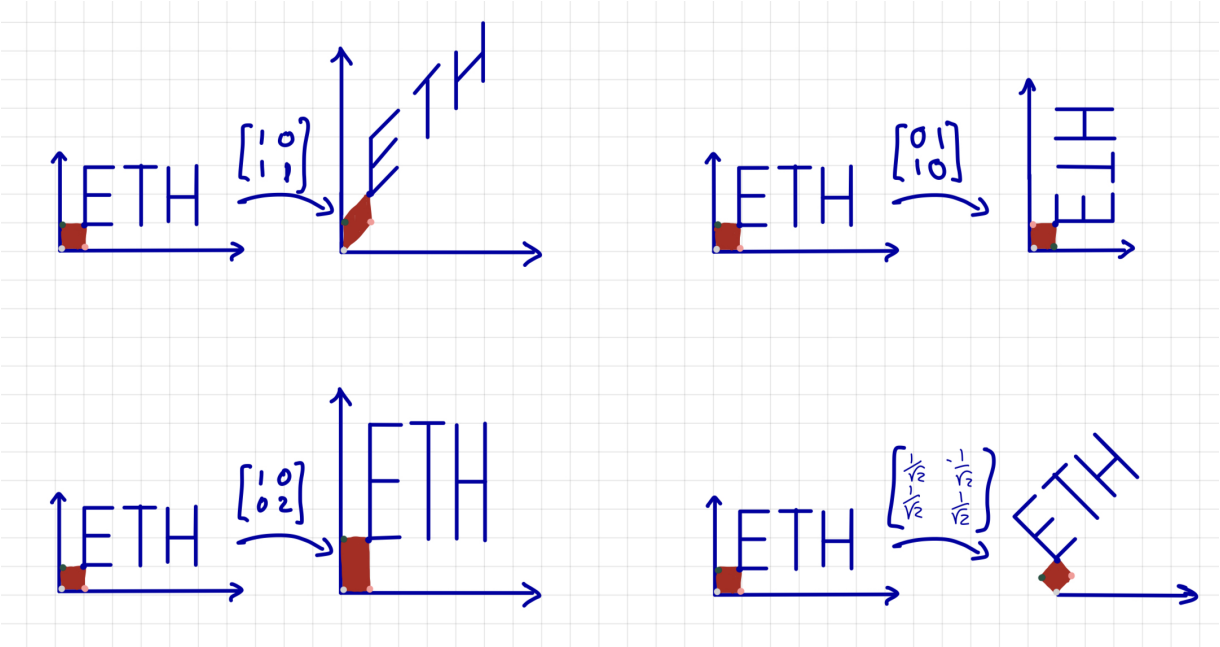
Linear Algebra

Fall 2023 (ETHZ 401-0131-00L)

Lecture Notes
Part II



Afonso S. Bandeira
ETH Zürich

Last update on November 28, 2023

"READ ME" FOR PART II

My webpage, with contact information, is: `https://people.math.ethz.ch/~abandeira`

These lecture notes serve as **a continuation**[1] **of Part I**, taught by Prof. Bernd Gärtner, available at `https://ti.inf.ethz.ch/ew/courses/LA23/notes_part_I.pdf`. Please read the Preface there. Please note there may be some changes in notation. Furthermore, we will try to stay close to the notation in [Str23], but there also be some differences.

[Str23] Gilbert Strang. Introduction to Linear Algebra. Wellesley - Cambridge Press, 6th ed., 2023.

The **course page** has relevant information for the course: `https://ti.inf.ethz.ch/ew/courses/LA23`.

I offer **office hours** (in HG G23.1) almost weekly, feel free to stop by, to chat about the course, Mathematics, Computer Science, or University in general. Most office hours visitors stop by to learn more about research in Mathematics and Computer or Data Science. You can see the schedule on my webpage, on the calendar applet on the left.

There are countless excellent Linear Algebra books with the material covered in this course. For Part II we will roughly continue to follow, in structure and content, [Str23], with some small deviations. I will try to keep the numbering of Chapters/Sections and Sections/Subsections consistent with [Str23] (as far as the deviations allow). See Appendix A for some important preliminaries and some remarks on notation.

Throughout the notes, and the lectures, I will try to motivate some of the material with **Guiding Question**s. For students who would like to explore the topic further, I will include some **Exploratory Challenge**s and **Further Reading**, these often will include difficult problems or topics. I will also take some opportunities to share some active **Research Question**s related to the topics we covered (we are still discovering new phenomena in Linear Algebra today and for many years to come!).

After deriving a result, we will often do some **Sanity Check**s, and some things I will leave as a **Challenge**: these should be accessible and of difficulty comparable to homework questions, a $\star$ indicates a harder problem (but still within the scope). On the other hand, **Exploratory Challenges** are generally outside the scope of the course or of substantial higher difficulty.

[1] If you are reading these notes and did not follow Part I, please read Appendix A.

Linear Algebra is a beautiful topic, connecting Algebra with Geometry[2], and has countless applications making it a key component of almost all quantitative pursuits. I sincerely hope you will enjoy the course as much as I enjoy teaching this subject!

MISCELLANEOUS THOUGHTS

I believe the Questions, Sanity Checks, Challenges, etc are very useful to learn the material, but **when you want to review the material**, or do a last read before the exam, you can focus on the Definitions, Propositions, Theorems, etc (and **focus less on the blue parts**).

In many of my side comments (usually in blue), and in some of the CS Lens Lectures, I do not include specific citations, and sometimes use technical terms that you might not have seen before. My goal is, quoting my collaborator Dustin Mixon, "to provide enough breadcrumbs for the interested reader to find more information online".[3] If you would like specific references, tell me a bit more about your interests and I would be happy to point you to some references (different references are better for different takes/interest on each of the topics). While CS Lens Lectures are not covered in the lecture notes, slides can be accessed in the course website: `https://ti.inf.ethz.ch/ew/courses/LA23/index.html`

As your mathematical level matures over the semester, the notes will have less illustrations and more definitions and mathematical statements. My recommendation is to read the notes with pen & paper next to you and to draw the picture yourself, this "translation" you will be doing — from mathematical statement to picture — will (I believe) help you greatly in the learning of Mathematics!

There are also countless high-quality videos and other content online about Linear Algebra, for example there is also an excellent series of videos by Gil Strang filmed ∼15 years ago: `https://www.youtube.com/playlist?list=PLE7DDD91010BC51F8`.

Strang actually retired just a few months ago, at almost 90 years of age! You can see his last lecture online: `https://www.youtube.com/watch?v=lUUte2o2Sn8`

[2] and Analysis, as you will likely see later in your academic life. For example, when Joseph Fourier invented Fourier Series to developed a theory of heat transfer he was essentially finding good orthonormal bases for functions.

[3] This itself also provides enough breadcrumbs for you to find the lecture notes I am quoting; they are excellent Linear Algebra lecture notes! (the order and content is somewhat different from our course)

Moreover, there are many excellent animations online giving lots of great intuition on several Linear Algebra topics and phenomena. While it is a great idea to take advantage of this, I would recommend first trying yourself to develop an intuition of the concept/phenomenon (e.g. by drawing a picture) and using these tools only after — use them to improve your intuition, not to create it!

