vec{x} \in \mathbb{R}^m \mid x = c_1\vec{v}_1 + \dots + c_n\vec{v}_n \quad \forall c_i \in \mathbb{R}\}.$$

For instance, the linear span of the $\vec{v}_i \in \mathbb{R}^m$ is $V \subset \mathbb{R}^m$.

Example 8.6.4. Is the row span of a matrix a vector space? Is the column span of a matrix a vector space?

Let V be a vector subspace of \mathbb{R}^n . The set of vectors orthogonal to every vector in V is also a vector subspace. We denote this space by V^\perp and call it the *orthogonal complement* of V :

$$V^\perp = \{\vec{x} \in \mathbb{R}^n \mid \vec{x} \cdot \vec{v} = 0 \forall \vec{v} \in V\}.$$

You should prove to yourself that V^\perp is closed under scalar multiplication and vector addition.

Example 8.6.5. How does this show that the set of all solutions to the equation $A\vec{y} = \vec{0}$ is a vector subspace of \mathbb{R}^n ? This vector subspace is called the *right null space* of the matrix A .

Example 8.6.6. How does this show that the set of all solutions to the equation $\vec{x}A = \vec{0}$ is a vector subspace of \mathbb{R}^n ? This vector subspace is called the *left null space* of the matrix A .

We define the following vocabulary:

- A function $f : A \rightarrow B$ is **one-to-one** if whenever $f(a_1) = f(a_2)$, we have $a_1 = a_2$. Example: $f(x) = x^3$, where $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$, is one-to-one. If $f(x) = 1$, you know $x = 1$ and there are no other choices. Non-example: $f(x) = x^2$ is not one-to-one, because if I tell you $f(x) = x^2 = 1$, you can reply, “Clearly x must equal either 1 or -1 ! Which would you like?”
- A function $f : A \rightarrow B$ is **onto** if for all b in B there’s an a in A so that $f(a) = b$. The function f maps on to its image, covering all the points in B . Example: $f(x) = x^3$ again (why?). Non-example: $f(x) = x^2$ again (why?). More subtle non-example: $f(x) = (x, x^3)$. Here $f : \mathbb{R}^1 \rightarrow \mathbb{R}^2$. Points like $(1, 1)$ are in the image of this function, but is $(1, 2)$? No! So not all points of \mathbb{R}^2 are covered by this function; f is not onto. Be careful of domain and range here.
- For a function $f : A \rightarrow B$, take a point b in B that is in the image of f . The **preimage** of this point b in B is the set of points in A that map to b under f . For instance, for $f(x) = x^2$, the preimage of 1 is 1, -1 , the set of all points whose square is 1. For a multivariable example, consider $f(x, y) = x^2 + y^2$. Here $f : \mathbb{R}^2 \rightarrow \mathbb{R}^1$. Take any point in \mathbb{R}^1 and try to “go backward”: the preimage of -1 is the empty set, the preimage of 0 is $(0, 0)$, and the preimage of 1 is the circle $x^2 + y^2 = 1$ in \mathbb{R}^2 .

8.7 Bases

If V is the linear span of a set of vectors $\vec{v}_1, \dots, \vec{v}_n$, we call these \vec{v}_i a *spanning set* for V . A *minimal spanning set* for V is such a set with as few elements as possible: if you remove any vector from a minimal spanning set, the remaining vectors will no longer span V . We call a minimal spanning set a *basis* for V .

Theorem 8.7.1. Let $\vec{v}_1, \dots, \vec{v}_n$ be a basis of V . Then the vectors \vec{v}_i are *linearly independent*: that is, no \vec{v}_i can be written as a linear combination of the remaining $n - 1$ vectors. Equivalently, the only way to write the zero vector $\vec{0}$ as a

linear combination of the \vec{v}_i ,

$$c_1\vec{v}_1 + \cdots + c_n\vec{v}_n = \vec{0}$$

is to take all coefficients $c_1 = c_2 = \cdots = c_n = 0$. In addition, every basis for V has exactly n vectors.

Example 8.7.1. Use contradiction to prove the first part: if we could write v_n (for instance) as a linear combination of the other vectors, then (what?)

Example 8.7.2. Challenge: prove that every basis for V must have the same number of vectors.

The *dimension* of a vector space is the number of vectors in a basis for the vector space.

We can easily work with the linear span of a set of vectors $\vec{v}_1, \dots, \vec{v}_m$ by writing the vectors as the rows of an $m \times n$ matrix. Go through the following questions and give your best answers:

- Example 8.7.3.**
- Does swapping two rows of a matrix change the row space of the resulting matrix?
 - Does replacing row i of a matrix with row i minus row j change the row space of the matrix?
 - Does multiplying a row in a matrix by a real number change the row space of the resulting matrix?

The answer to all of the above is no, because each of these is a linear combination of row vectors, and the row space (the span of the row vectors) is closed under linear combinations (scalar multiplication and vector addition). This means that we can use *row reduction* techniques to solve matrix equations of the form $A\vec{y} = \vec{b}$ on paper. The goal is to streamline old techniques for solving systems of equations by row reducing the augmented matrix $[A|\vec{b}]$ to have A in row echelon form.

Row echelon form is a special form of a matrix: at the bottom of the matrix, rows with only zeroes; all other rows have first nonzero entry 1, which we call

a “leading 1”; all entries above and below leading 1s are zero. Not only does this allow us to solve equations of the form $A\vec{y} = \vec{b}$, but it allows us to easily see the rank of a matrix A , and this form gives an easy way of determining the dimension of the right null space (just the number of rows of zeroes).

Example 8.7.4. Find the space of solutions to

$$(x_1 \ x_2 \ x_3) \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 4 \end{pmatrix} = (2 \ 1).$$

Hint: Using the symbol T for transpose (swapping rows and columns), we can change $\vec{x}A = \vec{b}$ to $A^T\vec{x}^T = \vec{b}^T$. This makes the equation easier to deal with using methods discussed in class.

(The solution space is one-dimensional – we’ve got three variables and two conditions on them, so one degree of freedom in solutions. Check that your work gives you something like $x_1 = 5x_3$, $x_2 = 1 - 4x_3$, x_3 free. We can write this answer parametrically as

$$(x_1 \ x_2 \ x_3) = (0 \ 1 \ 0) + x_3 (5 \ -4 \ 1).$$

Example 8.7.5. Show that

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}$$

has no solution.

(Find a contradiction.)

Example 8.7.6. Consider the following three investment opportunities, showing the net profit under four possible outcomes:

$$(-20, 20, 20, 20), (0, 20, -10, -10), (30, -40, -20, -30).$$

Show that these do not offer the possibility of arbitrage, and determine the price of the following investment opportunity: $(20, 0, 30, 0)$. Hint: Use the

theorem of no arbitrage! You can show there's no possibility of arbitrage ("free money") directly, by reducing the matrix and showing there's no solution, or you can find a probability vector \vec{p} that satisfies case (2) of the No Arbitrage theorem. The *existence* of this probability vector shows that there can be no arbitrage!

The fundamental theorem of linear algebra relates the four "fundamental linear subspaces" of an $m \times n$ matrix A . These four subspaces are the right and left null spaces, the column space, and the row space. I'll remind you that

$$L : \mathbb{R}^n \rightarrow \mathbb{R}^m \quad (8.1)$$

$$\vec{x} \mapsto A\vec{x} \quad (8.2)$$

and

$$R : \mathbb{R}^m \rightarrow \mathbb{R}^n \quad (8.3)$$

$$\vec{y} \mapsto \vec{y}A = A^T\vec{y}^T. \quad (8.4)$$

Theorem 8.7.2. The dimension of the row space of A plus the dimension of right null space of A equals n .

$$\dim(\text{im}(A^T)) + \dim(\text{null}(A)) = n.$$

The dimension of the column space of A plus the dimension of left null space of A equals m .

$$\dim(\text{im}(A)) + \dim(\text{null}(A^T)) = m.$$

Theorem 8.7.3. The dimension of row space of A is the same as the dimension of column space of A , and these are both equal to the *rank* of matrix A .

Example 8.7.7. Challenge: think about how you could get a set of orthogonal basis vectors for a vector space from any basis provided. Pick a vector to start with, then use the projection formula discussed a few classes ago to find the next one. How do you get the third basis vector? Remember it must be orthogonal to both of the previous vectors! Explain how to proceed in the case of n vectors.

Example 8.7.8. Challenge two: can you modify this process to produce an *orthonormal* basis? That means all basis vectors are orthogonal to each other, and moreover each is a unit vector.

This will be discussed more in the next chapter.

8.8 Applications to financial math

The principle of no arbitrage in finance has many different phrasings. One that I like is that if two assets have the same risk and the same cash flow, they've got to sell at the same price. Otherwise you could make money from exploiting the difference in price between the two. Another way people express the no arbitrage principle is to say that you can't make more money than the market without taking on more risk, and yet another is that there is "no free lunch."

Now, is the no arbitrage principle true? Well, not exactly. The "efficient market hypothesis" says that if there's an arbitrage opportunity (the chance to make money without risk) then the market will notice, people will take advantage of it, and prices will then adjust to eliminate that opportunity. This means that in an efficient market, prices reflect information accurately. (Look up strong, semi-strong, and weak efficiency elsewhere!) The capital asset pricing model (CAPM) and Black-Scholes options pricing model are both built on the principle of no arbitrage, and that's why we need to understand it. Moreover, CAPM and Black-Scholes are really useful out in the real world. But like all models, they're wrong, as is a really strict no-arbitrage statement. Robert Shiller, for instance, looked at changes in dividend prices and their effect on share prices of assets. He found that share prices "overreact." Look up the work of Fama, Schiller, and Hansen, who jointly won the Nobel Prize in Economics in 2013 for their (separate) work on asset pricing. You'll find that the truth about asset prices is a lot more complicated than no-arbitrage – but you can't understand what is really going on without knowing this basic principle.

We will formulate a "No Arbitrage" theorem via linear algebra. First, we assume that we can form an $m \times n$ matrix S of net profits, describing what will happen to m stocks under n outcomes or scenarios. (Entry S_{ij} is the net profit for stock i under scenario j .)

Example 8.8.1. What are the possible net profit vectors for all the portfolios I could create? Test yourself by translating this into the language of linear algebra. What vector space am I asking for?

Consider the situation given by

$$(x_1 \ x_2 \ x_3) \begin{pmatrix} -2 & -1 & 0 & 1 \\ 3 & 2 & -1 & -1 \\ 1 & 0 & 6 & -5 \end{pmatrix}$$

We have three stocks and four scenarios under examination.

Example 8.8.2. Can we invest in a way that produces the vector $(0 \ 0 \ 0 \ 5)$? How do you figure out the answer to this? How do you interpret the answer once you have it?

Example 8.8.3. Can we invest in a way that produces the vector $(4, 2, 12, -9)$?

Example 8.8.4. Let's add another investment opportunity, the "savings account." This opportunity is an idealized situation in which you put in a dollar and get a dollar. This of course ignores interest rates for the moment. How do you add this information to the matrix?

Example 8.8.5. Given the new matrix, can you invest in a way that produces the vector $(4, 2, 12, -9)$?

Example 8.8.6. What is the cost of such a portfolio? Interpret this carefully, looking at net profit vectors and the savings vector and considering the *cost* of each.

We can find the cost of such a portfolio by solving for a probability vector that satisfies

$$S\vec{p} = (0 \ 0 \ 0 \ 1)^T$$

and then taking the dot product of our portfolio vector \vec{x} with this probability vector \vec{p} .

By the axioms of probability, a probability vector has only non-negative entries and its entries sum to one. In particular, if the vector \vec{p} must have *non-positive* elements to satisfy the matrix equation given above, we can create a portfolio with no negative elements and at least one positive element which costs no money – and that’s arbitrage.

Let’s go back to vector spaces so that we can gather some techniques for more efficiently solving the matrix equations above. With new language, we can also restate the “no arbitrage” theorem in terms of linear algebra and provide some strategies for proof.

Using the new language of linear algebra, we can restate the no-arbitrage theorem as the following: EITHER

- the row space of S contains a non-negative vector with at least one positive element, OR
- the orthogonal complement of the row space of S (the right null space) contains a vector whose elements are all strictly positive.

I use incorrect capitalization here to emphasize that these outcomes are mutually exclusive. To test your understanding, ask yourself: which of these cases is the case with arbitrage? How do you know? In the no-arbitrage scenario, what’s true about the probabilities of the scenarios?

Initially we considered the situation with n scenarios and m stocks, so

- each portfolio \vec{x} has m entries,
- the matrix of prices S is an $m \times n$ matrix,
- and the probability vector \vec{p} has n entries.

Figuring out which case of the no-arbitrage theorem holds involves either finding $\vec{x}S$ with some strictly positive entry (making some money!) and the corresponding portfolio \vec{x} that guarantees us this risk-free money, or finding the probability vector \vec{p} that gives $S\vec{p} = \vec{0}$.

To consider cost or the “savings account” approach, add a row of ones to the bottom of S . Call the new matrix S' . Add an entry x_{m+1} to the end of

the portfolio vector \vec{x} , representing how much money you're putting in the savings account, and call the resulting vector \vec{x}' . If the "savings account" interpretation is distasteful to you (and it does have some drawbacks), consider this a mathematical way of requiring that \vec{p} be a probability vector: that is, $p_1 + \cdots + p_n = 1$. The reason you might not like the "savings account" interpretation is that it does not quite line up with the idea that the matrix entries of S are net change in stock price, or profits. The row of ones does not represent absolute profit (making \$3 per share of stock i) but instead leaves the amount x_{m+1} unchanged in each scenario.

Example 8.8.7. Prove to yourself that if we have a probability vector \vec{p} so that $S\vec{p} = \vec{0}$, then $\vec{x}'S'\vec{p} = x_{m+1}$.

Example 8.8.8. Show that if we have \vec{p} a probability vector with only positive entries, then we must have negative entries in the portfolio vector \vec{x} . (This is one part of the proof of the no arbitrage theorem.)

Example 8.8.9. Challenge: prove that if we have a vector \vec{x} whose entries are all non-negative, and at least one of whose entries is positive, so that $\vec{x}S$ is non-negative and has at least one strictly positive entry, then there can be no strictly positive probability vector \vec{p} so that $S\vec{p} = \vec{0}$.

8.9 Invertible transformations

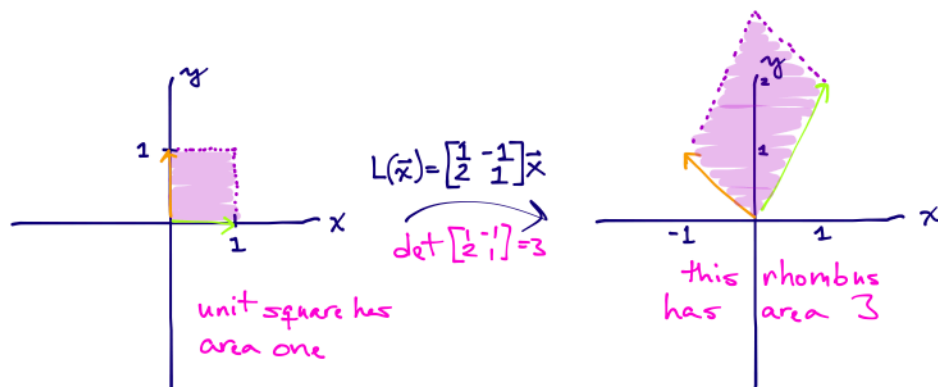
This section mainly emphasizes ideas of rank and determinant, and reinforces what we've learned about transformations. All matrices in this section are square matrices.

Theorem 8.9.1. If the rank of an $n \times n$ matrix A is n , then the linear transformation $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined by $L(\vec{x}) = A\vec{x}$ is both one-to-one and onto. If the rank of A is less than n , then the linear transformation L is neither one-to-one or onto. Thus the linear transformation L is invertible if and only if the matrix A has rank n .

