MIKE CUMMINGS

MATH 2LA3 COURSE NOTES

Contents

Introduction

Welcome to Math 2LA3! This is a course about using linear algebra to solve real-world problems. This document contains the notes for the course's lectures.

The main reference for this course is the textbook *Linear Algebra and its Applications* (6^{th} ed.) by Lay, Lay, and McDonald. Most of the material in these notes has been derived from this textbook, so if you want more detail, check there. The course schedule lists the textbook section(s) associated to each topic.

It is important to remember that reading mathematics is not a passive activity; it requires checking all of the details and constantly checking with yourself that you understand each little point. There is a common saying for reading mathematics: *Don't read it; fight it!*[1] Don't believe the author, you should be convincing yourself that every detail is correct and, when you have questions, ask the instructor!

[1] Attribution: Paul Halmos in his autobiography, *I Want to be a Mathematician, An Automathography.*

This document was last updated on August 6, 2024

A
Linear Programming

A.1 Introduction

Here's a problem, familiar from the world of calculus.

Find the maximal value of the function $f(x)$ on the interval $x \in [0,1]$.

This is an example of an optimization problem, where we try to either maximize or minimize a function on some domain. Here's another example of an optimization problem.

Example A.1. Assume that your favourite pastry shop sells two main products: doughnuts for \$1.25 and muffins for \$1.50.

- Each doughnut requires 30g of flour, 10g of sugar, and 20g of oil;
- Each muffin requires 40g of flour, 20g of sugar, and 10g of oil.

If the store has 6kg of flour, 2.8kg of sugar, and 3kg of oil, how many of each product should they make to maximize revenue?

Denote by D the number of doughnuts and M the number of muffins. Revenue is given by $R(D,M) = 1.25D + 1.50M$. We also have to bear in mind the amount of ingredients we have,

$$30D + 40M \leq 6000, \qquad \text{(flour)}$$
$$10D + 20M \leq 2800, \qquad \text{(sugar)}$$
$$20D + 10M \leq 3000. \qquad \text{(oil)}$$

Rewriting using matrix multiplication notation, we want to maximize the quantity

$$\begin{bmatrix} 1.25 & 1.50 \end{bmatrix} \begin{bmatrix} D \\ M \end{bmatrix} = 1.25D + 1.50M$$

Make sure you believe this matrix multiplication. If you don't, check your notes from your last linear algebra class!

while also satisfying the constraints

$$\begin{bmatrix} 30 & 40 \\ 10 & 20 \\ 20 & 10 \end{bmatrix} \begin{bmatrix} D \\ M \end{bmatrix} \leq \begin{bmatrix} 6000 \\ 2800 \\ 3000 \end{bmatrix},$$

as well as the implicit constraints

$$D \geq 0 \quad \text{and} \quad M \geq 0.$$

This is an example of a *linear programming problem*, so named because both the constraints and the function we wish to maximize are *linear* functions. We formalize this in the following definition.

Compare this definition with the previous example. Everything in this definition should have a counterpart in the example, and vice versa.

For instance, $\vec{c}^T\vec{x}$ is the generalized version of the revenue function $R(D, M)$ from the previous example.

Definition A.2. A **canonical linear programming problem** asks the following: find the vector $\vec{x}$ to

$$\begin{aligned} &\texttt{maximize} && \vec{c}^T\vec{x} \\ &\texttt{subject to} && \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq 0, \end{cases} \end{aligned}$$

where $\vec{b} \in \mathbb{R}^m$ and $\vec{c} \in \mathbb{R}^n$ are fixed (column) vectors and A is an $m \times n$ matrix with entries in $\mathbb{R}$. We call $\vec{c}^T\vec{x}$ the **objective function**.

In this definition, $\vec{c}^T\vec{x}$ is the linear function we wish to maximize, which is exactly the function $R(D, M)$ from the example.[1] Similarly, the condition $A\vec{x} \leq \vec{b}$ is exactly the system of constraints from the example for the flour, sugar, and oil.

[1] We have to take the transpose of $\vec{c}$ in order for the matrix multiplication $\vec{c}^T\vec{x}$ to make sense. If you're not convinced, go back to the example and try to compute $\vec{c}\vec{x}$ where $\vec{c} = \begin{bmatrix} 1.25 \\ 1.50 \end{bmatrix}$ and $\vec{x} = \begin{bmatrix} D \\ M \end{bmatrix}$.

Definition A.3. Consider the canonical linear programming problem:

$$\begin{aligned} &\texttt{maximize} && \vec{c}^T\vec{x} \\ &\texttt{subject to} && \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq 0. \end{cases} \end{aligned}$$

The **feasible region** of this problem is the set of vectors $\vec{x}$ that satisfy the constraints $A\vec{x} \leq \vec{b}$ and $\vec{x} \geq 0$. We denote by $\mathscr{F}$ the feasible region:

$$\mathscr{F} = \{\vec{x} \in \mathbb{R}^n \mid A\vec{x} \leq \vec{b} \text{ and } \vec{x} \geq 0\}.$$

Remember that when we write $\vec{x} \geq 0$, we are requiring that each component x_i of the vector $\vec{x}$ satisfies $x_i \geq 0$.

A vector $\vec{x}_*$ in $\mathscr{F}$ is called an **optimal solution** if it solves the linear programming problem. That is, it is optimal if

$$\vec{c}^T\vec{x}_* = \max_{\vec{x} \in \mathscr{F}} \left(\vec{c}^T\vec{x}\right).$$

Before discussing how to find solutions to linear programming problems, we record a fact that guarantees the existence of a solution in some cases. What are the cases where we have a solution? Well, the following two examples give us an idea of cases that *don't* have a solution.

Example A.4. Consider the following linear programming problem:

$$\begin{aligned} &\texttt{maximize} && 5x \\ &\texttt{subject to} && \begin{cases} x \leq 3 \\ -x \leq -4 \\ x \geq 0. \end{cases} \end{aligned}$$

Elements x of the feasible set must satisfy both $x \leq 3$ and also $x \geq 4$, which is impossible. Since the feasible region is empty, there is no solution.

Example A.5. Consider the following linear programming problem:

$$\begin{aligned} &\texttt{maximize} && 5x \\ &\texttt{subject to} && \begin{cases} -x \leq -4 \\ x \geq 0. \end{cases} \end{aligned}$$

The constraints only require that $x \geq 4$, so $5x$ can get arbitrarily large! So we have no hope of picking a single x that maximizes $5x$.

An **unbounded** linear programming problem is one where the objective function is unbounded on the feasible region, like the previous example. We say the problem is **bounded** otherwise.

Theorem A.6. *Consider a bounded linear programming problem with feasible set $\mathscr{F}$. If $\mathscr{F}$ is nonempty, then there is at least one solution to the problem.*

Moreover, at least one of the optimal solutions occurs at a vertex of the feasible region.

This theorem is extremely useful! It says that when solutions exist, all we need to do is to find the *vertices*. Because our constraints are a system of linear equations, this implies that the feasible region will have *corners* and these corners are exactly what we mean by an extreme point.

A vertex is also called an **extreme point** or a **corner**.

If you're not yet sure what we mean by a *vertex*, we will draw pictures in a minute that will clarify.

A.2 Geometric method

When the domain of linear programming problem is in low-dimension, such as $\mathbb{R}^2$, we can directly apply the result of Theorem A.6. That is, we can sometimes explicitly compute the feasible region to find the vertices.

We proceed with an example. Consider the line $x + 2y = 4$. This is exactly the line $y = -\frac{1}{2}x + 2$ sketched in Figure A.1.

Replacing the equality with an inequality yields $x + 2y \leq 4$. Performing the same rearrangement, this is equivalent to $y \leq -\frac{1}{2}x + 2$. We interpret this as the region in the xy-plane underneath the line $-\frac{1}{2}x + 2$. This region is also sketched in Figure A.1, where we additionally impose the constraint that $x \geq 0$ and $y \geq 0$.

The vertices corresponding to the region $x + 2y \leq 4$, $x \geq 0$, and $y \geq 0$ are the three vertices of the triangle. Theorem A.6 then guarantees that, if this were the feasible region of a linear programming problem, then a solution to this problem would occur at one of these vertices.

Example A.7. Let us continue Example A.1. Recall that our linear

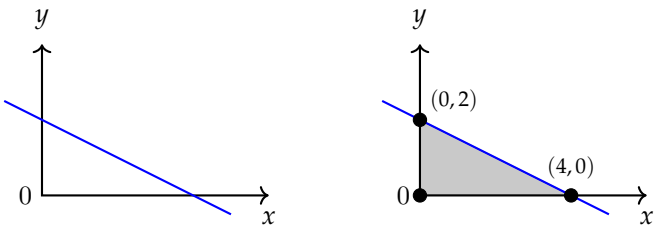


Figure A.1: The curve $x + 2y = 4$ (left) and the region $x + 2y \leq 4$ with $x \geq 0$ and $y \geq 0$ (right).

programming problems is:

$$\texttt{maximize} \quad R(D, M) = \begin{bmatrix} 1.25 & 1.50 \end{bmatrix} \begin{bmatrix} D \\ M \end{bmatrix}$$

$$\texttt{subject to} \quad \begin{bmatrix} 30 & 40 \\ 10 & 20 \\ 20 & 10 \end{bmatrix} \begin{bmatrix} D \\ M \end{bmatrix} \leq \begin{bmatrix} 6000 \\ 2800 \\ 3000 \end{bmatrix},$$

$$\texttt{and} \qquad D \geq 0, \quad M \geq 0$$

where D and M are the number of doughnuts and muffins, respectively, to produce. Let us construct the feasible region for this linear programming problem. That is, we are required to intersect the regions $30D + 40M \leq 6000$, with $10D + 20M \leq 2800$, and with $20D + 10M \leq 3000$. We sketch the feasible region in the figure below. It is also available on desmos.[2]

[2] Thanks, Prof. Junkins!

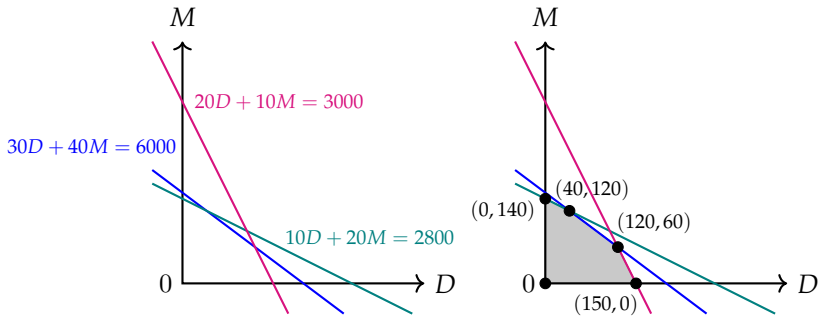


Figure A.2: Doughnuts and Muffins feasible region.

Theorem A.6 guarantees that a solution to the linear programming problem is one of the vertices, so we compute $R(D, M)$ for each vertex:

For the figure, we compute the vertices as points of intersections between pairs of constraints. For instance, the vertex $(40, 120)$ is the intersection of $10D + 20M = 2800$ with $30D + 40M = 6000$.

$$R(0, 0) = 0, \quad R(0, 140) = 210, \quad R(40, 120) = 230,$$

$$R(120, 60) = 240, \quad R(150, 0) = 187.50.$$

We conclude that the solution to the linear programming problem is $D = 120$ and $M = 60$. That is, the shop should produce 120 doughnuts and 60 muffins.

We summarize the geometric method as follows. Remember that the reason this method works is because of the result of Theorem A.6.

Geometric method.
Consider a linear programming problem in canonical form:

$$\begin{array}{ll} \texttt{maximize} & \vec{c}^{\,T}\vec{x} \\ \texttt{subject to} & \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq \vec{0}. \end{cases} \end{array}$$

1. Construct the feasible region $\mathscr{F}$, the set of $\vec{x}$ that satisfy both $A\vec{x} \leq \vec{b}$ and $\vec{x} \geq 0$.
2. Find the vertices of the feasible region.
3. For each vertex $\vec{v}$ of $\mathscr{F}$, compute $\vec{c}^{\,T}\vec{v}$.

   The solutions to the linear programming problem are the $\vec{v}$ that yield the largest value of $\vec{c}^{\,T}\vec{v}$.

Exercise 1. Consider the following linear programming problem.

$$\begin{array}{ll} \texttt{minimize} & 5x + 3y \\ \texttt{subject to} & \begin{cases} 2x + 5y \geq 10 \\ 3x + y \geq 6 \\ x + 7y \geq 7 \\ x \geq 0 \text{ and } y \geq 0. \end{cases} \end{array}$$

Notice that we are being asked to *minimize* our objective function rather than *maximize*. Also, notice that the first three inequalities are flipped and are in the wrong direction.

(a) Rewrite the linear programming problem as a *canonical* linear programming problem (as in Definition A.2).

Hints: (i) How can you flip an inequality? (ii) What did we do in Example A.4?

(b) Use the geometric method to solve the canonical linear programming problem from part (a).

A.3 Optional review: Solving linear systems using matrices

Unfortunately, the geometric method cannot always be used to solve a linear programming problem. We are only able to reasonably compute the feasible region in low dimension (in $\mathbb{R}^2$ and *maybe* $\mathbb{R}^3$), so in higher dimensions, we need a new tool. This new tool, called the simplex

method, will be discussed in the next section but we first need to recall some terminology and technology from first-year linear algebra.

One motivation for using matrices is that solving a large system of linear equations is quite unwieldily. For instance, we will solve the following system in two ways: (i) directly, and (ii) using matrices.

Example A.8. Find the solution(s) (x, y, z) to the system

$$\begin{cases} 3x + 5y - 4z = 7 \\ -3x - 2y + 4z = -1 \\ 6x + y - 8z = -4. \end{cases}$$

Method 1 (direct). Rearrange the third equation to solve for y,

$$y = -6x + 8z - 4.$$

Now replace y in each of the first two equations with this formula. This yields the system

$$\begin{cases} 3x + 5(-6x + 8z - 4) - 4z = 7 \\ -3x - 2(-6x + 8z - 4) + 4z = -1 \\ y = -6x + 8z - 4. \end{cases}$$

This simplifies to the following:

$$\begin{cases} -27x + 36z = 27 \\ 9x - 12z = -9 \\ y = -6x + 8z - 4 \end{cases} \iff \begin{cases} -3x + 4z = 3 \\ 3x - 4z = -3 \\ y = -6x + 8z - 4, \end{cases}$$

where we have divided by 3 each of the first two equations. It is now clear the first two equations are the same; they say that we must have that

$$x = \frac{4}{3}z - 1$$

and, plugging in to the third equation, this yields:

$$\begin{cases} x = \frac{4}{3}z - 1 \\ y = -6\left(\frac{4}{3}z - 1\right) + 8z - 4 \end{cases} \iff \begin{cases} x = \frac{4}{3}z - 1 \\ y = 2. \end{cases}$$

So the solutions to the system are of the form

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{4}{3}t - 1 \\ 2 \\ t \end{bmatrix} = \begin{bmatrix} 4/3 \\ 0 \\ 1 \end{bmatrix} t + \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix},$$

where t is a **free variable**. By that, we mean that we get a solution for any value of $t \in \mathbb{R}$.

Method 2 (Gauss-Jordan elimination). Start by writing the system as the augmented matrix

$$\left[\begin{array}{ccc|c} 3 & 5 & -4 & 7 \\ -3 & -2 & 4 & -1 \\ 6 & 1 & -8 & -4 \end{array}\right]$$

and we perform the following elementary row operations:

$$\left[\begin{array}{ccc|c} 3 & 5 & -4 & 7 \\ -3 & -2 & 4 & -1 \\ 6 & 1 & -8 & -4 \end{array}\right] \begin{matrix} R1-5R3 \\ R2+2R3 \\ \sim \end{matrix} \left[\begin{array}{ccc|c} -27 & 0 & 36 & 27 \\ 9 & 0 & -12 & -9 \\ 6 & 1 & -8 & -4 \end{array}\right] \begin{matrix} \frac{-1}{9}R1 \\ \frac{1}{3}R2 \\ \sim \end{matrix} \left[\begin{array}{ccc|c} 3 & 0 & -4 & -3 \\ 3 & 0 & -4 & -3 \\ 6 & 1 & -8 & -4 \end{array}\right] \begin{matrix} \\ R2-R1 \\ R3-2R1 \end{matrix}$$

$$\sim \left[\begin{array}{ccc|c} 3 & 0 & -4 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \end{array}\right] \begin{matrix} \frac{1}{3}R1 \\ R2\leftrightarrow R3 \\ \sim \end{matrix} \left[\begin{array}{ccc|c} 1 & 0 & -4/3 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

The row of zeros tells us that we have a free variable and, as a result, have infinitely-many solutions. We can read off the solutions satisfy

$$\begin{cases} x - \frac{4}{3}z = -1 \\ y = 2 \end{cases}$$

or equivalently,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} + \begin{bmatrix} 4/3 \\ 0 \\ 1 \end{bmatrix} t,$$

for any $t \in \mathbb{R}$. This computation agrees with Method 1 but was much cleaner. It is straightforward to see that Method 2 will be easier to perform for arbitrarily large systems of linear equations, compared to Method 1.

Exercise 2. Without repeating the computations, how many solutions would we have to the system in the previous example if:

(a) the second equation is replaced by $-3x - 2y + 4z = 0$.
(b) the second equation is replaced by $-3x - 2z + 3z = -1$.

This exercise can be done without repeating the computations and without plugging the system into an online calculator. Think about what it means for a system to have infinitely-many solutions vs exactly one solution vs no solutions.

We now review some terminology from first-year linear algebra relating Gauss-Jordan elimination (AKA row reduction) of matrices.

A matrix is said to be in **row echelon form (REF)** if:

(i) rows of all 0's are at the bottom; below all nonzero rows,
(ii) every leading entry of a row is in a column to the right of the leading entry of the row above it,
(iii) all entries in a column below a leading entry are zero.

A matrix is in **reduced row echelon form (RREF)** if it is in row echelon form and, moreover, satisfies:

(iv) the first nonzero entry in each row is a 1 (called a **leading 1**),
(v) each leading 1 is the only nonzero entry in its column.

Exercise 3. Give an example of a matrix that is in reduced row echelon form. Give an example of a matrix that is in row echelon form but not reduced row echelon form. Lastly, give an example of a matrix that is not in row echelon form.

Recall that we can read off the number of solutions to a system of linear equations from the REF of an augmented matrix. If the REF has a row of all zeros, then the system will have infinitely-many solutions. On the other hand, if the REF has a row of all zeros except for a nonzero entry in the last column, then there will be no solution. Otherwise, there is a unique solution.

A **pivot** of a matrix is an entry that corresponds to a leading 1 in the RREF of that matrix. The column containing a pivot is called a **pivot column**.

Pivots will play an important role in the simplex method that we will discuss in the next section.

Gauss-Jordan elimination is the process of applying elementary row operations to a matrix to reduce it to (reduced) row echelon form. **Elementary row operations** are:

- multiplying a row by a nonzero scalar,
- adding a multiple of a row to another row,
- reordering rows.

A.4 Simplex method

The simplex method is an algorithm that solves linear programming problems. Although it is more complicated than the geometric approach we discussed earlier, its benefit is that it does not require us to be able to draw the feasible region, something that is impractical in higher dimensions.

The simplex method for canonical linear programming problems

We first outline the simplex method in an example before stating the general algorithm. The key to this example, and the simplex method more generally, is translating the linear programming problem into one for which we can exploit the power of matrices we discussed in the previous section.

Example A.9. Consider the following linear programming problem:

$$\begin{aligned} &\texttt{maximize} && 2x + y \\ &\texttt{subject to} && \begin{cases} 3x + 2y \leq 15 \\ -x + 2y \leq 9 \\ x, y \geq 0. \end{cases} \end{aligned}$$

The inequality $3x + 2y \leq 15$ is equivalent to requiring that both $3x + 2y + s = 15$ and $s \geq 0$. Indeed, if $s \geq 0$, then $3x + 2y \leq 3x + 2y + s = 15$. So we will rewrite the linear programming problem as:

$$\begin{aligned} &\texttt{maximize} && 2x+y \\ &\texttt{subject to} && \begin{cases} 3x+2y+s=15 \\ -x+2y+t=9 \\ s,t,x,y \geq 0. \end{cases} \end{aligned}$$

The variables s and t are called **slack variables**, variables introduced to rewrite an inequality as an equality.

A **basic solution** is when $x=0$ and $y=0$, so $2x+y=0$. It is clear that this is not an optimal solution, so we must do some more work. Let $M=2x+y$, so maximizing the objective function is the same as maximizing M. We then have the following three equations:

$$\begin{cases} 3x+2y+s=15 \\ -x+2y+t=9 \\ -2x-y+M=0 \end{cases}$$

which we can write as the following augmented matrix,

$$\left[\begin{array}{ccccc|c} 3 & 2 & 1 & 0 & 0 & 15 \\ -1 & 2 & 0 & 1 & 0 & 9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right],$$

which we call the corresponding **simplex tableau**. The first column corresponds to x, the second to y, and so on for s, t, and M, respectively.

Our **basic solution** corresponds to setting equal to zero the variables corresponding to columns that are not from the identity matrix. That is, the third, fourth, and fifth columns are zero with one nonzero entry, so the other columns (the first and second) correspond to variables that we set to zero in the basic solution.

Equivalently, the variables that we set to zero for the basic solution are exactly those corresponding to nonzero entries in the bottom row.

Now, our goal is to increase M. Since $M=2x+y$, the largest increase in M will come from increasing x. This can be seen from the augmented matrix by identifying entry in the last row with the most negative entry.

To increase x, we cannot set it to zero in the basic solution. That means, we must perform row operations so that the column corresponding to x looks like a column from an identity matrix. We do so by pivoting the first column as follows:

$$\left[\begin{array}{ccccc|c} 3 & 2 & 1 & 0 & 0 & 15 \\ -1 & 2 & 0 & 1 & 0 & 9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right] \overset{\frac{1}{3}R1}{\sim} \left[\begin{array}{ccccc|c} 1 & 2/3 & 1/3 & 0 & 0 & 5 \\ -1 & 2 & 0 & 1 & 0 & 9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right] \begin{matrix} \\ R2+R1 \\ R3+2R1 \end{matrix} \sim \left[\begin{array}{ccccc|c} 1 & 2/3 & 1/3 & 0 & 0 & 5 \\ 0 & 8/3 & 1/3 & 1 & 0 & 14 \\ \hline 0 & 1/3 & 2/3 & 0 & 1 & 10 \end{array}\right].$$

The non-identity-matrix columns are the second and third columns, so our new basic solution corresponds to $y=0$ and $s=0$. This then says that $x=5$ and $M=10$.