As these Lecture Notes are being continuously updated, and sometimes the discussion in lectures leads us into proving an extra result, or suggests a remark, etc, I will try to add then and not change the numbering of things downstream, I do this by numbering them with $+1000$.

After each lecture, we post the handwritten notes from lecture on the course website `https://ti.inf.ethz.ch/ew/courses/LA23/index.html`. My suggestion would be to use the Lecture Notes to review the material, not the handwritten notes (which are mainly meant to support my oral exposition).

CONTENTS

"Read me" for Part II 2

Miscellaneous Thoughts 3

4. Orthogonality, Projections, and Least Squares 6

4.2. Projections 6

4.3. Least Squares Approximation 11

4.4. Orthonormal Bases and Gram Schmidt 15

4.5. The Pseudoinverse, also known as Moore–Penrose Inverse 20

5. Linear Transformations and Determinants 24

5.1. The Determinant 30

6. Eigenvalues and Eigenvectors 37

6.0. Complex Numbers 39

6.1. Introduction to Eigenvalues and Eigenvectors 41

Appendix A. Some Important Preliminaries and Remarks on Notation 48

Appendix B. Weekly Schedule 48

Appendix C. CS Lens Lectures 49

References 49

4. Orthogonality, Projections, and Least Squares

Guiding Question 1. If we have a system of linear equations that has no solution, how do we find the "solution" that has the smallest error? This question is central in countless applications[4].

Before diving into systems of equations, we will study Projections of vectors in a subspace.

4.2. **Projections.**

Definition 4.2.1 (Projection of a vector onto a subspace)**.** *The projection of a vector $b \in \mathbb{R}^m$ on a subspace S (of $\mathbb{R}^m$) is the point in S that is closest to b. In other words*

$$\text{proj}_S(b) = \underset{p \in S}{\text{argmin}} \|b - p\|. \tag{1}$$

Sanity Check 2. This is only a proper definition if the minimum exists and is unique. Can you show it exists and is unique? (perhaps at the end of the lecture?)

Let us build us some intuition by starting with projections to a line. Let S be the subspace corresponding to the line that goes through the vector a, i.e. $S = \text{Span}(a)$.

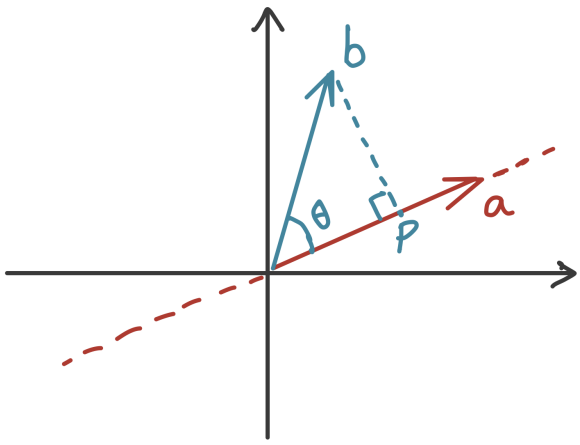


FIGURE 1. Projection on a line.

[4] as you will see later on, it is in a sense what Machine Learning is all about.

The projection p is the vector in the subspace S such that the "error vector" $e = b - p$ is perpendicular to a (i.e. $b - p \perp a$). Since $p \in S$ we have $p = \hat{x}a$, for some $\hat{x} \in \mathbb{R}$. Since $b - p \perp a$ we have $a^\top(b-p) = 0$. Substituting gives

$$a^\top(b - \hat{x}a) = 0 \iff \hat{x} = \frac{a^\top b}{a^\top a} \iff p = \frac{a^\top b}{a^\top a}a \iff p = \frac{aa^\top}{a^\top a}b.$$

Indeed, we have the following Proposition.

Proposition 4.2.2. *Let $a \in \mathbb{R}^m$ be a non-zero vector. The projection of a vector $b \in \mathbb{R}^m$ on $S = \mathrm{Span}(a)$ the span of a, is given by*

$$\mathrm{proj}_S(b) = \frac{aa^\top}{a^\top a}b.$$

Sanity Check 3. The projection of a vector that is already a multiple of a should be the identity operation. Check that this is the case! (and do it later, again, for general subspaces).

For general subspaces the idea is precisely the same. Let S be a subspace in $\mathbb{R}^m$ with dimension n. Let $a_1, \ldots, a_n$ be a basis of S, meaning that $S = \mathrm{Span}(a_1, \ldots, a_n)$ and $S = \mathrm{C}(A)$ where

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

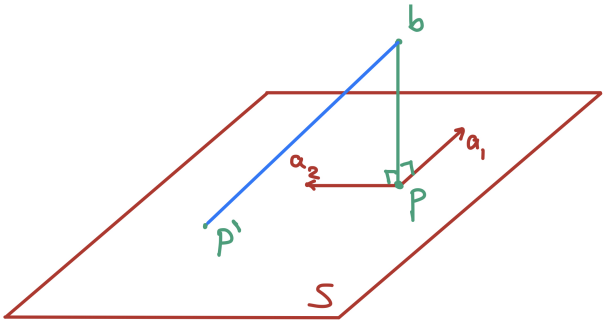


FIGURE 2. Projection on a subspace.

Similarly to the case of a line, it is easy to see (see Figure 2) that the projection p of a vector b on S is such that the error vector $e = b - p$ is perpendicular to each of the a_k's. To prove this fact rigorously we start by showing existence of such a vector p: take the orthogonal complement $S^\perp$ of S and write b as a sum of $e \in S^\perp$ and $p \in S$, then $e = b - p$ is orthogonal to the subspace S. Now, let us assume that there exists another point p' (as in Figure 2) and note that since $p' - p \in S$ we have that $b - p \perp p' - p$, and so, by Pythagoras' Theorem we have $\|p' - b\|^2 =$

$\|p-p'\|^2+\|p-b\|^2$, which implies that $\|p'-b\|^2 \geq \|p-b\|^2$ (with equality holding only when $p=p'$).[5] [6]

We just showed that $a_k^\top(b-p)=0$ for $k=1,\ldots,n$. In matrix-vector notation

$$\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}^\top (b-p)=0 \iff A^\top(b-p)=0.$$

Since $p \in \mathrm{C}(A)$ we have $p=A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$. This means that

$$A^\top(b-A\hat{x})=0 \iff A^\top A\hat{x}=A^\top b.$$

We just proved the following Proposition.

Proposition 4.2.3. *The projection p of a vector $b \in \mathbb{R}^m$ on a subspace S with a basis $a_1,\ldots,a_n$ can be written as $p=A\hat{x}$ where $\hat{x} \in \mathbb{R}^n$ satisfies the normal equations*

$$A^\top A\hat{x}=A^\top b,$$

where $A=\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}$ *is the matrix whose columns are a basis of S.*

If we can show that $A^\top A$ is invertible then we would have $p=A\hat{x}=A\left(A^\top A\right)^{-1}A^\top b$. Let's make a detour to show that it is indeed invertible.