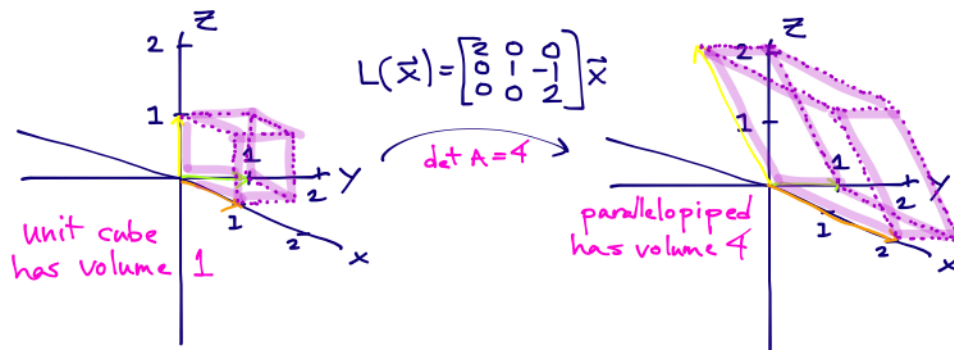
A square matrix is of full rank if and only if its determinant nonzero; a determinant of zero means the matrix is degenerate (the matrix gives a transformation that is not onto – the image is contained in a smaller linear subspace of the target space).

In much of linear algebra before this class, you've probably concentrated on full rank matrices, which give invertible transformations. That is because they give systems of equations easy to solve with matrix methods: if A is full rank and $A\vec{x} = \vec{y}$, then $\vec{x} = A^{-1}\vec{y}$. This is extraordinarily important, and also not that interesting!

Remember from earlier that the determinant of a matrix A gives the change in volume that the transformation induced by the transformation $L(\vec{x}) = A\vec{x}$. Maybe some illustrations will help: in two dimensions,



and in three dimensions



This holds for transformations $\mathbb{R}^n \rightarrow \mathbb{R}^n$, in fact. Think of it this way: take the unit hypercube $[0, 1]^n$ and look at its image under the transformation L . The image of the unit hypercube under L will have n -dimensional volume $\det A$. This also says something about transformations that are not invertible: if a square matrix is not invertible, then the determinant is zero – the transformation is not onto and one-to-one. That means that the image of the unit hypercube has zero n -volume, which means the unit hypercube in \mathbb{R}^n was *squashed* into a lower-dimensional subspace by the transformation L . Remember this when we start talking about principal component analysis, dimension reduction techniques, and singular value decomposition!!

Chapter 9

Spectral theorem and portfolio management

9.1 Orthogonal matrices and orthonormal bases

An $n \times n$ matrix A is called *orthogonal* if its rows are *orthonormal*: that is, all rows are perpendicular to each other and all have length one as vectors in \mathbb{R}^n . For instance,

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is an orthonormal matrix (check!). More generally, if A is an $n \times n$ matrix with rows $\vec{r}_1, \dots, \vec{r}_n$, then A is orthonormal if $\vec{r}_i \cdot \vec{r}_j = 0$ for all $i \neq j$ between 1 and n and $\vec{r}_i \cdot \vec{r}_i = 1$ for all $i = 1, 2, \dots, n$.

Geometrically, what does an orthogonal matrix do?

Think about the consequences of the comments above: if A is orthonormal, then

$$AA^T = I_n,$$

where I_n is the identity matrix.

Example 9.1.1. Prove this!

This means that $A^T = A^{-1}$: A^T is the inverse of A .

Example 9.1.2. What does this mean for $A^T A$?

Example 9.1.3. Prove the columns of A are orthonormal.

Example 9.1.4. Compare the length of a vector \vec{x} with the length of the vector $A\vec{x}$: carry out the calculations necessary to do so.

Example 9.1.5. Compare the angle between \vec{x} and \vec{y} to the angle between $A\vec{x}$ and $A\vec{y}$ by using the formula that relates dot products and angles between vectors.

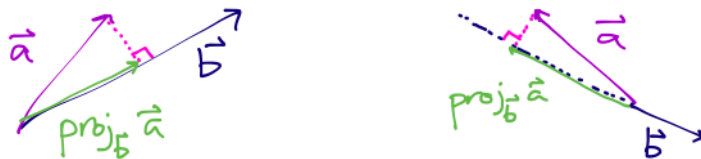
If you did all the questions above, you proved the following theorem:

Theorem 9.1.1. Let L be a linear transformation from \mathbb{R}^n to \mathbb{R}^n represented by an orthogonal $n \times n$ matrix A . Then L preserves the lengths of vectors and the angles between vectors. Equivalently, for all column vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$, $\vec{x} \cdot \vec{y} = (A\vec{x}) \cdot (A\vec{y})$.

9.2 Gram-Schmidt orthogonalization

To find an orthonormal basis from any collection of basis vectors, use Gram-Schmidt orthogonalization. The idea is elegant and simple: pick a first vector and make it length one. Pick the next vector and take the component of it orthogonal to the first, then normalize. Pick a third vector and take the component of it orthogonal to the first two, and so on.

First, of course, we need to define what the projection of a vector \vec{a} onto another vector \vec{b} is. Imagine that a projection is the shadow of \vec{a} on the line in direction \vec{b} if the sun is “directly overhead.”



The algebraic formulation is

$$\text{proj}_{\vec{b}}\vec{a} = \frac{\vec{a} \cdot \vec{b}}{|\vec{b}|^2} \vec{b}.$$

Notice that this vector has a direction (it goes in the direction of \vec{b}) and a magnitude (the magnitude of the projection is $|\vec{a}| \cos(\theta)$, where θ is the angle between the vectors \vec{a} and \vec{b}). You could figure out this definition of projection yourself by looking at the natural geometric expression $|\vec{a}| \cos(\theta) \frac{\vec{b}}{|\vec{b}|}$ and using $\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos(\theta)$ to prove the formula in terms of the dot product.

Why do we need projection for Gram-Schmidt orthogonalization? It turns out that $\vec{a} - \text{proj}_{\vec{b}}\vec{a}$ will give the component of \vec{a} that is perpendicular, or orthogonal, to \vec{b} . By subtracting off the part of \vec{a} that is in the direction of \vec{b} , we're left only with the part that is perpendicular to \vec{b} . This is what we want for “orthogonalization.”

Formalize our descriptions with mathematical language: To find a Gram-Schmidt orthogonal basis $\{\vec{u}_1, \dots, \vec{u}_n\}$ or orthonormal basis $\{\vec{b}_1, \dots, \vec{b}_n\}$ for the space V spanned by a set of vectors $\{\vec{v}_1, \dots, \vec{v}_n\}$,

- Pick one vector to start with: let's choose \vec{v}_1 . Normalize it (make it length one) via

$$\frac{\vec{v}_1}{|\vec{v}_1|}$$

if you want an orthonormal basis. Then the first vector in your orthogonal basis is $\vec{u}_1 = \vec{v}_1$, and the first vector in the orthonormal basis is $\vec{b}_1 = \frac{\vec{v}_1}{|\vec{v}_1|}$.

- Pick the next one to deal with: I choose \vec{v}_2 . Subtract off the component in the direction of \vec{u}_1 :

$$\vec{u}_2 = \vec{v}_2 - \text{proj}_{\vec{u}_1}\vec{v}_2.$$

Normalize the result if you want an orthonormal basis (that makes $\vec{b}_2 = \frac{\vec{u}_2}{|\vec{u}_2|}$).

- Take \vec{v}_3 and subtract off the components in the directions of \vec{u}_1 and \vec{u}_2 :

$$\vec{u}_3 = \vec{v}_3 - \text{proj}_{\vec{u}_1} \vec{v}_3 - \text{proj}_{\vec{u}_2} \vec{v}_3.$$

Normalize this if you want an orthonormal basis ($\vec{b}_3 = \frac{\vec{u}_3}{|\vec{u}_3|}$)

- Repeat until done!

Order sort of matters: no matter what order you go in, you'll get an orthogonal or orthonormal basis, but if you use a different order than a friend, you'll very often get a different basis. Check out the picture below:

add picture**

Example 9.2.1. Why does this work? Why doesn't the process lead to vectors outside the vector space spanned by the original set of vectors? Use the words "linear combination" in your solution.

9.3 Rotation and scaling

A particular class of invertible linear transformations is given by rotation. Yes, rotation: just pick up everything in your vector space and rotate it by θ radians in some particular plane. The rotation matrix that gives this transformation is

$$A(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

Check for yourself that this is an orthonormal matrix. What is its determinant?

More things to check for yourself:

- What is $A(-\theta)$? What is $A(\theta)^{-1}$? Yes, they happen to be the same! Why?
- What is $A(\theta_1 + \theta_2)$? What is $A(\theta_1)A(\theta_2)$? What is $A(\theta_2)A(\theta_1)$? Yes, they happen to be the same! Why? You'll need your angle-sum formulas from trigonometry to see this.

- What do you think $\sqrt{A(\theta)}$ should be? Can you come up with a definition?
- What is $A(\theta)^k$ for k a positive integer? Note that this denotes carrying out matrix multiplication $k - 1$ times ($A(\theta)$ times itself k times), not raising the elements in the matrix to the k th power. With this nice rotation matrix $A(\theta)$, what's a nice way to streamline this calculation? If you deduced $A(\theta)^k = A(k\theta)$, you'd be right.

Rotation matrices certainly extend to rotations in \mathbb{R}^n , but are more complicated to write out. No matter what, if we're rotating only by an angle θ , a single parameter, there will be some axis that is invariant under the rotation in \mathbb{R}^n . In \mathbb{R}^2 , you rotate by θ around a point (the origin). In \mathbb{R}^3 , you rotate by θ around some line. For instance,

$$\begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix}$$

rotates by θ around the y -axis – the $x - z$ plane is rotated but points $(0, y, 0)$ stay untouched.

You may have noticed that whether in \mathbb{R}^2 or \mathbb{R}^n it is dramatically easier to calculate $(A(\theta))^k$ for any rational number k than it would be for a general $n \times n$ matrix. This is not a coincidence. The beautiful geometry of a rotation matrix allows a very simple interpretation of these operations. We'll be able to put this to use in Taylor series of matrices, among other applications. But you know that not every matrix is a rotation – we need some wider techniques.

The next baby step is to consider *scaling* as well, and look at the matrix

$$A(r, \theta) = r \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} r \cos(\theta) & -r \sin(\theta) \\ r \sin(\theta) & r \cos(\theta) \end{bmatrix}.$$

Notice we've given this matrix a name, $A(r, \theta)$. Notice, too, that now we've introduced another bit of ambiguity: if you look carefully, $A(1, \pi) = A(-1, 0)$. This is the same equivalence that occurs in polar coordinates.

Consider what happens when two such matrices $A(r_1, \theta_1)$ and $A(r_2, \theta_2)$ are multiplied: geometric reasoning tells you that the angles add and the scalars r_1 and r_2 multiply. When you multiply out the matrices, though, what happens?

Example 9.3.1. Multiply out $A(r_1, \theta_1)$ and $A(r_2, \theta_2)$. How does this agree with your geometric reasoning?

Raising matrices to a power is also easy using our new notation:

$$A(r, \theta)^p = A(r^p, p\theta).$$

This formula essentially uses polar coordinates, if you remember those!

How does this work for fractional powers? Almost the same way! Notice that when you take a square root, you still have the choice of positive or negative that you would in the real numbers.

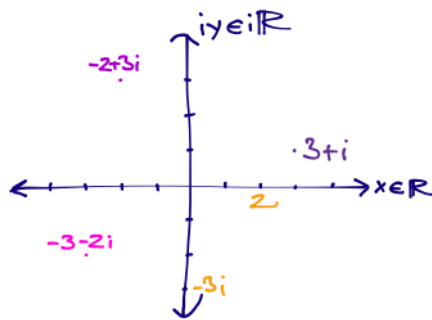
Example 9.3.2. Take the square root of $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. Make sure you get two different answers. Express those answers in $A(r, \theta)$ form.

This may be reminding you of complex numbers.....

9.4 Complex Numbers

At some point in your math education you encountered the frustration of wanting to understand the solution to $x^2 + 1 = 0$. Hopefully at about the same time you encountered the idea of complex and imaginary numbers. We define the imaginary number i to be $\sqrt{-1}$; notice that $-i$ is a fine imaginary number as well and $(-i)^2 = -1$ as well. (Remember that $f(x) = \sqrt{x}$ is only a *function* if it has a unique output y for each input x , and we define the unique output to be the “positive” number y that satisfies the equation $y^2 = x$.)

You may have also encountered the idea of the *complex plane* at about this time. The complex plane is a representation of \mathbb{C} that is topologically equivalent to \mathbb{R}^2 . The way to see this is that every number $x + iy \in \mathbb{C}$ can be represented as a pair $(x, y) \in \mathbb{R}^2$ by viewing the x -axis as the “real” axis and the y -axis as the “imaginary” axis.



Complex numbers add together just like vectors in \mathbb{R}^2 , as $x + iy + z + iw = (x + z) + i(y + w)$, and multiplication by real scalars also behaves the same. The difference between \mathbb{C} and \mathbb{R}^2 lies in their multiplicative structures. In \mathbb{R}^2 , we might pick out dot product as the way to multiply (x, y) and (z, w) , but in \mathbb{C} we've got a rule we want satisfied – that $i^2 = -1$ – and dot product clearly doesn't do the right thing!!

To illuminate the multiplicative structure of the complex numbers, we'll pick out two equivalent ways of thinking about complex numbers. First, the complex number $x + iy$ behaves like the matrix

$$\begin{bmatrix} x & -y \\ y & x \end{bmatrix}.$$

What I mean by “behaves like” is that if you carry out matrix multiplication here, you'll get a structure that is equivalent to complex multiplication:

$$(x + iy)(z + iw) = (xz - yw) + i(xw + yz)$$

corresponds with

$$\begin{bmatrix} x & -y \\ y & x \end{bmatrix} \begin{bmatrix} z & -w \\ w & z \end{bmatrix} = \begin{bmatrix} xz - yw & -(xw + yz) \\ xw + yw & xz - yw \end{bmatrix}.$$

Example 9.4.1. How can you write this matrix as $A(r, \theta)$? Challenge: think about the ways in which you can extract θ from the information you have, and how it may be constrained by the way in which you write it.

Example 9.4.2. Think of three ways in which you can find the *reciprocal* of the complex number represented by a matrix $A(r, \theta)$. Remember, the reciprocal of the number $z \in \mathbb{C}$ is $\frac{1}{z}$. It is the multiplicative inverse.

The questions above ask you to think about the second way to represent a complex number: instead of using Cartesian coordinates via $x + iy$, you can use polar coordinates via $A(r, \theta)$, or with Euler's formula,

$$re^{i\theta} = r \cos \theta + ri \sin \theta.$$

Euler's formula gives another way to link the geometry of \mathbb{C} with the matrix representation and the Cartesian representation of the same complex number.

9.4.1 Taylor series

Why are we bothering with series, especially right now, right after complex numbers? The use of power series representations for functions makes it very easy to extend these functions to the complex numbers. If I ask you what $\cos(3 + 2i)$ is, you will probably be unable to answer the question using your unit-circle understanding of trigonometry. You can just plug $z = 3 + 2i$ into the power series definition, though! Convergence is still a big deal, but we can prove that convergence in real numbers is the "same" as convergence in the complex numbers. That is, if $P(x)$ converges for $|x| < a$, then $P(z)$ converges for $|z| < a$ and $z \in \mathbb{C}$.

Example 9.4.3. Use your power series representation for e^x to write the series for $e^{i\theta}$. Separate the terms with odd degree from the terms with even degree to prove that

$$e^{i\theta} = \cos \theta + i \sin \theta.$$

9.5 Changes of basis and coordinates

9.5.1 Changing basis alone

We are used to working in the standard basis for \mathbb{R}^n , with vectors $\vec{e}_1 = (1, 0, \dots, 0)$, $\vec{e}_2 = (0, 1, 0, \dots, 0)$, through $\vec{e}_n = (0, \dots, 0, 1)$. However, there might be a sit-

uation in which we'd like to work in a different basis – maybe an eigenbasis.

9.5.2 Changing linear transformations into a new basis

Example 9.5.1. In \mathbb{R}^3 , call our standard basis \mathcal{E} and denote another basis by \mathcal{B} . Let \mathcal{B} consist of $\vec{b}_1 = (1, 2, 3)$, $\vec{b}_2 = (0, 1, 0)$, and $\vec{b}_3 = (0, 1, 1)$. Say we've got a linear transformation $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ that we've written as $T(\vec{x}) = A\vec{x}$ in terms of the standard basis \mathcal{E} . How can we write T in terms of this new basis? That is, what's the matrix for T in terms of the basis \mathcal{B} ?