Geometrically, we have moved from the basic solution of $(x,y) = (0,0)$ to our new basic solution of $(x,y) = (5,0)$. It turns out that if a basic solution lies inside the feasible region, then it occurs at a vertex of the feasible region.

Because, from the third row of the augmented matrix, we have that $M = 10 - \frac{1}{3}y - \frac{2}{3}s$, we cannot increase M anymore. We conclude that subject to the given constrains, the maximum of $2x + y$ is 10, which occurs when $(x,y) = (5,0)$.

We must have found a maximum because we cannot further increase our objective function.

Exercise 4. Use the geometric method to solve the previous linear programming problem. Verify that your two answers are the same.

There is a subtlety that we will discuss next about this method. But first, it should be clear that this approach, which can be completed by working only with matrices, will scale easily to higher-dimensional problems, unlike the geometric method.

Here is the subtlety: In Example A.9, we needed to pivot the first column and chose to use the first row to do so. However, as the following computation shows, we cannot use the second row for the pivot.

$$\left[\begin{array}{ccccc|c} 3 & 2 & 1 & 0 & 0 & 15 \\ -1 & 2 & 0 & 1 & 0 & 9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right] \begin{matrix} \\ -R2 \\ \\ \end{matrix} \sim \left[\begin{array}{ccccc|c} 3 & 2 & 1 & 0 & 0 & 15 \\ 1 & -2 & 0 & -1 & 0 & -9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right] \begin{matrix} R1-3R2 \\ \\ R3+2R2 \end{matrix} \sim \left[\begin{array}{ccccc|c} 0 & 8 & 1 & 3 & 0 & 42 \\ 1 & -2 & 0 & -1 & 0 & -9 \\ \hline 0 & -5 & 0 & -2 & 1 & -18 \end{array}\right].$$

The corresponding basic solution has $y = 0$ and $t = 0$, so $x = -9$. But x must be nonnegative, so we have a contradiction.

Question A.10. Which row(s) can we pick for pivoting?

Pivoting Procedure.
Consider the following simplex tableau:

$$\left[\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,m} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} & b_2 \\ & & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} & b_n \\ \hline c_1 & c_2 & \cdots & c_n & k \end{array}\right]$$

The column we pivot is the column whose entry c_j in the final row is the most negative among the columns corresponding to the non-slack variables. Suppose this is the j-th column.
To find the corresponding row:

1. For each entry $a_{i,j}$ appearing in the same column as c_j, compute the ratio $b_i/a_{i,j}$.
2. Pivot at the entry $a_{r,j}$ for which $b_r/a_{r,j}$ is the smallest *nonnegative* ratio computed in step 1.

Following this procedure avoids the issue from the previous example; doing this will guarantee that we will never find a negative answer.

It may happen that there may be more than one row with the smallest nonnegative ratio $b_k/a_{r,k}$. This can lead to a phenomenon called **cycling**, where the simplex method fails because the tableau remains constant after each pivot. This usually does not arise when our linear programming problems arise from the real-world.

Example A.11. Let's revisit the simplex tableau from example A.9:

$$\left[\begin{array}{ccccc|c} 3 & 2 & 1 & 0 & 0 & 15 \\ -1 & 2 & 0 & 1 & 0 & 9 \\ \hline -2 & -1 & 0 & 0 & 1 & 0 \end{array}\right].$$

Among the given variables x and y, corresponding to columns 1 and 2, the most negative entry in the last row is the -2, so we want to pivot the first column.

Above this -2, there is only one positive entry, which occurs in the first row, so we pivot with respect to the entry in the first row, first column.

Here's another simplex tableau:

$$\left[\begin{array}{ccccccc|c} 2 & 3 & 4 & 1 & 0 & 0 & 0 & 60 \\ 3 & 1 & 5 & 0 & 1 & 0 & 0 & 46 \\ 1 & 2 & 1 & 0 & 0 & 1 & 0 & 50 \\ \hline -25 & -33 & -18 & 0 & 0 & 0 & 1 & 0 \end{array}\right].$$

From the tableau, we can see that there are 3 given variables (corresponding to the first three columns) and 3 slack variables (columns four through six). Among those in the last row, the entry in the second column is the most negative entry, so we wish to pivot with respect to one of the entries in the second column, indicated here:

$$\left[\begin{array}{ccccccc|c} 2 & \textcircled{3} & 4 & 1 & 0 & 0 & 0 & 60 \\ 3 & \textcircled{1} & 5 & 0 & 1 & 0 & 0 & 46 \\ 1 & \textcircled{2} & 1 & 0 & 0 & 1 & 0 & 50 \\ \hline -25 & -33 & -18 & 0 & 0 & 0 & 1 & 0 \end{array}\right].$$

We compute the ratios $b_i/a_{i,2}$ for $i = 1, 2, 3$:

$$60/3 = 20; \quad 46/1 = 46; \quad 50/2 = 25.$$

So we pivot with respect to the first row.

Exercise 5. Give an example of a canonical linear programming problem whose initial simplex tableau dictates that the first pivot should occur in the third row and third column.

We now have all the tools to state the simplex algorithm for canonical linear programming problems. We do that next, then work through a final example, and then move on to the next topic.

Simplex algorithm for canonical linear programming problems.
Consider a canonical linear programming problem,

$$\begin{array}{ll} \texttt{maximize} & \vec{c}^T\vec{x} \\ \texttt{subject to} & \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq \vec{0}, \end{cases} \end{array}$$

where $\vec{b}$ is a vector with strictly positive entries.

1. Add **slack variables** to change the inequality $A\vec{x} \leq \vec{b}$ into an equality. Write $M = \vec{c}^T\vec{x}$.
2. Set up the initial **simplex tableau** as the following augmented block matrix:

$$\left[\begin{array}{c:ccc|c} A & & I & & \vec{b} \\ \hline -\vec{c}^T & 0 & \cdots & 0 & 0 \end{array}\right],$$

   where I are the columns of some identity matrix.
3. Check the last row to see if the current state is optimal. If all of the entries are nonnegative, then the solution is optimal. Otherwise, let k be the column for which the most negative entry in the last row occurs in the k-th spot. Apply the Pivoting Procedure.
4. Repeat step 3 until the solution is optimal, which occurs when all the entries in the last row are nonnegative.

Example A.12. We will the following canonical linear programming problem using the simplex method:

$$\begin{array}{ll} \texttt{maximize} & 25x + 33y + 18z \\ \texttt{subject to} & \begin{cases} 2x + 3y + 4z \leq 60 \\ 3x + y + 5z \leq 46 \\ x + 2y + z \leq 50 \\ x, y, z \geq 0. \end{cases} \end{array}$$

We start by introducing slack variables, so the constraints become:

$$\begin{cases} 2x + 3y + 4z + s_1 = 60 \\ 3x + y + 5z + s_2 = 46 \\ x + 2y + z + s_3 = 50 \\ x, y, z, s_1, s_2, s_3 \geq 0, \end{cases}$$

and construct the simplex tableau:

$$\left[\begin{array}{cccccccc} 2 & 3 & 4 & 1 & 0 & 0 & 0 & 60 \\ 3 & 1 & 5 & 0 & 1 & 0 & 0 & 46 \\ 1 & 2 & 1 & 0 & 0 & 1 & 0 & 50 \\ \hline -25 & -33 & -18 & 0 & 0 & 0 & 1 & 0 \end{array}\right],$$

where the last row corresponds to the objective function $M = 25x + 33y + 18z$.

The basic solution at the moment is to take $x = y = z = 0$ which corresponds to $M = 0$.

The most negative entry in the last row is the -33, so we want to pivot an entry in the second column. This is exactly the second tableau from Example A.11, so we know we must pivot with respect to the first row and second column.

$$\left[\begin{array}{cccccccc} 2 & 3 & 4 & 1 & 0 & 0 & 0 & 60 \\ 3 & 1 & 5 & 0 & 1 & 0 & 0 & 46 \\ 1 & 2 & 1 & 0 & 0 & 1 & 0 & 50 \\ \hline -25 & -33 & -18 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} \frac{1}{3}R1 \\ \\ \\ \\ \end{array} \sim \left[\begin{array}{cccccccc} 2/3 & 1 & 4/3 & 1/3 & 0 & 0 & 0 & 20 \\ 3 & 1 & 5 & 0 & 1 & 0 & 0 & 46 \\ 1 & 2 & 1 & 0 & 0 & 1 & 0 & 50 \\ \hline -25 & -33 & -18 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} \\ R2-R1 \\ R3-2R1 \\ R4+33R1 \end{array}$$

$$\sim \left[\begin{array}{cccccccc} 2/3 & 1 & 4/3 & 1/3 & 0 & 0 & 0 & 20 \\ 7/3 & 0 & 11/3 & -1/3 & 1 & 0 & 0 & 26 \\ -1/3 & 0 & -5/3 & -2/3 & 0 & 1 & 0 & 10 \\ \hline -3 & 0 & 26 & 11 & 0 & 0 & 1 & 660 \end{array}\right]$$

Now the basic solution is $x = z = s_1 = 0$, which forces $y = 20$ and, in turn, $M = 660$. We repeat the pivoting procedure, now with the first column. The ratios are

$$\frac{20}{2/3} = 30; \quad \frac{26}{7/3} = \frac{78}{7} \approx 11.142,$$

Recall that we do not need to compute the ratio for the $-1/3$ in the first column, because it is negative.

so we pivot with respect to the first column, second row.

$$\left[\begin{array}{cccccccc} 2/3 & 1 & 4/3 & 1/3 & 0 & 0 & 0 & 20 \\ 7/3 & 0 & 11/3 & -1/3 & 1 & 0 & 0 & 26 \\ -1/3 & 0 & -5/3 & -2/3 & 0 & 1 & 0 & 10 \\ \hline -3 & 0 & 26 & 11 & 0 & 0 & 1 & 660 \end{array}\right] \begin{array}{l} \\ \frac{3}{7}R2 \\ \\ \\ \end{array} \sim \left[\begin{array}{cccccccc} 2/3 & 1 & 4/3 & 1/3 & 0 & 0 & 0 & 20 \\ 1 & 0 & 11/7 & -1/7 & 3/7 & 0 & 0 & 78/7 \\ -1/3 & 0 & -5/3 & -2/3 & 0 & 1 & 0 & 10 \\ \hline -3 & 0 & 26 & 11 & 0 & 0 & 1 & 660 \end{array}\right] \begin{array}{l} R1-\frac{2}{3}R2 \\ \\ R3+\frac{1}{3}R2 \\ R4+3R2 \end{array}$$

$$\sim \left[\begin{array}{cccccccc} 0 & 1 & 2/7 & 3/7 & -2/7 & 0 & 0 & 88/7 \\ 1 & 0 & 11/7 & -1/7 & 3/7 & 0 & 0 & 78/7 \\ 0 & 0 & -8/7 & -5/7 & 1/7 & 1 & 0 & 96/7 \\ \hline 0 & 0 & 215/7 & 74/5 & 9/7 & 0 & 1 & 4854/7 \end{array}\right].$$

Because all of the coefficients in the last row are nonnegative, we are done. The solution is $z = s_1 = s_2 = 0$ and hence $x = \frac{78}{7}$ and $y = \frac{88}{7}$. This corresponds to a maximum of $M = \frac{4854}{7}$.

A.5 The simplex method beyond the canonical setup

Consider a canonical linear programming problem:

$$\begin{array}{ll} \texttt{maximize} & \vec{c}^T\vec{x} \\ \texttt{subject to} & \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq \vec{0}. \end{cases} \end{array}$$

In the previous section, we required that $\vec{b} > \vec{0}$, that is, that every entry of $\vec{b}$ is positive.

Question A.13. What happens if some entries of $\vec{b}$ are zero or negative?

If an entry of $\vec{b}$ is zero, then **cycling** may occur during the simplex method, where the tableau does not change during pivoting, which means the process will never terminate! Thankfully, this usually does not occur for problems arising in the real world.

Unfortunately, problems from the real world can have yield negative entries in $\vec{b}$. The issue is that the initial basic solution will not lie within the feasible region, so the first step will be to pivot variables so that we do lie inside the feasible region. If we are able to do this, then we can still apply the simplex method.

Example A.14.

$$\begin{array}{ll} \texttt{minimize} & x+2y \\ \texttt{subject to} & \begin{cases} -x-y \leq -14 \\ x-y \leq 2 \\ x,y \geq 0 \end{cases} \end{array}$$

Notice that this *maximization* problem is equivalent to *minimizing* the function $x+2y$ subject to the constraints $x+y \geq 14$ and $x-y \leq 4$, with $x,y \geq 0$.

Minimizing $x+2y$ is equivalent to maximizing $-x-2y$. We then introduce slack variables as usual and let $M = -x-2y$, or equivalently, $x+2y+M=0$. We thus have the following simplex tableau:

$$\left[\begin{array}{ccccc|c} -1 & -1 & 1 & 0 & 0 & -14 \\ 1 & -1 & 0 & 1 & 0 & 2 \\ \hline 1 & 2 & 0 & 0 & 1 & 0 \end{array}\right].$$

However, our basic solution $x=y=0$ results in the slack variable for the third column taking value $s=-14$, but we required all variables (including slack variables) to be nonnegative! This can be solved by pivoting in the first row at an entry with a negative value. Let's pivot at the first row, second column.[3]

[3] You can try pivoting at row 1, column 1. It also works!

$$\left[\begin{array}{ccccc|c} -1 & -1 & 1 & 0 & 0 & -14 \\ 1 & -1 & 0 & 1 & 0 & 2 \\ \hline 1 & 2 & 0 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} -R1 \\ \\ \\ \end{array} \sim \left[\begin{array}{ccccc|c} 1 & 1 & -1 & 0 & 0 & 14 \\ 1 & -1 & 0 & 1 & 0 & 2 \\ \hline 1 & 2 & 0 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} \\ R2+R1 \\ R3-2R1 \end{array} \left[\begin{array}{ccccc|c} 1 & 1 & -1 & 0 & 0 & 14 \\ 2 & 0 & -1 & 1 & 0 & 16 \\ \hline -1 & 0 & 2 & 0 & 1 & -28 \end{array}\right]$$

The basic solution is now $x = s = 0$, where s is the slack variable in the third column. Also, $y = 14$ and the slack variable in the fourth column is $t = 16$, so the basic solution lies in the feasible region and we can begin the simplex method. We now want to pivot the first column, and we have no choice but to pivot the second row.

$$\left[\begin{array}{ccccc|c} 1 & 1 & -1 & 0 & 0 & 14 \\ 2 & 0 & -1 & 1 & 0 & 16 \\ \hline -1 & 0 & 2 & 0 & 1 & -28 \end{array}\right] \begin{matrix} \\ \frac{1}{2}R2 \\ \\ \end{matrix} \sim \left[\begin{array}{ccccc|c} 1 & 1 & -1 & 0 & 0 & 14 \\ 1 & 0 & -1/2 & 1/2 & 0 & 8 \\ \hline -1 & 0 & 2 & 0 & 1 & -28 \end{array}\right] \begin{matrix} R1-R2 \\ \\ R3+R2 \end{matrix} \sim \left[\begin{array}{ccccc|c} 0 & 1 & -1/2 & -1/2 & 0 & 6 \\ 1 & 0 & -1/2 & 1/2 & 0 & 8 \\ \hline 0 & 0 & 3/2 & 1/2 & 1 & -20 \end{array}\right]$$

This solution is now optimal; we have introduced both x and y to the solution. Both the slack variables are 0 which forces $x = 8$ and $y = 6$, corresponding to a maximum value of $-x - 2y$ of -20. Hence, the minimum of $x + 2y$ is 20.

Remark A.15. The upshot is that the usual simplex method can still be used beyond the canonical setup. However, before starting, we must always ensure that the basic solution lies in the feasible region. In general, to do this, we have to follow the following pivoting procedure.

Infeasible region pivoting procedure. Consider the following simplex tableaux.

$$X = \left[\begin{array}{c:ccc|c} & & & & \\ A & & I & & \vec{b} \\ & & & & \\ \hline -\vec{c}^{\,T} & 0 & \cdots & 0 & 0 \end{array}\right].$$

If one of the entries in $\vec{b}$ is negative, causing a basic solution to lie outside the feasible region, we can fix this by pivoting as follows.

1. Let i be the row that we wish to pivot. That is, the i-th entry of $\vec{b}$ is negative.
2. For each entry $x_{i,j}$ in the same row as b_i, compute the ratio $b_i / x_{i,j}$.
3. Pivot in row i, column k, where k is the index for which the ratio $b_k / x_{i,k}$ is the **smallest nonnegative** ratio computed in Step 2.

Remark A.16. We now have two pivoting procedures: the first to find optimal values and the second to move within the feasible region. In both, the pivot is determined by picking the smallest nonnegative ratio. However,

- *for the first pivoting procedure (to move closer to an optimal value),*

we first fix the column we wish to pivot, then use the ratios to pick the corresponding row

- *for the second pivoting procedure (to move within the feasible region)*, we first fix the row we wish to pivot, then use the ratios to pick the corresponding column.

Simplex algorithm beyond the canonical setup.
Consider a canonical linear programming problem,

$$\begin{aligned} &\texttt{maximize/minimize} && \vec{c}^T\vec{x} \\ &\texttt{subject to} && \begin{cases} A\vec{x} \leq \vec{b} \\ \vec{x} \geq \vec{0}, \end{cases} \end{aligned}$$

where $\vec{b}$ is now any vector.

1. If this is a minimization problem, then replace it with the corresponding maximization problem.
2. Add **slack variables** to change the inequality $A\vec{x} \leq \vec{b}$ into an equality. Write $M = \vec{c}^T\vec{x}$.
3. Set up the initial **simplex tableau** as before.
4. If the initial basic solution does not lie within the feasible region, pivot variables until it is. (Equivalently, pivot variables until the last column only contains nonnegative entries.) To do this, follow the **infeasible pivoting procedure**.
5. Now run the canonical simplex algorithm.

Exercise 6. Solve the exercises from Section 9.3 of the textbook until you are comfortable with the simplex method (both for the canonical problems, and minimization problems/problems with negative entries).

A.6 FAQs

Q. What are the differences between feasible solutions, optimal solutions, and basic solutions?
A. These are defined as follows:

- A **feasible solution** is a vector that lies within the feasible region.
- An **optimal solution** is a vector that maximizes/minimizes the linear programming problem.
- A **basic solution** is a solution that is "trivial" in some sense. These solutions can be identified from the simplex tableau, which dictates which variables to set to 0, which forces exact values on the other variables (rather than inequalities). The variables that are set to 0 are called **basic variables** and the others are called **free variables**.

Every optimal solution is also a feasible solution. Similarly, basic solutions *should* also be feasible. If they are not, then we are in the setting of Section A.5. Feasible solutions and basic solutions are not necessarily optimal.

Q. What happens if we have a negative entry in the bottom row that corresponds to a slack variable?
A. We are still allowed to pivot in this column as we would a non-slack variable. For an example, see Example 6 in Chapter 9.3 of the course textbook.

B
Dynamical Systems

B.1 Introduction

Suppose we want to model the populations of rabbits and wolves with respect to time. An increase in the population of wolves will lead to a decrease in the population of rabbits, while an increase in the population of rabbits will lead to an increase in the population of wolves. So, we can model the populations using the following system of differential equations:

$$\begin{cases} R'(t) = \alpha R(t) + \beta W(t) \\ W'(t) = \gamma R(t) + \delta W(t). \end{cases}$$

In this class, we will be taking the xkcd approach to modelling populations. That is, we will discuss mathematical tools for population modelling but not think about the biological feasibility of our models. This aspect of modelling is discussed in Math 3MB3.

The simplest case is when β and γ are both zero. In this case, we can solve each equation separately, as we will review in Section B.2. If either β or γ are nonzero, we will make use of linear algebra to simplify our system. In this case, we will use diagonalization to perform a change of variables into a new coordinate system, with respect to which, the system is one that we can solve directly. Then, we will undo the change of coordinates to recover the solution in the given terms.

This process is intentionally hand-wavy for the time being. We will get into the precise details in Section B.4, so do not worry about them yet.

B.2 Differential equations, very briefly

In this section, we will review some basic facts about differential equations. What is discussed here should be familiar from your Calc II course, so our treatment will be brief.

Recall that $\frac{d}{dx}e^x = e^x$ and $\frac{d}{dx}ke^x = ke^x$ for any k. More generally, if $f(x)$ is a real-valued function of x, then the only solutions to $f'(x) = f(x)$ are when $f(x) = ke^x$ (which includes the trivial solution of $f(x) = 0$). Similarly, the only solutions to $f'(x) = Cf(x)$ are of the form $f(x) = ke^{Cx}$. Indeed, we have that

$$\frac{d}{dx}ke^{Cx} = ke^{Cx} \cdot \frac{d}{dx}(Cx) = C \cdot ke^{Cx} = Cf(x).$$

So if we have the following system of differential equations,

This is exactly the easy case discussed in Section B.1.

$$\begin{cases} f'(x) = \alpha f(x) \\ g'(x) = \delta g(x), \end{cases}$$

then the solutions are $f(x) = k_1 e^{\alpha x}$ and $g(x) = k_2 e^{\delta x}$. However, if we have cross-terms, when the system has the following form,

$$\begin{cases} f'(x) = \alpha f(x) + \beta g(x) \\ g'(x) = \gamma f(x) + \delta g(x), \end{cases}$$

then this approach does not work. Not all is lost, however. Write this system in the form $\vec{y}' = A\vec{y}$,

$$\underbrace{\begin{bmatrix} f'(x) \\ g'(x) \end{bmatrix}}_{\vec{y}'} = \underbrace{\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} f(x) \\ g(x) \end{bmatrix}}_{\vec{y}},$$

and suppose that there exists matrices P and D such that P is invertible, D is diagonal, and $A = PDP^{-1}$. Define a new variable by $\vec{z} := P^{-1}\vec{y}$, so $\vec{y} = P\vec{z}$. Hence, the differential equation $\vec{y}' = A\vec{y}$ becomes

$$\frac{d}{dx}(P\vec{z}) = A(P\vec{z}) = (PDP^{-1})P\vec{z} = PD\vec{z},$$

while the left-hand side simplifies to $P\vec{z}'$. Multiplying both sides by P^{-1} yields the differential equation $\vec{z}' = D\vec{z}$, which is of the form we know how to solve. Our desired solution is thus given by $\vec{y} = P\vec{z}$.