Proposition 4.2.4. *$A^\top A$ is invertible if and only if A has linearly independent columns.*

Proof. We show this by showing that $A^\top A$ and A have the same nullspace. This is enough because, since $A^\top A$ is a square matrix it is invertible if and only if its nullspace only has the 0 vector, and A has linearly independent columns if and only if its nullspace only has the 0 vector.[7]

If $x \in \mathrm{N}(A)$ then $Ax=0$ and so $A^\top Ax=0$, thus $x \in \mathrm{N}(A^\top A)$. The other implication is more interesting.

If $x \in \mathrm{N}(A^\top A)$ then $A^\top Ax=0$. This implies that $x^\top A^\top Ax=x^\top 0=0$. But $x^\top A^\top Ax=(Ax)^\top(Ax)=\|Ax\|^2$ so Ax must be a vector with norm 0 which implies that $Ax=0$ and so $x \in \mathrm{N}(A)$.

□

[5] We have also, as a byproduct, answered the question in Sanity Check 2.

[6] Sometimes the projection is simply defined as the point on the subspace such that the error vector is orthogonal to the subspace, here we showed the two possible definitions are equivalent.

[7] We usually call a nullspace with only the zero vector, a trivial nullspace.

Corollary 4.2.5. *If A has linearly independent columns then $A^\top A$ is a square matrix, it is invertible and symmetric.*[8]

Now back to deriving a formula for projections: Since the columns of A are a basis they are linearly independent and so $A^\top A$ is indeed invertible. We just proved the following.

Theorem 4.2.6. *Let S be a subspace in $\mathbb{R}^m$ and A a matrix whose columns are a basis of S. The projection of $b \in \mathbb{R}^m$ to S is given by*

$$\text{proj}_S(b) = Pb,$$

where $P = A\left(A^\top A\right)^{-1} A^\top$ is the projection matrix.

The matrix $P = A\left(A^\top A\right)^{-1} A^\top$ is known as a Projection Matrix, it maps a vector b to its projection Pb on a subspace S. For the case of lines, P was given by $P = \frac{aa^\top}{a^\top a} = a\frac{1}{a^\top a}a^\top$.

Caution! 4. The matrix A (and $A^\top$) are not necessarily square, and so they don't have inverses. The expression $A\left(A^\top A\right)^{-1} A^\top$ **cannot** be simplified by expanding $\left(A^\top A\right)^{-1}$ (which would yield $I = P$, this would only make sense if S was all of $\mathbb{R}^m$ and note that, unsurprisingly, this would correspond exactly to the case when A is invertible).

Just as with the "sanity check" above, we should have $P^2 = P$, because if we project a point twice, the second time should not do anything as the point is already in S and indeed

$$P^2 = \left(A\left(A^\top A\right)^{-1} A^\top\right)^2 = A\left(A^\top A\right)^{-1} A^\top A\left(A^\top A\right)^{-1} A^\top = A\left(A^\top A\right)^{-1} A^\top = P.$$

Challenge 5. Is $I - P$ a projection? If so, which projection does it correspond to?

Challenge 6. How does the rank of P depend on properties of the subspace S?

Exploratory Challenge 7. We derived all of the formulas for projections using geometry. If you have taken Analysis/Calculus (I know many of you haven't, but you will in a few months) you can try to re-derive everything using the fact that derivatives at the minimum should be zero. You will see that you will get exactly the same answers.

In lecture, when discussing Figure 2 we explicitly proved the following proposition.

Proposition 4.2.1006. *Let S be a subspace in $\mathbb{R}^m$ with a basis $a_1, \ldots, a_n$. For $v \in \mathbb{R}^m$, v being orthogonal to all vectors in S is equivalent to being orthogonal to a_k for all $1 \leq k \leq n$.*

[8] Corollary is like a Theorem or a Proposition but one that follows directly from another one, this one follows directly from the Proposition above.

Proof. Since $a_1, \ldots, a_n$ are in S, if v is perpendicular to all vectors in S, it is in particular perpendicular to $a_1, \ldots, a_n$. On the other hand, any $w \in S$ can be written as $w = \alpha_1 a_1 + \cdots + \alpha_n a_n$ and $w^\top v = \alpha_1 a_1^\top v + \cdots + \alpha_n a_n^\top v = 0$. □

4.3. **Least Squares Approximation.** We go back to the guiding question of what to do when we want to "solve" a linear system that does not have an exact solution. More precisely let us suppose we have a linear system

$$Ax = b,$$

for which no solution x exists (for example, with too many equations, which would happen if $A \in \mathbb{R}^{m\times n}$ and $m > n$). A natural approach is to try to find x for which Ax is as close as possible to b

$$\min_{\hat{x}\in\mathbb{R}^n} \|A\hat{x} - b\|^2. \tag{2}$$

Further Remark 8. This seemingly simple observation is key to countless technologies. Measurement systems often have errors and so it is impossible to find the target object/signal x that satisfies them all exactly, and we look for the one that satisfies them the best possible. In Data Science and Learning Theory we often want to find a predictor that best describes a set of *training data*, but usually no predictor described the data exactly, so we look for the best possible, etc etc. We'll see a couple of applications later.

We can solve this problem using the ideas we developed above. What we are looking for is a vector $\hat{x}$ for which the error $e = b - A\hat{x}$ is as small as possible. Since the set of possible vectors $y = A\hat{x}$ is exactly $\mathrm{C}(A)$, $A\hat{x}$ is precisely the projection of b on $\mathrm{C}(A)$. As we saw in the Section above, this means that

$$A^\top(b - A\hat{x}) = 0.$$

These are known as the *normal equations* and can be rewritten as

$$A^\top A\hat{x} = A^\top b. \tag{3}$$

Remark 4.3.1. *For this to make sense it must be that* (3) *always has a solution. If we think geometrically, it is relatively easy to see that it must, because of how we constructed the normal equations. Can you give a rigorous algebraic proof of this fact? Note that essentially what you are proving is the Proposition below.*

Proposition 4.3.2. *For any matrix A, $\mathrm{C}(A^\top) = \mathrm{C}(A^\top A)$.*

Challenge 9. Try to prove this Proposition. This can be done in a few different ways. I suggest starting by trying to show that $\mathrm{rank}(A) = \mathrm{rank}(A^\top) = \mathrm{rank}(A^\top A) = \mathrm{rank}(AA^\top)$.

We know that if A has linearly independent columns, then $A^\top A$ is invertible and so we can write $\hat{x} = (A^\top A)^{-1} A^\top b$. We will address the case in which A has dependent columns shortly.

Fact 4.3.3. *A minimizer of* (2) *is also a solution of* (3). *When A has independent columns the unique minimizer $\hat{x}$ of* (2) *is given by*

$$\hat{x} = (A^\top A)^{-1} A^\top b \tag{4}$$

Exploratory Challenge 10. Similarly to the projections derivation, this derivation can also be done by differentiating (2). Try it.