Introduce a transition matrix S that relates the two bases. S has entries S_{ji} defined by

$$\vec{b}_i = \sum_{j=1}^3 S_{ji} \vec{e}_j.$$

In our example, since $\vec{b}_1 = \vec{e}_1 + 2\vec{e}_2 + 3\vec{e}_3$, $\vec{b}_2 = \vec{e}_2$, and $\vec{b}_3 = \vec{e}_2 + \vec{e}_3$, we get the matrix

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 1 \\ 3 & 0 & 1 \end{pmatrix}.$$

Notice the columns here!!

A few calculations show, then, that if our transformation in the standard basis is written $T(\vec{x}) = A\vec{x}$ and the same transformation in the new basis \mathcal{B} is written $B\vec{x}$, then

$$AS = SB,$$

or in particular

$$S^{-1}AS = B.$$

This operation is called conjugation.

Example 9.5.2. To continue our example, say that

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

and S is as above. Then the matrix B is

$$B = \begin{pmatrix} 2 & 0 & 0 \\ -5 & 1 & -2 \\ 3 & 0 & 3 \end{pmatrix}.$$

Check this!

This means that B takes the vector $\vec{b}_1 = \vec{e}_1 + 2\vec{e}_2 + 3\vec{e}_3$ to the vector $2\vec{b}_1 - 5\vec{b}_2 + 3\vec{b}_3$. Notice that this is $(2, 2, 9)$ in the standard basis!

Example 9.5.3. Say we want to write a linear transformation $L : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ in terms of the new basis \mathcal{B} given by

$$\vec{b}_1 = (\sqrt{2}/2, \sqrt{2}/2)$$

$$\vec{b}_2 = (-\sqrt{2}/2, \sqrt{2}/2),$$

an orthogonal basis for \mathbb{R}^n . Let L be given by

$$A\vec{x} = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix} \vec{x}.$$

What is S ? What is the matrix for L in the new basis?

If two matrices A and B are conjugates of each other, that is, there's an S so that $S^{-1}AS = B$ or so that $A = SBS^{-1}$, then we say that A and B are *similar matrices*.

Ideally, we can find a *diagonal* matrix D so that $D = S^{-1}AS$. This is very nice because

- Diagonal matrices are easy to multiply: they simply scale rows or columns, depending on which side of the product they're on.
- In particular, D^p is very easy to calculate – you just raise each diagonal entry to the p th power.
- Calculations like A^p are suddenly easy too. Since $A = SDS^{-1}$, $A^p = (SDS^{-1})(SDS^{-1}) \cdots (SDS^{-1})$, p times. Notice most of the S 's cancel with S^{-1} .

- The concept of diagonalizability (being able to find a diagonal similar matrix) is going to be very useful when we look at eigenvalues and eigenvectors!

Example 9.5.4. Draw a picture in \mathbb{R}^2 of the standard basis and of the basis given by $\vec{b}_1 = (1, 1)$ and $\vec{b}_2 = (-1, 1)$. Draw a picture of the vector $\vec{c} = -2\vec{e}_1 - 3\vec{e}_2$, and express it in terms of \vec{b}_1 and \vec{b}_2 .

Example 9.5.5. If \vec{v} has coordinates of 3, -2 with respect to the basis \mathcal{B} given in the previous problem, what are the coordinate of \vec{v} in the standard basis?

Example 9.5.6. Using again the basis \mathcal{B} given by $\vec{b}_1 = (1, 1)$ and $\vec{b}_2 = (-1, 1)$, what matrix will change a vector in standard coordinates to a vector in coordinates with respect to \mathcal{B} ? Use that matrix to find the \mathcal{B} coordinates of $(2, 2)$.

9.6 Eigenvalues and eigenvectors

Let L be a linear transformation from \mathbb{R}^n to \mathbb{R}^n . A nonzero vector $\vec{w} \in \mathbb{R}^n$ is an *eigenvector* of L if

$$L(\vec{w}) = \lambda\vec{w}$$

for some scalar λ . The number λ is the *eigenvalue* of L associated to the eigenvector \vec{w} . For an $n \times n$ matrix, there are up to n distinct eigenvectors and eigenvalues. As you'll see, we'll care a lot how many distinct eigenvalues a matrix has.

The geometric meaning of the eigenvectors and eigenvalues of a real matrix is that an eigenvector gives a *direction* in which a matrix transformation gives a “pure stretch,” and the associated eigenvalues gives the *magnitude* of this stretch. At least, this is true if the eigenvalue is a real number and the eigenvector is a real vector. Direction and magnitude – sounds like a vector – why do we need the pair? The eigenvector truly is just a basis vector for a particular vector subspace, while the eigenvalue can be either positive or negative, indicating pure stretch when positive and a stretch and reflection when negative.

Example 9.6.1. Find the eigenvalues and eigenvectors of the matrix

$$S = \begin{pmatrix} 2 & 0 \\ 0 & -4 \end{pmatrix}$$

solely by thinking (no calculating allowed!!). Link this understanding visually or geometrically with scaling in the x and y directions.

Example 9.6.2. Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}.$$

To do this, use the characteristic polynomial $P(\lambda) = \det(A - \lambda I_2)$ to find out for what values of λ the matrix $A - \lambda I_2$ is singular. The solutions to

$$P(\lambda) = 0$$

are the eigenvalues of A . In this situation, you'll have two or fewer eigenvalues: check that you get $\lambda = -1$ and $\lambda = -2$ as solutions. Write $\lambda_1 = -1$ and $\lambda_2 = -2$ to keep track of them, and then find the associated eigenvectors using the definition. We know that we must have $A\vec{v}_1 = \lambda_1\vec{v}_1$, so our first equation is

$$0x + 1y = -1x,$$

implying $y = -x$. A solution to this is $x = 1, y = -1$; does this also satisfy the second equation,

$$-2x - 3y = -1y?$$

Check: $-2 - 3(-1) = 1$ is true! Thus

$$\vec{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

is a fine eigenvector for $\lambda_1 = -1$. In fact, any scalar multiple of this would work; check that

$$\vec{v}_1 = c \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

satisfies the definition of eigenvector for any $c \in \mathbb{R}$. I'll let you calculate \vec{v}_2 . Check your answer in the footnote.¹

Geometric and *algebraic* multiplicity of eigenvalues are useful and subtly different ideas that come into play when you have fewer than n eigenvalues for an $n \times n$ matrix. Eigenvectors of a matrix A are a basis of the null space of $A - \lambda I_n$ – that's why finding the roots of the characteristic polynomial works to find eigenvalues. We know there is a relationship between the dimension of the (right) nullspace and the rank of the matrix; let's explore this relationship and geometric and algebraic multiplicities by example.

Example 9.6.3. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}.$$

Check to see that the two eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$, and that each of $A - (-1)I_2$ and $A - 2I_2$ have a nullspace of dimension 1.

Example 9.6.4. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}.$$

The characteristic polynomial is $(1 - \lambda)^2$. The only root, $\lambda = 1$, occurs twice – the *algebraic multiplicity* is two because the exponent of the term $(1 - \lambda)$ is two. The *geometric multiplicity* is the dimension of the nullspace of $A - (1)I_2$. Since

$$A - \lambda I_2 = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$$

is rank one, its nullspace has dimension one by the rank-nullity theorem. That means the geometric multiplicity of the eigenvalue $\lambda = 1$ is one.

These multiplicities play into whether we can find a *basis of eigenvectors*. If an $n \times n$ matrix has n eigenvectors, we have a basis of eigenvectors and we

¹The second eigenvector can be any multiple of $(1, -2)^T$.

can diagonalize the matrix. Else we can't! Let's think through this: Consider a specific transformation $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and there's a basis \mathcal{B} consisting of n distinct eigenvectors \vec{w}_i . Let's find the matrix of L with respect to basis \mathcal{B} . Thinking through this carefully, we see that the matrix of L in the matrix \mathcal{B} *must* be the matrix

$$S = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}.$$

Call this diagonal matrix $D(\lambda_1, \dots, \lambda_n)$. Notice that if we use our knowledge from earlier, this matrix is easy to find:

$$D(\lambda_1, \dots, \lambda_n) = S^{-1}AS$$

for A the matrix of L in the standard basis and S the transition matrix defined above.

Example 9.6.5. Silly example: consider the matrix

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

The only eigenvalue of this matrix is two, with both algebraic and geometric multiplicity two: that means that even though $\lambda_1 = \lambda_2 = 2$, we can pick linearly independent eigenvectors

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Can you diagonalize the matrix A ? Not a trick question!

Example 9.6.6. Use the characteristic polynomial to find the eigenvalues of

$$A = \begin{pmatrix} 3 & 0 \\ 1 & 1 \end{pmatrix}.$$

Find the corresponding eigenvectors of A . Then find the matrix S that satisfies $D = S^{-1}AS$.

Example 9.6.7. Use the characteristic polynomial to find the eigenvalues of

$$A = \begin{pmatrix} 0 & 1 \\ -2 & 2 \end{pmatrix}.$$

Find the corresponding eigenvectors of A . Then find the matrix S that satisfies $D = S^{-1}AS$.

How are these two examples different?

Example 9.6.8. Last, look at the matrix

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Check that you can find three linearly independent eigenvectors even though one of the eigenvalues has algebraic multiplicity two.

Remarkably, if the eigenvalues of a matrix are all distinct, then the corresponding eigenvectors are automatically linearly independent, thus providing a basis for

So far we've ignored one sticky case: the case of complex eigenvalues. Look at the matrix

$$A = \begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}.$$

Check that you get roots of $P(\lambda)$ that are complex: $\lambda = 1 \pm 2i$. No coincidence that they occur in a conjugate pair! The fundamental theorem of algebra guarantees that any degree n polynomial with real coefficients will have exactly n solutions *when* you count with multiplicity (explored above) and count complex solutions, and that complex solutions will always come in conjugate pairs because $(x - (a + ib))(x - (a - ib)) = x^2 - 2ax + (a^2 + b^2)$ has real coefficients, while any other product $(x - (a + ib))(x - (c + id))$ will have complex coefficients. But how can this make sense if we are considering “real-world” applications?

First, go through the work of finding the eigenvectors for the matrix A we just considered. You'll find eigenvectors

$$\vec{v}_1 = c \begin{pmatrix} 1 \\ 1 - i \end{pmatrix}$$

for some $c \in \mathbb{C}$ (so if you got a different-looking answer, just check that it's a multiple by some c) and

$$\vec{v}_2 = c \begin{pmatrix} 1 \\ 1 - i+ \end{pmatrix},$$

again for $c \in \mathbb{C}$. Notice that these eigenvectors also come in a conjugate pair. That's also not coincidence. Second, let's try to interpret this as a rotation and scaling using properties of complex numbers. *****

9.7 Quadratic forms and definiteness

Having eigenvalues and eigenvectors also allows us to discuss quadratic forms and *positive definite*, *negative definite*, and *indefinite* matrices much more easily.

A quadratic form is a homogeneous degree-two polynomial with real coefficients. Homogeneous means that every term of the polynomial has the same degree; for example, $x^2 + y^2 - 2xy$ is a homogeneous polynomial of degree two and thus a quadratic form. Using summation symbols, we could write a quadratic form in n variables x_1, \dots, x_n as

$$q(x_1, \dots, x_n) = \sum_{i=1}^n a_i x_i^2 + \sum_{1 \leq i < j \leq n} b_{ij} x_i x_j.$$

More excitingly, though, we can write any quadratic form using vectors and matrices: let $\vec{x} = (x_1, \dots, x_n)$, and convince yourself that you can write

$$q(\vec{x}) = \vec{x}^T S \vec{x}$$

for a unique real symmetric matrix S . What must the entries of S be? Compare terms between our two expressions: we must have $s_{ii} = a_i$ and we must have $s_{ij} = s_{ji} = \frac{b_{ij}}{2}$.

We call a quadratic form and its associated matrix S *positive definite* if

$$q(\vec{x}) = \vec{x}^T S \vec{x} > 0$$

for all non-zero vectors \vec{x} in \mathbb{R}^n . If instead we have the weaker condition

$$q(\vec{x}) = \vec{x}^T S \vec{x} \geq 0$$

then both the form and the matrix are *positive semidefinite*. Likewise, if

$$q(\vec{x}) = \vec{x}^T S \vec{x} < 0$$

for all non-zero $\vec{x} \in \mathbb{R}^n$, we call both q and S *negative definite*, and if

$$q(\vec{x}) = \vec{x}^T S \vec{x} \leq 0$$

they're *negative semidefinite*. If none of the above, call q and S *indefinite*.

9.8 Power series of matrices

Using the diagonalization of a matrix we can more easily work with power series. For instance, if $A = SDS^{-1}$, then for $f(x) = e^x$,

$$e^A = e^{SDS^{-1}} = \sum_{k=0}^{\infty} \frac{1}{k!} SD(\lambda_1^k, \dots, \lambda_n^k)S^{-1}.$$

We can make this into a matrix of power series, and then if the power series converges to $f(\lambda)$ for each λ , we get

$$f(A) = SD(f(\lambda_1), \dots, f(\lambda_n))S^{-1}.$$

This is easy, and works for real and complex eigenvalues!

We don't always get convergence, though – if some of the eigenvalues are too big for convergence this won't work. For instance, the power series expansion of $(I_n - M)^{-1}$ only converges if the eigenvalues of M are of absolute value less than one.

Example 9.8.1. Calculate $\cos(A)$ using power series where

$$A = \begin{pmatrix} \pi/3 & -6 \\ 0 & 0 \end{pmatrix}$$

9.9 Applications to financial math

Often you might be interested in how two quantities vary together – their *covariance*. If you have a number of quantities whose covariances are of interest, you can assemble a *covariance matrix*. Each entry σ_{ij} is the covariance between quantity i and quantity j . Since $\sigma_{ij} = \sigma_{ji}$, the covariance matrix is symmetric.

Principal component analysis is mathematically simply the act of finding the eigenvectors and eigenvalues of the covariance matrix. I think it's fair to say that the first step in principal component analysis is simply linear regression – finding the best-fit line through the data.

Since the covariance matrix is symmetric, the Spectral Theorem will tell us (soon!) that when all the eigenvalues are distinct, the eigenvectors are actually all orthogonal. Since we can choose eigenvectors of whatever length we like (it's direction that is not up to us), that means it's easy to find an orthonormal basis of eigenvectors for a real symmetric matrix.

- Eigenvectors come up in analyzing components of Brownian motion in many dimensions.
- Power series and their truncations come up a lot when looking at normal distributions of prices, for instance.

9.10 Complex vectors

Complex vectors are very similar to real vectors, but there are a few key differences. First, the similarities:

- Multiplication by a scalar works the same.
- Addition of vectors works the same, and is commutative.
- The formula for length is analogous (but not quite the same!).

Here is the first difference: We want a formula for length or magnitude of a complex vector $\vec{v} \in \mathbb{C}^n$ so that

$$\|\vec{v}\|^2 = \|(z_1, z_2, \dots, z_n)\|^2 = |z_1|^2 + |z_2|^2 + \dots + |z_n|^2,$$

in analogy with real vectors. But what is $|z_j|$ for a complex number?

We want $|z_j|$ to be the “absolute value” or length of $|z_j|$, and looking at $z_j = x + iy$ that magnitude would be $\sqrt{x^2 + y^2}$. An easy way to get this is by using the *complex conjugate* $\bar{z}_j = x - iy$ to get the length:

$$|z_j|^2 = \bar{z}_j z_j = (x - iy)(x + iy) = x^2 - i^2 y^2 = x^2 + y^2.$$

This means that the right way to define length of a complex vector is using the *inner product*

$$\vec{v} \cdot \vec{v} = \|\vec{v}\|^2 = |z_1|^2 + \dots + |z_n|^2$$

using

$$|z_j|^2 = \bar{z}_j z_j.$$

This also allows us to define the inner product of two different vectors:

$$\vec{r} \cdot \vec{s} = \bar{r}_1 s_1 + \bar{r}_2 s_2 + \dots + \bar{r}_n s_n.$$

BIG CHANGE here: we are now talking about the **inner product** of two vectors, rather than the dot product, and this inner product is **not commutative**. Check it yourself!