So now we know theoretically how to solve systems of linear equations of the form $\vec{y}' = A\vec{y}$, provided we can write $A = PDP^{-1}$ for convenient matrices P and D. We thus have the following questions:

Question B.1. When can we write $A = PDP^{-1}$? Moreover, when we can write A in this form, how do we compute P and D?

As you may already know from first-year linear algebra, the answer is given by the theory of diagonalization. This is discussed in Section B.4 but we must first take a detour through the theory of eigenvalues and eigenvectors.

B.3 Eigenvalues and eigenvectors

Definition B.2. Let A be a square matrix. An **eigenpair** $(\lambda, \vec{v})$ of A is a scalar λ and nonzero vector $\vec{v}$ satisfying $A\vec{v} = \lambda v$. We say that λ is an **eigenvector** of A with corresponding **eigenvector** $\vec{v}$.

An eigenvector of A is a vector $\vec{v}$ such that $A\vec{v}$ is rescaling. From a computational perspective, it is easier to compute a rescaling $\lambda\vec{v}$ compared to a matrix-vector product, as fewer operations are needed.[1]

[1] If you are interested in learning about algorithms for linear algebra that are more/less efficient, consider taking Math 3NA3.

Finding eigenvectors

Finding eigenpairs of a matrix A is a two-step process: we must first find the eigenvalues of A and then find the corresponding eigenvectors for each eigenvalue. We recall these processes, in the reverse order.

Example B.3. An eigenvalue of the following matrix is $\lambda = 2$.

$$A = \begin{bmatrix} 4 & -1 & 6 \\ 2 & 1 & 6 \\ 2 & -1 & 8 \end{bmatrix}$$

An eigenvector $\vec{v}$ of A must satisfy $A\vec{v} = 2\vec{v}$, or equivalently, $A\vec{v} - 2\vec{v} = \vec{0}$. Rewrite this as $(A - 2I)\vec{v} = \vec{0}$, we now have a system of linear equations to solve which we can represent as an augmented matrix. First, write

We cannot write $A - 2$ because we cannot subtract a scalar from a vector. However, $2\vec{v} = 2I\vec{v}$, where I is the identity matrix, so $A\vec{v} - 2\vec{v} = A\vec{v} - 2I\vec{v} = (A - 2I)\vec{v}$.

$$A - 2I = \begin{bmatrix} 4 & -1 & 6 \\ 2 & 1 & 6 \\ 2 & -1 & 8 \end{bmatrix} - \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 6 \\ 2 & -1 & 6 \\ 2 & -1 & 6 \end{bmatrix}.$$

Thus the system $(A - 2I)\vec{v} = \vec{0}$ can be written as the following augmented matrix

$$\left[\begin{array}{ccc|c} 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \end{array}\right]$$

so we need only row reduce to find the eigenvector $\vec{v}$.

$$\left[\begin{array}{ccc|c} 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \end{array}\right] \begin{matrix} \\ R2-2R1 \\ R3-2R1 \end{matrix} \sim \left[\begin{array}{ccc|c} 2 & -1 & 6 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Eigenvectors $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$ thus satisfy $2v_1 = v_2 - 6v_3$, or equivalently, $v_1 = \frac{1}{2}v_2 - 3v_3$. Because v_2 and v_3 are both free variables, eigenvectors of A with eigenvalue 2 are of the form

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}v_2 - 3v_3 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix} v_2 + \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix} v_3$$

Hence eigenvectors of A with respect to $\lambda = 2$ are

$$\begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}.$$

Notice that eigenvectors are not unique. We found these eigenvectors by setting (i) $v_2 = 1$ and $v_3 = 0$, and (ii) $v_2 = 0$ and $v_3 = 1$ in the

previous equation for $\vec{v}$. For example, the following are also eigenvectors for A with respect to $\lambda = 2$:

$$\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 0 \\ -1/3 \end{bmatrix}.$$

Exercise 7. Exercises 9−15 from Chapter 5.1 of the course textbook are similar to the previous example: they give you a matrix and its eigenvalue(s), and ask you to compute the corresponding eigenvectors. Complete as many of these problems as you need until you are comfortable computing eigenvectors. Check your answer(s) by verifying that the equation $A\vec{v} = \lambda\vec{v}$ holds.

Definition B.4. Let A be a square matrix with eigenvalue λ. The **eigenspace** of A corresponding to λ is the set of vectors $\vec{v}$ for which $A\vec{v} = \lambda\vec{v}$. That is, it is the set of all eigenvectors corresponding to λ, together with the zero vector.

Fix a square matrix A and one of its eigenvalues λ. Suppose that we find that $\vec{v}_1, \ldots, \vec{v}_r$ are linearly independent eigenvectors of A corresponding to λ. Then, the eigenspace of A corresponding to λ is the set of all linear combinations of $\vec{v}_1, \ldots, \vec{v}_r$. We say that $\vec{v}_1, \ldots, \vec{v}_r$ is a **basis** for the eigenspsace.

Recall that a list of vectors $\vec{v}_1, \ldots, \vec{v}_r$ is **linearly dependent** if, for some i, $\vec{v}_i$ can be expressed as a linear combination of the other vectors $\vec{v}_1, \ldots, \vec{v}_{i-1}, \vec{v}_{i+1}, \ldots, \vec{v}_r$. A list of vectors is **linearly independent** if it is not linearly dependent.

In particular, two vectors are linearly dependent if and only if one is a scalar multiple of the other. Hence, a list of two vectors is linearly independent if neither is a scalar multiple of the other.

We will discuss linear independence and bases more in Chapter C.

Example B.5. Continuing the previous example, a basis for the eigenspace of A corresponding to $\lambda = 2$ is given by

$$\begin{bmatrix} 1/2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}.$$

Given a (square) matrix and an eigenvalue, we saw in the previous example that to find corresponding eigenvectors, we need only perform row reduction. We next turn our attention to the problem of finding eigenvalues.

Finding eigenvalues

Recall that eigenvalues and eigenvectors satisfy $A\vec{v} = \lambda\vec{v}$, or equivalently, $(A - \lambda I)\vec{v} = \vec{0}$. Because eigenvectors $\vec{v}$ are nonzero, the equation $(A - \lambda I)\vec{v} = \vec{0}$ says that $A - \lambda I$ is a noninvertible matrix, because it maps a nonzero vector to $\vec{0}$. Moreover, recall that any matrix B is noninvertible if and only if its determinant is zero. So, to find eigenvalues, we are required to find the λ such that

$$\det(A - \lambda I) = 0.$$

Because λ is an unknown, we treat it as a variable, so the expression $\det(A - \lambda I)$ is a polynomial in the variable λ. We call this polynomial the **characteristic polynomial** of A.

Thus to compute eigenvalues, we must first compute a determinant. We review this process now, first with two special cases, followed by the general formula.

Fact B.6. The determinant of a 2×2 matrix is given by

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

For a 3×3 matrix, there is a slightly more complicated formula. To compute

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix},$$

first repeat the first two columns as follows:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix} \begin{matrix} a & b \\ d & e \\ g & h \end{matrix}$$

and compute the products corresponding to the diagonals,

$$aek,\ bfg,\ cdh$$

and corresponding to the antidiagonals,

$$ceg,\ afh,\ bdk.$$

The determinant is the difference of these lists:

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aek + bfg + cdh - ceg - afh - cdh.$$

Definition B.7. Let A be an $n \times n$ matrix. Denote by $a_{i,j}$ the (i,j)-th entry of A, and denote by $A_{i,j}$ the submatrix of A obtained by deleting the i-th row and j-th column. Then, the **determinant** of A is given by the following recursive formula,

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det(A_{i,j}).$$

This is called the **cofactor expansion** of A along the i-th row. The above formula gives the same answer for any row $i = 1, \ldots, n$. Similarly, we could instead compute the determinant using a cofactor expansion along the j-th column, for any $j = 1, \ldots, n$,

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{i,j} \det(A_{i,j}).$$

Example B.8. We will compute the determinant of the 4×4 matrix

$$A = \begin{bmatrix} 1 & -2 & 5 & 0 \\ 2 & 0 & 4 & -1 \\ 3 & 1 & 0 & 7 \\ 0 & 4 & -2 & 0 \end{bmatrix}.$$

It is convenient to perform cofactor expansions along rows/columns with many zeros. Indeed, these zeros will appear as $a_{i,j}$ in the summation definition of the determinant, meaning we will have to compute fewer of the $\det(A_{i,j})$.

We will perform a cofactor expansion along the fourth row. That is, we will use the first formula above, with $i = 4$.

$$\begin{aligned} \det A &= \sum_{j=1}^{4} (-1)^{4+j} a_{4,j} \det(A_{4,j}) \\ &= (-1)^{4+1}(0)\det(A_{4,1}) + (-1)^{4+2}(4)\det(A_{4,2}) \\ &\quad + (-1)^{4+3}(-2)\det(A_{4,3}) + (-1)^{4+4}(0)\det(A_{4,4}) \end{aligned}$$

$$= 4\begin{vmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 3 & 0 & 7 \end{vmatrix} + 2\begin{vmatrix} 1 & -2 & 0 \\ 2 & 0 & -1 \\ 3 & 1 & 7 \end{vmatrix}$$

$$\begin{aligned} &= 4\big[(28 + (-15) + 0) - (0 + 0 + 70)\big] \\ &\quad - (-2)\big[(0 + 6 + 0) - (0 + (-1) + (-28))\big] \\ &= -228 + 70 = -158 \end{aligned}$$

Exercise 8. Compute $\det A$ for the same matrix as the previous example using cofactor expansion along the fourth column. Verify that your answer is the same.

We conclude this section with an eigenvalue-finding example, then recall some other facts about eigenvalues and determinants.

Example B.9. We will find the eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}.$$

The eigenvalues are the roots of the characteristic polynomial,

$$\begin{aligned} \det(A - \lambda I) &= \det\left(\begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}\right) \\ &= \det\begin{bmatrix} -\lambda & 0 & -2 \\ 1 & 2-\lambda & 1 \\ 1 & 0 & 3-\lambda \end{bmatrix} \\ &= \big[(-\lambda)(2-\lambda)(3-\lambda) + 0 + 0\big] \\ &\quad - \big[(-2)(2-\lambda) + 0 + 0\big] \\ &= -\lambda^3 + 5\lambda^2 - 8\lambda + 4 \\ &= -(\lambda - 1)(\lambda - 2)^2. \end{aligned}$$

So the eigenvalues are the roots: $\lambda = 1$ and $\lambda = 2$.

The steps were not shown, but polynomial long division was used to factor out the root at $\lambda = 1$. Then, we noticed that the remaining quadratic was a perfect square.

Finding eigenvalues is difficult, because finding roots of polynomials can be difficult. Fortunately, in some special cases, we can find the eigenvalues without any computation.

Exercise 9. If A is a 3×3 triangular matrix (either lower or upper triangular), show that the eigenvalues of A are exactly its diagonal entries. Convince yourself that the result holds for a triangular matrix of any size.

We also have the following.

Fact B.10. The determinant of a matrix is equal to the product of its eigenvalues. The sum of the eigenvalues is the trace of the matrix.

B.4 Matrix decomposition: Diagonalization

Recall that in Section B.2 we asked: Given a matrix A, when is it possible to write $A = PDP^{-1}$ for some invertible matrix P and diagonal matrix D? When such matrices P and D exist, we say that A is **diagonalizable**. The following theorem tells us exactly when the matrix A is diagonalizable.

More generally, matrices A and B are **similar** if there is an invertible matrix P such that $A = PBP^{-1}$. So the matrix A is diagonalizable if it is similar to a diagonal matrix.

Matrices that are similar represent the same linear transformation, just with respect to different coordinate systems. The matrix P is giving the change of coordinates to go from one linear transformation to the other. Loosely, P is performing a combination of stretches and rotations to go between coordinate systems.

Theorem B.11. *An $n \times n$ matrix is diagonalizable if and only if it has n linearly independent eigenvectors.*

If such eigenvectors exist, then they form the columns of P and the corresponding eigenvalues form the entries of D.

Recall from first-year linear algebra that a list of vectors $\vec{v}_1, \ldots, \vec{v}_r$ is **linearly independent** if no vector in the list can be written as a linear combination of the other vectors. For example, two vectors $\vec{v}, \vec{u}$ are linearly independent neither is a scalar multiple of the other.

Exercise 10. Give an example of three vectors in $\mathbb{R}^3$ that are linearly independent, and three that are linearly dependent. Can you find a list of four linearly independent vectors in $\mathbb{R}^3$? Lastly, show that any list containing the zero vector is linearly dependent.

Proof of Theorem B.11. First suppose A is diagonalizable, so $A = PDP^{-1}$, or equivalently, $AP = PD$. Write:

$$P = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix},$$

so the vectors $\vec{v}_1, \ldots, \vec{v}_n$ form the columns of P. We then compute both AP and PD.

$$AP = A \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ A\vec{v}_1 & A\vec{v}_2 & \cdots & A\vec{v}_n \\ | & | & & | \end{bmatrix}$$

$$PD = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda_n\vec{v}_n \\ | & | & & | \end{bmatrix}.$$

Because $AP = PD$, comparing the above two equations column-by-column tells us that $A\vec{v}_i = \lambda_i\vec{v}_i$ for each $i = 1, \ldots, n$. Hence each λ_i is an eigenvalue with eigenvector $\vec{v}_i$. Each $\vec{v}_i$ is nonzero because P was assumed to be invertible.[2] Moreover, the eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ are linearly independent, again, because P is invertible. (See the Invertible Matrix Theorem in the course textbook.)

[2] A matrix with a row/column of zeros must have determinant zero. Indeed, perform cofactor along that row/column. Thus, a matrix with a row or column of all zeros cannot be invertible.

Now we will assume that A has n linearly independent eigenvectors and we will show that A is diagonalizable. Denote by $\vec{v}_1, \ldots, \vec{v}_n$ the linearly independent eigenvectors. Construct a matrix P whose columns are the eigenvectors $\vec{v}_i$,

$$P = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix},$$

and diagonal matrix whose entries are the corresponding eigenvalues,

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n. \end{bmatrix}$$

Notice that the eigenvalues in the list $\lambda_1, \ldots, \lambda_n$ may appear with repetition. Then, to show that $A = PDP^{-1}$, it is equivalent to show that $AP = PD$. Indeed,

$$AP = A \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ A\vec{v}_1 & A\vec{v}_2 & \cdots & A\vec{v}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda_n\vec{v}_n \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n. \end{bmatrix} = PD.$$

■

This theorem says that to diagonalize a matrix, we need to compute the corresponding eigenvalues and eigenvectors to construct D and P, respectively. We next recall a sufficient condition for a matrix to be diagonalizable.

Corollary B.12. *If an $n \times n$ matrix has n distinct eigenvalues, then it is diagonalizable.*

The converse to this corollary is false! An $n \times n$ matrix may be diagonalizable even if it does not have n eigenvalues.

Proof. Eigenvectors corresponding to distinct eigenvalues are linearly independent. Having n distinct eigenvalues thus implies that we have n linearly independent eigenvalues, so apply Theorem B.11. ■

Our original motivation for diagonalization was to solve dynamical systems. Let us recall this process, then work through an example of diagonalization to solve a dynamical system. The same procedure works for larger (e.g., 3×3) systems as well.

This process is called **decoupling**, as we rewrite the system in such a way that each differential equation is independent from the other.

Diagonalization for solving dynamical systems.
Consider the dynamical system represented as

$$\begin{cases} f'(x) = \alpha f(x) + \beta g(x) \\ g'(x) = \gamma f(x) + \delta g(x). \end{cases}$$

1. Write the system in the form $\vec{y}' = A\vec{y}$, where

$$\underbrace{\begin{bmatrix} f'(x) \\ g'(x) \end{bmatrix}}_{\vec{y}'} = \underbrace{\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} f(x) \\ g(x) \end{bmatrix}}_{\vec{y}}.$$

2. Diagonalize A, if possible, writing $A = PDP^{-1}$. (If this is not possible, then we cannot solve the dynamical system using diagonalization.)

3. Define a new variable $\vec{z} = P^{-1}\vec{y}$, so $\vec{y} = P\vec{z}$. The differential equation $\vec{y}' = A\vec{y}$ becomes

$$\frac{d}{dx}(P\vec{z}) = A(P\vec{z}) = (PDP^{-1})P\vec{z} = PD\vec{z}.$$

   The left-hand side is $P\vec{z}'$, so after multiplying both sides by P^{-1}, the system $\vec{y}' = A\vec{y}$ has become

$$\vec{z}' = D\vec{z}.$$

4. Solve the system $\vec{z}' = D\vec{z}$.

5. Substitute $\vec{y} = P\vec{z}$ to solve the original system.

Example B.13. Suppose that the populations of rabbits and wolves can be modelled over time with by the following system of differential equations:

$$\begin{cases} R'(t) = -2R(t) - 5W(t) \\ W'(t) = 1R(t) + 4W(t). \end{cases}$$

We will find explicit equations for $R(t)$ and $W(t)$, given the initial conditions $R(0) = 100$ and $W(0) = 4$.

First, write the system in matrix form,

$$\begin{bmatrix} R'(t) \\ W'(t) \end{bmatrix} = \begin{bmatrix} -2 & -5 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} R(t) \\ W(t) \end{bmatrix}',$$

so $\vec{y} = \begin{bmatrix} R(t) \\ W(t) \end{bmatrix}$ and $A = \begin{bmatrix} -2 & -5 \\ 1 & 4 \end{bmatrix}$. We start by finding the eigenvalues and eigenvectors of A. For the eigenvalues, we compute the characteristic polynomial:

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{bmatrix} -2-\lambda & -5 \\ 1 & 4-\lambda \end{bmatrix} \\ &= (-2-\lambda)(4-\lambda) + 5 \\ &= \lambda^2 - 2\lambda - 8 + 5 \\ &= \lambda^2 - 2\lambda - 3 \\ &= (\lambda - 3)(\lambda + 1). \end{aligned}$$

So the eigenvalues are $\lambda = 3$ and $\lambda = -1$. We next compute corresponding eigenvectors.

- $[\lambda = 3]$ Eigenvectors are the solutions to the system $(A - 3I)\vec{v} = \vec{0}$.

$$\left[\begin{array}{cc|c} -5 & -5 & 0 \\ 1 & 1 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

  Hence, eigenvectors $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ corresponding to $\lambda = 3$ satisfy $v_1 = -v_2$. So, an eigenvector is

$$\vec{v} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

- $[\lambda = -1]$ Eigenvectors are the solutions to the system $(A + 1I)\vec{v} = \vec{0}$.

$$\left[\begin{array}{cc|c} -1 & -5 & 0 \\ 1 & 5 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 5 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

  Hence, eigenvectors $\vec{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ corresponding to $\lambda = -1$ satisfy $u_1 = -5u_2$. So, an eigenvector is

$$\vec{u} = \begin{bmatrix} -5 \\ 1 \end{bmatrix}.$$

We thus construct

$$D = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}, \quad \text{and} \quad P = \begin{bmatrix} -1 & -5 \\ 1 & 1 \end{bmatrix}.$$

Recall that the columns of P are the eigenvectors. Moreover, they must appear in the same order as their corresponding eigenvalues appear in D.

We now introduce our new variable $\vec{z} = P^{-1}\vec{y}$ and corresponding decoupled system $\vec{z}' = D\vec{z}$. Write $\vec{z} = \begin{bmatrix} f(t) \\ g(t) \end{bmatrix}$, so we have the system

Notice that we never actually need to compute the inverse of P!

$$\begin{cases} f'(t) = 3f(t) \\ g'(t) = -g(t) \end{cases} \quad \text{which implies} \quad \begin{cases} f(t) = c_1 e^{3t} \\ g(t) = c_2 e^{-t}. \end{cases}$$

Then, $\vec{y} = P\vec{z}$, so

$$\begin{bmatrix} R(t) \\ W(t) \end{bmatrix} = \begin{bmatrix} -1 & -5 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 e^{3t} \\ c_2 e^{-t} \end{bmatrix} = \begin{bmatrix} -c_1 e^{3t} - 5c_2 e^{-t} \\ c_1 e^{3t} + c_2 e^{-t} \end{bmatrix}.$$

If we were not given initial conditions, then we would be done. Our final answer would be $R(t) = -c_1 e^{3t} - 5c_2 e^{-t}$ and $W(t) = c_1 e^{3t} + c_2 e^{-t}$.

The last step is to impose the additional given conditions that $R(0) = 100$ and $W(0) = 4$. Substituting these conditions yields the system

$$\begin{bmatrix} R(0) \\ W(0) \end{bmatrix} = \begin{bmatrix} 100 \\ 4 \end{bmatrix} = \begin{bmatrix} -c_1 - 5c_2 \\ c_1 + c_2 \end{bmatrix},$$

which we can solve using an augmented matrix as follows.

$$\left[\begin{array}{cc|c} -1 & -5 & 100 \\ 1 & 1 & 4 \end{array}\right]_{R2+R1} \sim \left[\begin{array}{cc|c} -1 & -5 & 100 \\ 0 & -4 & 104 \end{array}\right]_{-\frac{1}{4}R2}^{-R1} \sim \left[\begin{array}{cc|c} 1 & 5 & -100 \\ 0 & 1 & -26 \end{array}\right]_{R1-5R2} \sim \left[\begin{array}{cc|c} 1 & 0 & 30 \\ 0 & 1 & -26 \end{array}\right].$$

So the solution to our system is

$$\begin{cases} R(t) = -30e^{3t} + 130e^{-t} \\ W(t) = 30e^{3t} - 26e^{-t}. \end{cases}$$

Exercise 11. Suppose that a trio of populations are modelled by the differential equations

$$\begin{cases} x'(t) = -x(t) + 4y(t) - 2z(t) \\ y'(t) = -3x(t) + 4y(t) \\ z'(t) = -3x(t) + y(t) + 3z(t), \end{cases}$$

with initial conditions

$$x(0) = 1, \quad y(0) = 2, \quad z(0) = 3.$$

Use diagonalization to solve the differential equations.

C
Minimizing Distance

Suppose that we are given a matrix A and a vector $\vec{b}$ for which there is no solution $\vec{x}$ to the equation $A\vec{x} = \vec{b}$. Since we are unable to find a solution, we can cut our losses and instead ask for the vector $\vec{x}$ that best approximates $A\vec{x} \approx \vec{b}$.

In general, when does $A\vec{x} = \vec{b}$ have solutions? How does this change if A is a square matrix versus a rectangular matrix?