4.3.2. *Linear Regression — fitting a line to data points.* One of the most common tasks in data analysis is linear regression, to fit a line through data points. Let us consider data points

$$(t_1, b_1), (t_2, b_2), \ldots, (t_m, b_m),$$

perhaps representing some attribute b over time t. If the relation between t and b is (at least partly) explained by a linear relationship then it makes sense to search for constants $\alpha_0 \in \mathbb{R}$ and $\alpha_1 \in \mathbb{R}$ such that

$$b_k \approx \alpha_0 + \alpha_1 t_k.$$

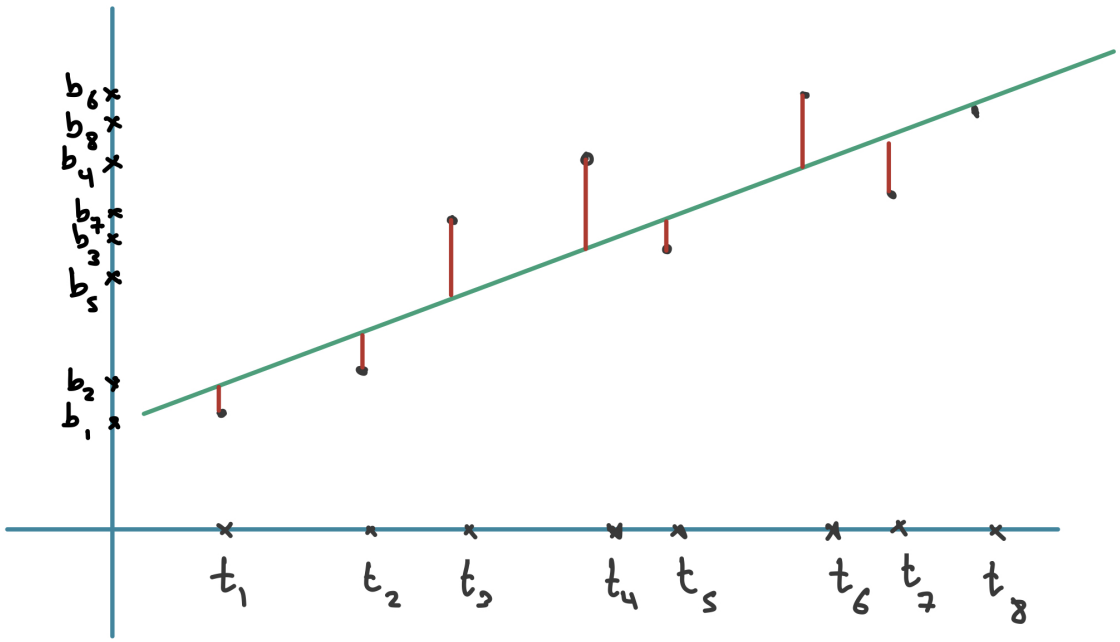


FIGURE 3. Fitting a line to points

See Figure 3. In particular, it is natural to search for α_0 and α_1 that minimize the sum of squares of the errors ("least squares"),

$$\min_{\alpha_0,\alpha_1} \sum_{k=1}^{m} (b_k - [\alpha_0 + \alpha_1 t_k])^2 .$$

In matrix-vector notation

$$\min_{\alpha_0,\alpha_1} \left\| b - A \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \right\|^2 , \tag{5}$$

where

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{m-1} \\ b_m \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{m-1} \\ 1 & t_m \end{bmatrix} .$$

As long as A has independent columns (see Remark 4.3.4) the solution to (5) is given by

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = (A^\top A)^{-1} A^\top b = \begin{bmatrix} m & \sum_{k=1}^m t_k \\ \sum_{k=1}^m t_k & \sum_{k=1}^m t_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^m b_k \\ \sum_{k=1}^m t_k b_k \end{bmatrix}$$

Remark 4.3.4. *It is worth working out what it means for the columns of A, in this example, to be linearly dependent. It essentially corresponds to all points t_k being the same, which is clearly a degenerate case of linear regression.*

Remark 4.3.5. *If the columns of A are pairwise orthogonal, then $A^\top A$ is a diagonal matrix, which is easy to invert. In this example, the columns of A being orthogonal corresponds to $\sum_{k=1}^m t_k = 0$. We could simply do a change of variables to a new time $t_k^{new} = t_k - \frac{1}{m}\sum_{i=1}^m t_i$ to achieve this. If indeed $\sum_{k=1}^m t_k = 0$ then the equation above could be easily simplified:*

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} m & 0 \\ 0 & \sum_{k=1}^m t_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^m b_k \\ \sum_{k=1}^m t_k b_k \end{bmatrix} = \begin{bmatrix} \frac{1}{m} & 0 \\ 0 & \frac{1}{\sum_{k=1}^m t_k^2} \end{bmatrix} \begin{bmatrix} \sum_{k=1}^m b_k \\ \sum_{k=1}^m t_k b_k \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{m}\sum_{k=1}^m b_k \\ \left(\sum_{k=1}^m t_k b_k\right) / \left(\sum_{k=1}^m t_k^2\right) \end{bmatrix},$$

this is an instance where having orthogonal vectors is beneficial. In this next Section we will see how to build orthonormal basis for subspaces, and some of the many benefits they have.

Challenge 11. Try to work out the actual change of variables that makes the t_k's add up to zero and derive a formula for fitting a line to points without the assumption in Remark 4.3.5

Example 4.3.6 (Fitting a Parabola)**.** *We can use Linear Algebra to do fits of many other curves (or functions), not just lines. If we believe the relationship between t_k and b_k is quadratic we could attempt to fit a Parabola:*

$$b_k \approx \alpha_0 + \alpha_1 t_k + \alpha_2 t_k^2.$$

While this isn't a linear function in t_k, this is still a linear function on the coefficients α_0, α_1, and α_2, and this is what is important. Similarly as with linear regression, it is natural to attempt to minimze

$$\min_{\alpha_0,\alpha_1,\alpha_2} \left\| b - A \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} \right\|^2, \tag{6}$$

where

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{m-1} \\ b_m \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \\ 1 & t_{m-1} & t_{m-1}^2 \\ 1 & t_m & t_m^2 \end{bmatrix},$$

and we can use the technology we developed in this section to solve this problem as well.

Challenge 12. Try to work out the example of fitting a parabola further. What is $A^\top A$? When is $A^\top A$ diagonal?

Further Reading 13. There is a whole (beautiful) area of Mathematics related to studying so-called *Orthogonal Polynomials*. The basic idea can be already hinted at from these examples: In the example of the parabola we wrote a function of t as a linear combination of the polynomials 1, t, and t^2. But we could have picked other polynomials, we could have e.g. written something like $b \approx \alpha_0' + \alpha_1'(t-2023) + \alpha_2(t^2+t)$, and a particularly good choice (that would depend on the distribution of the points t_k) might have resulted in a diagonal matrix $A^\top A$... *search "orthogonal polynomials" online to learn more.*

Further Reading 14. A lot of Machine Learning includes Linear Regression as a key component. The idea is to create, find, or *learn* features of the data points. Given n data points $t_1, \ldots t_n$ (which now can be perhaps pixel images, rather than just timepoints) we might want to do classification (for example, in the case of images, maybe we want a function that is large when the picture has a dog in it and small when it has a cat in it). It is hard to imagine that this can be done with a linear fit, but if we build good feature vectors $\varphi(t_k) \in \mathbb{R}^p$ for very large p then the function can depend on all coordinates of $\varphi(t_k)$ (the p features) and this is incredible powerful. There are several ways to construct features, a bit over a decade ago they were sometimes handmade, now

they are often learned (this is in a sense what Deep Learning does). Another important way to build (or compute with) features are the so-called Kernel Methods, you can see more in the CS Lens Lecture (Appendix C).