Example 9.10.1. Given $\vec{r} = (3 + i, 2 - 2i)^T$ and $\vec{s} = (i, -1 - 2i)^T$, compare $\vec{r} \cdot \vec{s}$ and $\vec{s} \cdot \vec{r}$. What is their relationship?

It turns out that

$$\vec{r} \cdot \vec{s} = \overline{\vec{s} \cdot \vec{r}}.$$

9.11 Complex matrices

In the previous section, we determined that to take the inner product of two complex vectors we need to use complex conjugates: basically,

$$\vec{r} \cdot \vec{s} = \overline{\vec{r}^T \vec{s}}$$

in matrix notation. Here, we want to maintain some of the *structure* we had in real matrix land — to check if a real matrix A is orthonormal, for instance, we just see if $A^T A$ is equal to the identity matrix. This relies on dot products, so needs to be modified for the complex world.

Define the **Hermitian transpose** M^H of a matrix M by

$$M^H = \bar{M}^T.$$

That is, we take the complex conjugate of every entry and take the transpose of the whole matrix. You can do this to any matrix: for example, the Hermitian transpose of

$$\begin{bmatrix} 1 - i & 2 \\ i & -3 + i \\ 0 & 2i \end{bmatrix}$$

is

$$\begin{bmatrix} 1 - i & 2 \\ i & -3 + i \\ 0 & 2i \end{bmatrix}^H = \begin{bmatrix} 1 + i & -i & 0 \\ 2 & -3 - i & -2i \end{bmatrix}.$$

Also notice that

$$(AB)^H = B^H A^H$$

if the product AB of complex matrices made sense in the first place. You can prove this by using properties of the transpose.

We call M a “complex orthogonal” matrix if it is square ($n \times n$) and has rows \vec{v}_i which satisfy

$$\vec{v}_i \cdot \vec{v}_j = 0$$

for $i \neq j$ and

$$\vec{v}_i \cdot \vec{v}_i = 1.$$

Another name for this is “unitary.” Such a matrix satisfies $M^H M = I_n$, so we have

$$M^{-1} = M^H.$$

(Notice that you can prove $MM^H = I_n$ as well, as $(M^H M)^H = I_n^H$.)

9.12 The spectral theorem

An $n \times n$ complex matrix M is called **Hermitian** if $M = M^H$. Any real symmetric matrix is Hermitian; all the diagonal entries of a Hermitian matrix must be real, even if the rest of the entries are complex. (Check this yourself! Notice that the diagonal entries of the matrix must all be equal to their conjugates.)

Hermitian matrices are very special. Following is a non-exhaustive list of their special properties.

Proposition 9.12.1. Let M be an $n \times n$ Hermitian matrix. Then for column vectors $\vec{v}, \vec{w} \in \mathbb{C}^n$,

$$(M\vec{v}) \cdot \vec{w} = \vec{v} \cdot (M\vec{w}).$$

Proposition 9.12.2. Every eigenvalue of a Hermitian matrix is a real number.

Proposition 9.12.3. Every eigenvalue of a real symmetric matrix is a real number, and the eigenvectors of such a matrix can always be chosen to be real eigenvectors.

Proposition 9.12.4. If \vec{v} and \vec{w} are eigenvectors of a Hermitian matrix, and the corresponding eigenvalues are distinct, then \vec{v} and \vec{w} are orthogonal.

Proposition 9.12.5. If M is an $n \times n$ Hermitian matrix with n distinct eigenvalues $\lambda_1, \dots, \lambda_n$, then all of these eigenvalues are real and we can choose the corresponding eigenvectors so that they are an orthonormal basis for \mathbb{C}^n . If A is the matrix whose columns are those eigenvectors, then

$$A^H M A = D(\lambda_1, \dots, \lambda_n).$$

If M is real and symmetric, then the eigenvectors can be taken to be real, and they form an orthonormal basis for \mathbb{R}^n .

Theorem 9.12.1. If M is an $n \times n$ Hermitian matrix with n eigenvalues $\lambda_1, \dots, \lambda_n$ listed according to multiplicity, then all of these eigenvalues are real. If λ is one of these eigenvalues and has multiplicity k , then there are k corresponding eigenvectors that are an orthonormal to each other. There is an orthonormal

basis for \mathbb{C}^n consisting of eigenvectors of M . If A is the matrix whose columns are those eigenvectors, then

$$A^H M A = D(\lambda_1, \dots, \lambda_n).$$

If M is real and symmetric, then the eigenvectors can be taken to be real, and they form an orthonormal basis for \mathbb{R}^n .

9.13 Singular value decomposition

We have talked a lot about eigenanalysis of square matrices. For any square matrix A , you can now find the eigenvalues and eigenvectors and have some intuitive notion about how these relate to the geometry of the linear transformation given by the matrix A . In particular, you know that you can think of the linear transformation as “scaling” by the eigenvalue λ_i in the direction of the corresponding eigenvector \vec{v}_i .

This point of view works well in many situations. (You will use it to analyze systems of linear first-order differential equations and to classify critical points of surfaces in three-dimensional space, for instance.) Even if one’s data does not appear in a square matrix, relationships between different quantities can often be analyzed in this way by carrying out eigenanalysis of a corresponding *covariance matrix*, which will be square.

However, we need a new tool when we are looking at transformations from \mathbb{R}^m to \mathbb{R}^n . For instance, we may want to look at a rectangular matrix of stock prices under different scenarios, or we might want to analyze how economic factors influence GDP of geographically related provinces or countries. Singular value decomposition can even be used to analyze voting patterns in the US Senate, showing interesting information about party affiliation. All of these involve rectangular arrays of data, and invite interpretation based on the meaning of the domain space and the range space.

Singular value decomposition of a rectangular matrix A will factor A into three matrices:

$$A = U \Sigma V^T.$$

Σ will be an almost-diagonal matrix – it will have the same dimensions as A , so need not be square, but will be zero everywhere but the diagonal. On the diagonal will appear the *singular values* of A , explained below. U and V will be orthogonal matrices that give rotation or reflection on \mathbb{R}^n or \mathbb{R}^m , as appropriate.

Here's the fast version of the process of SVD:

- To find the orthonormal basis $\vec{v}_1, \dots, \vec{v}_n$ of \mathbb{R}^n that will give the columns of the matrix V , we find the eigenvectors of $A^T A$.
- We order the \vec{v}_i by the magnitude of their eigenvalues: that means we have $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, and the matrix V is

$$V = \begin{bmatrix} \vdots & & \vdots \\ \vec{v}_1 & \dots & \vec{v}_n \\ \vdots & & \vdots \end{bmatrix}.$$

- The *singular values* that give the diagonal entries of Σ are $\sigma_j = \sqrt{\lambda_j}$.
- We can find the columns \vec{u}_i of U , which also give (part of?) an orthonormal basis of \mathbb{R}^m , by solving $A\vec{v}_i = \sigma_i\vec{u}_i$ for all the $i = 1, \dots, \min(m, n)$. If $m > n$, we can fill in the remaining \vec{u}_i by Gram-Schmidt orthogonalization.
- Then the matrix U is

$$U = \begin{bmatrix} \vdots & & \vdots \\ \vec{u}_1 & \dots & \vec{u}_n \\ \vdots & & \vdots \end{bmatrix}.$$

Since we define the singular values σ_j of A by $A\vec{v}_j = \sigma_j\vec{u}_j$, for \vec{v}_j an eigenvector of $A^T A$, we can follow the calculations to prove some of the claims made above. For \vec{v}_j an eigenvector of $A^T A$ with eigenvalue λ_j ,

$$(A\vec{v}_i)(A\vec{v}_j) = \vec{v}_i^T A^T A \vec{v}_j \quad (9.1)$$

$$= \vec{v}_i^T \lambda_j \vec{v}_j. \quad (9.2)$$

At the same time, using the definition of singular value we've got

$$(A\vec{v}_i)(A\vec{v}_j) = (\sigma_i \vec{u}_i) \cdot (\sigma_j \vec{u}_j) \quad (9.3)$$

$$= \sigma_i \sigma_j \vec{u}_i \cdot \vec{u}_j. \quad (9.4)$$

Since the \vec{v}_i are orthonormal, by the Spectral Theorem, this tells us that either $\sigma_j^2 |\vec{u}_j|^2 = \lambda_j |\vec{v}_j|^2$ (when $i = j$ in the calculation above) or $\sigma_i \sigma_j \vec{u}_i \cdot \vec{u}_j = 0$, when $i \neq j$. That implies the \vec{u}_j are also orthonormal, which we claimed above but didn't justify.

This gives you a way of finding the SVD decomposition of a matrix A , but what does it all *mean*?

9.14 Applications of SVD

You may have read about principal component analysis (PCA). It's a way of analyzing data by looking at the principal (most predictive, or highest variance) directions in the data, in a way that can be made mathematically precise. PCA is essentially a singular value decomposition on a correlation matrix – that's one reason we cover SVD in my course and this textbook. A correlation matrix is always symmetric, so its SVD matrix factorization ends up being a straightforward diagonalization $C = Q^T D Q$.****

SVD is more versatile than PCA, though. Singular value decomposition examines the “principal components” in both the source space and the target space, allowing two views of the same set of data. This can be very useful when you're trying to get a qualitative as well as quantitative understanding of the data.

Chapter 10

Joint distributions

10.1 Jointly distributed discrete random variables

For simplicity, we'll start by considering two jointly distributed random variables at a time. Once we've established these definitions, we can easily extend to n jointly distributed random variables.

We say that X and Y are *jointly distributed* discrete random variables if

- both X and Y are defined on the same sample space Ω , and
- there is a probability measure P that satisfies the axioms of probability, so that
- we have a function $p(x, y) = P(X = x, Y = y)$ called the joint probability mass function of X and Y .

Notice that $p(x, y)$ is a function defined on \mathbb{R}^2 . Remember that a random variable X or Y is a function from sample space Ω to \mathbb{R} that assigns numerical values to outcomes of probability experiments. In particular, then, $P(X = x, Y = y) = P(A \cap B)$ where $A = \{\omega \in \Omega | X(\omega) = x\}$ and $B = \{\omega \in \Omega | Y(\omega) = y\}$. This seems a little pedantic but it's nice to avoid confusion by trying precision from the start!

We can call (X, Y) or $\begin{bmatrix} X \\ Y \end{bmatrix}$ a *random vector*. You can see how linear algebra will start emerging...

10.1.1 Marginal probability mass function

The marginal probability mass functions are what we get by looking at only one random variable and letting the other roam free. For jointly distributed random variables (X, Y) , we sum over all possible values of Y to get the marginal for X :

$$p_X(x) = \sum_y P(X = x, Y = y).$$

Similarly, sum over all possible values of X to get the marginal for Y :

$$p_Y(y) = \sum_x P(X = x, Y = y).$$

You can think of these as collapsing back to single-variable probability. This definition has its foundation in one of the initial rules of probability we learned: if A_i are mutually exclusive events and $\cup_i A_i = A$, then $\sum_i P(A_i) = P(A)$. Think about how this gives the marginal probability mass functions above.

10.1.2 Examples of discrete jointly distributed random variables

Let's do a pretty simple example first: we'll return to our coin-tossing roots.

Toss a quarter and a dime into the air. We'll let D and Q be the indicator random variables for heads on the dime and the quarter, respectively. That means that $D = 1$ if the dime shows heads and $D = 0$ if the dime shows tails, and $Q = 1$ for the quarter showing heads, $Q = 0$ for tails.

The space Ω of outcomes is pretty limited here: just four outcomes (H, H) , (H, T) , (T, H) , (T, T) with equal probability. So

$$p(d, q) = 1/4 \text{ for } d, q \in \{0, 1\}.$$

Moreover, D and Q are independent. Check for yourself and see that $p(d, q) = p_Q(q)p_D(d)$.

Here is a transformation of the two random variables D and Q . Let X be the number of heads showing when you've tossed the two coins, and let Y be

the amount of money showing tails. For instance, if the dime shows heads and the quarter shows tails, $X = 1$ and $Y = 25$ cents. What is the joint pmf of X and Y ?

It's nice to draw a picture here:

Q \ D	H	T	$f_Q(q)$
H	1/4	1/4	1/2
T	1/4	1/4	1/2
$f_D(d)$	1/2	1/2	1

X \ Y	0	10	25	35	$f_Y(y)$
0	0	0	0	1/4	1/4
1	0	1/4	1/4	0	2/4
2	1/4	0	0	0	1/4
$f_X(x)$	1/4	1/4	1/4	1/4	1

Notice how this table allows us to see the joint pmf for X and Y and also shows the marginal pmfs nicely.

Are X and Y independent? Definitely not – simply by reasoning about the problem, we know that how much money shows tails is related to how many coins show heads. We'll return to this example in the next few sections.

Our coin-tossing problem didn't have any nice formulas to work with. For a contrasting discrete probability problem, we can return to dice!

Roll two dice. Let X be the maximum roll and Y the minimum roll. What is the joint probability mass function for X and Y ?

Again, let's draw a table illustrating the joint pmf and the marginal pmfs:

*****add picture*****

Notice that we can come up with nice formulas for the marginal pmfs:

$$p_X(x) = \frac{2x - 1}{36}$$

and

$$p_Y(y) = \frac{13 - 2y}{36}$$

for $x, y \in \{1, \dots, 6\}$.

10.1.3 Conditional probability mass function

Rather than looking at the probability of given outcomes for Y regardless of the value of X (or vice versa), maybe we'd like to look at the probability of a given outcome of Y *given* a particular value of X . That is, maybe we want $P(Y = y|X = x)$. Using our previous rules for conditional probability, we know that

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

as long as $P(X = x)$ is not zero. Following through the definitions, we can find a conditional probability mass function as well:

$$p(y|x) = \frac{p(x, y)}{p_X(x)}$$

for $p_X(x) > 0$. Likewise,

$$p(x|y) = \frac{p(x, y)}{p_Y(y)}$$

for $p_Y(y) > 0$.

This also allows us to define conditional expected value:

$$E(Y|X = x) = \sum_y yp(y|x).$$

Let's apply this to the example of rolling two dice that we considered a few paragraphs ago. Remember X is the max of our roll of two dice, while Y is the minimum of the roll of two dice. We can find a nice equation for conditional probability of X given any value $Y = y$:

$$p(x|y) = \begin{cases} 0 & x < y \\ \frac{1}{13-2y} & x = y \\ \frac{2}{13-2y} & x > y \end{cases}.$$

For example, if we want the probability mass function for X conditioned on $Y = 3$, we'd evaluate and find:

$$p(x|Y = 3) = \begin{cases} 0 & x < 3 \\ \frac{1}{7} & x = 3 \\ \frac{2}{7} & x > 3 \end{cases}.$$

Check this result using techniques from the past.

10.1.4 Independence

Build off of our old rule for independence: events A and B are independent if $P(A \cap B) = P(A)P(B)$, so X and Y are independent (jointly distributed) discrete random variables if and only if

$$p(x, y) = p_X(x)p_Y(y)$$

for all $(x, y) \in \mathbb{R}^2$. Again, you can often use reasoning to guide you to whether two random variables are independent or not.

10.1.5 Multivariate versions

Extend everything I just said by looking at random vectors

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

taking values in \mathbb{R}^n .

Notice that we can write the expected value of a random vector as the vector of expected values:

$$E(\vec{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix} = \vec{\mu}.$$

10.2 Jointly distributed continuous random variables

Again, we have the concept of random vector (X, Y) , and again we'll do everything in the two-dimensional case first. The main difference from the last section is that continuous random variables have a probability density function and a cumulative distribution function to define, and so that's what we'll do!