To make this precise, let us denote by $\|\cdot\|$ the length of a vector. So, if $A\vec{x} = \vec{b}$, then $A\vec{x} - \vec{b} = \vec{0}$, and $\|A\vec{x} - \vec{b}\| = \|\vec{0}\| = 0$. But if no solution exists, then we want to find the $\vec{x}$ that minimizes the distance between $A\vec{x}$ and $\vec{b}$. That is, we want to find the following:

$$\min_{\vec{x}\in\mathbb{R}^n} \left\{ \|A\vec{x} - \vec{b}\| \right\}.$$

Definition C.1. The vector $\vec{v}$ for which $\|A\vec{v} - \vec{b}\| \leq \|A\vec{x} - \vec{b}\|$ for all other $\vec{x}$ is called a **least squares solution** to the system $A\vec{x} = \vec{b}$.

C.1 Dot product and distance

To find our solution to the problem discussed above, we first need to make rigorous what we mean by a "distance". The way we do this is via the *dot product*, which will also allow us discuss angles and orthogonality. These topics should be familiar from first-year linear algebra, so we will be brief.

The dot product

Let $\vec{x}$ and $\vec{y}$ be vectors in $\mathbb{R}^n$, writing

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

The **dot product** of $\vec{x}$ and $\vec{y}$ is

$$\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Notice that $\vec{x} \cdot \vec{x} = x_1^2 + x_2^2 + \cdots + x_n^2$ so $\vec{x} \cdot \vec{x} \geq 0$ for all vectors $\vec{x} \in \mathbb{R}^n$ and, in particular, $\vec{x} \cdot \vec{x} = 0$ if and only if $\vec{x} = \vec{0}$. We have the following.

Fact C.2 (Properties of the dot product)**.** Let $\vec{x}, \vec{y}, \vec{z}$ be vectors in $\mathbb{R}^n$ and $c \in \mathbb{R}$ a scalar. Then,

(i) $\vec{x} \cdot \vec{x} \geq 0$ and $\vec{x} \cdot \vec{x} = 0$ if and only if $\vec{x} = 0$,
(ii) $\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x}$,
(iii) $(\vec{x} + \vec{y}) \cdot \vec{z} = \vec{x} \cdot \vec{z} + \vec{y} \cdot \vec{z}$,
(iv) $c(\vec{x} \cdot \vec{y}) = (c\vec{x}) \cdot \vec{y} = \vec{x} \cdot (c\vec{y})$.

Exercise 12. Show that Fact C.2 holds for vectors in $\mathbb{R}^2$.

Norms

A **norm** is a function that measures distance. Recall that the distance between two points (and hence, between two vectors) $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$ is given by

$$\mathsf{dist}(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

Meanwhile, notice that

$$\vec{x} - \vec{y} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \cdots \\ x_n - y_n \end{bmatrix},$$

so the expression beneath the square root in the formula for $\mathsf{dist}(\vec{x}, \vec{y})$ looks like the dot product $(\vec{x} - \vec{y}) \cdot (\vec{x} - \vec{y})$. That is,

$$\mathsf{dist}(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y}) \cdot (\vec{x} - \vec{y})}.$$

Because the dot product is always nonnegative, it makes sense to take a square root. We call the quantity $\mathsf{dist}(\vec{x}, \vec{y})$ the **norm** of the vector $\vec{x} - \vec{y}$ and write

$$\|\vec{x} - \vec{y}\| = \mathsf{dist}(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y}) \cdot (\vec{x} - \vec{y})}.$$

This also allows us to compute the norm of a single vector. Indeed, notice that the length of a vector is $\|\vec{x}\| = \|\vec{x} - \vec{0}\|$, so we have that

$$\|\vec{x}\| = \sqrt{\vec{x} \cdot \vec{x}}.$$

You may remember that one interpretation of the dot product is a measure of angle. Indeed, in $\mathbb{R}^2$ and $\mathbb{R}^3$, the angle between $\vec{u}$ and $\vec{v}$ is

$$\theta = \arccos\left(\frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|}\right).$$

Exercise 13. Let $\vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Compute $\|\vec{x}\|$. Find a vector $\vec{y} \in \mathbb{R}^2$ which is distance 1 away from $\vec{x}$. Find a vector $\vec{z} \in \mathbb{R}^2$ that is distance $\sqrt{2}$ away from $\vec{x}$.

C.2 Subspaces and bases

Recall that our goal for Chapter C is to find the least squares solution vector $\vec{x}$ that minimizes the norm $\|A\vec{x} - \vec{b}\|$ for fixed A and $\vec{b}$. The purpose of this subsection is to reformulate this problem in a way that is more conducive to computations. To that end, we develop a little bit of linear algebra theory in this section and Section C.3, before coming to our solution in Section C.4.

So fix some matrix A and vector $\vec{b}$, with a view towards finding the $\vec{x}$ that best approximates $A\vec{x} \approx \vec{b}$. Write A and $\vec{x}$ as follows:

Here, we are taking A to be an $m \times n$ matrix, meaning it has m rows and n columns. This means that $\vec{x}$ is an $n \times 1$ matrix and $\vec{b}$ is an $m \times 1$ matrix.

$$A = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix},$$

where the vectors $\vec{v}_i$ are the column vectors of A, and the c_i are real numbers, the components of $\vec{x}$. Then, the product $A\vec{x}$ can be computed as follows:

$$A\vec{x} = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n.$$

Thus, for any $\vec{x} \in \mathbb{R}^n$, the vector $A\vec{x}$ is a *linear combination of the columns of* A.

The **span** of a list of vectors is the set of all linear combinations of these vectors. For instance, given a matrix A as above, the span of its columns is the set of vectors of the form

$$a_1\vec{v}_1 + \cdots + a_n\vec{v}_n,$$

for any scalars $a_1, \ldots, a_n \in \mathbb{R}$. Because, in this example, the vectors are the columns of a matrix, we call this the **column span** of A. We denote by $\mathrm{col}(A)$ the column span of A.

Example C.3. The span of the vector $(2, 1)$ in $\mathbb{R}^2$ is the set of all vectors of the form $\lambda(2, 1) = (1\lambda, \lambda)$. Geometrically, it is the set of all scalar multiples of the vector $(2, 1)$, or equivalently, any element on the line in Figure C.1.

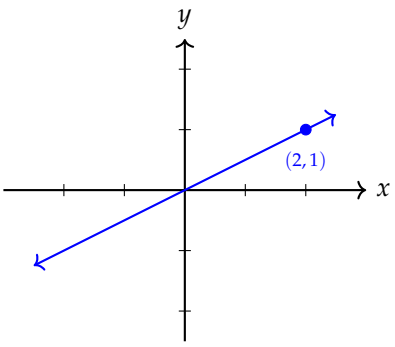


Figure C.1: The set of vectors of the form $(2\lambda, \lambda)$ in $\mathbb{R}^2$.

Fix two vectors $\vec{u}_1$ and $\vec{u}_2$ in the span of $\vec{v}_1, \ldots, \vec{v}_n$, writing

$$\vec{u}_1 = a_1\vec{v}_1 + a_2\vec{v}_2 + \cdots + a_n\vec{v}_n,$$
$$\vec{u}_2 = b_1\vec{v}_1 + b_2\vec{v}_2 + \cdots + b_n\vec{v}_n.$$

Notice that the sum $\vec{u}_1 + \vec{u}_2$ is also in the span of $\vec{v}_1, \ldots, \vec{v}_n$, because:

$$\begin{aligned}\vec{u}_1 + \vec{u}_2 &= a_1\vec{v}_1 + a_2\vec{v}_2 + \cdots + a_n\vec{v}_n \\ &\quad + b_1\vec{v}_1 + b_2\vec{v}_2 + \cdots + b_n\vec{v}_n \\ &= (a_1 + b_1)\vec{v}_1 + (a_2 + b_2)\vec{v}_2 + \cdots + (a_n + b_n)\vec{v}_n,\end{aligned}$$

so $\vec{u}_1 + \vec{u}_2$ is a linear combination of $\vec{v}_1, \ldots, \vec{v}_n$. Similarly, if $k \in \mathbb{R}$ is any scalar, then,

$$k\vec{u}_1 = k(a_1\vec{v}_1 + \cdots + a_n\vec{v}_n) = (ka_1)\vec{v}_1 + \cdots + (ka_n)\vec{v}_n,$$

so scalar multiplication remains within the span. This tells us that the span of a fixed collection of vectors is a special type of subset.

Definition C.4. A subset S of $\mathbb{R}^m$ is called a **vector subspace** if it is closed under vector addition, closed under scalar multiplication, and contains the zero vector.

That is, S is a subspace if:

(i) $\vec{0} \in S$,
(ii) if $\vec{u}_1$ and $\vec{u}_2$ are in S, then so is $\vec{u}_1 + \vec{u}_2$,
(iii) if $\vec{u} \in S$ and $k \in \mathbb{R}$, then $k\vec{u}$ is also in S.

We thus have the following.

Fact C.5. The column span of a matrix is a vector subspace.

Proof. Let A be an $m \times n$ matrix. The column span is a subset of $\mathbb{R}^m$ and we showed early it is closed under vector addition and closed under scalar multiplication. It also contains the zero vector because, if the columns of A are $\vec{v}_1, \ldots, \vec{v}_n$, then $0\vec{v}_1 + \cdots + 0\vec{v}_n = \vec{0}$ is an element of the column span. ■

Any subspace of $\mathbb{R}^m$ can be viewed as the span of a finite list of vectors. However, not all spanning lists are created equal.

For example, consider the subspace S of $\mathbb{R}^2$ consisting of vectors of the form (x,x).[1] This set can be viewed as the span of $(1,1)$, but it can also be viewed as the span of $(1,1)$ and $(2,2)$. The second list is less desirable because it is not minimal; the vector $(2,2)$ is not telling us any new information. The following defines a "good" spanning set for a subspace.

[1] The set S is a subspace because it contains the zero vector and is closed under vector addition and scalar multiplication. Indeed, $(x,x)+(y,y) = (x+y,x+y)$ and $k(x,x) = (kx,kx)$.

Definition C.6. Let S be a subspace of $\mathbb{R}^m$. A **basis** for S is a *linearly independent* list of vectors whose span is S.

Example C.7. Let S be the subspace of $\mathbb{R}^2$ of points of the form (x,x). We saw earlier that $(1,1),(2,2)$ is *not* a basis of S because the list $(1,1),(2,2)$ is linearly dependent. However, the list $(1,1)$ is linearly dependent and spans S, hence it is a basis of S.

The list $(1,0,0),(0,1,0),(0,0,1)$ is called the **standard basis** of $\mathbb{R}^3$.

Exercise 14. Find a basis of $\mathbb{R}^3$ that is not the standard basis. Find a list of vectors that spans $\mathbb{R}^3$ but is not linearly independent. Then find a list of vectors in $\mathbb{R}^3$ that is linearly independent but whose span is not $\mathbb{R}^3$. Then look back at Exercise 10.

Exercise 15. Find a basis for the column span of the following matrices.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

You should recall from first-year linear algebra how to find a basis for the column space of a given matrix. If you want to review, here is a video explanation for the column space, or you can see Chapter 4.2 of the course textbook.

C.3 Orthonormal bases and Gram-Schmidt

In this section, we continue developing the theory needed for our approximation question $A\vec{x} \approx \vec{b}$. Intuitively, we will say that two vectors are orthogonal if their intersection forms a right-angle. Formally, we say the following.

Definition C.8. Vectors $\vec{u}$ and $\vec{v}$ are **orthogonal** if $\vec{u} \cdot \vec{v} = 0$.

For instance, notice that $(1,0)$ and $(0,1)$ are orthogonal because $(1,0)\cdot(0,1) = (1)(0)+(0)(1) = 0$.

Exercise 16. Show that the vector (x,x) is orthogonal to $(y,-y)$, for any values of x and y. Also, sketch these two vectors to convince yourself that they always intersect at a right angle.

Theorem C.9 (Pythagoras)**.** *Suppose that $\vec{u}$ and $\vec{v}$ are orthogonal. Then, $\|\vec{u}+\vec{v}\|^2 = \|\vec{u}\|^2 + \|\vec{v}\|^2$.*

Do you see how this theorem generalizes the usual Pythagorean Theorem? Draw a picture!

Proof. $\|\vec{u}+\vec{v}\|^2 = (\vec{u}+\vec{v})\cdot(\vec{u}+\vec{v}) = \|\vec{u}\|^2 + 2(\vec{u}\cdot\vec{v}) + \|\vec{v}\|^2$ and $\vec{u}$ and $\vec{v}$ are orthogonal, so $\vec{u}\cdot\vec{v} = 0$. ■

In the previous section, we saw that not all spanning lists of subspaces are created equal: some are bases, while some are not. It turns out that not every basis is created equal either. Here are two even nicer types of bases.

Definition C.10. Let S be a subspace of $\mathbb{R}^m$. A basis $\vec{v}_1, \ldots, \vec{v}_n$ of S is called an **orthogonal basis** of S if $\vec{v}_i$ and $\vec{v}_j$ are orthogonal, for every $i \neq j$. Moreover, a basis $\vec{v}_1, \ldots, \vec{v}_n$ of S is called an **orthonormal basis** of S if it is both an orthogonal basis and $\|v_i\| = 1$ for every i.

Exercise 17. The standard basis of $\mathbb{R}^2$, consisting of the vectors $(1,0)$ and $(0,1)$ is an orthonormal basis. Find:

- a basis of $\mathbb{R}^2$ that is not orthogonal;
- an orthogonal basis of $\mathbb{R}^2$ that is not orthonormal;
- an orthonormal basis of $\mathbb{R}^2$ that is not the standard basis.

Fix a nonzero vector $\vec{v} \in \mathbb{R}^n$. Given any other vector $\vec{u}$ in $\mathbb{R}^n$, we wish to write $\vec{u} = k\vec{v} + \vec{w}$ where $\vec{v}$ and $\vec{w}$ are orthogonal and k is a scalar. This is called an *orthogonal decomposition* of $\vec{u}$.

Proposition C.11. *As above, let $\vec{u}$ and $\vec{v}$ be vectors in $\mathbb{R}^n$ with $\vec{v}$ nonzero. Then, $\vec{u} = k\vec{v} + \vec{w}$, where $\vec{v}$ and $\vec{w}$ are orthogonal, when:*

$$k = \frac{\vec{u}\cdot\vec{v}}{\|\vec{v}\|^2} \quad \text{and} \quad \vec{w} = \vec{u} - k\vec{v}$$

Proof. Because $\vec{w} = \vec{u} - k\vec{v}$, we have that $\vec{u} = k\vec{v} + \vec{w}$ as desired. So we need only show that $\vec{v}$ and $\vec{w}$ are orthogonal. We compute:

$$\vec{v}\cdot\vec{w} = \vec{v}\cdot(\vec{u} - k\vec{v}) = \vec{v}\cdot\vec{u} - k(\vec{v}\cdot\vec{v}) = \vec{v}\cdot\vec{u} - k\|\vec{v}\|^2.$$

Then, by definition of k, we have

$$\vec{v}\cdot\vec{w} = \vec{v}\cdot\vec{u} - \frac{\vec{u}\cdot\vec{v}}{\|\vec{v}\|^2}\|\vec{v}\|^2 = \vec{v}\cdot\vec{u} - \vec{u}\cdot\vec{v} = 0,$$

where we use the fact that $\vec{x}\cdot\vec{y} = \vec{y}\cdot\vec{x}$ for all vectors $\vec{x}$ and $\vec{y}$. ■

In fact, we can generalize the previous theorem as follows. Although these results may seem rather abstract at the moment, they will play a key role in answering our minimization problem.

Theorem C.12. *Let S be a subspace of $\mathbb{R}^n$. Any vector $\vec{u}$ of $\mathbb{R}^n$ can be written in the form $\vec{u} = \vec{v} + \vec{w}$ where $\vec{v} \in S$ and $\vec{v}\cdot\vec{w} = 0$.*

Moreover, if $\vec{v}_1, \ldots, \vec{v}_m$ is an orthonormal basis of S, then $\vec{v}$ and $\vec{w}$ are given by

$$\vec{v} = (\vec{u}\cdot\vec{v}_1)v_1 + (\vec{u}\cdot\vec{v}_2)v_2 + \cdots + (\vec{u}\cdot\vec{v}_m)v_m \quad \text{and} \quad \vec{w} = \vec{u} - \vec{v}.$$

In this case, we call $\vec{v}$ the **projection** of $\vec{u}$ onto S. We will write $\vec{v} = \text{proj}_S(\vec{u})$. It turns out that the vector in S that is closest to $\vec{u}$ is $\text{proj}_S(\vec{u})$, but we are not yet ready to show this. We proceed as follows: Next, we will show that Theorem C.12 is true. Then, we will show that given a subspace, we can always find an orthogonal basis, so we can always apply Theorem C.12. Then, in Section C.4, we will prove the minimization result.

Proof. As before, our choice of $\vec{w}$ guarantees that $\vec{u} = \vec{v} + \vec{w}$, so it remains to show that $\vec{v} \cdot \vec{w} = 0$. We will first show that $\vec{v}_i \cdot \vec{w} = 0$ for each i. Indeed, for $i = 1$ we have

$$\begin{aligned}
\vec{v}_1 \cdot \vec{w} &= \vec{v}_1 \cdot (\vec{u} - \vec{v}) \\
&= \vec{v}_1 \cdot \left[\vec{u} - \left(\frac{\vec{u} \cdot \vec{v}_1}{\|v_1\|^2} v_1 + \frac{\vec{u} \cdot \vec{v}_2}{\|\vec{v}_2\|^2} v_2 + \cdots + \frac{\vec{u} \cdot \vec{v}_m}{\|\vec{v}_m\|^2} v_m \right) \right] \\
&= \vec{v}_1 \cdot \vec{u} - \left(\frac{\vec{u} \cdot \vec{v}_1}{\|v_1\|^2} \right) (\vec{v}_1 \cdot \vec{v}_1) - \left(\frac{\vec{u} \cdot \vec{v}_2}{\|v_2\|^2} \right) (\vec{v}_1 \cdot \vec{v}_2) - \cdots - \left(\frac{\vec{u} \cdot \vec{v}_m}{\|v_m\|^2} \right) (\vec{v}_1 \cdot \vec{v}_m).
\end{aligned}$$

Because $\vec{v}_1, \ldots, \vec{v}_m$ is an orthogonal list, we have $\vec{v}_i \cdot \vec{v}_j = 0$ for all $i \neq j$. Also, recall that $\vec{v}_1 \cdot \vec{v}_1 = \|v_1\|^2$. Hence, the above simplifies to

$$\vec{v}_1 \cdot \vec{w} = \vec{v}_1 \cdot \vec{u} - \vec{u} \cdot \vec{v}_1 = 0.$$

A similar argument shows that $\vec{v}_i \cdot \vec{w} = 0$ for all i. This then implies that $\vec{v} \cdot \vec{w} = 0$, as you should show. ■

Exercise 18. In the above proof, show that $\vec{v} \cdot \vec{w} = 0$. Your argument will be similar to the argument to conclude that $\vec{v}_1 \cdot \vec{w} = 0$ and you should make use of properties of the dot product, listed in Fact C.2.

We conclude this section with Gram-Schmidt, a procedure to find an orthonormal basis for any given subspace.[2]

Theorem C.13 (Gram-Schmidt)**.** *Let S be a subspace of $\mathbb{R}^n$ with basis $\vec{b}_1, \ldots, \vec{b}_m$. An orthonormal basis $\vec{v}_1, \ldots, \vec{v}_m$ of S is given by $\vec{v}_1 = \vec{b}_1 / \|b_1\|$ and, for $i \geq 2$,*

$$\begin{aligned}
\vec{v}_i &= \frac{\vec{b}_i - (\vec{b}_i \cdot \vec{v}_1)\vec{v}_1 - (\vec{b}_i \cdot \vec{v}_2)\vec{v}_2 - \cdots - (\vec{b}_i \cdot \vec{v}_{i-1})\vec{v}_{i-1}}{\left\| \vec{b}_i - (\vec{b}_i \cdot \vec{v}_1)\vec{v}_1 - (\vec{b}_i \cdot \vec{v}_2)\vec{v}_2 - \cdots - (\vec{b}_i \cdot \vec{v}_{i-1})\vec{v}_{i-1} \right\|} \\
&= \frac{\vec{b}_i - \text{proj}_{\text{span}(\vec{v}_1, \ldots, \vec{v}_{i-1})} \vec{b}_i}{\left\| \vec{b}_i - \text{proj}_{\text{span}(\vec{v}_1, \ldots, \vec{v}_{i-1})} \vec{b}_i \right\|}.
\end{aligned}$$

That is, $\text{span}(\vec{v}_1, \ldots, \vec{v}_m) = \text{span}(\vec{b}_1, \ldots, \vec{b}_m)$ *and* $\vec{v}_i \cdot \vec{v}_j = 0$ *for all* $i \neq j$.

The idea behind the Gram-Schmidt process is to repeatedly apply the orthogonal decompositions from Proposition C.11 and Theorem C.12 to transform a given basis into an orthonormal one. The wikipedia page for Gram-Schmidt includes an animation of the procedure at work, which you should watch.

[2] The Gram-Schmidt procedure presented here is a little different to the textbook's. The difference is that the version presented here returns an *orthonormal* basis, while the textbook's returns only an *orthogonal* basis.

Exercise 19 (Optional)**.** Use induction to prove Gram-Schmidt.

C.4 Closest vector to a subspace

Theorem C.12 tells us that if we fix a subspace S of $\mathbb{R}^n$ and a vector $\vec{u}$ of $\mathbb{R}^n$, then we can find vectors $\vec{v}$ and $\vec{w}$ such that:

- $\vec{u} = \vec{v} + \vec{w}$,
- $\vec{v} \in S$,
- $\vec{v}$ and $\vec{w}$ are orthogonal.

It turns out that this gives us everything we need to solve our minimization problem.

Theorem C.14 (Best approximation to a subspace)**.** *Let S be a subspace of $\mathbb{R}^n$. Fix a vector $\vec{u}$ of $\mathbb{R}^n$. The orthogonal projection $\vec{v} = \text{proj}_S(\vec{u})$ is the vector in S closest to $\vec{u}$. That is, for any $\vec{y} \in S$ with $\vec{y} \neq \vec{v}$,*

$$\|\vec{u} - \vec{v}\| < \|\vec{u} - \vec{y}\|.$$

Remark C.15. Recall that our original problem was: Given a matrix A and vector $\vec{b}$ for which the system $A\vec{x} = \vec{b}$ has no solutions, find the vector $\vec{v}$ for which $A\vec{v}$ best approximates $\vec{b}$. The above theorem produces the vector $\vec{y} \in \text{col}(A)$ that minimizes $\|\vec{b} - \vec{y}\|$. Because $\vec{y} \in \text{col}(A)$, we can then solve the system $A\vec{x} = \vec{y}$ for $\vec{x}$. This $\vec{x}$ will satisfy our desired minimization problem.