4.4. **Orthonormal Bases and Gram Schmidt.** When we think of (or draw) a basis of a subspace, we tend to think of (or draw) vectors that are orthogonal (have an angle of 90°) and that have the same length (length 1). Indeed, these bases have many advantages, this section is about these bases, some of their advantages, and how to find them.

Definition 4.4.1 (Orthonormal vectors)**.** *We say n vectors $q_1, \ldots, q_n \in \mathbb{R}^m$ are orthonormal if they are orthogonal and have norm 1. In other words, for all $i, j = 1, \ldots, n$*

$$q_i^T q_j = \delta_{ij},$$

where δ_{ij} is the Kronecker delta

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{7}$$

If Q is the matrix whose columns are the vectors q_i's, then the condition that the vectors are orthonormal can be rewritten as $Q^\top Q = I$.

Caution! 15. Q may not be a square matrix, and so it is not necessarily the case that $QQ^\top = I$.

Example 4.4.2. *A classical example of an orthonormal set of vectors is the canonical basis, $e_1, \ldots, e_n \in \mathbb{R}^n$ where e_i is the vector with a 1 in the i-th entry and 0 in all other entries: $(e_i)_j = \delta_{ij}$.*

When Q is a square matrix then $Q^\top Q = I$ implies also that $QQ^\top = I$ and so $Q^{-1} = Q^\top$. We call such matrices *orthogonal matrices*. This corresponds to the case when the q_i's are an orthonormal basis for all of $\mathbb{R}^n$.

Definition 4.4.3 (Orthogonal Matrix)**.** *A square matrix $Q \in \mathbb{R}^{n \times n}$ is an Orthogonal Matrix when $Q^\top Q = I$. In this case, $QQ^\top = I$, $Q^{-1} = Q^\top$, and the columns of Q form an orthonormal basis for $\mathbb{R}^n$.*

Remark 4.4.4. *It is often useful to think of an $m \times n$ matrix A as a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, that takes $x \in \mathbb{R}^n$ to $Ax \in \mathbb{R}^m$.*

$$\begin{aligned} A: \quad & \mathbb{R}^n \to \mathbb{R}^m \\ & x \to Ax \end{aligned}$$

Later in the course, when we discuss Linear Transformations*, we will, among other things, discuss which functions can be described by a matrix this way (and some properties of these functions/transformations). For now, let us just keep in mind that a matrix can be thought of as a function. It is also worth noting that this explains why in some Linear Algebra books the Nullspace is called the Kernel (it is the set of vectors x that are mapped to 0 by this function) and the Column Space is called Image, or Range, as it is the set of vectors in $\mathbb{R}^m$ that is the image of this function.*

Example 4.4.5. *The 2×2 matrix Q that corresponds to rotating, counterclockwise, the plane by θ,*

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is an orthogonal matrix.

Challenge 16. Prove that the rotation matrices R_θ are orthogonal matrices. Are there other 2×2 orthogonal matrices? If so, can you describe them all?

Example 4.4.6. *Permutation matrices are another example of orthogonal matrices.*

Challenge 17. Show that indeed permutation matrices are orthogonal matrices

Exploratory Challenge 18. One of the most important structures in Algebra is that of a group. The set of Permutations of n elements is an example of a group, two permutations can be composed to form another permutation and for every permutation there is one corresponding to undoing it (called the inverse). Permutation matrices represent the permutations, composing corresponds to matrix multiplication and the inverse permutation corresponds to the matrix inverse of the permutation matrix. There is a whole field of Mathematics, called Representation Theory, that studies matrix representations of groups (and in many important cases the matrices involved are orthogonal). Can you come up with a matrix representation of addition modulo 2? What about addition modulo 5?

Challenge 19 ($\star$). Show that for every permutation matrix P there exists a positive integer k such that $P^k = I$.

Proposition 4.4.7. *Orthogonal matrices preserve norm and inner product of vectors. In other words, if $Q \in \mathbb{R}^{n \times n}$ is orthogonal then, for all $x, y \in \mathbb{R}^n$*

$$\|Qx\| = \|x\| \text{ and } (Qx)^\top (Qy) = x^\top y$$

Proof. To show the second inequality note that, for $x, y \in \mathbb{R}^n$ we have that $(Qx)^\top (Qy) = x^\top Q^\top Qy = x^\top Iy = x^\top y$. To show the first equality note that, since for $x \in \mathbb{R}^n$ we have that $\|Qx\| \geq 0$ and $\|x\| \geq 0$, it suffices to show that the squares are equal and indeed $\|Qx\|^2 = (Qx)^\top (Qx) = x^\top x = \|x\|^2$. □

4.4.1. *Projections with Orthonormal Basis.* One of the advantages of orthonormal basis is that projections become much simpler. The reason is simple: when we discussed projections and least squares, many of the expressions we derived included $A^\top A$, but in the case when A has orthonormal columns, these all simplify as $A^\top A = I$. We collect these observations in the following proposition.

Proposition 4.4.8. *Let S be a subspace of $\mathbb{R}^m$ and $q_1, \ldots, q_n$ be an orthonormal basis for S. Let Q be the $m \times n$ matrix whose columns are the q_i's; $Q = \begin{bmatrix} q_1 & , & \cdots & , & q_n \end{bmatrix}$. Then the Projection Matrix that projects to S is given by $QQ^\top$ and the Least Squares solution to $Qx = b$ is given by $\hat{x} = Q^\top b$.*

Remark 4.4.9. *When Q is a square matrix then the projection $QQ^\top$ is simply the identity (corresponding to projecting to the entire ambient space $\mathbb{R}^n$. Even in this seemingly trivial instance, it is useful to look closer at what this operation does: For a vector $x \in \mathbb{R}^n$ it gives*

$$x = q_1 \left(q_1^\top x\right) + q_2 \left(q_2^\top x\right) + \cdots + q_1 \left(q_n^\top x\right).$$

It is writing x as a linear combination of the orthonormal basis $\{q_i\}_{i=1}^n$ (as we will see later this is sometimes referred to as a change of basis*).*[9]

4.4.2. *Gram-Schmidt Process.* Hopefully by now you are convinced that orthonormal basis are useful, now we discuss how to construct them. Fortunately, there is a relatively simple process to construct orthonormal bases, that will also suggest a new matrix factorization.