Definition 10.2.1. The continuous random variables X and Y have a joint probability density function $f(x, y)$ if there exists a probability measure P that satisfies

$$P(X \leq a, Y \leq b) = \int_{x=-\infty}^a \int_{y=-\infty}^b f(x, y) dy dx$$

with $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ and

$$\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f(x, y) dy dx = 1.$$

A great fact: for a “nice” region $C \subset \mathbb{R}^2$, we have

$$P((X, Y) \in C) = \iint_C f(x, y) dx dy.$$

Remember that in single-variable probability we were able to find the pdf from the cdf by differentiating. Now we need partial derivatives, because we have two variables to consider. Using the multivariate fundamental theorem of calculus, we can see that

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y).$$

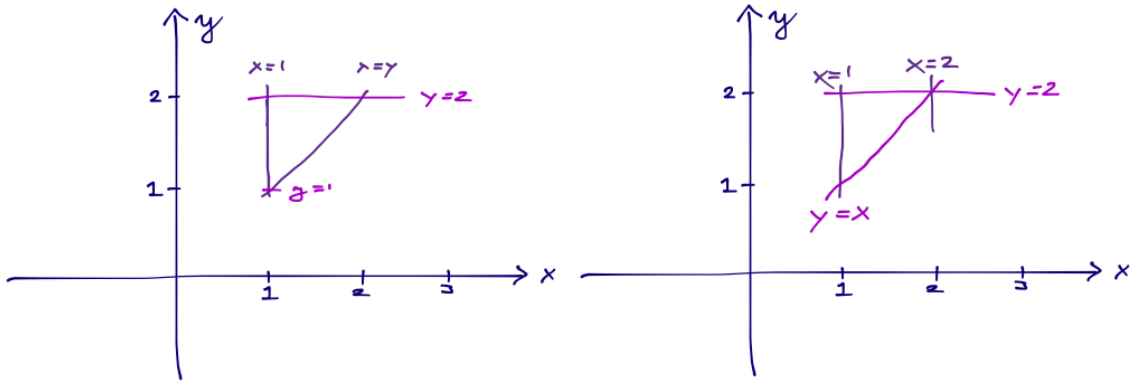
Let's do an example to see how this works in practice:

Example 10.2.1. I will give you a probability density function with a mystery parameter c :

$$f(x, y) = \begin{cases} c(x + y) & 1 < x < y < 2 \\ 0 & \text{else} \end{cases}.$$

10.2. JOINTLY DISTRIBUTED CONTINUOUS RANDOM VARIABLES 195

The restrictions on the domain mean that the random vector (X, Y) takes values in a triangle:



Let's find the parameter c that ensures that this pdf respects the laws of probability: we need

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = \int_{y=1}^{y=2} \int_{x=1}^{x=y} c(x + y) dx dy = 1.$$

Notice the bounds of integration here: why are they what they are? We want to integrate only over a triangle $1 < x < y < 2$, not over the square $1 < x, y < 2$. Do the integration to find that this implies $1.5c = 1$, so $c = 2/3$. Thus the pdf is

$$f(x, y) = \begin{cases} \frac{2}{3}(x + y) & 1 < x < y < 2 \\ 0 & \text{else} \end{cases}.$$

10.2.1 Marginal and conditional probability density functions

We have a familiar pattern: take everything you learned from discrete random variables, turn the sum into an integral, and you've got properties of continuous random variables if correctly interpreted. So, we've got marginal pdfs

$$f_X(x) = \int_{y=-\infty}^{\infty} f(x, y) dy$$

and

$$f_Y(y) = \int_{x=-\infty}^{\infty} f(x, y) dx.$$

Example 10.2.2. Let's revisit the example of X and Y jointly distributed with pdf

$$f(x, y) = \begin{cases} \frac{2}{3}(x + y) & 1 < x < y < 2 \\ 0 & \text{else} \end{cases}.$$

Integrate to find the marginal pdf $f_X(x)$, and think carefully what bounds you want:

$$f_X(x) = \int_{y=x}^{y=2} \frac{2}{3}(x + y) dy = \frac{2}{3} \left(xy + \frac{y^2}{2} \right)$$

for $1 < x < 2$ and $f_X(x) = 0$ else. Likewise,

$$f_Y(y) = \int_{x=1}^{x=y} \frac{2}{3}(x + y) dx = \frac{2}{3} \left(\frac{x^2}{2} + xy \right)$$

for $1 < y < 2$ and $f_Y(y) = 0$ else. These are both functions of a single variable, and so only that variable can appear in the expression for the marginal pdf.

Remember that the marginal pdfs must also satisfy the laws of probability, so we have to have $f_X(x) \geq 0$ and $f_Y(y) \geq 0$ and

$$\int_{-\infty}^{\infty} f_X(x) dx = 1,$$

$$\int_{-\infty}^{\infty} f_Y(y) dy = 1.$$

Again, if (X, Y) is a random vector with a joint density function $f(x, y)$, then we can define the conditional probability density function for a particular value of X as

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

when $f_X(x) > 0$ and zero elsewhere. If we do this, we can define conditional expected value:

$$E(Y|X = x) = \int_{-\infty}^{\infty} yf(y|x)dy.$$

This is the expected value of Y given that $X = x$. However, if we are very careful here we may find some problems, and one of them is a very famous “paradox” in probability, the Borel-Kolomogorov paradox.

10.3 Covariance and correlation, again

Covariance and correlation were mentioned briefly earlier. Remember that we looked at the variance of a sum of independent random variables in Chapter 5. Now, we care about non-independent random variables (jointly distributed random variables!). So...

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

is one way to *define* covariance between X and Y . There are of course more convenient formulas. Another definition of covariance is

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Use linearity of expectation to prove for yourself that this implies

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y].$$

If you need to compute the covariance of two random variables, this is often the easiest way to do it.

Covariance thus defined for two (one-dimensional) random variables is great, but in the multivariate situation we’ll want to look at the covariance matrix! (This is exciting! Trust me!) For a random vector

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

where all the X_i are jointly distributed, we define the covariance matrix $\Sigma = \text{Cov}(\vec{X})$ as having entries $\Sigma_{ij} = \text{cov}(X_i, X_j)$ for $1 \leq i, j \leq n$. This is always a square matrix, and because of the symmetry of covariance, it's always a symmetric matrix as well. Be aware that terminology varies from field to field and book to book: some people call this a variance-covariance matrix, as you can see that the terms on the diagonal are all $\text{cov}(X_i, X_i) = \text{var}(X_i)$. Notice that if all the variables X_i are independent of each other, $\Sigma = \text{Cov}(\vec{X})$ will be a diagonal matrix.

Another cute way to use linear algebra to write the covariance matrix of a random vector $X \in \mathbb{R}^n$ is to say that

$$\Sigma = E[(\vec{X} - E(\vec{X}))(\vec{X} - E(\vec{X}))^T].$$

Again, you can use properties of expectation to write

$$\Sigma = E(\vec{X}\vec{X}^T) - \vec{\mu}\vec{\mu}^T.$$

Last, we can prove that covariance matrices are always positive semidefinite by using the following trick. Remember one characterization of a matrix S being positive semidefinite was that $\vec{v}^T S \vec{v} \geq 0$ for all non-zero vectors $\vec{v} \in \mathbb{R}^n$. Well, for any random vector \vec{X} with mean $\vec{\mu} \in \mathbb{R}^n$,

$$\vec{v}^T \Sigma \vec{v} = \vec{v}^T E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T] \vec{v} \tag{10.1}$$

$$= E[\vec{v}^T (\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T \vec{v}] \tag{10.2}$$

$$= E[(\vec{v} \cdot (\vec{X} - \vec{\mu}))((\vec{X} - \vec{\mu}) \cdot \vec{v})] \tag{10.3}$$

$$= E[s^2] \geq 0. \tag{10.4}$$

The quantity $\vec{v} \cdot (\vec{X} - \vec{\mu})$ will be some scalar, which we call s in the last line of our calculation, and since s^2 is always positive or at worst zero, we're set!

Covariance has the drawback of being intimately related to the units and magnitudes of the random variables under consideration. Correlation is the unitless sister to covariance, and more easily allows comparisons between different types of information. Write $\sigma_{X_i} = \sqrt{\text{var}(X_i)}$. Then as before we write

$\rho(X_i, X_j)$ for the correlation of X_i and X_j ,

$$\rho(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}.$$

Remember that you can prove

$$-1 \leq \rho(X_i, X_j) \leq 1,$$

and in fact you can now prove that using the Cauchy-Schwarz inequality! Moreover, we can define the correlation matrix $\text{Corr}(\vec{X})$ as the matrix whose entries at spot i, j are $\rho(X_i, X_j)$.

Mentioning the Cauchy-Schwarz inequality should give you flashbacks to the section on vector inequalities (Section 8.4). In fact, covariance is conceptually a lot like an inner product, and in particular like the dot product for real vectors. Remember that the dot product for real vectors satisfied a few properties:

- **Symmetry:** just as for $\vec{v}, \vec{w} \in \mathbb{R}^n$ we have $\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$, we have $\text{cov}(X, Y) = \text{cov}(Y, X)$.
- **Linearity:** just as $(a\vec{v} + b\vec{w}) \cdot \vec{u} = a\vec{v} \cdot \vec{u} + b\vec{w} \cdot \vec{u}$, we have $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$.
- **Positive-definiteness:** just as $\vec{v} \cdot \vec{v} \geq 0$ and $\vec{v} \cdot \vec{v} = 0$ if and only if $\vec{v} = 0$, we have $\text{cov}(X, X) = \text{var}(X) \geq 0$ and $\text{var}(X) = 0$ if and only if X is (more or less) constant.¹

This is actually useful for financial applications.

10.4 Multivariate change of variables

In the linear algebra sections of the book, we spent significant effort on changing bases. Now we're going to change variables. This is conceptually similar

¹To be precise here, I should actually say “almost everywhere constant” or “almost surely constant” – constant everywhere but a set of measure zero. But you need measure theory for this.

and will use some linear algebra, but the huge difference is that we don't have to make linear changes of variables! We can make all sorts of changes of variables!

First, the technical details; then some examples. Consider n continuous random variables X_1, \dots, X_n with a continuous joint distribution. Call their joint probability density function $f(x_1, \dots, x_n)$ and let it be defined on a domain S . Say we want to transform these random variables into n new random variables, Y_1, \dots, Y_n , using some *one-to-one differentiable* functions r_1, \dots, r_n from S to T :

$$Y_i = r_i(X_1, \dots, X_n)$$

for $i = 1, \dots, n$. Here T is some new domain (could be the same as S or different). Since these r_i are one-to-one and differentiable, there is an inverse transformation that has $x_i = s_i(y_1, \dots, y_n)$ for each $i = 1, \dots, n$. This takes points in T to points in S . (Think about how this is the analogue of the “monotonically increasing” or “monotonically decreasing” condition in the single-variable case.)

Along with the inverse of the transformation, we need to know how the transformation changes the variables infinitesimally. We understand this change through using the Jacobian,

$$J = \det \begin{bmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{bmatrix}.$$

Determinants are intimately related to volume – remember the determinant of a three by three matrix is the volume of the parallelepiped spanned by the three columns of the matrix, for instance. (This was discussed in [section 8.9](#).) The partial derivatives $\frac{\partial s_i}{\partial y_j}$ measures how much s_i is changing with respect to y_j . The combination of the determinant of the matrix of derivatives can be thought of as measuring the change of volume that is forced by the transformation.

Given all this set-up, the joint probability density function $g(y_1, \dots, y_n)$ of the random variables Y_1, \dots, Y_n is

$$g(y_1, \dots, y_n) = \begin{cases} f(s_1, \dots, s_n)|J| & (y_1, \dots, y_n) \in T \\ 0 & \text{else} \end{cases}$$

There are essentially two actions here: you need to substitute in the new variables y_1, \dots, y_n by plugging in each $x_i = s_i(y_1, \dots, y_n)$, and then you need to compensate for “stretching” caused by the transformation using $|J|$.

10.5 Bivariate and multivariate normal

The bivariate normal distribution is a great place to start exploring multivariate normal distributions, as we can actually draw pictures. Let’s start with a definition and a derivation.

Definition 10.5.1. A random vector $\vec{X} = (X, Y)$ of jointly distributed random variables has the bivariate normal distribution if $aX + bY$ is normally distributed for every pair $(a, b) \in \mathbb{R}^2$ with $(a, b) \neq (0, 0)$.

While a bit abstract, this is actually a very useful characterisation of bivariate normal distributions. You might wonder, though, what the probability density function is. That’s fair. Let’s start, as in the single-variable case, with the simplest situation – here, two independent though jointly-distributed standard normal random variables. Thus we consider $Z_1 \sim \mathcal{N}(0, 1)$ and $Z_2 \sim \mathcal{N}(0, 1)$, with $\rho(Z_1, Z_2) = 0$. Then the pdf for their bivariate normal distribution is

$$f(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2 + z_2^2)}$$

for $(z_1, z_2) \in \mathbb{R}^2$. Notice this is no big surprise; since Z_1 and Z_2 are independent, their joint density function is just the product of their individual density functions.

As in the single-variable case, we can transform our way from this straightforward density function to any other bivariate normal density function. How would we get to the probability density function for $\vec{X} = (X, Y)$ where $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and $\rho(X, Y) = \rho$?

One way to look at this is to transform the standard normal Z_i to have the desired parameters, by using affine linear transformations:

$$X = \sigma_X Z_1 + \mu_X$$

$$Y = \sigma_Y[\rho Z_1 + \sqrt{1 - \rho^2} Z_2] + \mu_Y.$$

We shift Z_1 to get the desired mean μ_X and stretch to get the desired variance $\text{var}(X) = \sigma_X^2$. Check for yourself that Y has the desired mean and variance! Since X and Y are linear combinations of normal random variables Z_1 and Z_2 , X and Y will also be normally distributed because of our definition previously. Notice that $\text{cov}(X, Y) = \sigma_X \sigma_Y \rho$, as well.

How do we find the joint density function of X and Y from this? Use the multivariate change of variables formula discussed earlier. In our current situation, we have inverse functions

$$s_1(x, y) = \frac{x - \mu_X}{\sigma_X}$$

and

$$s_2(x, y) = \frac{1}{\sqrt{1 - \rho^2}} \left[\frac{y - \mu_Y}{\sigma_Y} - \rho \frac{x - \mu_X}{\sigma_X} \right].$$

These give the Jacobian

$$J = \det \begin{bmatrix} \frac{1}{\sigma_X} & 0 \\ \frac{-\rho}{\sigma_X \sqrt{1 - \rho^2}} & \frac{1}{\sigma_Y \sqrt{1 - \rho^2}} \end{bmatrix} = \frac{1}{\sigma_X \sigma_Y \sqrt{1 - \rho^2}}.$$

Putting this all together, we can get the horrible formula

$$f(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} e^{\frac{-1}{2(1 - \rho^2)} \left(\frac{(x - \mu_X)^2}{\sigma_X^2} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} - 2\rho \frac{(x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} \right)}.$$

Why do I call this horrible? Because it takes a long time to type, and some people find it easy to mess up the recall of the formula on exams due to its length. Use the power of linear algebra to simplify this expression! If we write

$$\vec{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix},$$

then we can write instead

$$f(\vec{x}) = \frac{1}{2\pi \sqrt{\det(\sigma)}} e^{-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1}(\vec{x} - \vec{\mu})}.$$

The best part is that this formula generalizes to a k -dimensional multivariate normal: if \vec{x} is k -dimensional, then

$$f(\vec{x}) = \frac{1}{(2\pi)^{k/2} \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})}.$$

The only change is in the power of 2π ; the structure of the density function remains the same. Of course, here Σ is the covariance matrix with $\Sigma_{i,j} = \text{cov}(X_i, X_j)$.

You need to familiarize yourself with the probability density function for the bivariate and multivariate normal distributions to call yourself a financial mathematician. They're foundational. However, in practice you won't use these equations explicitly all that often. You'll often be able to rely on computer implementations of the multivariate normal to simulate returns or other quantities that you're modeling with a normal distribution, and you'll be able to use properties of the multivariate normal distribution to solve other problems without going to the pdf. For instance, you might want to do a basic calculation like this:

Example 10.5.1. You're considering two stocks and modeling their returns using a joint bivariate distribution. Based on historical data, stock A has returns with mean $\mu_A = 0.03$ and variance $\sigma_A^2 = 0.02$, while stock B has returns with mean $\mu_B = 0.1$ and variance $\sigma_B^2 = 0.01$. What is the probability that a portfolio that invests equally in both stocks will have returns greater than 0.08?

Using standardization and z-tables will be a fine technique for solving many multivariate normal problems, and I would be negligent not to discuss these problems here. On the other hand, you can find this material in many probability texts and I urge you to visit them for examples (see for instance Rosenkrantz).

So as not to duplicate commonly-available standard materials, I'll turn instead to visualization and some topics I feel haven't been covered very well in the texts I've looked at.

10.6 Visualizing the bivariate normal distribution

We'll concentrate on the bivariate normal distribution just because it's easy to graph, but these ideas again generalize to higher dimensions!