In Section C.5, we will see how *QR factorization* can help us answer the minimization problem in one step.

To compute this projection, or equivalently, to apply Theorem C.12, we first need an orthonormal basis for $\text{col}(A)$. So, we have the following procedure.

Best approximation for systems $A\vec{x} = \vec{b}$ with no solutions.

1. Find a basis for $\text{col}(A)$.
2. Apply Gram Schmidt (Theorem C.13) to transform this into an orthonormal basis for $\text{col}(A)$.
3. Use this orthonormal basis and Theorem C.12 to find the orthogonal projection of $\vec{b}$ onto $\text{col}(A)$.

Theorem C.14 says that this projection is the vector in $\text{col}(A)$ that is closest to $\vec{b}$.

Proof of Theorem C.14. We already know that $\vec{v}$ is in S, because it is defined as the projection of $\vec{u}$ onto S. So we need only show that if $\vec{x} \in W$ satisfies $\vec{x} \neq \vec{v}$, then $\|\vec{u} - \vec{v}\| < \|\vec{u} - x\|$.

So suppose that $\vec{x}$ is in S and is distinct from $\vec{v}$. We first claim that $(\vec{u}-\vec{v})\cdot(\vec{v}-\vec{x}) = 0$. Indeed, $\vec{v}-\vec{x}$ is in S and $\vec{u}-\vec{v}$ is orthogonal to every vector in S. We then apply the Pythagorean Theorem (Theorem C.9) to compute that:

$$\|\vec{u}-\vec{x}\|^2 = \|(\vec{u}-\vec{v})+(\vec{v}-\vec{x})\|^2 = \|\vec{u}-\vec{v}\|^2 + \|\vec{v}-\vec{x}\|^2.$$

Since $\vec{x} \neq \vec{v}$, we have that $\|\vec{v}-\vec{x}\| > 0$, so we conclude that

$$\|\vec{u}-\vec{x}\|^2 = \|\vec{u}-\vec{v}\|^2 + \|\vec{v}-\vec{x}\|^2 > \|\vec{u}-\vec{v}\|^2,$$

as desired. ■

Example C.16. We will find the vector $\vec{v}$ in $\mathbb{R}^3$ that best approximates $A\vec{v} \approx \vec{b}$ for

$$A = \begin{bmatrix} 1 & 0 & \sqrt{2} \\ 0 & 1/\sqrt{2} & 2 \\ 0 & 1/\sqrt{2} & 2 \end{bmatrix} \quad \text{and} \quad \vec{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

An orthonormal basis for $\mathrm{col}(A)$ is given by

$$\vec{c}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{c}_2 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

Then by the result of Theorem C.12, the vector $\vec{y} \in \mathrm{col}(A)$ closest to $\vec{b}$ is given by

$$\begin{aligned} \vec{y} &= \mathrm{proj}_{\mathrm{col}(A)}\,\vec{b} \\ &= (\vec{b}\cdot\vec{c}_1)\vec{c}_1 + (\vec{b}\cdot\vec{c}_2)\vec{c}_2 \\ &= \left(\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right)\vec{c}_1 + \left(\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}\right)\vec{c}_2 \\ &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \frac{5}{\sqrt{2}}\begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 \\ 5/2 \\ 5/2 \end{bmatrix}. \end{aligned}$$

We are then required to solve for the vector $\vec{v}$ that satisfies the system $A\vec{v} = \vec{y}$. We row reduce as follows:

$$\left[\begin{array}{ccc|c} 1 & 0 & \sqrt{2} & 1 \\ 0 & 1/\sqrt{2} & 2 & 5/2 \\ 0 & 1/\sqrt{2} & 2 & 5/2 \end{array}\right] \underset{R3-R2}{\sim} \left[\begin{array}{ccc|c} 1 & 0 & \sqrt{2} & 1 \\ 0 & 1/\sqrt{2} & 2 & 5/2 \\ 0 & 0 & 0 & 0 \end{array}\right] \sqrt{2}R2 \sim \left[\begin{array}{ccc|c} 1 & 0 & \sqrt{2} & 1 \\ 0 & 1 & 2\sqrt{2} & 5\sqrt{2}/2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Thus any vector $\vec{v}$ of the following form minimizes the distance $\|A\vec{v}-\vec{b}\|$,

$$\vec{v} = \begin{bmatrix} 1 \\ 5\sqrt{2}/2 \\ 0 \end{bmatrix} + t\begin{bmatrix} -\sqrt{2} \\ -2\sqrt{2} \\ 1 \end{bmatrix}, \quad t \in \mathbb{R}.$$

C.5 Matrix decomposition: QR factorization

In the previous example, the columns of A were not linearly independent. However, when the columns are linearly independent, we have the following result.

Theorem C.17. *If A is a matrix whose columns are linearly independent, then there exist matrices Q and R such that $A = QR$ and*

- *Q is a matrix of the same shape as A and whose columns form an orthonoromal basis for $\mathrm{col}(A)$, and,*
- *R is a square, upper triangular invertible matrix with positive entries on its diagonal.*

This theorem says that matrices Q and R exist, but does not say how to compute them. This question, of how to compute Q and R, will be answered when we prove the previous theorem, but first, we discuss an application of QR factorization.

Theorem C.18. *Suppose that A is a matrix with linearly independent columns. Write $A = QR$ be a QR factorization of A. Then, for any vector $\vec{b}$, the vector $\vec{v} = R^{-1}Q^T\vec{b}$ satisfies*

$$\|A\vec{v} - \vec{b}\| \leq \|A\vec{x} - \vec{b}\|$$

for all vectors $\vec{x}$.

This is an example of a *least squares problem* and we will discuss these further in the following section, including real-world examples of such problems. For now, we return our attention to the problem of computing Q and R.

Proof of Theorem C.17. Let A be an $m \times n$ matrix. We will construct an $m \times n$ matrix Q and an $n \times n$ matrix R that satisfy the conditions of the theorem. Write

$$A = \begin{bmatrix} | & | & & | \\ \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \\ | & | & & | \end{bmatrix},$$

so the columns of A are the vectors $\vec{c}_1, \ldots, \vec{c}_n$. Because we assumed that the columns of A are linearly independent, apply the Gram-Schmidt procedure (Theorem C.13) to $\vec{c}_1, \ldots, \vec{c}_n$ to get an orthonormal basis $\vec{v}_1, \ldots, \vec{v}_n$ for $\mathrm{col}(A)$. Let

$$Q = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix}$$

Now for each $i = 1, \ldots, n$, it is clear that $\vec{c}_i \in \mathrm{span}(\vec{c}_1, \ldots, \vec{c}_i)$. But by Gram-Schmidt, $\mathrm{span}(\vec{c}_1, \ldots, \vec{c}_i) = \mathrm{span}(\vec{v}_1, \ldots, \vec{v}_i)$, so there exist constants $r_{1,i}, \ldots, r_{i,i}$ such that

$$\vec{c}_i = r_{1,i}\,\vec{v}_1 + \cdots + r_{i,i}\vec{v}_i + 0\vec{v}_{i+1} + \cdots + 0\vec{v}_n.$$

Encode these coefficients in the vector $\vec{r}_i = (r_{1,i}, r_{2,i}, \ldots, r_{i,i}, 0, \ldots, 0)$. Then, for each i,

$$\begin{aligned}\vec{c}_i &= r_{1,i}\vec{v}_1 + \cdots + r_{i,i}\vec{v}_i + 0\vec{v}_{i+1} + \cdots + 0\vec{v}_n \\ &= \begin{bmatrix} | & & | & & | \\ \vec{v}_1 & \cdots & \vec{v}_i & \cdots & \vec{v}_n \\ | & & | & & | \end{bmatrix} \begin{bmatrix} r_{1,i} \\ \vdots \\ r_{i,i} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = Q\vec{r}_i\end{aligned}$$

Repeating this for each i, we have that

$$A = \begin{bmatrix} | & | & & | \\ \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Q\vec{r}_1 & Q\vec{r}_2 & \cdots & Q\vec{r}_n \\ | & | & & | \end{bmatrix} = QR,$$

where

$$R = \begin{bmatrix} | & | & & | \\ \vec{r}_1 & \vec{r}_2 & \cdots & \vec{r}_n \\ | & | & & | \end{bmatrix}$$

By construction, we have that Q is a matrix whose columns form an orthonormal basis of $\mathsf{col}(A)$ and that R is a square, upper triangular invertible matrix. Indeed, notice that by definition, in the i-th column of R, only the first i entries are nonzero. It remains to show that R is invertible with positive entries on the diagonal.

To see that R is invertible, we will show that all the diagonal entries are nonzero.[3] Assume towards a contradiction that there is a diagonal entry $r_{i,i}$ that is zero. That is,

$$\vec{c}_i = r_{1,i}\vec{v}_1 + \cdots + r_{i-1,i}\vec{v}_{i-1},$$

so $\vec{c}_i \in \mathsf{span}(\vec{v}_1, \ldots, \vec{v}_{i-1})$. But $\mathsf{span}(\vec{v}_1, \ldots, \vec{v}_{i-1}) = \mathsf{span}(\vec{c}_1, \ldots, \vec{c}_{i-1})$, so it follows that $\vec{c}_i \in \mathsf{span}(\vec{c}_1, \ldots, \vec{c}_{i-1})$. Yet this implies that $\vec{c}_1, \ldots, \vec{c}_{i-1}, \vec{c}_i$ is a linearly dependent list, which is a contradiction as we assumed that the columns of A were linearly independent. We conclude that R is invertible because all of the diagonal entries are nonzero.

Lastly, assume that a diagonal entry is negative, say $r_{i,i} < 0$. Then, replace $\vec{v}_i$ in the orthonormal basis with $-\vec{v}_i$ and correspondingly replace $r_{i,i}$ with $-r_{i,i}$. ■

[3] Recall that for an upper-triangular matrix, the determinant is the product of the diagonal entries. So R is invertible if and only if the product of the diagonal entries is nonzero, which occurs if and only if every diagonal entry is nonzero.

QR factorization algorithm. Suppose A is a matrix with linearly independent columns $\vec{c}_1, \ldots, \vec{c}_n$.

1. Apply Gram-Schmidt to $\vec{c}_1, \ldots, \vec{c}_n$ to produce an orthonormal basis $\vec{v}_1, \ldots, \vec{v}_n$ for $\text{col}(A)$.
2. For each i, find constants $r_{1,i} \ldots, r_{i,i}$ that satisfy
$$\vec{c}_i = r_{1,i}\,\vec{v}_1 + \ldots + r_{i,i}\,\vec{v}_i.$$
3. If $r_{i,i} < 0$, replace $\vec{v}_i$ in the orthonormal basis with $-\vec{v}_i$ and correspondingly replace $r_{i,i}$ with $-r_{i,i}$ in the above equation.
4. Set Q be the matrix with columns $\vec{v}_1, \ldots, \vec{v}_n$ and R the matrix whose (i,j)-th entry is zero if $i > j$ and $r_{i,j}$ otherwise.

C.6 Least squares problems

Recall that given a matrix A and vector $\vec{b}$, a least squares solution to $A\vec{x} = \vec{b}$ is a vector $\vec{v}$ such that $\|A\vec{v} - \vec{b}\| \leq \|A\vec{x} - \vec{b}\|$ for all $\vec{x}$. Theorem C.18 gives a solution to the least squares. In this section, we will prove this theorem and other results related to least squares problems.

Definition C.19. The **normal equations** to the system $A\vec{x} = \vec{b}$ are those specified by the system $A^TA\vec{x} = A^T\vec{b}$.

Theorem C.20. *The set of solutions $\vec{x}$ to the least squares problem $A\vec{x} \approx \vec{b}$ is exactly the solution set to the normal equations $A^TA\vec{x} = A^T\vec{b}$. Moreover, this solution set is nonempty.*

This theorem, along with QR factorization will be the main tools for the proof of Theorem C.18.

Proof. Consider the least squares problem arising from $A\vec{x} = \vec{b}$. We showed in the previous sections that there is a vector $\vec{y}$ in the column space of A that is closest to $\vec{b}$, meaning that the solution set to the least squares problem is nonempty. It remains to show that this solution also satisfies the normal equations.

Let $\vec{v}$ be the vector such that $A\vec{v} = \vec{y}$. That is, $\|A\vec{v} - \vec{b}\| < \|A\vec{x} - \vec{b}\|$ for all $\vec{x} \neq \vec{v}$. We will show that $\vec{v}$ satisfies the normal equations.

By Theorem C.12, it follows that $\vec{b} - \vec{y}$ is orthogonal to every vector in $\text{col}(A)$. In particular, $\vec{b} - \vec{y} = \vec{b} - A\vec{v}$ is orthogonal to every column $\vec{c}_i$ of A. Then, because of how the dot-product and matrix-vector multiplication are defined, we have that
$$\vec{c}_i^{\,T}(\vec{b} - A\vec{v}) = \vec{c}_i \cdot (\vec{b} - A\vec{v}) = 0.$$

Moreover, each $\vec{c}_i^T$, appearing in the left-hand side above, is a row of A^T. As a result, we have that $A^T(\vec{b} - A\vec{v}) = \vec{0}$. After distributing, we have that $A^T\vec{b} - A^TA\vec{v} = \vec{0}$, or equivalently, $A^TA\vec{v} = A^T\vec{b}$, as desired.

Each of these steps can be reversed to show that any vector satisfying the normal equations must also be a least squares solution.[4] ■

[4] For more details, see the textbook's proof of Theorem 13 in Chapter 6.5.

The following tells us when there is a unique solution to the least squares problem. The condition of the rows being linearly independent should be familiar from the previous sections, for instance in Example C.16 and Theorem C.17.

Theorem C.21. *Let A be a matrix. The following are equivalent:*

(i) for any $\vec{b}$, the equation $A\vec{x} = \vec{b}$ has a unique least squares solution;
(ii) the columns of A are linearly independent;
(iii) the matrix A^TA is invertible.

When (any of) these statements are true, the least squares solution is the vector $\vec{v} = (A^TA)^{-1}A^T\vec{b}$.

This theorem says that QR factorization and least squares problems are two sides of the same coin: we have a least squares solution if and only if we have a least squares factorization. Also, it turns out that Theorems C.18 and C.21 are saying the same thing. Recall:

Notice that if the columns of A are linearly dependent, then we cannot apply Gram-Schmidt, and thus, cannot find a QR factorization!

> **Theorem C.18.** *If A is a matrix with linearly independent columns, write $A = QR$ for its QR factorization. For any vector $\vec{b}$, the solution to the least squares problem $A\vec{x} \approx b$ is given by the vector $\vec{v} = R^{-1}Q^T\vec{b}$.*

Write $A = QR$ for the QR factorization of A and compute the matrix $(A^TA)^{-1}A^T$ in the statement of Theorem C.21:

$$\begin{aligned}(A^TA)^{-1}A^T &= \left((QR)^TQR\right)^{-1}(QR)^T \\ &= \left(R^TQ^TQR\right)^{-1}R^TQ^T \\ &= \left(R^{-1}Q^{-1}(Q^T)^{-1}(R^T)^{-1}\right)R^TQ^T \\ &= \left(R^{-1}Q^{-1}\right)\left((Q^T)^{-1}(R^T)^{-1}R^TQ^T\right) \\ &= R^{-1}Q^{-1}\end{aligned}$$

Because the columns of Q are orthonormal, it is an example of an *orthogonal matrix*. It turns out that orthorgonal matrices $\mathcal{O}$ satisfy $\mathcal{O}^{-1} = \mathcal{O}^T$. Hence, the above computation shows that the matrix $(A^TA)^{-1}A^T$ from Theorem C.21 is exactly the matrix $R^{-1}Q^T$ from Theorem C.18.

Here, we are not requiring orthogonal matrices to be square.

Proof (transpose of an orthogonal matrix is its inverse — optional). Let $\mathcal{O}$ be a 3×3 orthogonal matrix; the proof of the general case is similar. Denote by $\vec{u}, \vec{v}, \vec{w}$ the orthonormal columns of $\mathcal{O}$. To show that

$\mathcal{O}^{-1} = \mathcal{O}^T$, it suffices to show that $\mathcal{O}^T\mathcal{O} = I$, where I is the identity matrix. So,

$$\mathcal{O}^T\mathcal{O} = \begin{bmatrix} \vec{u}^T \\ \vec{v}^T \\ \vec{w}^T \end{bmatrix} \begin{bmatrix} \vec{u} & \vec{v} & \vec{w} \end{bmatrix} = \begin{bmatrix} \vec{u}^T\vec{u} & \vec{u}^T\vec{v} & \vec{u}^T\vec{w} \\ \vec{v}^T\vec{u} & \vec{v}^T\vec{v} & \vec{v}^T\vec{w} \\ \vec{w}^T\vec{u} & \vec{w}^T\vec{v} & \vec{w}^T\vec{w} \end{bmatrix}.$$

Then, because $\vec{x}^T\vec{y} = \vec{x} \cdot \vec{y}$, and using the assumption that $\vec{u}$, $\vec{v}$, and $\vec{w}$ are orthogonal, we have that

$$\mathcal{O}^T\mathcal{O} = \begin{bmatrix} \|\vec{u}\|^2 & 0 & 0 \\ 0 & \|\vec{v}\|^2 & 0 \\ 0 & 0 & \|\vec{w}\|^2 \end{bmatrix} = I,$$

as $\vec{u}$, $\vec{v}$, and $\vec{w}$ are orthonormal, so have unit length. ■

QR factorization, revisited

In Section C.5 we learnt how to compute a QR factorization of a given matrix A. This involved manually computing the entries of R. However, because we can now easily compute Q^{-1}, we have a second method to find R.

QR factorization algorithm (version 2). Suppose A is a matrix with linearly independent columns $\vec{c}_1, \ldots, \vec{c}_n$.

1. Apply Gram-Schmidt to $\vec{c}_1, \ldots, \vec{c}_n$ to produce an orthonormal basis $\vec{v}_1, \ldots, \vec{v}_n$ for $\text{col}(A)$.
2. Let Q be the matrix whose columns are $\vec{v}_1, \ldots, \vec{v}_n$.
3. Set $R = Q^T A$.
4. If a diagonal entry $r_{i,i}$ of R is negative, replace it with $-r_{i,i}$ and multiply the i^{th} column of Q by -1.

C.7 Lines of best fit

Consider the following dataset:

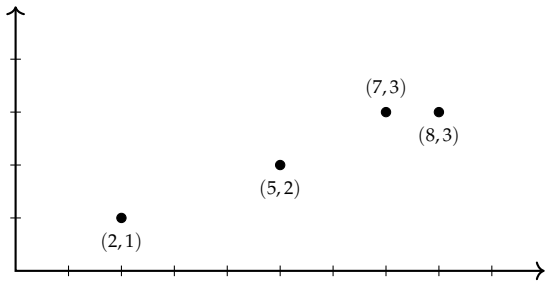


Figure C.2: A sample dataset.

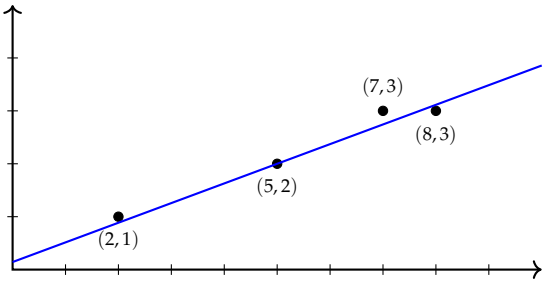


Figure C.3: A sample dataset with a *line of best fit*.

The data appears to lie approximately on a straight line, something akin to the following.

Suppose we have a line $y = \beta_0 + \beta_1 x$ that approximates our data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. That is, we have the system of approximations,

$$\begin{cases} \beta_0 + \beta_1 x_1 \approx y_1 \\ \beta_0 + \beta_1 x_2 \approx y_2 \\ \quad \vdots \\ \beta_0 + \beta_1 x_m \approx y_m \end{cases} \iff \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \approx \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

This now looks like a least squares problem! We wish to find the matrix $\vec{\beta} = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}^T$ that gives the best approximation to the system $X\vec{\beta} \approx \vec{y}$, where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \quad \text{and} \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

We will solve for $\vec{\beta}$ using the normal equations and Theorem C.20. That is, the desired vector $\vec{\beta}$ is given by the solution to the normal equations $X^T X\vec{\beta} = X^T\vec{y}$. We compute:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix},$$

and,

$$X^T\vec{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

We then encode the normal equations $X^T X\vec{\beta} = X^T\vec{y}$ in the following

augmented matrix and row reduce,

$$\left[\begin{array}{cc|c} 4 & 22 & 9 \\ 22 & 142 & 57 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 11/2 & 9/4 \\ 22 & 142 & 57 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 11/2 & 9/4 \\ 0 & 21 & 15/2 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 11/2 & 9/4 \\ 0 & 1 & 5/14 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 2/7 \\ 0 & 1 & 5/14 \end{array}\right].$$

We conclude that when $\beta_0 = 2/7$ and $\beta_1 = 5/14$, the line of best fit is given by $y = \beta_0 + \beta_1 x = 2/7 + (5/14)x$.

In Stats 3A03, you learn many generalizations of this least-squares approach for lines of best fit.

Exercise 20. Use QR factorization instead to solve for $\vec{\beta}$ in the above example. Verify that you get the same answer.

D
Constrained Optimization

In this chapter, we continue with optimization problems. The problems that we will discuss here are of the following form:

$$\begin{aligned} &\texttt{maximize} && \vec{x}^T A \vec{x} \\ &\texttt{subject to} && \|\vec{x}\| = 1 \end{aligned}$$

where A is a symmetric matrix. Notice that the condition that $\|\vec{x}\| = 1$ says that the vector $\vec{x}$ must lie on the unit circle in $\mathbb{R}^2$, or the unit sphere in $\mathbb{R}^3$, and so on in higher dimensions.

Recall that a matrix A is **symmetric** if $A = A^T$.

If A is diagonal, then the answer is straightforward.

Example D.1. Let $A = \begin{bmatrix} 3 & 0 \\ 0 & 7 \end{bmatrix}$ and find the maximal value of $\vec{v}^T A \vec{v}$ subject to the constraint that $\|\vec{v}\| = 1$. Write $\vec{v} = (x, y)$ (a column vector) so the objective function is given as follows:

$$\vec{v}^T A \vec{v} = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 3x^2 + 7y^2.$$

Then,

$$3x^2 + 7y^2 \leq 7x^2 + 7y^2 = 7(x^2 + y^2) = 7,$$

where for the last step, we have used the fact that $\|\vec{v}\| = 1$ so $\|\vec{v}\|^2 = 1$ also. This gives us an upper bound on the objective function. Now, notice that $\vec{v} = (0, 1)$ satisfies $\|\vec{v}\| = 1$ and $3(0)^2 + 7(1)^2 = 7$, so the vector $(0, 1)$ maximizes the objective function, and we are done.