The idea is simple: If we have 2 linearly independent vectors a_1 and a_2 which span a subspace S, it is straightforward to transform them into an orthonormal basis of S: we first normalize a_1: $q_1 = \frac{a_1}{\|a_1\|}$, then subtract from a_2 a multiple of q_1 so that it becomes orthogonal to q_1, followed by a normalization step:

$$q_2 = \frac{a_2 - (a_2^\top q_1) q_1}{\left\| a_2 - (a_2^\top q_1) q_1 \right\|}.$$

[9] There are countless instances in which doing this operation is beneficial, for example one of the most important algorithms, the *Fast Fourier Transform*, is an instance of this operation.

Let us check that indeed these vectors are orthonormal: By construction they have unit norm, and

$$q_1^\top q_2 = q_1^\top \frac{a_2 - (a_2^\top q_1)q_1}{\left\|a_2 - (a_2^\top q_1)q_1\right\|} = \frac{q_1^\top a_2 - (a_2^\top q_1)q_1^\top q_1}{\left\|a_2 - (a_2^\top q_1)q_1\right\|} = \frac{0}{\left\|a_2 - (a_2^\top q_1)q_1\right\|} = 0.$$

Note that the denominator is not zero because a_1 and a_2 are linearly independent; and that, since q_1 has unit norm, $(a_2^\top q_1)q_1 = \text{proj}_{\text{Span}(q_1)}(a_2)$.

For more vectors, the idea is to this process recursively, by removing from a vector a_{k+1} the projection of it on the subspace spanned by the k vectors before it. More formally:

Algorithm 4.4.10. *[Gram-Schmidt Process] Given n linearly independent vectors $a_1, \ldots, a_n$ that span a subspace S, the Gram-Schmidt process constructs $q_1, \ldots q_n$ the following way:*

- $q_1 = \frac{a_1}{\|a_1\|}$.
- *For $k = 2, \ldots, n$ do*
  $q'_k = a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_i$
  $q_k = \frac{q'_k}{\|q'_k\|}$.

Theorem 4.4.11 (Correctness of Gram-Schmidt)**.** *Given n linearly independent vectors $a_1, \ldots, a_n$, the Gram-Schmidt process outputs an orthonormal basis for the span of $a_1, \ldots, a_n$.*

Proof. [10] We prove this by induction.[11] Let S_k be the subspace spanned by $a_1, \ldots, a_k$. Then $S = S_n$. We will show, by induction, that $q_1, \ldots, q_k$ are an orthonormal basis for S_k. It is enough to show that they are orthonormal and are in S_k since orthonormality implies linearly independence and S_k has dimension k.

For the base case, note that $\|q_1\| = 1$ and q_1 is a multiple of a_1 and so $q_1 \in S_1$.

Now we assume the hypothesis for $i = 1, \ldots k-1$ and prove it for k. By the hypothesis $q_1, \ldots, q_{k-1}$ are orthonormal, so we have to show that $\|q_k\| = 1$ and that $q_i^\top q_k = 0$ for all $1 \leq i \leq k-1$.

- Since a_k is linearly independent from the other original vectors it is not in S_{k-1} and so $q'_k \neq 0$. Thus $\|q_k\| = 1$.
- By construction $a_k \in S_k$ and so $q_k \in S_k$.
- Let $1 \leq j \leq k-1$. Since $q_1, \ldots, q_{k-1}$ are orthonormal, we have

  $$q_j^\top \left(a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_i \right) = q_j^\top a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_j^\top q_i = q_j^\top a_k - (a_k^\top q_j) = 0,$$

  and $q_j^\top q_k = \frac{1}{\|q'_k\|} q_j^\top q'_k = 0$.

[10] This is a good Theorem to try to prove yourself before reading the proof.

[11] Since this is our first proof by Induction, we will do it slowly.

□

Challenge 20. Try to do the Gram-Schmidt process for the columns of

$$\begin{bmatrix} 1 & 2 & 3 & 0 \\ 0 & 4 & 5 & 6 \\ 0 & 0 & 7 & 8 \\ 0 & 0 & 0 & 9 \end{bmatrix}.$$

Is it the case that the Gram-Schmidt process of the columns of an upper triangular matrix (with non-zero diagonal elements) is always a subset of the canonical basis? Can you come up with an example of a set of vectors for which Gram-Schmidt does not output elements of the canonical basis?

Gram-Schmidt actually provides us with a new matrix factorization. Let A be an $m \times n$ matrix with linearly independent columns $a_1, \ldots, a_n$ and Q the $m \times n$ matrix whose columns are $q_1, \ldots, q_n$ as outputted by Algorithm 4.4.10. Let $R = Q^\top A$, since each q_k is orthogonal to every a_i for $i < k$ we have that R is upper triangular. Q is not necessarily a square matrix, and so not necessarily invertible. But $QQ^\top$ is the projection on the span of the q_i's and thus also on the a_i's, this means that $QQ^\top A = A$ and so we have that $QR = QQ^\top A = A$. We call $A = QR$ the QR decomposition.

Definition 4.4.12 (QR decomposition)**.** *Let A be an $m \times n$ matrix with linearly independent columns the QR decomposition is given by*

$$A = QR,$$

where Q is an $m \times n$ matrix with orthonormal columns (they are the output of Gram Schmidt, Algorithm 4.4.10, on the columns of A) and R is an upper triangular matrix given by $R = Q^\top A$.

Remark 4.4.1012. *Note that R is a square matrix ($n \times n$), and since the columns of A are linearly independent we have $N(A) = \{0\}$ and so, since $A = QR$, we have also $N(R) = \{0\}$ and so both R and $R^\top$ are invertible.*

Fact 4.4.13. *The QR decomposition greatly simplifies calculations involving Projections and Least Squares.*

- *Since the $\mathrm{C}(A) = \mathrm{C}(Q)$ then projections on $\mathrm{C}(A)$ can be done with Q which means they are given by $\mathrm{proj}_{\mathrm{C}(A)}(b) = QQ^\top b$.*
- *The least squares solution to $Ax = b$ is $\hat{x}$ solution of the normal equations (recall (3))*

$$A^\top A\hat{x} = A^\top b.$$

Furthermore, $A^\top A = (QR)^\top (QR) = R^\top Q^\top QR = R^\top R$, and so we can write

$$R^\top R\hat{x} = R^\top Q^\top b. \tag{8}$$

Since R has independent columns (is full column rank) then $\mathrm{N}(R) = \{0\}$ and so we can simplify (8) to

$$R\hat{x} = Q^\top b, \tag{9}$$

which can be efficiently solved by back-substitution since R is a triangular matrix.

4.5. **The Pseudoinverse, also known as Moore–Penrose Inverse.** The goal of this Section is to construct an analogue to the inverse of a matrix A for matrices that have no inverse, this is

called the Pseudoinverse, or the Moore-Penrose Inverse, and we will denote it by $A^\dagger$. It is also commonly denoted by A^+.

Guiding Question 21. While not all matrices are A invertible, we saw that we can still aim to find the (or a) vector x such that Ax is as close as possible to a target vector b. Can we develop this idea to define a "pseudoinverse" for any matrix A, a matrix that is, in a sense, closest to being an inverse for A? What should "closest to being an inverse" even mean?