You can graph the probability density function of a bivariate normal as a surface, with $f(x, y) = z$:

add pic**

This gives you a nice visual representation of the fact that if X and Y have a bivariate normal distribution then any $aX + bY$ is also normally distributed (for $(a, b) \neq \vec{0}$).

****add pic*****

Contour plots are a nice way of visualizing this pdf: we can take slices for constant values c of $f(x, y) = z$ and project them onto the xy plane, using different colors to show different values of c :

*****add pic*****

At first glance this all seems a bit obvious, but the power of mathematics only comes into play if you dig deeper and question the obvious. Why do we always get ellipses? Do we always get ellipses? If we have ellipses, what does their shape tell us? (Any ideas? Ponder these questions for three minutes and then read on, or if you're truly dedicated to understanding, get the Python worksheet or grab your pencil and experiment.)

Do we always get ellipses? Well, yes and no. You don't get an ellipse if X and Y are equal, for instance. More generally, you don't get an ellipse if the covariance matrix is singular, but in that case you also can't use the usual equation for the probability density function because Σ^{-1} is not defined and you have a divide-by-zero error at the very beginning (the $\frac{1}{\det(\Sigma)}$ factor). If the covariance matrix is full rank, which must be the case when you're using the pdf given, you get an ellipse.

Why an ellipse? To answer this, we need quadratic forms. Each of these level sets is the set of points (x, y) that satisfy $f(x, y) = c$. Let's solve:

$$\frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{m}\mu)^T \Sigma^{-1}(\vec{x}-\vec{m}\mu)} = c$$

$$e^{-\frac{1}{2}(\vec{x}-\vec{m}\vec{u})^T\Sigma^{-1}(\vec{x}-\vec{m}\vec{u})} = 2\pi\sqrt{\det(\Sigma)}c = \tilde{c}$$

I don't want to write out all the parts of this constant, so make a new constant \tilde{c} .

$$(\vec{x} - \vec{m}\vec{u})^T \Sigma^{-1} (\vec{x} - \vec{m}\vec{u}) = -2 \ln \tilde{c}$$

This is a quadratic form!! Bet you thought you wouldn't need [Section 9.7](#) again. Hah. The matrix for this quadratic form is Σ^{-1} . What do we know about Σ^{-1} ? Well, we know it's positive definite. How do we know that? Σ is positive definite, with eigenvalues $\lambda_i > 0$ and corresponding eigenvectors \vec{v}_i . Then Σ^{-1} has eigenvalues $1/\lambda_i$ for corresponding eigenvectors \vec{v}_i , and all those eigenvalues are also positive – so Σ^{-1} must also be positive definite.

Speaking of these eigenvectors, they provide the directions of the major and minor axes of the ellipses under consideration. (In higher dimensions, they give all the axes of the “covariance ellipsoids”.) The eigenvector corresponding to the biggest eigenvalue of Σ gives the major axis of the ellipse (the long dimension of the ellipse) and the eigenvector corresponding to the smaller eigenvalue gives the minor axis of the ellipse, in this bivariate case.

Chapter 11

Optimization and Newton's method

In the single-variable portion of the course, we emphasized short- and long-term predictions (differentiation and integration) along with single-variable probability. Then we learned about linear algebra with real and complex numbers, mixing that up with joint distributions of random variables. Now we're going to head toward different kinds of approximation: approximating solutions to equations via Newton's method and approximating scalar-valued functions themselves using power series and Taylor polynomials.

11.1 Single-variable optimization

As a bit of motivation and a setting for these techniques, let's start with optimization for functions $f : \mathbb{R} \rightarrow \mathbb{R}$. The single-variable case is very familiar to you, as it's what a first calculus course emphasizes, but I'd like to call out the parts that are particularly useful when we move to functions $f(\vec{x})$ defined on \mathbb{R}^n for $n > 1$.

11.1.1 Single-variable unconstrained minima and maxima

Long ago and possibly far far away in your first calculus class, you learned how to find local and global maxima and minima of functions $f : \mathbb{R} \rightarrow \mathbb{R}$:

- find the points x_i at which $f'(x_i) = 0$ or at which $f'(x_i)$ does not exist (critical points)
- classify such points as local maxima, minima, or saddles using either the first or second derivative tests,
- then compare all the values and the behavior of the function to determine global maxima and minima (if they exist).

In most calculus class homework problems, finding x_i such that $f'(x_i) = 0$ is possible algebraically or using trigonometry. If it is not straightforward to solve the equations, you can use numerical methods like the bisection method or Newton's method to find values for x_i . We'll discuss Newton's method in the next section. To classify the critical points as maxima, minima, or saddle points, you can look at the first derivative on either side of the point x_i (does the function change from increasing to decreasing, decreasing to increasing, or not change direction?) or you can use the second derivative (check if the function is concave up, concave down, or neither at the point x_i). Let's do a few examples.

Example 11.1.1. Here's an example that won't exactly come up in finance, but which is an illustration of the analytical methods you can use to find local extrema. Think about the function $f(x) = x + \sin x$.

Identify critical points first: $f'(x) = 1 + \cos x = 0$ when $\cos x = -1$, so for all $x = \pi + 2\pi k$ for $k \in \mathbb{Z}$. How can we classify the critical points?

The first derivative test would tell us to look at where f is increasing (where $f'(x) > 0$) and where f is decreasing (where $f'(x) < 0$). Well, we know $0 \leq 1 + \cos x \leq 2$ for all x , so $f(x)$ is always non-decreasing – so all those critical points are actually saddles by the first derivative test.

Another check: if we look at $f''(x) = -\sin x$, we see that f does alternate between being concave up ($f'' > 0$) and concave down ($f'' < 0$). But at

$x = \pi + 2\pi k$, $f''(x) = 0$. So what does this tell us? Actually, nothing. The second derivative test is indeterminate, because each critical point is an inflection point as well.

A slight modification of the function is very different. If you consider $g(x) = \frac{1}{2}x + \sin x$, you'll find $g'(x) = \frac{1}{2} + \cos x = 0$ forces $\cos x = -\frac{1}{2}$. Looking at this geometrically, you can see that two sets of values solve this: $x = \frac{2\pi}{3} + 2\pi k$ and $x = -\frac{2\pi}{3} + 2\pi k$ for $k \in \mathbb{Z}$. So we again have an infinite set of critical points.

How do we classify these critical points? Since $g'(x) > 0$ on the intervals where $\cos x > -\frac{1}{2}$, we have g increasing on $(-\frac{2\pi}{3}, \frac{2\pi}{3})$ and all 2π translates, and g decreasing on $(\frac{2\pi}{3}, \frac{4\pi}{3})$ and its 2π translates. Increasing to decreasing... that's a max, and so each $x = \frac{2\pi}{3} + 2\pi k$ is a max. Decreasing to increasing makes a minimum, and so $x = -\frac{2\pi}{3} + 2\pi k$ are all minima.

If you don't want to think like that, check the second derivative. $g''(x) = -\sin x$, and so $g''(x) > 0$ for all $x = -\frac{2\pi}{3} + 2\pi k$, which then are all minima. Similarly, $g''(x) = -\sin x < 0$ for $x = \frac{2\pi}{3} + 2\pi k$, so those are maxima.

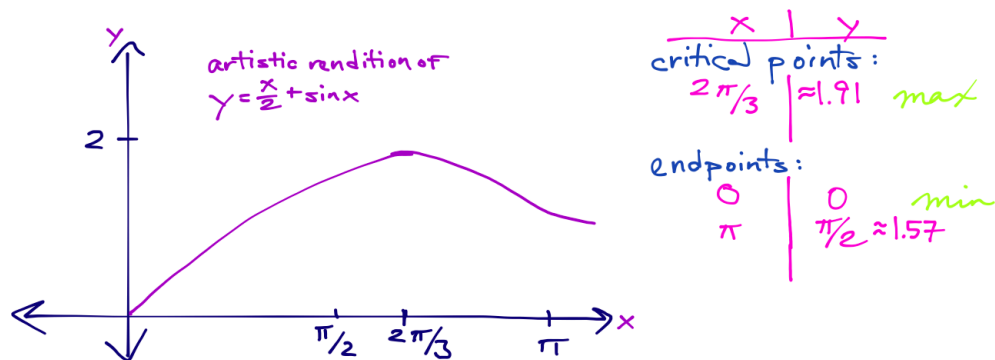
To know whether your extrema are local or global extrema,

- Check whether the function goes to infinity or negative infinity as x goes to infinity or negative infinity, and check to see if the function has any asymptotes.
- Make a list of the value of the function at the maxes and mins.
- Compare the two parts above and pick out the local and global maxes and mins.

Remember that a global or absolute maximum point x_0 for $f : \mathbb{R} \rightarrow \mathbb{R}$ is a point such that for all other $x \in \mathbb{R}$, $f(x) \leq f(x_0)$. For instance, a function whose highest-degree term ax^{2k} has a positive coefficient a will go to positive infinity as x goes to ∞ or $-\infty$, because of the even degree. That means it cannot have a global maximum at any particular value of x in \mathbb{R} , because you can always find a bigger x and a bigger output.

11.1.2 Single-variable constrained optimization

In the single-variable case, it's actually usually really easy to deal with constraints, because they just can't be that complicated: constraints in \mathbb{R} take the form of restricting to an interval or a point. For instance, you can find the mins and maxes of $f(x) = \frac{x}{2} + \sin(x)$ on the interval $[0, \pi]$ by using our previous work (which tells us $x = \frac{2\pi}{3}$ is a critical point in this interval) and then comparing the output of f there with the value of f at the ends of the interval:



With a picture it's very easy to see what's going on. Even in the absence of a picture we can follow the following procedure as long as $f(x)$ is continuous on $[a, b]$.

To find the absolute maximum and minimum of $f(x)$ on $[a, b]$:

- Find the critical points of $f(x)$ in the interior (a, b) . (You don't need to classify them.)
- Evaluate $f(x)$ at the critical points and evaluate $f(a)$ and $f(b)$. Write these down in a list.
- Circle the biggest value, circle the smallest value. These give you your absolute min and absolute max!

There is a theorem that applies here:

Theorem 11.1.1 (Extreme Value Theorem). As long as $f(x)$ is continuous on $[a, b]$, $f(x)$ will attain an absolute maximum and an absolute minimum on the interval $[a, b]$.

The idea that you should look at the interior of the constrained region and then look at the boundaries will carry over into the multivariable setting.

11.2 Newton's method, single-variable

Many equations can't be solved exactly. Even many polynomials can't be solved exactly! Newton's method provides a way of approximating roots of any differentiable function. You might think that Newton's method is thus not very versatile... but you just spent a few hours (I hope) finding $f'(x) = 0$ for various functions $f(x)$. That means you were finding roots of the function f' . Newton's method can be generalized into a variety of optimization techniques, even inspiring methods in machine learning. Again, we'll start with the familiar single-variable version (and look at some failures of the method) and then generalize.

Newton's method relies on our old idea of short-term approximation or linear approximation, in essence "running it backward" to find roots by assuming that there is some x close to our initial guess x_1 such that $f(x) = 0$. If we have an approximate solution x_1 , so $f(x_1) \approx 0$, we want a better solution $x_2 = x_1 + \Delta x$ (one so that $f(x_2)$ is ideally closer to zero than $f(x_1)$). Use the basic equation for short-term prediction to estimate this new x_2 :

$$0 \approx f(x_2) = f(x_1 + \Delta x) \approx f(x_1) + f'(x_1)\Delta x.$$

This suggests we consider

$$\Delta x = -\frac{f(x_1)}{f'(x_1)},$$

which we can use in $x_2 = x_1 + \Delta x$. In fact, this often works! And if it works once, why not try again and find an even better x_3 ?

Newton's method, then, is an iterative process that solves $f(x) = 0$ by finding a sequence $x_1, x_2, \dots, x_n, \dots$ using

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}.$$

When successful, this sequence of x_i s converges to a root x . In general, you don't have to iterate too many times. If the root you are seeking to approximate is of multiplicity one and your initial guess was a good one, then the number of correct digits after the decimal roughly doubles after every iteration. This is proved using *power series*, which we will discuss further.

Example 11.2.1. Find an approximate value for $\sqrt{50}$. (Hint: this is the same as solving the equation $x^2 - 50 = 0$.)

Example 11.2.2. Find an approximate value for the solution of $(\cos x)^2 = x$.

Newton's method is fairly straightforward when it works. It fails to work for several reasons. Two important "failure modes" are when $f'(x_i) = 0$ for some x_i in the iterative process, and when the sequence of x_i fails to converge.

Visualize the first failure mentioned above: when $f(x_i) = 0$, the graph of f has a horizontal tangent line, and so the next step of Newton's method (finding the intersection of the horizontal tangent line with the x -axis) cannot be carried out.

The second failure is very interesting mathematically, though outside the scope of this class. It leads directly into what's called "dynamical systems theory" and "chaos theory".

A note on conditions for convergence of Newton's method: You can prove that for a function f there exists an open interval $U \in \mathbb{R}^1$ around a root of f so that for any x_0 in U , Newton's method converges if f is continuously differentiable over U and $f'(x) \neq 0$ for all x in U . Basically, if f is "nice" and U is small, we can get convergence.

Example 11.2.3. Analyze $f(x) = x^{1/3}$ and explain why $x_0 = 1$, for instance, leads to a sequence x_i that does not converge. We have a theorem that guarantees conditions under certain conditions. Why doesn't the theorem apply?

Example 11.2.4. Analyze $f(x) = x^3 - 2x + 2$. Start with $x_0 = 0$. What happens? Draw a picture.

This is an example of a period-two solution. Newton's method oscillates between two solutions forever. Instead of starting with $x_0 = 0$, start with $x_0 = 0.001$ or another similarly small value. Do you get a different solution? What happens? You should see an example of "sensitivity to initial conditions." While this seems somewhat theoretical, this comes up a lot in financial modeling. If you're creating a model more complicated than a linear regression, you may come up with something that gives you pretty different results depending on your initial conditions. "Robustness" and "sensitivity analysis" are both words that come up in this context. Often professionals in mathematics prefer robust models, models that don't give wildly different results given slightly different inputs, because so many inputs coming from the real world are estimates or adjustments.

Example 11.2.5. Consider the function $f(x) = \frac{9}{2\sqrt{3}}(x^3 - x)$ and try any initial value $x_0 \neq 0$ between -1 and 1 . What happens? Draw a picture if you can!

This is an example of "chaotic" behavior. Chaos is avoided by most financial mathematicians, but some instead lean into it: Benoit Mandelbrot, for instance, has put forward the view that traditional stochastic calculus is fundamentally inaccurate for financial mathematics and that chaos theory and in particular fractal self-similarity are better for describing the behavior of financial markets. Check out his book, "The Mis-behavior of Markets," to learn more.

11.2.1 Newton's method for single-variable optimization

Instead of using Newton's method to find $f(x) = 0$, we can use it to find $f'(x) = 0$ and thus find extrema more directly. Check for yourself that now the iteration is

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}.$$

Since you'll have computed $f''(x)$ in the process, this will give a quick and streamlined way of finding and classifying extrema as long as $f''(x) \neq 0$. If $f''(x) = 0$, you won't be able to use this method at all – the iteration will diverge quickly!

Put this into practice by finding the absolute minimum of $f(x) = x^2 + \sin x$ to four decimal places without first finding the critical points of $f(x)$.

11.3 Multivariate Taylor approximations

You will notice that everything above used $f'(x)$ and maybe $f''(x)$, first and second derivatives of single-variable functions. These pieces of information can be encapsulated in linear approximations and quadratic approximations of the function $f(x)$. To generalize the methods above to multiple variables, we need multivariate versions.

We will start with degree-two approximations of single-variable functions to establish our notation and the ideas, then move to multivariable functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Our primary multivariable examples will be functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ because these can be drawn on paper!

How do you write the degree-two polynomial approximation to any function $f : \mathbb{R}^1 \rightarrow \mathbb{R}$? Dredge up memories from Taylor approximation and remind yourself now. Remember that the Taylor series of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ or $f : \mathbb{C} \rightarrow \mathbb{C}$ at a point a is given by the power series

$$T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

By just truncating this power series at some point, we can get fine degree- k approximations near the input a :

$$f(x) \approx \sum_{n=0}^k \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

This includes our linear approximation,

$$f(x) \approx f(a) + f'(a)x$$

and the quadratic approximation

$$f(x) \approx f(a) + f'(a)x + \frac{1}{2}f''(a)x^2.$$

We'd like to do something similar for a multivariable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Write $f(x_1, \dots, x_n)$ or $f(\vec{x})$ for the (scalar) output of the function.