Exercise 21. Notice that the maximum value in Example D.1 is exactly the largest eigenvalue of the matrix A. Given a generic diagonal matrix $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, argue that the maximal value of $\vec{x}^T D \vec{x}$, for any $\vec{x} \in \mathbb{R}^n$ satisfying $\|\vec{x}\| = 1$, is the largest eigenvalue.

By largest eigenvalue, we mean the most positive eigenvalue, not the eigenvalue with the largest absolute value.

In the rest of this chapter, we discuss solutions to optimization problems of this form when A is a symmetric—but not necessarily diagonal—matrix. This chapter is organized as follows: In Section D.1, we

discuss properties of symmetric matrices and in Section D.2 we discuss the objective functions of these optimization problems, called quadratic forms. We conclude the chapter in Section D.3 with the solutions to the optimization problems.

D.1 Matrix decomposition: Orthogonal diagonalization

We begin by discussing properties of symmetric matrices. In Chapter B, Theorem B.11 told us that an $n \times n$ matrix is diagonalizable if and only if it has n linearly independent eigenvectors. It turns out that symmetric matrices are always diagonalizable and, in fact, something stronger is true.

Recall from Section C.6 that a matrix P is **orthogonal** if its columns are orthonormal. Also recall from the same section that orthogonal matrices P satisfy $P^{-1} = P^T$.

Definition D.2. The square matrix A is **orthogonally diagonalizable** if there is a diagonal matrix D and orthogonal matrix P such that $A = PDP^T$.

The set of eigenvalues of a matrix is sometimes called the *spectrum*, hence the name, *Spectral Theorem*.

Theorem D.3 (Spectral Theorem)**.** *Let A be a symmetric matrix. Then,*

(i) A is orthogonally diagonalizable,[1]

(ii) Eigenvectors of A are orthogonal when they correspond to different eigenvalues,

(iii) Every eigenvalue of A is real,[2]

(iv) The dimension of each eigenspace equals the multiplicity of the corresponding root in the characteristic polynomial.

[1] As you will argue on Assignment 4, it turns out that a matrix A is symmetric if and only if it is orthogonally diagonalizable.

[2] That is, none of the eigenvalues are complex numbers.

Proof of (ii). Let $\vec{v}_1$ and $\vec{v}_2$ be eigenvectors corresponding to distinct eigenvalues λ_1 and λ_2. We are required to show that $\vec{v}_1 \cdot \vec{v}_2 = 0$. Because $\lambda_1 \neq \lambda_2$, we have that $\lambda_1 - \lambda_2 \neq 0$, so it suffices to show that $(\lambda_1 - \lambda_2)(\vec{v}_1 \cdot \vec{v}_2) = 0$. This is equivalent to show that

$$\lambda_1(\vec{v}_1 \cdot \vec{v}_2) = \lambda_2(\vec{v}_1 \cdot \vec{v}_2).$$

So we compute:

$$\begin{aligned}\lambda_1(\vec{v}_1 \cdot \vec{v}_2) &= (\lambda_1\vec{v}_1) \cdot \vec{v}_2 = (\lambda_1\vec{v}_1)^T\vec{v}_2 = (A\vec{v}_1)^T\vec{v}_2 \\ &= \vec{v}_1^T(A^T\vec{v}_2) = \vec{v}_1^T(A\vec{v}_2),\end{aligned}$$

where we use the fact that A is symmetric. Continuing,

$$= \vec{v}_1^T(\lambda_2\vec{v}_2) = \lambda_2(\vec{v}_1^T\vec{v}_2) = \lambda_2(\vec{v}_1 \cdot \vec{v}_2),$$

as desired. ■

Exercise 22. Complete Exercises 27 and 28 from Chapter 5.5 of the course textbook to provide justification for item (iii) of the Spectral Theorem.

Example D.4. We will orthogonally diagonalize the symmetric matrix

$$A = \begin{bmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{bmatrix},$$

with characteristic polynomial $-(\lambda-7)^2(\lambda+2)$. An eigenvector associated to $\lambda = -2$ is $\begin{bmatrix} -1 \\ -1/2 \\ 1 \end{bmatrix}$ and eigenvectors associated to $\lambda = 7$ are $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1/2 \\ 1 \\ 0 \end{bmatrix}$.

Notice that we have two linearly independent eigenvalues associated to the eigenvalue $\lambda = 7$. The corresponding root in the characteristic polynomial has degree two, as predicted by the Spectral Theorem.

An orthonormal basis for the eigenspace corresponding to $\lambda = -2$ is given by normalizing the above vector,

$$\|(-1, -1/2, 1)\| = \sqrt{1 + \frac{1}{4} + 1} = \frac{3}{2},$$

$$\frac{1}{3/2}\begin{bmatrix} -1 \\ -1/2 \\ 1 \end{bmatrix} = \begin{bmatrix} -2/3 \\ -1/3 \\ 2/3 \end{bmatrix}.$$

Meanwhile, applying Gram-Schmidt to the linearly independent eigenvectors corresponding to $\lambda = 7$ yields the orthonormal list

$$\begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}, \quad \begin{bmatrix} -1/\sqrt{18} \\ 4/\sqrt{18} \\ 1/\sqrt{18} \end{bmatrix}.$$

We conclude that

$$A = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{18} & -2/3 \\ 0 & 4/\sqrt{18} & -1/3 \\ 1/\sqrt{2} & 1/\sqrt{18} & 2/3 \end{bmatrix} \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{18} & -2/3 \\ 0 & 4/\sqrt{18} & -1/3 \\ 1/\sqrt{2} & 1/\sqrt{18} & 2/3 \end{bmatrix}^T$$

$$= \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{18} & -2/3 \\ 0 & 4/\sqrt{18} & -1/3 \\ 1/\sqrt{2} & 1/\sqrt{18} & 2/3 \end{bmatrix} \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{18} & 4/\sqrt{18} & 1/\sqrt{18} \\ -2/3 & -1/3 & 2/3 \end{bmatrix}$$

$$= PDP^T$$

is the orthogonal diagonalization of A.

The procedure outlined in the above example produces an orthogonal diagonalization for any symmetric matrix.

Orthogonal diagonalization procedure for symmetric matrices. Let A be a symmetric matrix.

1. Find the eigenvalues of A.
2. For each eigenvalue λ_i, find d_i linearly independent associated eigenvectors, where d_i is the multiplicity of the root in the characteristic polynomial corresponding to λ_i.
3. Apply Gram-Schmidt (Theorem C.13) to each linearly independent list to produce an orthonormal basis for each eigenspace.
4. Use these orthonormal lists to construct the matrices P and D in the normal diagonalization way.

Because of item (ii), it follows that vectors from one eigenspace are orthogonal to the vectors in other eigenspaces, so P is an orthogonal matrix.

D.2 Quadratic forms

In this section we will discuss quadratic forms, the family of objective functions in this chapter's optimization problems. They earn this name as each term in a quadratic form has degree 2.

Definition D.5. A function $Q : \mathbb{R}^n \to \mathbb{R}$ is a **quadratic form** if there is a symmetric $n \times n$ matrix A such that $Q(\vec{x}) = \vec{x}^T A\vec{x}$ for all $\vec{x} \in \mathbb{R}^n$. We call A the **matrix of the quadratic form**.

Exercise 23. Show that each of the following functions Q are quadratic forms on $\mathbb{R}^3$ by, for each, finding a matrix A such that $Q = \vec{v}^T A\vec{v}$, where $\vec{v} = (x, y, z)$.

(a) $5x^2 + 3y^2 + 2z^2 - xy + 8yz$

(b) xz

(c) $x^2 + 11xy + 121y^2$

Give an example of a function $P : \mathbb{R}^2 \to \mathbb{R}$ that is not a quadratic form.

We saw in Example D.1 that our optimization problems are much simpler when the matrix of the quadratic form is diagonal. In this case, we say that the quadratic form has **no cross-product terms**. That is, if $\vec{x} = (x_1, \ldots, x_n)$, then the quadratic form $\vec{x}^T A\vec{x}$ has no cross-product terms if no monomial $x_i x_j$ appears in $\vec{x}^T A\vec{x}$ for which $i \neq j$.

Definition D.6. A **change of variables** is an equation of the form $\vec{x} = P\vec{y}$ where P is an invertible matrix.

In the above definition, the matrix P changes from $\vec{y}$ variables to $\vec{x}$ variables. The following theorem says that we can perform a change of variables from any quadratic form to one whose matrix is diagonal!

Theorem D.7 (Principal Axes Theorem)**.** *Let A be symmetric. There exists an orthogonal matrix P whose change of coordinates $\vec{x} = P\vec{y}$ transforms the quadratic form $\vec{x}^T A\vec{x} = \vec{y}^T D\vec{y}$, and the right-hand side has no cross-product terms.*

Recall that orthogonal matrices are necessarily invertible; see page 47.

Definition D.8. In the above theorem, the columns of P are called the **principal axes** of the quadratic form $\vec{x}^T A\vec{x}$.

Proof of Theorem D.7. Because A is symmetric, it is orthogonally diagonalizable, $A = PDP^T$. Let $\vec{x} = P\vec{y}$, or equivalently, $\vec{y} = P^T\vec{x}$. Then,

$$\vec{x}^T A\vec{x} = (P\vec{y})^T A(P\vec{y}) = \vec{y}^T(P^T AP)\vec{y} = \vec{y}^T D\vec{y},$$

where we use the fact that, because $A = PDP^T$, we have $D = P^T AP$. ■

D.3 Optimal values of quadratic forms

The tools in the previous section give us the following answer to the constrained optimization problem.

Theorem D.9. *Let A be symmetric and $Q(\vec{x}) = \vec{x}^T A\vec{x}$. Denote by $\lambda_1 \geq \lambda \geq \cdots \geq \lambda_n$ the eigenvalues of A. Then,*

$$\max\{Q(\vec{x}) \mid \|\vec{x}\| = 1\} = \lambda_1$$

and

$$\min\{Q(\vec{x}) \mid \|\vec{x}\| = 1\} = \lambda_n$$

In particular, we have $\lambda_1 = Q(\vec{v})$ where $\vec{v}$ is a unit eigenvector corresponding to λ_1, and similarly, $\lambda_n = Q(\vec{u})$ for a unit eigenvector $\vec{u}$ corresponding to λ_n.

That is, the solutions to the constrained maximization and minimization problems are the largest and smallest eigenvalues, respectively, and are attained at their corresponding unit eigenvectors.

Proof. Assume that A is a 3×3 matrix with eigenvalues $a \geq b \geq c$ and so orthogonally diagonalizes as $A = PDP^{-1}$ where

The argument is the same for larger matrices, but with more cumbersome notation.

$$P = \begin{bmatrix} | & | & | \\ \vec{u}_1 & \vec{u}_2 & \vec{u}_3 \\ | & | & | \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}.$$

Here, the vectors $\vec{u}_i$ are orthonormal eigenvectors associated to the eigenvalues a, b, and c. Because $a \geq b \geq c$, for any unit vector $\vec{y} = (y_1, y_2, y_3)$ we must have

$$\vec{y}^T D\vec{y} = ay_1^2 + by_2^2 + cy_3^2 \leq ay_1^2 + ay_2^2 + ay_3^2 = a(y_1^2 + y_2^2 + y_3^2) = a.$$

Hence $M = \max\{\vec{y}^T D\vec{y} \mid \|\vec{y}\| = 1\} \leq a$. Moreover, this upper bound is attained by plugging in $\vec{e}_1 = (1,0,0)$ for $\vec{y}$.

Returning to the $\vec{x}$'s, we have

$$M = a = \vec{e}_1^T D\vec{e}_1 = \vec{e}_1^T (P^T AP)\vec{e}_1 = (P\vec{e}_1)^T A(P\vec{e}_1) = \vec{u}_1^T A\vec{u}_1.$$

The minimization proof is similar. ■

Example D.10. We will find the maximal value of the quadratic form $Q(\vec{x}) = \vec{x}^T A\vec{x}$ when $\|\vec{x}\| = 1$ and

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}.$$

The characteristic polynomial of A is $(1-\lambda)(3-\lambda)(6-\lambda)$ so the eigenvalues are 6, 3, and 1. Hence the maximal value of the quadratic form $Q(\vec{x})$ is 6 and is attained when $\vec{x}$ is a unit eigenvector associated to $\lambda = 6$. Eigenvectors corresponding to $\lambda = 6$ satisfy

$$\left[\begin{array}{ccc|c} -3 & 2 & 1 & 0 \\ 2 & -3 & 1 & 0 \\ 1 & 1 & -2 & 0 \end{array}\right] \sim \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and so are of the form $\vec{x} = k(1,1,1)$. For $\vec{x}$ to be a unit vector, we take $\vec{x} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$. We check our answer; the quadratic form is

$$Q(x,y,z) = 3x^2 + 3y^2 + 4z^2 + 4xy + 2xz + 2yz$$

and so,

$$\begin{aligned} Q(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}) &= 3\left(\frac{1}{\sqrt{3}}\right)^2 + 3\left(\frac{1}{\sqrt{3}}\right)^2 + 4\left(\frac{1}{\sqrt{3}}\right)^2 + 4\left(\frac{1}{\sqrt{3}}\right)\left(\frac{1}{\sqrt{3}}\right) + 2\left(\frac{1}{\sqrt{3}}\right)\left(\frac{1}{\sqrt{3}}\right) + 2\left(\frac{1}{\sqrt{3}}\right)\left(\frac{1}{\sqrt{3}}\right) \\ &= 1 + 1 + \frac{4}{3} + \frac{4}{3} + \frac{2}{3} + \frac{2}{3} \\ &= 2 + \frac{12}{3} \\ &= 6, \end{aligned}$$

as expected.

Exercise 24. In the previous example, find the minimal value of the quadratic form and a vector $\vec{v}$ that attains this minimum.

We conclude this section with a modification of Theorem D.9 to a setting with additional constraints.

Theorem D.11. *Let $Q(\vec{x})$ be a quadratic form with associated symmetric $n \times n$ matrix A. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the eigenvalues of A and for each i, denote by $\vec{u}_i$ a unit eigenvector corresponding to λ_i.*

Then, the maximal value of $Q(\vec{x})$, subject to the constraints $\|\vec{x}\| = 1$ and $\vec{x} \cdot \vec{u}_1 = 0, \ldots, \vec{x} \cdot \vec{u}_k = 0$, is given by $\lambda_{k+1} = Q(\vec{u}_{k+1})$.

Exercise 25. Find the maximal value of the quadratic form $Q(\vec{x})$ given in Example D.10 subject to the constraints that $\|\vec{x}\| = 1$ and $\vec{x} \cdot (1,1,1) = 0$.

D.4 Classification of quadratic forms

Definition D.12. A quadratic form $Q(\vec{x})$ is called

(i) **positive definite** if $Q(\vec{x}) > 0$ for all $\vec{x} \neq \vec{0}$,
(ii) **positive semidefinite** if $Q(\vec{x}) \geq 0$ for all $\vec{x} \neq 0$.

Any quadratic form that is positive definite is also positive semidefinite. For example, $Q(x,y) = x^2 + y^2$ is both positive definite and positive semidefinite. However, the surface $\widetilde{Q}(x,y) = x^2$ is positive semidefinite but not positive definite, because $\widetilde{Q}(0,1) = 0$.

We have analogous negative versions of positive (semi)definiteness: A quadratic form $Q(\vec{x})$ is called

(iii) **negative definite** if $Q(\vec{x}) < 0$ for all $\vec{x} \neq \vec{0}$,
(iv) **negative semidefinite** if $Q(\vec{x}) \leq 0$ for all $\vec{x} \neq \vec{0}$.

If a quadratic form $Q(\vec{x})$ takes on both positive and negative values then we say that $Q(x)$ is **indefinite**.

Exercise 26. Explain why the quadratic form $z(x,y) = x^2 - y^2$ is indefinite. Sketch this surface as well as the examples given within Definition D.12.

It can be helpful to visualize a quadratic form that we are working with, and the classification above tells us what simpler surface (e.g., the paraboloid) "looks like" our quadratic form

In the examples we have seen so far, it has been relatively easy to classify However, this is not always the case.

Example D.13. The quadratic form

$$Q(x,y,z) = 3x^2 + 2y^2 + z^2 + 4xy + 4yz.$$

looks like it should be positive definite, or at least positive semidefinite, because of all the +'s. However, notice that

$$\begin{aligned} Q(1,-2,2) &= 3(1)^2 + 2(-2)^2 + 2^2 + 4(-2) + 4(-2)(2) \\ &= 3 + 8 + 4 - 8 - 16 = 7 - 16 \\ &= -9, \end{aligned}$$

so $Q(x,y,z)$ is **indefinite**.

The choice of $(1,-2,2)$ appeared to come out of nowhere, but there was a trick to find it! Denote by A the matrix associated to the above

quadratic form $Q(\vec{x}) = \vec{x}^T A\vec{x}$. Then, you can check that -1 is an eigenvalue of A with eigenvector $(1,-2,2)$. That $Q(x,y,z)$ takes on negative values was detected by the negative eigenvalue.

Theorem D.14. *If Q is a quadratic form associated to symmetric matrix A, then Q is...*

(i) positive definite if and only if all the eigenvalues of A are positive;
(ii) positive semidefinite if and only if all the eigenvalues of A are nonnegative;
(iii) negative definite if and only if all the eigenvalues of A are negative;
(iv) negative semidefinite if and only if all the eigenvalues of A are nonpositive;
(v) indefinite if and only if A has positive and negative eigenvalues.

Sketch of proof. Use the Principal Axes Theorem (Theorem D.7) to perform a change of coordinates such that $\vec{x}^T A\vec{x} = \vec{y}^T D\vec{y}$ so

$$Q(\vec{x}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$

The result then follows because the vectors $\vec{y}$ are in one-to-one correspondence with the vectors $\vec{x}$. ■

Exercise 27. Classify the quadratic forms from Exercise 23.

E
Three Applications of Singular Value Decomposition

We have seen matrix decompositions a few time throughout this course: diagonalization, QR factorization, and orthogonal diagonalization. In this chapter, we add another decomposition to our toolkit in singular value decomposition (SVD). We also discuss its applications to least squares problems and image compression.

E.1 Matrix decomposition: Singular value decomposition

If A is an $m \times n$ matrix, then A^TA is necessarily symmetric. Indeed,

$$(A^TA)^T = A^T(A^T)^T = A^TA.$$

Thus, by the results in the last section, we know that A^TA can be orthogonally diagonalized. That is, there are is an orthonormal basis of $\mathbb{R}^n$, say $\vec{v}_1, \ldots, \vec{v}_n$, consisting of eigenvectors of A^TA. Denote by $\lambda_1, \ldots, \lambda_n$ the eigenvalues associated to $\vec{v}_1, \ldots, \vec{v}_n$. Then, notice that for each $i = 1, \ldots, n$,

$$\begin{aligned} \|A\vec{v}_i\|^2 &= (A\vec{v}_i) \cdot (A\vec{v}_i) = (A\vec{v}_i)^T(A\vec{v}_i) \\ &= \vec{v}_i^T(A^TA)\vec{v}_i = \vec{v}_i^T(\lambda_i\vec{v}_i) = \lambda_i(\vec{v}_i^T \cdot \vec{v}_i) \\ &= \lambda_i, \end{aligned}$$

where the last equality holds because $\vec{v}_i$ is a unit vector.[1] Because $\|A\vec{v}_i\|^2 \geq 0$, it follows that each eigenvalue λ_i is also nonnegative.

[1] Recall that $\vec{v}_1, \ldots, \vec{v}_n$ is an orthonormal list.

Definition E.1. The **singular values** $\sigma_1, \ldots, \sigma_n$ of a matrix A are the square roots of the eigenvalues of A^TA, in decreasing order.

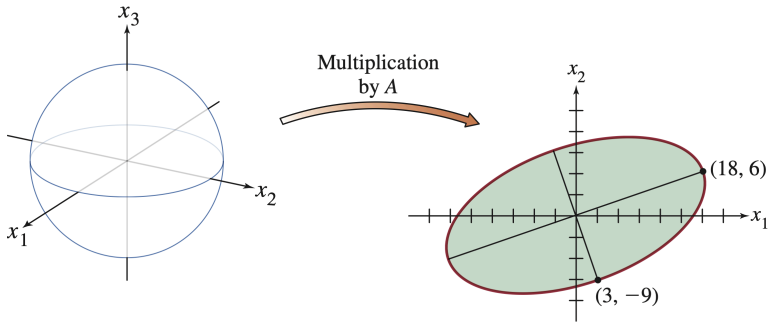
That is, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A^TA, then $\sigma_i = \sqrt{\lambda_i}$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

Remark E.2. From the above computation, we have that

$$\sigma_i = \sqrt{\lambda_i} = \sqrt{\|A\vec{v}_i\|^2} = \|A\vec{v}_i\|.$$

That is, the i-th singular value of A is the length of the vector $\|A\vec{v}_i\|$.

Example E.3. Let $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$. This represents a linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^2$. Suppose that we restrict the input of our linear transformation to only the vectors $\vec{x} \in \mathbb{R}^3$ satisfying $\|\vec{x}\| = 1$, that is, those vectors lying on the unit sphere. Multiplying by A any vector on the unit sphere yields an ellipse in $\mathbb{R}^2$.



Using the results of last chapter, we can find the maximal value of $\|A\vec{x}\|$ when $\vec{x}$ is a unit vector. Indeed,

- The unit vector $\vec{x}$ giving the maximal value of $\|A\vec{x}\|$ will also give the maximal value of $\|A\vec{x}\|^2$;
- $\|A\vec{x}\|^2 = \vec{x}^T(A^TA)x^T$, and the right-hand side is a quadratic form.

So, Theorem D.9 says that the maximum of $\|A\vec{x}\|^2$ is the largest eigenvalue of A^TA. Equivalently, we have that the maximum of $\|A\vec{x}\|$ is the largest singular value of A. We now compute the singular values of A. First, A^TA is given by

$$A^TA = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix} \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}$$

which has eigenvalues $\lambda_1 = 360$, $\lambda_2 = 90$, and $\lambda_3 = 0$ and corresponding unit eigenvectors

$$\vec{v}_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} -2/3 \\ -1/3 \\ 2/3 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 2/3 \\ -2/3 \\ 1/3 \end{bmatrix}.$$

Thus the maximal value of $\|A\vec{x}\|$, among unit vectors $\vec{x}$ is $\sigma_1 = \sqrt{\lambda_1} = \sqrt{360} = 6\sqrt{10}$ and is attained when $\vec{x} = \vec{v}_1$. Notice that

$$A\vec{v}_1 = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix} = \begin{bmatrix} 18 \\ 6 \end{bmatrix},$$

which is exactly a point on the ellipse furthest away from the origin.