There are (at least) three issues we need to overcome to try to define a *pseudoinverse* for a non-invertible matrix A: (i) For some vectors b there might not be a vector x such that $Ax = b$, (ii) For some vectors b there may be more than one x such that $Ax = b$ and we would have to pick one, and (iii) even if we make such choices, it is not clear that such operation will correspond to multiplying by a matrix $A^\dagger$.

Let $A \in \mathbb{R}^{m\times n}$ be an $m \times n$ matrix. There are a couple of different ways we could try to define a *pseudoinverse* $A^\dagger$ for a non-invertible matrix A. Let us start by building on what we discussed on Section 4.3 (Least Squares Approximations), if the columns of A are linearly independent that it would make sense to build $A^\dagger$ such that $A^\dagger b$ is the Least Squares Solution $\hat{x} = (A^\top A)^{-1}A^\top b$ (the vector $\hat{x}$ such that $A\hat{x}$ is as close as possible to b), and so for matrices A with independent columns we will define $A^\dagger = (A^\top A)^{-1}A^\top$. This motivates the following definition.

Definition 4.5.1 (Pseudoinverse for matrices with full column rank)**.** *For $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) = n$ we define the pseudo-inverse $A^\dagger \in \mathbb{R}^{n\times m}$ of A as*

$$A^\dagger = (A^\top A)^{-1}A^\top.$$

Proposition 4.5.2. *For $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) = n$, the pseudoinverse $A^\dagger$ is a left inverse of A, meaning that $A^\dagger A = I$.*

Proof. Since $\operatorname{rank}(A) = n$, $A^\top A$ is invertible. Furthermore, $A^\dagger A = (A^\top A)^{-1}A^\top A = I$. □

Let us know consider the case for which the rows are linearly independent (in other words, $A \in \mathbb{R}^{m\times n}$ is full row rank; or equivalently $\operatorname{rank}(A) = m$). One natural way to define pseudoinverse is by noting that $A^\top$ is full column rank and to define $A^\dagger$ as

$$\left(\left(A^\top\right)^\dagger\right)^\top = \left(\left(\left(A^\top\right)^\top\left(A^\top\right)\right)^{-1}\left(A^\top\right)^\top\right)^\top = \left(\left(AA^\top\right)^{-1}A\right)^\top = A^\top\left(AA^\top\right)^{-1}.$$

Definition 4.5.3 (Pseudoinverse for matrices with full row rank)**.** *For $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) = m$ we define the pseudo-inverse $A^\dagger \in \mathbb{R}^{n\times m}$ of A as*

$$A^\dagger = A^\top(AA^\top)^{-1}.$$

Proposition 4.5.4. *For $A \in \mathbb{R}^{m\times n}$ with $\text{rank}(A) = m$, the pseudoinverse $A^\dagger$ is a right inverse of A, meaning that $AA^\dagger = I$.*

Proof. Since $\text{rank}(A) = m$, $AA^\top$ is invertible. Furthermore, $AA^\dagger = AA^\top(AA^\top)^{-1} = I$. □

Let us try to understand what $A^\dagger$ is achieving for full row rank matrices A. Since A is full row rank, for all $b \in \mathbb{R}^m$, there exists $x \in \mathbb{R}^n$ such that $Ax = b$. The issue is that there are potentially many such vectors. A natural strategy in this case is to pick, among all such vectors, the one with smallest norm.[12] In other words to solve

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|^2 \\ s.t. \quad & Ax = b, \end{aligned} \tag{10}$$

where s.t. stands for "subject to" or "such that". If x_1 and x_2 are vectors such that $Ax_1 = Ax_2 = b$ then $x_1 - x_2 \in \text{N}(A)$, and conversely, if $Ax = b$ and $y \in \text{N}(A)$ then $A(x+y) = b$. Thus, given one vector x_1 such that $Ax_1 = b$ the set of solutions to $Ax = b$ are all vectors of the form $x_1 + y$ where $y \in \text{N}(A)$. So we would like to find the minimum $\|x_1 + y\|$ among all vectors $y \in \text{N}(A)$. Let us write $x_1 = \left(x_1 - \text{proj}_{\text{N}(A)}(x_1)\right) + \text{proj}_{\text{N}(A)}(x_1)$. Since $y \in \text{N}(A)$ we have that $\left(x_1 - \text{proj}_{\text{N}(A)}(x_1)\right) \perp \left(y + \text{proj}_{\text{N}(A)}(x_1)\right)$ and so, by Pythagoras,

$$\begin{aligned} \|x_1 + y\|^2 &= \left\| \left(x_1 - \text{proj}_{\text{N}(A)}(x_1)\right) + \text{proj}_{\text{N}(A)}(x_1) + y \right\|^2 \\ &= \left\| x_1 - \text{proj}_{\text{N}(A)}(x_1) \right\|^2 + \left\| \text{proj}_{\text{N}(A)}(x_1) + y \right\|^2, \end{aligned}$$

and so picking $y = -\text{proj}_{\text{N}(A)}(x_1)$ yields the smallest norm solution. Since the vectors orthogonal to $\text{N}(A)$ are precisely the vectors that are in the row space of A, $\text{C}(A^\top)$. We just proved:

Proposition 4.5.5. *For a full row rank matrix A, the (unique) solution to* (10) *is given by the vector $\hat{x} \in \text{C}(A^\top)$ that satisfies the constraint $A\hat{x} = b$.*

$A^\dagger$ is precisely the matrix that "takes b to $\hat{x}$ solution of (10)".

Proposition 4.5.6. *For a full row rank matrix A, the (unique) solution to* (10) *is given by the vector $\hat{x} = A^\dagger b$.*

Proof. By using Proposition 4.5.5 we just need to show that $\hat{x} = A^\dagger b$ satisfies $A\hat{x} = b$ and that $\hat{x} = A^\dagger b$ is in $\text{C}(A^\top)$. Both these are easy to verify: $A\hat{x} = AA^\dagger b = AA^\top(AA^\top)^{-1}b = b$ and

[12] This idea, of picking the smallest (or simplest) solution among many possibilities goes far beyond Linear Algebra and is known as "regularization" in Statistics, Machine Learning, Signal Processing, and Image Processing, etc. It can be viewed as a mathematical version of the famous "Occam's razor" principle in Philosophy.

$\hat{x} = A^\dagger b = A^\top \left((AA^\top)^{-1} b\right)$ and so $\hat{x} \in \mathrm{C}(A^\top)$. □

Guiding Question 22. We would like to define $A^\dagger$ for all matrices, not just full rank matrices. A natural construction would be to try to define $A^\dagger$ to be the matrix that takes a vector b to the smallest norm solution of the normal equations (3).

To define pseudoinverse of a non full rank matrix A we can do it via de $A = CR$ decomposition (recall from Part I of the course and/or see Appendix A(2)). For $A \in \mathbb{R}^{m\times n}$, with $\mathrm{rank}(A) = r$, the CR decomposition writes $A = CR$ where $C \in \mathbb{R}^{m\times r}$ has the first r linearly independent columns of A and $R \in \mathbb{R}^{r\times n}$ is upper triangular. Note that C is full column rank and R is full row rank.