What is the multivariable equivalent of the first derivative?

It is called the *gradient*, and it assembles the first derivatives of $f(\vec{x})$ with respect to every variable in one vector of partial derivatives: the gradient of $f(\vec{x})$ is

$$\vec{\nabla} f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

You can evaluate at a point $\vec{x} = \vec{a}$ by plugging in \vec{a} :

$$\vec{\nabla} f(\vec{x})|_{\vec{a}} = \vec{\nabla} f(\vec{a}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{a}) \\ \frac{\partial f}{\partial x_2}(\vec{a}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{a}) \end{bmatrix}.$$

At a point \vec{a} , this gradient vector $\vec{\nabla} f(\vec{a})$ indicates the direction of fastest change – the direction in which you'd walk from \vec{a} to experience the greatest “slope.”

What is the multivariable equivalent to the second derivative? Instead of a vector, we now need a matrix. This matrix $Hf(\vec{x})$ is called the *Hessian*, and it's a delightful matrix of the second derivatives of f with respect to each pair of coordinates.

$$Hf(\vec{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial^2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{bmatrix}.$$

What is true about the Hessian matrix $Hf(\vec{a})$?

- If f is a real-valued function of real variables, $Hf(\vec{a})$ is a real matrix.
- It is symmetric by Clairaut's theorem, as long as f is second-differentiable around \vec{a} .
- Its eigenvalues and eigenvectors at the point \vec{a} tell us about the properties of the function f near \vec{a} . In particular, we can classify critical points of f using the Hessian evaluated at the critical point.

The multivariable quadratic approximation to $f(x_1, \dots, x_n)$ at $a = (a_1, \dots, a_n)$ can be written

$$q(\vec{x}) \approx f(\vec{a}) + \vec{\nabla} f(\vec{a}) \cdot (\vec{x} - \vec{a}) + \frac{1}{2}(\vec{x} - \vec{a})^T Hf(\vec{a})(\vec{x} - \vec{a}),$$

or in a slightly different notation where \vec{h} is the small change from point \vec{a} ,

$$f(\vec{a} + \vec{h}) \approx f(\vec{a}) + \vec{\nabla} f(\vec{a}) \cdot \vec{h} + \frac{1}{2}\vec{h}^T Hf(\vec{a})\vec{h}.$$

Notice that $\vec{x} - \vec{a}$ corresponds to \vec{h} .

Try applying this yourself:

Example 11.3.1. Find the quadratic approximation to the function $f(\vec{x}) = x_1^2 \sin(x_2 - 2x_3)$ at the point $\vec{x} = (-2, -\pi/2, \pi/2)$.

These quadratic approximations are useful in a variety of contexts, not just optimization. In finance, option pricing methods include delta-approximations (Δ -approximation) which is a linear approximation, delta-gamma approximation ($\Delta\Gamma$ -approximation), and then delta-gamma-theta methods. These are all developed to estimate option price movements when the price of the underlying is also changing. For an option on a single stock, single-variable methods suffice... but of course you want to work with portfolios!

11.3.1 Zeroes of multivariate functions

You might ask yourself if you could find the roots of a function $f(\vec{x})$ using something like Newton's method. Well, kind of. The problem with finding

roots of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is that often zeroes of such a function are not isolated. For instance, what do the zeroes of $f(x, y) = x^2 + y^2 - 1$ look like? Any points in the unit circle $x^2 + y^2 = 1$ give $f(x, y) = 0$. Or if you want to know where $f(x, y, z) = x^2 + y^2 - z^2$ is zero – it’s an entire surface in \mathbb{R}^3 . When $f(\vec{x})$ is a polynomial, the study of sets of points that give $f(\vec{x}) = 0$ is an entire area of mathematics called algebraic geometry (the field in which I got my PhD). It’s very beautiful mathematics, and there are even practical applications in optimization, cryptography (elliptic curves, for instance), and genetics. However, algebraic geometry is really beyond the scope of this course.

One idea I will pull out from algebraic geometry is an idea that starts in multivariable calculus and linear algebra. In linear algebra, we saw that a system of n linear equations in n variables has a unique solution if the corresponding matrix is full rank. If you have more equations than variables, you may have an “overdetermined” system – too many conditions may mean no solutions. If you have fewer equations than variables, you may have an “underdetermined” system – too much freedom means that solutions are not unique. You’ve got a positive-dimensional space of solutions, in linear algebra. When you have nonlinear equations, it’s not quite so clear, but a similar idea carries through. Looking at $f(x, y, z) = x^2 + y^2 - z^2 = 0$, you’ve got one condition or constraint in a three-dimensional ambient space. That leaves two degrees of freedom, and so you’ve got two-dimensional solution set. This reasoning can be extended.

Anyhow, we’ll leave this discussion here. A multivariate Newton’s method for finding zeroes of $f(\vec{x}) = 0 \in \mathbb{R}$ just is not so easy because that zero set won’t just be a point. However, we *can* use Newton’s method more fruitfully in the setting of optimization. Isolated maxes and mins of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ are much more common, and very useful in finance. Maximizing the return of a portfolio, minimizing the variance – classic applications.

11.4 Multivariate optimization

11.4.1 Unconstrained optimization

A *critical point* of the function f is a point \vec{a} at which $\vec{\nabla} f(\vec{a}) = 0$, or at which f is not differentiable. For a differentiable function f , how can you interpret this in terms of “steepest ascent”?¹ (What does this look like for a differentiable function $f(x_1, x_2)$, for instance?) Maxima and minima of f can appear only at critical points.

We can classify a critical point \vec{a} of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as a maximum, minimum, or saddle by looking at $Hf(\vec{a})$. This is the multivariate version of the second derivative test. Specifically, if this matrix is

- positive definite, then \vec{a} is a strict minimum
- negative definite, then \vec{a} is a strict maximum
- positive/negative semidefinite, then \vec{a} is a non-strict minimum/maximum
- indefinite, then \vec{a} is a saddle point

Wonderfully, we can get the definiteness of a matrix from looking at its eigenvalues!

Quiz yourself:

- If the eigenvalues of a matrix are positive, the matrix is _____.
- If the eigenvalues of a matrix are negative, the matrix is _____.
- If one of the eigenvalues is zero, then the matrix is _____ or _____, depending on _____
- If the matrix has eigenvalues of opposite signs, then the matrix is _____.

I'll write this as a theorem so it's official (and this can carry over to constrained optimization):

¹If you said that $\vec{\nabla} f(\vec{a}) = 0$ means there is some sort of a “flat point” at \vec{a} , you may have the right idea.

Theorem 11.4.1. If $f(\vec{x})$ is a function with continuous second partial derivatives on a set $D \subset \mathbb{R}^n$, and \vec{a} is an interior point of D that is also a critical point of f , then the eigenvalues of the matrix $Hf(\vec{a})$ determine whether \vec{a} is a (local) maximum, minimum, or saddle point of f .

Again, if you'd like to check whether points are globally optimal, you'll need to understand the behavior of the function (continuous or not? bounded or not?) and then examine the values of the function at the critical points and as inputs go off to infinity in various directions.

11.4.2 Multivariate Newton's method for optimization

Let's return to the consideration of Newton's method for optimization. Why is Newton's method useful when we're considering optimization? Look at the linear and quadratic approximations to $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at \vec{a} :

$$L(\vec{x}) = f(\vec{a}) + \nabla f|_{\vec{a}} \cdot (\vec{x} - \vec{a})$$

$$Q(\vec{x}) = f(\vec{a}) + \nabla f|_{\vec{a}} \cdot (\vec{x} - \vec{a}) + \frac{1}{2}(\vec{x} - \vec{a})^T Hf|_{\vec{a}}(\vec{x} - \vec{a})$$

Now, if \vec{a} happens to be a critical point of the function f , then $\nabla f|_{\vec{a}} = \vec{0}$. But the point of Newton's method is that we create an iterative process that starts with a point that is *not* the minimum (if we knew the minimum, we wouldn't need to optimize!). So, rather than set $\nabla f|_{\vec{x}} = \vec{0}$ and solve, we want to be slightly cleverer. Instead set $\nabla Q(\vec{x}) = \vec{0}$ and use that to give us a hint.

$$\nabla Q(\vec{x}_0) = Hf|_{\vec{x}_0}\vec{x} + \nabla f|_{\vec{x}_0}$$

If we set this equal to zero and solve, we'd get

$$\vec{x}_1 = -Hf^{-1} + \nabla f|_{\vec{x}_0}.$$

Iterate!

11.4.3 Constrained optimization

Either you want maxes/mins on the boundary of a region, or in a bounded region.....

Chapter 12

Differential equations

Goals of this chapter

- find equilibrium solutions of basic differential equations
- classify equilibrium points of systems of first-order equations, at least roughly!

This topic is mostly qualitative, but involves elements of review from earlier topics as well. Keep an eye out for:

- short-term prediction
- long-term prediction
- Euler's method
- a possible application of Newton's method
- optimization
- eigenvalues and eigenvectors

12.1 Equilibrium: the concept

Differential equations allow us to study how quantities change in relation to each other, as they're equations that set out the relationships between given derivatives. *Equilibrium* is a useful concept to group our exploration: it's defined as "a state in which opposing forces or influences are balanced" according to Google's source on August 27, 2018. In the world of differential equations, think of equilibrium as a solution to a differential equation in which quantities are not increasing or decreasing, but instead are constant... a solution in which derivatives are zero!

12.2 A single ordinary differential equation

Consider a basic differential equation like

$$\frac{dx}{dt} = 0.02x.$$

This is the sort of equation we considered at the beginning of the semester: we were able to solve such an equation using the method of long-term approximation or using analytical methods:

$$\int \frac{dx}{x} = \int 0.02dt$$

so

$$\ln|x| = 0.02t + C$$

and thus

$$x(t) = e^C e^{0.02t} = P e^{0.02t}.$$

In general, humans are more interested in gaining or losing money than keeping it the same, but it's actually extraordinarily useful for mathematicians and analysts to understand *equilibria*, points at which the rate of change of a quantity is zero. For this equation under consideration, the equilibrium solution would be when $\frac{dx}{dt} = 0$, which is when $0.02x = 0$ (by the design of the equation), which occurs only when the function $x(t) = 0$.

Equilibrium analysis is useful because if we consider a smoothly-changing $\frac{dx}{dt}$, then $\frac{dx}{dt}$ can't change signs without passing through zero! (When I say "smoothly-changing" I mean that $f(x)$ in $\frac{dx}{dt} = f(x, t)$ is a continuous and differentiable function. This is stricter than necessary but will serve our purposes.) In fact, for a single differential equation like this which depends only on x , we can draw a *phase line* that illustrates the behavior of $x(t)$ by looking at the sign of $\frac{dx}{dt}$.

Example 12.2.1. Consider the example

$$\frac{dx}{dt} = 0.02x(1 - x).$$

For what values of $x(t)$ is the system at equilibrium? That is, when is $\frac{dx}{dt} = 0$? Draw a phase line and indicate where dx/dt is positive and where it is negative.

The equation above is an example of a *logistic* differential equation. The logistic equation is very useful in population ecology, and there is written

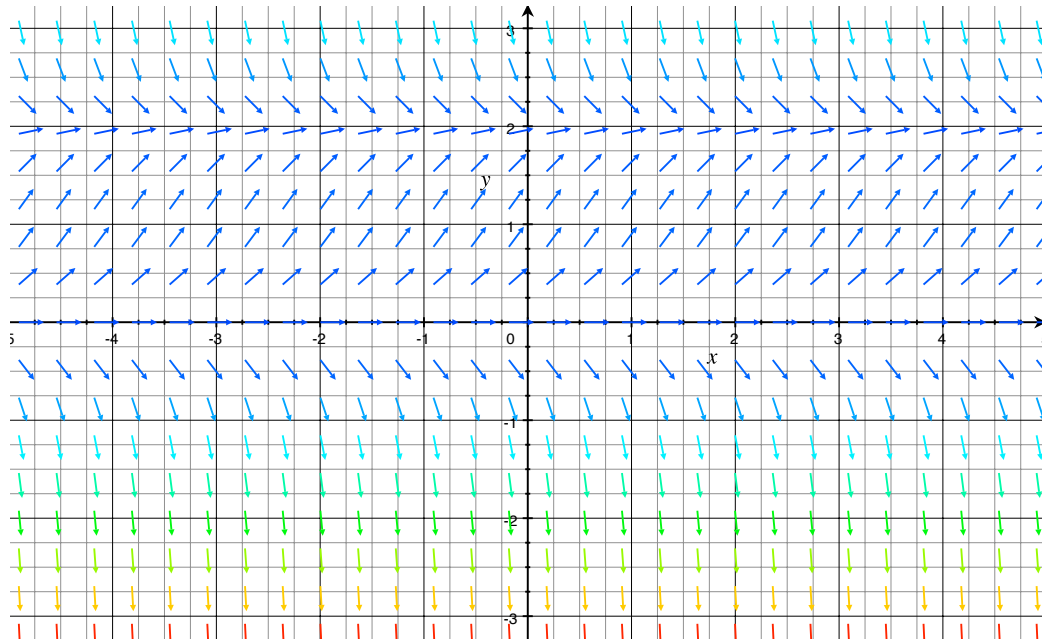
$$\frac{dP}{dt} = rP \left(1 - \frac{P}{K} \right).$$

In this situation $P(t)$ is population as a function of time, r is the growth rate of the population, and K is the carrying capacity of the environment.

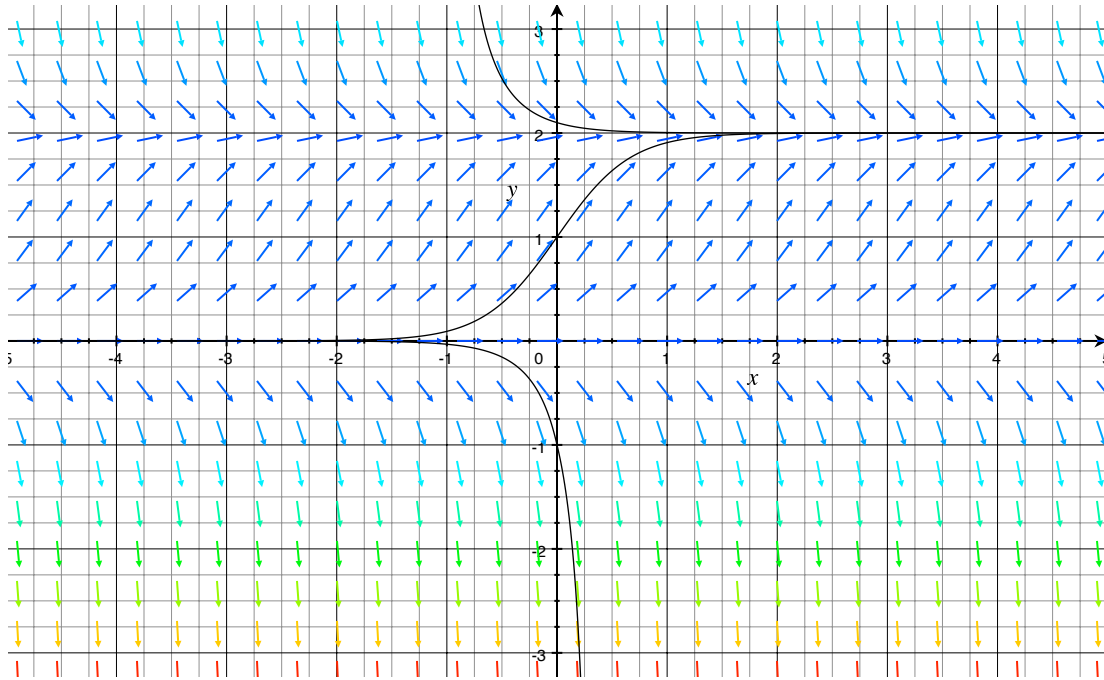
Example 12.2.2. Practice reading the differential equation: change in population is proportional to the rate parameter times the population times one minus the ratio of population to carrying capacity. What are the equilibrium values for the differential equation? Draw a phase line and characterize the behavior of the population with various *initial conditions*. **Review: Remember initial value problems were introduced in September.**

To carry out heuristic analysis of these differential equations, another great graphical tool is the *slope field* or *direction field*. Slope fields can be generated by computer or drawn by hand. To draw a slope field for a single differential equation of the form $\frac{dx}{dt} = f(x, t)$, draw the t - and x -axes and then for every point (t, x) draw a small tick mark with slope $f(x, t)$.