Exercise 28. Use Theorem D.11 conclude that the unit vector $\vec{u}$ that maximizes $\|A\vec{u}\|$ subject to $\vec{u} \cdot \vec{v}_1$ is $\vec{u} = \vec{v}_2$.

Notice that $A\vec{v}_1$ and $A\vec{v}_2$, the vectors corresponding to the maximal lengths in the example and exercise, are orthogonal. This is no accident, as we record in the following result.

Theorem E.4. *Let A be a matrix, so A^TA is symmetric with orthonormal eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ corresponding to eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Suppose that $\lambda_{r+1} = \cdots = \lambda_n = 0$. Then, $A\vec{v}_1, \ldots, A\vec{v}_r$ is an orthogonal basis for $\mathsf{col}(A)$ and $\mathsf{rank}(A) = r$.*

That $\lambda_{r+1} = \cdots = \lambda_n = 0$ says that we have r nonzero eigenvalues. Or equivalently, that we have r nonzero singular values.

Recall that $\mathsf{col}(A)$ is the vector subspace given by the span of the columns of A. The rank of A is the dimension of this subspace and we denote it by $\mathsf{rank}(A)$. Equivalently, it is the number of linearly independent columns of A.

Proof. To see that $A\vec{v}_i$ and $A\vec{v}_j$ are orthogonal, compute

$$(A\vec{v}_i) \cdot (A\vec{v}_j) = \vec{v}_i^T(A^TA)\vec{v}_j = \lambda_j(\vec{v}_i \cdot \vec{v}_j)$$

and recall that $\vec{v}_i$ and $\vec{v}_j$ were defined to be orthonormal. Moreover for any $i = 1, \ldots, r$, because $A\vec{v}_i = \lambda_i\vec{v}_i$ and $\lambda_i \neq 0$, it follows that $A\vec{v}_1, \ldots, A\vec{v}_r$ is linearly independent.[2] ■

[2] We omit the proof that this list also spans $\mathsf{col}(A)$ and, hence, is a basis for $\mathsf{col}(A)$. If you are interested, read the proof of Theorem 9 in Section 7.4 of the course textbook.

Theorem E.5 (Singular Value Decomposition)**.** *Let A be an $m \times n$ matrix with rank r. Then, there exist matrices U, V, and Σ such that $A = U\Sigma V^T$ and,*

- *Σ is an $m \times n$ matrix whose diagonal entries are $\sigma_1, \ldots, \sigma_r, 0, \ldots, 0$,*
- *U is an $m \times m$ orthogonal matrix, and*
- *V is an $n \times n$ orthogonal matrix.*

The columns of U are called the **left singular vectors** of A and the columns of V are the **right singular vectors**.

Singular Value Decomposition (SVD) Procedure.

1. Orthogonally diagonalize $A^TA = P\widetilde{D}P^T$ using the method in Section D.1, where the diagonal entries of D are listed in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.
2. Construct Σ as the $m \times n$ matrix whose diagonal entries are the nonzero singular values $\sigma_1 \geq \cdots \geq \sigma_r$ and has zeros elsewhere.
3. Set $V = P$.
4. Construct an orthonormal basis $\vec{u}_1, \ldots, \vec{u}_r$ for $\mathsf{col}(A)$ by normalizing each $A\vec{v}_i$. That is, $\vec{u}_i = \frac{1}{\|A\vec{v}_i\|}(A\vec{v}_i) = \frac{1}{\sigma_i}(A\vec{v}_i)$.
5. Extend $\vec{u}_1, \ldots, \vec{u}_r$ to an orthonormal basis $\vec{u}_1, \ldots, \vec{u}_m$ for $\mathbb{R}^m$. Define U to be the matrix whose i-th column is $\vec{u}_i$.

The final step is the most difficult as it is less algorithmic; we need to find the right vectors to include to extend our list to an orthonormal

basis for $\mathbb{R}^m$. In general, we need to extend the list to a basis of $\mathbb{R}^m$ by including linearly independent vectors, and then applying Gram-Schmidt.

Example E.6. We will find a singular value decomposition of

$$A = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

Here is a third, yet less efficient, method of finding U. We find V by orthogonally diagonalizing $A^TA = P\widetilde{D}P^T$ and setting $V = P$. Just as A^TA is symmetric, so is AA^T, and it turns out that a suitable choice for U is the orthogonal matrix P' that orthogonally diagonalizes $AA^T = P'\widetilde{D}'(P')^T$.

First, we have that

$$A^TA = \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix}$$

which has eigenvalues $\lambda_1 = 18$ and $\lambda_2 = 0$. Unit vectors corresponding to λ_1 and λ_2 are, respectively,

$$\vec{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \vec{v}_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

The singular values are $\sigma_1 = \sqrt{\lambda_1} = \sqrt{18} = 3\sqrt{2}$ and $\sigma_2 = 0$. We thus have that

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$V = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Our next step is to start with $A\vec{v}_1, A\vec{v}_2$ and transform it into an orthonormal basis for $\mathbb{R}^3$.

$$A\vec{v}_1 = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 2/\sqrt{2} \\ -4/\sqrt{2} \\ 4/\sqrt{2} \end{bmatrix}$$

$$A\vec{v}_2 = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

So our first column of U is

$$\vec{u}_1 = \frac{1}{\sigma_1} A\vec{v}_1 = \frac{1}{3\sqrt{2}} \begin{bmatrix} 2/\sqrt{2} \\ -4/\sqrt{2} \\ 4/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix}.$$

We are required to extend this to an orthonormal basis $\vec{u}_1, \vec{u}_2, \vec{u}_3$ of $\mathbb{R}^3$. Vectors $\vec{x} = (x_1, x_2, x_3)$ that are orthogonal to $\vec{u}_1$ must satisfy

$$\begin{aligned} \vec{u}_1 \cdot \vec{x} = 0 &\iff \frac{1}{3}x_1 - \frac{2}{3}x_2 + \frac{2}{3}x_3 = 0 \\ &\iff x_1 - 2x_2 + 2x_3 = 0, \end{aligned}$$

so a basis for the subspace of vectors orthogonal to $\vec{u}_1$ is given by

$$\vec{w}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{w}_2 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}.$$

Applying Gram-Schmidt to the list $\vec{u}_1, \vec{w}_2, \vec{w}_3$ yields the orthonormal list $\vec{u}_1, \vec{u}_2, \vec{u}_3$, where

$$\vec{u}_1 = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{bmatrix}, \quad \text{and} \quad \vec{u}_3 = \begin{bmatrix} -2/\sqrt{45} \\ 4/\sqrt{45} \\ 5/\sqrt{45} \end{bmatrix}.$$

We take these vectors to be the columns of U. We thus have the following singular value decomposition:

$$A = U\Sigma V^T = \begin{bmatrix} 1/3 & 2/\sqrt{5} & -2/\sqrt{45} \\ -2/3 & 1/\sqrt{5} & 4/\sqrt{45} \\ 2/3 & 0 & 5/\sqrt{45} \end{bmatrix} \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Exercise 29. Find a singular value decomposition of the matrix in Example E.3.

E.2 Pseudoinverses and least squares problems

For the remainder of this chapter, we will discuss applications of the singular value decomposition $A = U\Sigma V^T$. We start by revisiting least squares problems. Recall that given a matrix A and vector b, the least squares problem asks to find the vector $\vec{x}$ so that $A\vec{x}$ best approximates $\vec{b}$. That is, the vector $\vec{x}$ that minimizes the norm $\|A\vec{x} - \vec{b}\|$.

Our main result is that a singular value decomposition of A gives rise to a *pseudoinverse* of A, denoted by A^+. Suppose that A is an $m \times n$ matrix and has rank r, so A has r nonzero singular values. Then we can write Σ as the block matrix

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix},$$

where $D = \mathsf{diag}(\sigma_1, \ldots, \sigma_r)$ is an $r \times r$ diagonal matrix. Then,

$$A = U\Sigma V^T = \begin{bmatrix} | & | & & | \\ \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} — & \vec{v}_1^T & — \\ — & \vec{v}_2^T & — \\ & \vdots & \\ — & \vec{v}_n^T & — \end{bmatrix}.$$

Because of the 0's in Σ, the columns $\vec{u}_{r+1}, \ldots, \vec{u}_m$ of U and rows $\vec{v}_{r+1}, \ldots, \vec{v}_n$ of V^T do not contribute anything to the product $U\Sigma V^T$. That is, we

The pseudoinverse A^+ of A satisfies $A = AA^+A$ and $A^+ = A^+AA^+$. In the right-hand sides of these equations, we a factor of A^+A behaving like the identity matrix. If A is invertible, then $A^+ = A^{-1}$, however, these equations hold even when A is not invertible.

can equivalently write

$$A = U\Sigma V^T = \begin{bmatrix} | & | & & | \\ \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} — & \vec{v}_1^T & — \\ — & \vec{v}_2^T & — \\ & \vdots & \\ — & \vec{v}_n^T & — \end{bmatrix}$$

$$= \begin{bmatrix} | & & | & | & & | \\ \vec{u}_1 & \cdots & \vec{u}_r & \vec{0} & \cdots & \vec{0} \\ | & & | & | & & | \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} — & \vec{v}_1^T & — \\ & \vdots & \\ — & \vec{v}_r^T & — \\ — & \vec{0} & — \\ & \vdots & \\ — & \vec{0} & — \end{bmatrix}.$$

Denote by U_r the $m \times r$ matrix obtained by taking the first r columns of U. Similarly, denote by V_r^T the $r \times n$ matrix obtained by taking the first r rows of V^T. The computation above shows that

$$A = U_r D V_r^T.$$

Definition E.7. With notation as above, the **pseudoinverse** of A is

$$A^+ := V_r D^{-1} U_r^T.$$

The psudoeinverse is also known as the *Moore-Penrose inverse*, named after a pair of the mathematicians discovered it. Independently of one another, E. Hastings Moore (American) discovered it in 1920, Arne Bjerhammar (Swedish) in 1951, and Roger Penrose (British) in 1955.

Because U_r and V_r^T are not square matrices, they do not admit inverses. However, because U and V^T are orthogonal matrices, we have that $U^{-1} = U^T$ and $(V^T)^{-1} = V$, which alludes to the name *pseudoinverse*.

Theorem E.8. *A vector $\vec{x}_*$ that solves the least squares problem $A\vec{x} \approx \vec{b}$ is given by $\vec{x}_* = A^+\vec{b}$.*

Proof. Compute:

$$\begin{aligned} A\vec{x}_* &= A(A^+\vec{b}) = (U_r D V_r^T)(V_r D^{-1} U_r^T \vec{b}) \\ &= (U_r D)(V_r^T V_r)(D^{-1} U_r^T)\vec{b}. \end{aligned}$$

Because V is orthogonal, it follows that $V_r^T V_r$ is the $r \times r$ identity matrix. Thus,

$$= U_r (D D^{-1}) U_r^T \vec{b} = U_r U_r^T \vec{b}.$$

Recall that, by construction of U and U_r, the columns $\vec{u}_1, \ldots, \vec{u}_r$ form an orthonormal basis for $\mathsf{col}(A)$. Then,

$$U_r U_r^T \vec{b} = U_r \left(\begin{bmatrix} — & \vec{u}_1 & — \\ & \vdots & \\ — & \vec{u}_r & — \end{bmatrix} \vec{b} \right) = \begin{bmatrix} | & & | \\ \vec{u}_1 & \cdots & \vec{u}_r \\ | & & | \end{bmatrix} \begin{bmatrix} \vec{u}_1 \cdot \vec{b} \\ \vec{u}_2 \cdot \vec{b} \\ \vdots \\ \vec{u}_r \cdot \vec{b} \end{bmatrix} = (\vec{u}_1 \cdot \vec{b})\vec{u}_1 + (\vec{u}_2 \cdot \vec{b})\vec{u}_2 + \cdots + (\vec{u}_r \cdot \vec{b})\vec{u}_r.$$

The right-hand side is exactly the projection formula given in Theorem C.12. That is, $A\vec{x}_*$ is the projection of $\vec{b}$ onto $\mathrm{col}(A)$ and hence, $\vec{x}_*$ is a least squares solution to the least squares problem. ■

Example E.9. We will solve the least squares problem

$$A\vec{x} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \vec{x} \approx \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}.$$

Recall from Example E.6 that we have a singular value decomposition

$$A = U\Sigma V^T = \begin{bmatrix} 1/3 & 2/\sqrt{5} & -2/\sqrt{45} \\ -2/3 & 1/\sqrt{5} & 4/\sqrt{45} \\ 2/3 & 0 & 5/\sqrt{45} \end{bmatrix} \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

From Σ, we see that we have $r = 1$. Thus, Theorem E.8 says that a least squares solution is given by $\vec{x}_* = A^+\vec{b}$. First, compute

$$\begin{aligned} A^+ &= (V_r D^{-1} U_r^T) \\ &= \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/3\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/3 & -2/3 & 2/3 \end{bmatrix} \\ &= \begin{bmatrix} 1/6 \\ -1/6 \end{bmatrix} \begin{bmatrix} 1/3 & -2/3 & 2/3 \end{bmatrix} \\ &= \begin{bmatrix} 1/18 & -1/9 & 1/9 \\ -1/18 & 1/9 & -1/9 \end{bmatrix}. \end{aligned}$$

Then,

$$\vec{x}_* = A^+\vec{b} = \begin{bmatrix} 1/18 & -1/9 & 1/9 \\ -1/18 & 1/9 & -1/9 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 11/18 \\ -11/18 \end{bmatrix}$$

The best approximation is

$$A\vec{x}_* = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 11/18 \\ -11/18 \end{bmatrix} = \begin{bmatrix} 11/9 \\ -22/9 \\ 22/9 \end{bmatrix}.$$

E.3 Mean, variance, and covariance

In this section, we introduce several statistical tools to discuss principal component analysis in the next section. The main is idea is the following.

Suppose that you have a large dataset where each observation has multiple datapoints. For instance, if you wanted to study the output variable of precipitation, you might have input variables temperature, humidity, pressure, wind speed, and time of day. If you have N observations, then you would store the data in an $5 \times N$ matrix, where each column is an observation and each of the 5 rows correspond to the variables temperature, humidity, and so on.

Definition E.10. A **matrix of observations** A is a $p \times N$ matrix, where each of the N columns are observation vectors. Denote the columns of A by $\vec{X}_i$ as follows,

$$A = \begin{bmatrix} | & | & & | \\ \vec{X}_1 & \vec{X}_2 & \cdots & \vec{X}_N \\ | & | & & | \end{bmatrix},$$

that is, each $\vec{X}_i$ is an **observation**. The **sample mean** of $\vec{X}_1, \ldots, \vec{X}_N$ is the vector

$$\vec{M} = \frac{1}{N}\left(\vec{X}_1, \ldots, \vec{X}_N\right).$$

The mean of a set of real numbers $x_1, \ldots, x_n$ is $\frac{x_1+\cdots+x_n}{n}$, so the sample mean is exactly the same where we replace real numbers with vectors.

For each i, define $\hat{\vec{X}}_i = \vec{X}_i - \vec{M}$, the difference between $\vec{X}_i$ and the mean $\vec{M}$. Define a matrix B by

$$B = \begin{bmatrix} | & | & & | \\ \hat{\vec{X}}_1 & \hat{\vec{X}}_2 & \cdots & \hat{\vec{X}}_N \\ | & | & & | \end{bmatrix},$$

Geometrically, $\vec{M}$ is the centre of the dataset $\vec{X}_1, \ldots, \vec{X}_N$. Now for B and $\hat{\vec{X}}_1, \ldots, \hat{\vec{X}}_N$, the centre is at the origin; we have applied a shift by subtracting the mean from each vector.

which we say is in **mean-deviation form**. Lastly, the **sample covariance matrix** is following the $p \times p$ matrix,

$$S = \frac{1}{N-1} BB^T.$$

The sample covariance matrix S keeps track of two important pieces of information: variance of individual variables (on the diagonal of S) and correlation between pairs of variables (on the off-diagonal).

First, consider a diagonal entry of S. Then, using the above formula for S, it follows that

$$[S]_{i,i} = \frac{1}{N-1}(\vec{r}_i \cdot \vec{r}_i),$$

where $\vec{r}_i$ is the i-th row of B. Recall that rows of A, and hence rows of B, correspond to input variables (such as temperature in our running example). This value $[S]_{i,i}$ is the called the **variance** of the i-th input variable. Variance is a measure of spread, so the larger the value $[S]_{i,i}$, the wider the range of the corresponding observed datapoints.

Variance is the key to principal component analysis. Indeed, modelling a line or plane of best fit is simply trying to find a linear approximation that best explains the variance (or spread) of a given dataset.

Meanwhile, the off-diagonal entries of S are **covariance**, or correlation between two input variables (say, temperature and humidity).

$$[S]_{i,j} = \frac{1}{N-1}(\vec{r}_i \cdot \vec{r}_j).$$

A value of 0 indicates that the i-th and j-th variables (e.g., time of day and pressure) are uncorrelated; one has no effect on the other. The larger the value, the more they are correlated.

Example E.11. We will compute the sample mean and sample covariance matrix of the following dataset:

$$A = \begin{bmatrix} | & | & | & | \\ \vec{X}_1 & \vec{X}_2 & \vec{X}_3 & \vec{X}_4 \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 4 & 7 & 8 \\ 2 & 2 & 8 & 4 \\ 1 & 13 & 1 & 5 \end{bmatrix}.$$

The sample mean is

$$\vec{M} = \frac{1}{4}\left(\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix} + \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix} + \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}\right) = \frac{1}{4}\begin{bmatrix} 20 \\ 16 \\ 20 \end{bmatrix} = \begin{bmatrix} 5 \\ 4 \\ 5 \end{bmatrix}$$

To write the data in mean-deviation form, we compute,

$$\hat{\vec{X}}_1 = \vec{X}_1 - \vec{M} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 5 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} -4 \\ -2 \\ -4 \end{bmatrix}, \quad \hat{\vec{X}}_2 = \vec{X}_2 - \vec{M} = \begin{bmatrix} -1 \\ -2 \\ 8 \end{bmatrix}, \quad \hat{\vec{X}}_3 = \begin{bmatrix} 2 \\ 4 \\ -4 \end{bmatrix}, \quad \hat{\vec{X}}_4 = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

which yields that

$$B = \begin{bmatrix} | & | & | & | \\ \hat{\vec{X}}_1 & \hat{\vec{X}}_2 & \hat{\vec{X}}_3 & \hat{\vec{X}}_4 \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} -4 & -1 & 2 & 3 \\ -2 & -2 & 4 & 0 \\ -4 & 8 & -4 & 0 \end{bmatrix}.$$

Thus, the sample covariance matrix is

$$S = \frac{1}{N-1}BB^T = \frac{1}{3}\begin{bmatrix} -4 & -1 & 2 & 3 \\ -2 & -2 & 4 & 0 \\ -4 & 8 & -4 & 0 \end{bmatrix}\begin{bmatrix} -4 & -2 & -4 \\ -1 & -2 & 8 \\ 2 & 4 & -4 \\ 3 & 0 & 0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 30 & 18 & 0 \\ 18 & 24 & -24 \\ 0 & -24 & 96 \end{bmatrix} = \begin{bmatrix} 10 & 6 & 0 \\ 6 & 8 & -8 \\ 0 & -8 & 32 \end{bmatrix}.$$

Exercise 30. In the previous example, read off from S the variance of each of the three variables. Also read off from S the covariance for each of the three pairs of variables (1 and 2; 1 and 3; 2 and 3).

Definition E.12. The **total variance** of a dataset is the trace of the sample covariance matrix.

Example E.13. Continuing the previous example, the total variance of the data is $\mathrm{tr}(S) = 10 + 8 + 32 = 50$.

E.4 Principal component analysis

The goal in a multivariable linear regression is to explain the total variance of the dataset using a best linear approximation (e.g., a line or plane of best fit). However, often most of the input variables are not needed, as one or two or three of them may explain the vast majority of the total variance. The goal of this section is to make this precise and understand which variables explain the most variance.

Convention E.14. Throughout this section, we will assume that the dataset matrix A is **already in mean-deviation form**. (That is, its columns sum to zero.)

First notice that S is symmetric. Indeed,

$$S^T = \left(\frac{1}{N-1}AA^T\right)^T = \frac{1}{N-1}(A^T)^T A^T = \frac{1}{N-1}AA^T = S.$$

The Spectral Theorem (Theorem D.3) thus guarantees that S is orthogonally diagonalizable. That is, $S = PDP^T$ for some diagonal matrix D and orthogonal matrix P. Using a similar argument to that given at the beginning of Section E.1, it follows that the eigenvalues of S are all nonnegative. So, we may write

$$P = \begin{bmatrix} | & | & & | \\ \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_p \\ | & | & & | \end{bmatrix} \quad \text{and} \quad D = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p),$$

where $\vec{u}_1, \ldots, \vec{u}_p$ are orthonormal and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Definition E.15. The vectors $\vec{u}_1, \ldots, \vec{u}_p$ are called the **principal components** of the data (in the matrix of observations).

Specifically, we say that $\vec{u}_1$, the unit eigenvector for the largest eigenvalue, is the **first principal component**, $\vec{u}_2$ is the **second principal component**, and so on.

Denote by $\vec{X}$ a vector of input variables. That is, if are making p observations for each datapoint, then

$$\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}.$$

Continuing the discussion at the end of the previous section, the variance of x_i is $[S]_{i,i}$. Similarly, the covariance of x_i and x_j is $[S]_{i,j}$.

Our orthogonal matrix P defines a change of coordinates $\vec{X} = P\vec{Y}$ (or equivalently, $\vec{Y} = P^T\vec{X}$). It then follows that $y_i = \vec{u}_i^T\vec{X}$. Or, if $\vec{u}_i^T = \begin{bmatrix} c_1 & c_2 & \cdots & c_p \end{bmatrix}$, then

$$y_i = \begin{bmatrix} c_1 & c_2 & \cdots & c_p \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = c_1x_1 + c_2x_2 + \cdots + c_px_p.$$

Remark E.16. Because the trace of a matrix is the sum of its eigenvalues, it follows that the total variance of the data satisfies

$$\text{total variance} = \mathsf{tr}(S) = \mathsf{tr}(D) = \sum_{i=1}^{p} \lambda_i.$$

Fact E.17. The variance of y_i is λ_i.