Here is an example: the slope field for $\frac{dx}{dt} = 3x \left(1 - \frac{x}{2}\right)$ is



To use a slope field for qualitative analysis of solutions to a differential equation, start at an initial value (t_0, x_0) of your choice and simply follow the arrows! This traces a solution curve, and this curve is the graph of a solution to the differential equation.



Recall that we discussed Euler's method for numerically solving these differential equations. Euler's method puts together many short-term predictions to make a long-term prediction. How does that work again? Review!

Example 12.2.3. Use Euler's method or the method of long-term predictions to estimate $x(2)$ if you know $\frac{dx}{dt} = 3x(1 - x/2)$ and you start at $x(0) = 1$. Use two steps.

Example 12.2.4 (Challenge). Solve $\frac{dx}{dt} = 3x(1 - x/2)$ analytically.

12.3 So I can model a caribou population: what about money?

A good qualitative understanding of differential equations is essential to progressing toward the Ito calculus and Black-Scholes equation. Let's take a look at the Black-Scholes equation for a moment:

$$\frac{\partial C}{\partial t} + rS \frac{\partial C}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC.$$

Here $C(S, t)$ is the price, when the stock price is S and time is t , of a European call option struck at price K with an expiration date of T . The parameter σ is the volatility of the stock's returns. This equation is a *partial differential equation* (PDE) as opposed to the *ordinary differential equations* (ODEs) we're considering in this document. But the principles of "reading" a differential equation remain the same! A very first analysis can go as follows:

- If we considered a simplified $C(S, t)$ under some extraordinary condition in which $C(S, t)$ did not vary according to S at all (imagine simply holding S constant but letting time run) we'd end up with the equation

$$\frac{\partial C}{\partial t} = rC.$$

Look familiar? Yes, it's just the exponential growth from the first page of this document!

- If we let t stay constant and just let S vary, as a thought experiment, we'd get

$$rS \frac{\partial C}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC.$$

This is a *second-degree* differential equation, as it has those second derivatives. Ordinary autonomous second-degree differential equations can be split into two ordinary first-degree differential equations; check out this concept in the next chapter.

Understanding the uses and pitfalls of the Black-Scholes equation is a significant endeavor in financial math! Having an acquaintance with ordinary differential equations gives a mental context for dealing with Black-Scholes.

12.4 Systems of Differential Equations

Systems of differential equations: use everything you know about linear algebra and transfer it to the differential equation setting! Let's start with something simple to get an idea of why this might work.

Example 12.4.1. Consider the system of equations

$$\begin{aligned}\frac{dx}{dt} &= 3x \\ \frac{dy}{dt} &= -2y.\end{aligned}$$

Alone, you'd be happy to solve either of these equations. You would get $x(t) = P_x e^{3t}$ and $y(t) = P_y e^{-2t}$. This is still perfectly reasonable.

We can rewrite the system of differential equations using the language of matrices:

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Then the solution is

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} P_x e^{3t} \\ P_y e^{-2t} \end{pmatrix} = P_x e^{3t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + P_y e^{-2t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Solutions to this differential equation, then, are curves in xy -space: they depend on t (they are *parametrized* by t).

To draw a slope field or *phase plane* for a two-dimensional system of equations, take a point (x, y) in the plane and draw a tiny vector in the direction of

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix}.$$

***Notice something: no equation I write today has a t appearing explicitly on the right-hand side. Instead, they are of the form $f(x, y)$. These are called *autonomous* differential equations. It is possible in math to deal with non-autonomous systems

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} f(x, y, t) \\ g(x, y, t) \end{pmatrix}$$

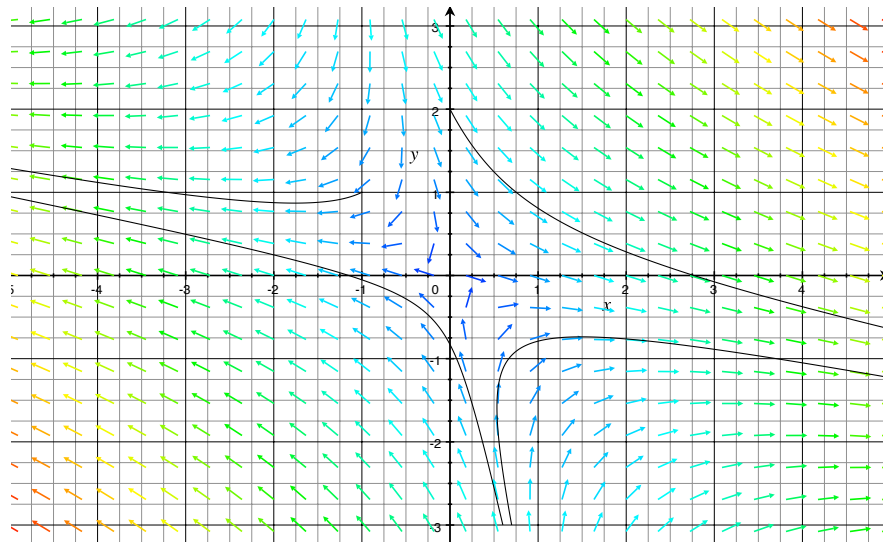
but that is definitely outside the scope of this class!!

Example 12.4.2. Systems of differential equations can be linear (the functions describing the derivatives are linear and of form $ax + by$, for a and b real

numbers at least for today) or non-linear. Here's an example of a linear system:

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

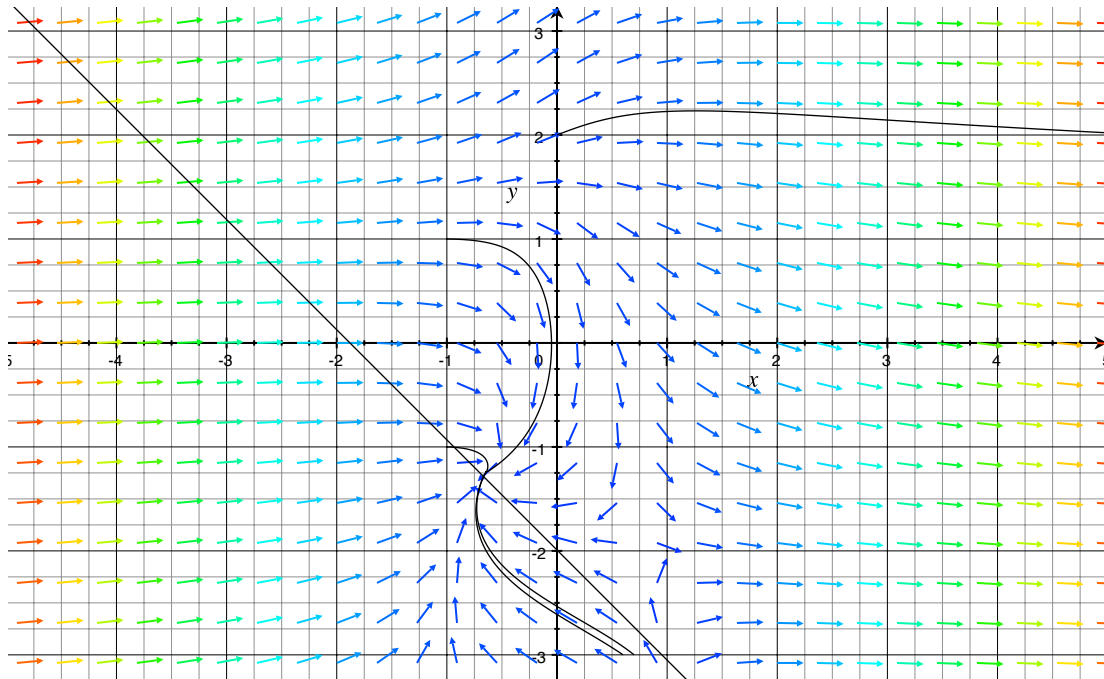
Here is a phase plane with some solutions included (for positive t only):



By contrast, here's a nonlinear system:

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 3x^2 + y \\ -x - 2\cos(y) \end{pmatrix}.$$

Notice that I can't write it using the matrix format! Here's the slope field, with some solutions sketched. *Which solution is wrong? Why is it wrong? Where are common sources of error when using Euler's method?*



12.5 Equilibria

We can find “equilibrium points” (we now call them *fixed points* instead) for these differential equations just as we found equilibrium solutions for single differential equations. Just solve for the points where $\frac{dx}{dt}$ and $\frac{dy}{dt}$ are both zero! For a linear system of differential equations, the origin $(0,0)$ will always be an/the only fixed point. (Why?) For nonlinear systems you’ve got to use whatever techniques you can to find these solutions.

Example 12.5.1. For

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 3x^2 + y \\ -x - 2\cos(y) \end{pmatrix},$$

solve

$$3x^2 + y = 0$$

and get the curve

$$y = -3x^2.$$

(A whole curve...! Can you see this on the slope field? What would it mean?)
Then solve

$$-x - 2 \cos(y) = 0$$

to get

$$x = -2 \cos(y).$$

Put these together: if both are true,

$$y = -12 \cos(y).$$

This has many solutions, but $y \approx -1.44969$ is one of them and fits into our picture. If that is true, then $x \approx 0.695147..$ and that gives us our fixed point. Messy! This is just one of many fixed points since $\cos(y)$ is periodic.

Look at this fixed point in the picture: how do the solutions relate to the fixed points? How do they behave together? It looks like solutions “tend toward” the fixed point. When solutions tend toward a fixed point, the fixed point is called a *sink*. When solutions tend away from a fixed point, the fixed point is called a *source*. When nearby solutions zoom past a fixed point, careening toward it and then rushing away, the fixed point is called a *saddle*.¹

Let's deal with a simpler situation:

Example 12.5.2. Consider

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 3x - y^2 \\ \sin(y) - x \end{pmatrix}$$

This has two equilibria. Solve $3x - y^2 = 0$ to get $3x = y^2$, and $\sin(y) - x = 0$ to get $x = \sin(y)$. Then $\sin(y) = y^2/3$, so $3 \sin(y) = y^2$. This has the solutions of $y = 0$ and $y \approx 1.72213$. That means $(0, 0)$ is an equilibrium point and $(.98857\dots, 1.72213\dots)$ is also an equilibrium point.

¹How does this remind you of minima, maxima, and saddles when we look at optimizing functions like $z = f(x, y)$?

12.6 Straight-line solutions

Looking at phase planes of systems of linear first-order differential equations, you can see that there are often *straight-line solutions* – solution curves that follow a straight line into or out of the origin. This is no accident. Turns out they run along eigenvectors. In fact, the solutions to a system of two linear DEs

$$\frac{d}{dt}\vec{x} = A\vec{x}$$

where A has real eigenvalues λ_1, λ_2 with corresponding eigenvectors \vec{w}_1, \vec{w}_2 are completely classified by

$$\vec{x}(t) = c_1 e^{\lambda_1 t} \vec{w}_1 + c_2 e^{\lambda_2 t} \vec{w}_2.$$

(Actually, we can do the same if the eigenvectors are complex, but we have to pick off the real parts of the solution by using the identity

$$e^{it} = \cos(t) + i \sin(t)$$

and combining this with the complex eigenvectors.)

Example 12.6.1. Consider the system

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} 2 & 7 \\ -1 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

The matrix A has eigenvalues $\lambda_1 = -5, \lambda_2 = 1$:

$$(2 - \lambda)(-6 - \lambda) + 7 = 0 \tag{12.1}$$

$$\lambda^2 + 4\lambda - 5 = 0 \tag{12.2}$$

$$(\lambda + 5)(\lambda - 1) = 0. \tag{12.3}$$

It's got eigenvectors

$$w_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

and

$$w_2 = \begin{pmatrix} -7 \\ 1 \end{pmatrix}$$

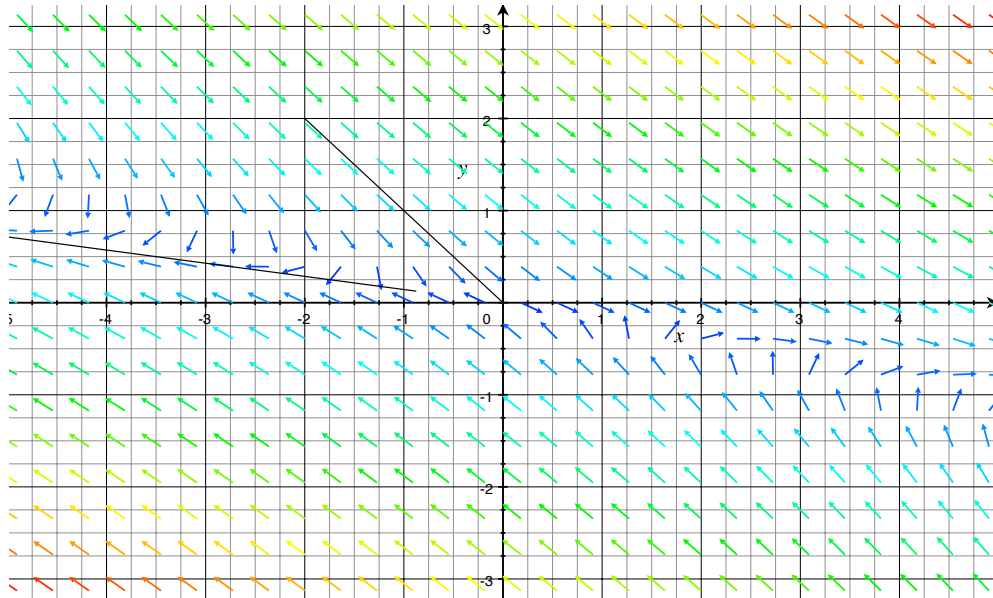
(check yourself!). So the straight-line solutions are

$$c_1 e^{-5t} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

and

$$c_2 e^{1t} \begin{pmatrix} -7 \\ 1 \end{pmatrix}.$$

Look at the phase plane:



Here I did not draw the whole straight line solution in either situation: I drew only the solution starting (with $t = 0$) at an initial value and continuing for positive t . I did this because then you can see that the solutions associated with λ_1 go “in” toward the origin (as e^{-5t} decreases when t increases) and the solutions associated with λ_2 go “out” from the origin (as e^t increases when t increases).

12.7 Back to Black-Scholes for a minute

In the first chapter I mentioned that second-degree differential equations can often be split into two first-degree differential equations. The most classic example of this is the motion of a spring with a mass on the end, possibly with some external forcing. Back to physics for a moment!

The equation for the displacement $u(t)$ of a spring in this situation is

$$m \frac{d^2 u}{dt^2} + \gamma \frac{du}{dt} + ku = F(t),$$

where m is the mass at the end of the spring, γ is some damping coefficient (is your spring moving through air, oil, peanut butter?), k is the spring constant (\approx stiffness), and $F(t)$ describes the external forcing.² Use a very simple trick to rewrite this second-order equation as two first-order equations:

$$\frac{du}{dt} = v.$$

Look: I made up a variable name for $\frac{du}{dt}$! Now we get

$$\begin{pmatrix} \frac{du}{dt} \\ \frac{dv}{dt} \end{pmatrix} = \begin{pmatrix} v \\ \frac{1}{m} (F(t) - \gamma v - ku) \end{pmatrix}.$$

This is not quite linear, but there are standard mathematical tools to deal with this. Moreover, if $F(t) = 0$ (in other words, there's no external forcing) you've got all the information you need to solve

$$\begin{pmatrix} \frac{du}{dt} \\ \frac{dv}{dt} \end{pmatrix} = \begin{pmatrix} v \\ \frac{1}{m} (-\gamma v - ku) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{-k}{m} & \frac{-\gamma}{m} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}.$$

²Compare it to Black-Scholes:

$$\frac{\partial C}{\partial t} + rS \frac{\partial C}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC$$

There are some similarities; the big difference is that Black-Scholes has derivatives with respect to time *and* stock price.