Proof. The variance of y_i is the (i,i)-th entry of the sample covariance matrix in the new coordinate system. This matrix is D, so we have that $[D]_{i,i} = \lambda_i$. ■

Putting this together, the i-th principal vector $\vec{u}_i$ explains λ_i-much of the total variance. As a percentage, the vector $\vec{u}_i$ explains

$$\frac{\lambda_i}{\text{total variance}} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

of the total variance.

Principal Component Analysis.
Consider a $p \times N$ dataset A, written in mean-deviation form.

1. Construct the sample covariance matrix
$$S = \frac{1}{N-1}(AA^T).$$
2. Orthogonally diagonalize the sample covariance matrix $S = PDP^T$, where P is orthogonal and D is diagonal and its entries are nonnegative and listed in decreasing order (largest first).
3. The total variance T is
$$T = \text{tr}(S) = \text{tr}(D) = \sum_{j=1}^{p} \lambda_j.$$
4. The columns $\vec{u}_i$ of P are the principal component vectors. The i-th principal component vector $\vec{u}_i$ explains λ_i / T of the total variance.

Example E.18. Consider a dataset whose covariance matrix is

$$S = \begin{bmatrix} 2382.78 & 2611.84 & 2136.2 \\ 2611.84 & 3106.47 & 2553.9 \\ 2136.2 & 2553.9 & 2650.71 \end{bmatrix}$$

which is orthogonally diagonalized as

$$S = PDP^T = \begin{bmatrix} 0.5417 & -0.4894 & 0.6834 \\ 0.6295 & -0.3026 & -0.7157 \\ 0.5570 & 0.8179 & 0.1441 \end{bmatrix} \begin{bmatrix} 7614.23 & 0 & 0 \\ 0 & 427.63 & 0 \\ 0 & 0 & 98.1 \end{bmatrix} \begin{bmatrix} 0.5417 & 0.6295 & 0.5570 \\ -0.4894 & -0.3026 & 0.8179 \\ 0.6837 & -0.7157 & 0.1441 \end{bmatrix}.$$

The total variance, the trace of either S or D, is

$$T = 8139.96.$$

The first principal component vector (the first column of P) explains

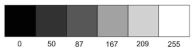
$$\frac{7614.23}{8139.96} = 93.5\%$$

of the total variance.

Exercise 31. In the above example, compute the amount of variance explained by the second and third principal vectors.

E.5 Image compression

Colour images are comprised of *pixels*, each specifying a value on red, green, and blue sliders. For simplicity, we will work with greyscale images, where each pixel is specified by a single number between 0 and 255 corresponding to black (0), white (255), or somewhere in between.



Computers store numbers in binary, also known as base 2. We will need to store $256 = 2^8$ different greyscale values, so we will need 8 "bits" to store our greyscale values.

A "bit" is the smallest unit of computer storage. Your phone or laptop might have some number of gigabytes of storage. Now, there are 8 bits in a byte (B), 1000 bytes in a kilobyte (KB), 1000 kilobytes in a megabyte (MB), and 1000 kilobytes in a gigabyte (GB). So, a phone with 128GB of storage can store 1,099,511,627,776 bits.

If your phone has 128GB of storage takes pictures that are approximately 3000×4000 pixels (width $\times$ height), then your phone can only store about 10'000 greyscale images—and nothing else!

This might sound like a lot, but colour images (naïvely) require three times the storage (for red, green, and blue), and videos consist of many images per second.

Image compression is just one technique software developers use to optimize our phones' storage!

Consider a greyscale image with $m \times n$ pixels whose values we store in a matrix A. Now, compute the reduced singular value decomposition of A,

$$\begin{aligned} A = U_r D V_r^T &= \begin{bmatrix} | & & | \\ \vec{u}_1 & \cdots & \vec{u}_r \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} — & \vec{v}_1^T & — \\ & \vdots & \\ — & \vec{v}_r^T & — \end{bmatrix} \\ &= \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \cdots + \sigma_r \vec{u}_r \vec{v}_r^T. \end{aligned}$$

Because the singular values are listed in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, each successive term in the above sum contributes less and less; we have a diminishing returns effect. So, the **best rank $k \leq r$ approximation of** A is truncating this sum after the first k terms:

$$A \approx A_k := \sigma_1 \vec{u}_1 \vec{v}_1^T + \cdots + \sigma_k \vec{u}_k \vec{v}_k^T.$$

Example E.19. The following is a singular value decomposition $A = U\Sigma V^T$,

$$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/3\sqrt{2} & -1/3\sqrt{2} & 4/3\sqrt{2} \\ -2/3 & 2/3 & 1/3 \end{bmatrix}.$$

Notice that $\mathsf{rank}(A) = 2$. The best rank 1 approximation to A is:

$$\begin{aligned} A_1 = \sigma_1 \vec{u}_1 \vec{v}_1^T &= 5 \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} \\ &= 5 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} = \begin{bmatrix} 5/2 & 5/2 & 0 \\ 5/2 & 5/2 & 0 \end{bmatrix}. \end{aligned}$$

Recall that the storage required for an uncompressed image is $b \cdot mn$ bits, where m and n are the width and height (in pixels) and b is the number of bits per pixel. This image corresponds to an $m \times n$ matrix A. The best rank k approximation of A consists of the following data:

- k singular values,
- k vectors $\vec{u}_1, \ldots, \vec{u}_k$ in $\mathbb{R}^m$,
- k vectors $\vec{v}_1, \ldots, \vec{v}_k$ in $\mathbb{R}^n$,

which requires the following storage:

$$b(k + km + kn) = b \cdot k(1 + m + n) \text{ bits},$$

where again b is the number of bits per pixel.

Definition E.20. The **compression ratio** is the ratio of original pixels to compressed pixels,

$$\frac{b \cdot mn}{b \cdot k(1 + m + n)} = \frac{mn}{k(1 + m + n)}.$$

Exercise 32. (a) Compute the compression ratio from Example E.19.

(b) Also, compute the compression ratio of an arbitrary 4032×3024 image.[3] (Your answer will depend on k).

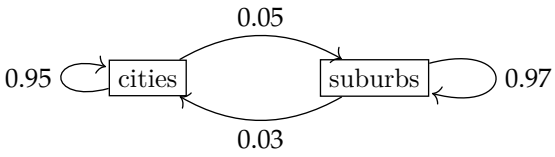
[3] This is the size of the most recent image on the instructor's phone. (A picture of some math on a blackboard.)

F
Markov Chains

Markov chains model systems with a finite number of states given probabilities of moving between each state. For instance, suppose we group a population into those who live in cities versus suburbs. Let's say that at start time $t = 0$, we have the following initial populations,

$$c_0 = 600'000, \qquad \text{(cities)}$$
$$s_0 = 400'000. \qquad \text{(suburbs)}$$

Also, assume that every year on average, 5% of those in cities move to suburbs and the remaining 95% remain in the city. Meanwhile, every year 3% of those in suburbs move to cities and the remaining 97% remain in the suburbs. We can equivalently write these probabilities in the following diagram.



The populations after 1 year are given by

$$\begin{cases} c_1 = 0.95c_0 + 0.03s_0 \\ s_1 = 0.05c_0 + 0.97s_0 \end{cases} \iff \begin{bmatrix} c_1 \\ s_1 \end{bmatrix} = \begin{bmatrix} 0.95 & 0.03 \\ 0.05 & 0.97 \end{bmatrix} \begin{bmatrix} c_0 \\ s_0 \end{bmatrix}$$

$$\begin{bmatrix} c_1 \\ s_1 \end{bmatrix} = \begin{bmatrix} 0.95 & 0.03 \\ 0.05 & 0.97 \end{bmatrix} \begin{bmatrix} 600'000 \\ 400'000 \end{bmatrix} = \begin{bmatrix} 582'000 \\ 418'000 \end{bmatrix}$$

Denote by

$$\vec{p}_t = \begin{bmatrix} c_t \\ s_t \end{bmatrix}$$

the vector of populations in the cities and suburbs after t years. That is, $\vec{p}_t = M\vec{p}_{t-1}$, where M is the above 2×2 matrix of probabilities. Notice that

$$\vec{p}_2 = M\vec{p}_1 = M(M\vec{p}_0) = M^2\vec{p}_0$$

Historical sidenote. Andrey Markov began the study of what we now know as "Markov chains" in the early 1900s. One of his original applications was applying this theory to understand the poetry of Alexander Pushkin by looking at patterns of vowels and consonants.

Today, Markov chains are similar to how generative AI tools, like ChatGPT, generate text. Given a prompt and the words it has already written, ChatGPT produces the word most likely to come next through a Markov process.

There is a slight notational difference here and what we did, say, for dynamical systems. For dynamical systems, we used $\vec{p}(t)$ to denote population after time t, while here we are using $\vec{p}_t$. The difference is that for dynamical systems, our models were continuous functions, so we could take about fractions of units time. While here, we are iterating matrix multiplication so the time units are *discrete*.

and

$$\vec{p}_3 = M(\vec{p}_2) = M(M^2\vec{p}_0) = M^3\vec{p}_0.$$

This pattern continues, so it follows that

$$\vec{p}_t = M^t\vec{p}_0.$$

So to find, say, $\vec{p}_{100}$, rather than having to compute each of $\vec{p}_1, \ldots, \vec{p}_{100}$, we can compute M^{100} instead, which, in some cases, can be less computationally intensive.

The main question we will answer is the following.

Question F.1. What is the long-term behaviour of a Markov chain? Do we converge to some limit vector?

F.1 Stochastic matrices and long-term behaviour

The following definition should look familiar from the columns of M in the introduction of this chapter.

Definition F.2. A **probability vector** is a vector with nonnegative entries which sum to 1. A **stochastic matrix** is a square matrix whose columns are probability vectors.

Example F.3. The columns of M in the introduction,

$$\begin{bmatrix} 0.95 \\ 0.05 \end{bmatrix}, \quad \begin{bmatrix} 0.03 \\ 0.97 \end{bmatrix}$$

have nonnegative entries. Moreover, the sum of each's entries is 1, so these are both probability vectors. As a result, M is a stochastic matrix.

On the other hand, the population vector

$$\begin{bmatrix} 600'000 \\ 400'000 \end{bmatrix}$$

is clearly not a probability vector. However, if we replace each population with its corresponding percentage of the total population,[1] we get the vector

$$\begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix},$$

which is a probability vector.

[1] That is, the total population is $1'000'000$. So we replace $600'000$ with $600'000/1'000'000 = 0.6$ and similarly, $400'000$ becomes 0.4.

Recall that Question F.1 asks to describe the long-term behaviour. In particular, we want to know about the existence of a vector of the following type.

Definition F.4. Let M be a stochastic matrix. An **equilibrium vector** of M is a probability vector $\vec{q}$ for which $M\vec{q} = \vec{q}$.

We make the following two observations.

- Such a vector $\vec{q}$ is a state vector $\vec{p}_t$ for which $\vec{p}_{t+1} = M\vec{p}_t$. In such a case, $\vec{p}_t = \vec{p}_T$ for all time T later than t (that is, $T \geq t$). If we reach such a vector, then after every unit time, the state vector does not change.

- Equilibrium vectors $\vec{q}$, by definition, satisfy $M\vec{q} = \vec{q}$. That is, they are probability vectors that are also eigenvectors of M with corresponding eigenvalue 1.

The following theorem guarantees that any stochastic matrix has a corresponding equilibrium vector.

Theorem F.5. *Every stochastic matrix has* 1 *as an eigenvalue. Equivalently, every stochastic matrix has an equilibrium vector.*

Proof in the 3×3 *case.* We will make use of the fact that a matrix M and its transpose M^T have the same eigenvalues. To see this, notice that it is enough to argue that they have the same characteristic polynomial. Using properties of transpose,

$$(M - \lambda I)^T = M^T - (\lambda I)^T = M^T - \lambda I,$$

because λI is diagonal. Then because the determinant does not change under taking transposes, we have that

$$\det(M - \lambda I) = \det\left((M - \lambda I)^T\right) = \det(M^T - \lambda I).$$

Now let M be a 3×3 stochastic matrix, writing

$$M = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}.$$

Because M is stochastic, its columns sum to 1. Consequently, the rows of M^T sum to 1,

$$M^T = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & k \end{bmatrix}.$$

It then follows that

$$M^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & k \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a+d+g \\ b+e+h \\ c+f+k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

so 1 is an eigenvalue of M^T and thus is also an eigenvalue of M. We conclude that M has an equilibrium vector. ■

Example F.6. We will find a steady state vector for the stochastic matrix

$$M = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}.$$

We are required to find an eigenvector of M corresponding to the eigenvalue 1, whose entries sum to 1. First, the eigenvectors of M are the solutions to $(M-I)\vec{x}=\vec{0}$. That is, they are the solutions to the system

$$\left[\begin{array}{cc|c} -0.4 & 0.3 & 0 \\ 0.4 & -0.3 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 4/10 & -3/10 & 0 \\ 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & -3/4 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

So eigenvectors of M are of the form $(x,y)=(3y/4,y)$, or equivalently, the scalar multiples of $(3/4,1)$ or $(3,4)$. Such a vector, whose entries sum to 1 is $(3/7,4/7)$.

You can check that $\vec{q}=(3/7,4/7)$ is an equilibrium vector by checking its entries sum to 1 and that $\vec{q}=M\vec{q}$.

Finding equilibrium vectors of stochastic matrices. Let M be a stochastic matrix.

1. Find an eigenvector $\vec{v}$ of M corresponding to the eigenvalues 1, such that $\vec{v}$ has nonnegative entries.
2. An equilibrium vector $\vec{q}$ is given by dividing $\vec{v}$ by the sum of the entries of $\vec{v}$.

This theorem guarantees the existence of an equilibrium vector but this need not be the unique equilibrium vector. In some cases, this equilibrium vector is unique.

Definition F.7. A stochastic matrix M is **regular** if there is some positive integer k for which M^k contains only strictly positive entries.

Stochastic matrices have nonnegative entries, some of which could be zero. Every regular matrix is a stochastic matrix, but the reverse is not true.

Example F.8. The following matrix M is a regular matrix.

$$M=\begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.8 & 0.3 \\ 0.2 & 0 & 0.4 \end{bmatrix}$$

It is regular because M^2 contains only positive entries.

$$M^2=\begin{bmatrix} 0.37 & 0.26 & 0.33 \\ 0.45 & 0.7 & 0.45 \\ 0.18 & 0.04 & 0.22 \end{bmatrix}$$

Theorem F.9. *If M is a regular matrix, then it has a* unique *equilibrium vector $\vec{q}$. Moreover, given any initial probability vector $\vec{x}_0$,*

$$\lim_{t\to\infty}\vec{x}_t=\lim_{t\to\infty}M^t\vec{x}_0=\vec{q}.$$

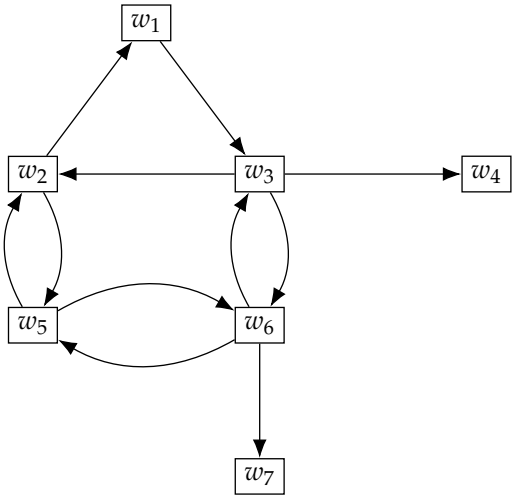
F.2 PageRank

In this section, we will discuss how to use Markov chains to answer the following question.

Question F.10. How does one rank all the webpages on the internet?

Our answer comes in the form of Google's original approach, which originated as a school project in 1996. Developed by Google co-founders Sergey Brin and Larry Page, graduate students at Stanford University, PageRank orders webpages by the number of links to this webpage from other sites.

Consider a toy model with only six webpages. In the following figure, an arrow from one webpage to another denotes a link from the source webpage that points to the target.



This gives rise to a Markov chain in the following way. Suppose we start at webpage w_3. This webpage has links to w_2, w_4, and w_6, so we choose between these webpages with equal probability and move to the chosen page. We then repeat ad nauseam. There is an underlying stochastic matrix here,

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 1/3 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1 \end{bmatrix}.$$

Notice that w_4 and w_7 do not link to any webpages, so if we end up on either of these pages, we cannot leave! In M, we see this as the fourth and seventh columns having only one nonzero entry.

Given any starting probability vector $\vec{x}_0$, we have a Markov chain in the sequence $\vec{x}_0, M\vec{x}_0, \ldots, M^t\vec{x}_0, \ldots$, which models the probabilities of being on any given webpage after time t. We then rank the pages

using the equilibrium vector: the page with the greatest probability is the highest rank webpage; the page with the lowest probability receives the lowest rank.

However, M is not a regular matrix, so there is not a unique equilibrium vector. We make the following two adjustments: one to fix the issue of landing on pages which do not link to other pages, and one to fix this irregularity issue.

(i) If we land on a page that does not link to any other pages, move to any page with uniform probability. That is, we replace the 4^{th} and 7^{th} columns in M with the vectors $(1/7, \ldots, 1/7)$. The 7 in the denominator is the same as the total number of webpages in our model. So now, M is the following matrix.

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 1/7 & 0 & 0 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 1/2 & 0 & 1/7 \\ 1 & 0 & 0 & 1/7 & 0 & 1/3 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 0 & 0 & 1/7 \\ 0 & 1/2 & 0 & 1/7 & 0 & 1/3 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 1/2 & 0 & 1/7 \\ 0 & 0 & 0 & 1/7 & 0 & 1/3 & 1/7 \end{bmatrix}.$$

(ii) Pick some probability p. That is, p is a real number satisfying $0 < p < 1$. At every step, we will move with probability p from the current webpage to one linked from the current page. With probability $1 - p$, we will instead move to a random webpage, picked uniformly.

Definition F.11. The **Google matrix** is the matrix constructed in this way after making a choice of p.

Example F.12. Let $p = 0.85$. Then the Google matrix for the above example is

$$G = pM + (1 - p) \cdot \left(\frac{1}{7}\mathbf{1}\right),$$

where $\mathbf{1}$ is the 7×7 matrix with a 1 in every entry. Explicitly, and using the fact that $0.85 = 17/20$, we have that

$$G = \frac{17}{20}\begin{bmatrix} 0 & 1/2 & 0 & 1/7 & 0 & 0 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 1/2 & 0 & 1/7 \\ 1 & 0 & 0 & 1/7 & 0 & 1/3 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 0 & 0 & 1/7 \\ 0 & 1/2 & 0 & 1/7 & 0 & 1/3 & 1/7 \\ 0 & 0 & 1/3 & 1/7 & 1/2 & 0 & 1/7 \\ 0 & 0 & 0 & 1/7 & 0 & 1/3 & 1/7 \end{bmatrix} + \frac{3}{20}\begin{bmatrix} 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 17/40 & 0 & 17/140 & 0 & 0 & 17/140 \\ 0 & 0 & 17/60 & 17/140 & 17/40 & 0 & 17/140 \\ 17/20 & 0 & 0 & 17/140 & 0 & 17/60 & 17/140 \\ 0 & 0 & 17/60 & 17/140 & 0 & 0 & 17/140 \\ 0 & 17/40 & 0 & 17/140 & 0 & 17/60 & 17/140 \\ 0 & 0 & 17/60 & 17/140 & 17/40 & 0 & 17/140 \\ 0 & 0 & 0 & 17/140 & 0 & 17/60 & 17/140 \end{bmatrix} + \begin{bmatrix} 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \\ 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 & 3/140 \end{bmatrix}$$

$$= \begin{bmatrix} 3/140 & 25/56 & 3/140 & 1/7 & 3/140 & 3/140 & 1/7 \\ 3/140 & 3/140 & 32/105 & 1/7 & 25/56 & 3/140 & 1/7 \\ 61/70 & 3/140 & 3/140 & 1/7 & 3/140 & 32/105 & 1/7 \\ 3/140 & 3/140 & 32/105 & 1/7 & 3/140 & 3/140 & 1/7 \\ 3/140 & 25/56 & 3/140 & 1/7 & 3/140 & 32/105 & 1/7 \\ 3/140 & 3/140 & 32/105 & 1/7 & 25/56 & 3/140 & 1/7 \\ 3/140 & 3/140 & 3/140 & 1/7 & 3/140 & 32/105 & 1/7 \end{bmatrix}.$$

You can verify that this is a regular matrix (that is, because its entries are all strictly positive, you need only see that every column sums to 1).

We can then find an equilibrium vector. Using your favourite symbolic calculator[2], you can find that the eigenvectors must satisfy the system

$$\left[\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{matrix} -193'640/153'877 \\ -280'680/153'877 \\ -318'471/153'877 \\ -3'291'689/3'077'540 \\ -273'166/153'877 \\ -280'680/153'877 \\ 0 \end{matrix},$$

or equivalently, are of the form

$$t \begin{bmatrix} 193'640/153'877 \\ 280'680/153'877 \\ 318'471/153'877 \\ 3'291'689/3'077'540 \\ 273'166/153'877 \\ 280'680/153'877 \\ 1 \end{bmatrix}$$

for any $t \in \mathbb{R}$. Choosing $t = 3'077'540$ yields the eigenvector

$$\begin{bmatrix} 3'872'800 \\ 5'613'600 \\ 6'369'420 \\ 3'291'689 \\ 5'463'320 \\ 5'613'600 \\ 3'077'540 \end{bmatrix},$$

[2] Actually, some symbolic calculators (including WolframAlpha) refuse to do this computation. However you can find one that does. I used *Macaulay2*. You should use something else.

from which we find the equilibrium vector by dividing by the sum of its entries:

$$\vec{q} = \begin{bmatrix} 3'872'800/33'301'969 \\ 5'613'600/33'301'969 \\ 6'369'400/33'301'969 \\ 3'291'689/33'301'969 \\ 5'463'320/33'301'969 \\ 5'613'600/33'301'969 \\ 3'077'540/33'301'969 \end{bmatrix} \approx \begin{bmatrix} 0.11629 \\ 0.16857 \\ 0.19126 \\ 0.09884 \\ 0.16405 \\ 0.16857 \\ 0.09241 \end{bmatrix}.$$

So the ranking of the webpages is:

1. w_3
2. tied between w_2 and w_6
4. w_5
5. w_1
6. w_4
7. w_7.