Definition 4.5.7 (Pseudoinverse for all matrices)**.** *For $A \in \mathbb{R}^{m\times n}$, with $\mathrm{rank}(A) = r$, with CR decomposition $A = CR$ we define the pseudoinverse $A^\dagger$ as*

$$A^\dagger = R^\dagger C^\dagger,$$

which can be rewritten as

$$A^\dagger = R^\top \left(RR^\top\right)^{-1} \left(C^\top C\right)^{-1} C^\top = R^\top \left(C^\top CRR^\top\right)^{-1} C^\top = R^\top \left(C^\top AR^\top\right)^{-1} C^\top.$$

The following proposition shows that indeed this definition achieves what was asked in Guiding Question 22.

Proposition 4.5.8. *Given $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^n$, the (unique) solution to*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|^2 \\ s.t. \quad & A^\top Ax = A^\top b, \end{aligned} \tag{11}$$

is given by $\hat{x} = A^\dagger b$.

Proof. Let r be the rank of A and $A = CR$ with $C \in \mathbb{R}^{m\times r}$ and $R \in \mathbb{R}^{r\times n}$. Then $\hat{x} = A^\dagger b = R^\top \left(C^\top AR^\top\right)^{-1} C^\top b$. Thus,

$$A^\top A\hat{x} = A^\top AR^\top \left(C^\top AR^\top\right)^{-1} C^\top b = R^\top C^\top AR^\top \left(C^\top AR^\top\right)^{-1} C^\top b = R^\top C^\top b = A^\top b.$$

Using Proposition 4.5.5, to show that it is the smallest norm solution we just need to show that $\hat{x} \in \mathrm{C}(A^\top A)$, but by Proposition 4.3.2 it is enough to show that $\hat{x} \in \mathrm{C}(A^\top)$ and since $\mathrm{C}(A^\top) = \mathrm{C}(R^\top)$ we have that $\hat{x} = R^\top \left(C^\top AR^\top\right)^{-1} C^\top b \in \mathrm{C}(A^\top)$. □

In this proof, the only property of the matrices CR we used is that $A = CR$ and both C and R are full rank. So we have actually shown that we can compute the pseudoinverse from any full rank factorization, not just specifically the CR decomposition. We write it here as a proposition.

Proposition 4.5.9. *For $A \in \mathbb{R}^{m\times n}$, with $\operatorname{rank}(A) = r$, and let $S \in \mathbb{R}^{m\times r}$ and $T \in \mathbb{R}^{r\times n}$ such that $A = ST$. Then,*

$$A^\dagger = T^\dagger S^\dagger.$$

Remark 4.5.1009. *Note that If $A = ST$ and $\operatorname{rank}(A) = r$ then rank$(S) \geq r$ and $\operatorname{rank}(T) \geq r$ and so the matrices ST in Proposition 4.5.9 are indeed full rank (either full column rank or full row rank).*

Proposition 4.5.10. *Given $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, we have*

(1) $(AB)^\dagger = B^\dagger A^\dagger$, *as long as* $\operatorname{rank}(A) = \operatorname{rank}(B) = n$.
(2) $\left(A^\top\right)^\dagger = \left(A^\dagger\right)^\top$,
(3) $AA^\dagger$ *is symmetric, and is the projection matrix for projection on* $\mathrm{C}(A)$,
(4) $A^\dagger A$ *is symmetric, and is the projection matrix for projection on* $\mathrm{C}(A^\top)$.

Challenge 23. Prove Proposition 4.5.10. (Hint: use Proposition 4.5.9).

Challenge 1023. Given $A \in \mathbb{R}^{m\times n}$, show that

$$AA^\dagger A = A \quad \text{and} \quad A^\dagger AA^\dagger = A^\dagger.$$

Proposition 4.5.11. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix and recall that $\mathrm{C}(A)$ and $\mathrm{C}(A^\top)$ denote respectively its column and row spaces. When $A : x \to Ax$ is viewed as a function from $\mathrm{C}(A^\top)$ to $\mathrm{C}(A)$ it is a bijection. In other words, for all $b \in \mathrm{C}(A)$ there is one and only one $x \in \mathrm{C}(A^\top)$ such that $Ax = b$.*

Challenge 24. Prove Proposition 4.5.11

Further Remark 1024. A different way to define the Pseudo-Inverse of a matrix A is to ask for a matrix $A^\dagger$ that satisfies the conditions in the Challenge 1023 and that both $AA^\dagger$ and $A^\dagger A$ are symmetric. It is nontrivial, but it turns out these conditions are enough to define $A^\dagger$.

5. Linear Transformations and Determinants

Further Remark 25. In this part of the course we slightly deviate from [Str23] and will introduce Linear Transformations, before Determinants. To keep the numbering compatible [Str23] we number the section on Linear Transformations as 5.0.2 (this material is in Chapter 8 of [Str23]).

As we pointed out in Remark 4.4.4, we can view an $m \times n$ matrix A as a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, that takes $x \in \mathbb{R}^n$ to $Ax \in \mathbb{R}^m$

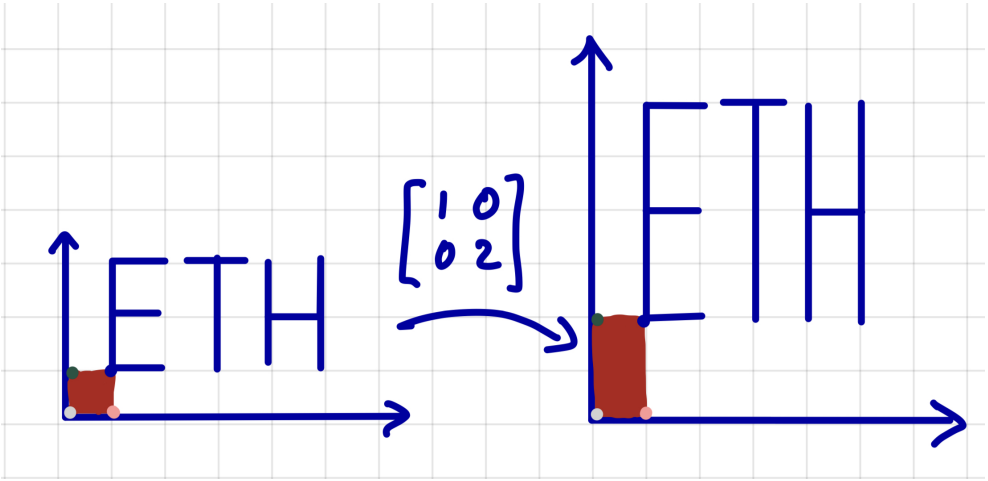
$$\begin{aligned} A: \quad & \mathbb{R}^n \to \mathbb{R}^m \\ & x \to Ax. \end{aligned}$$

These functions have a very important property, they are linear: for all $x_1, x_2 \in \mathbb{R}^n$ and any $\alpha \in \mathbb{R}$ we have $A(x_1 + x_2) = Ax_1 + Ax_2$ and $A(\alpha x_1) = \alpha A x_1$. We will call functions satisfying these properties *Linear Transformations*.

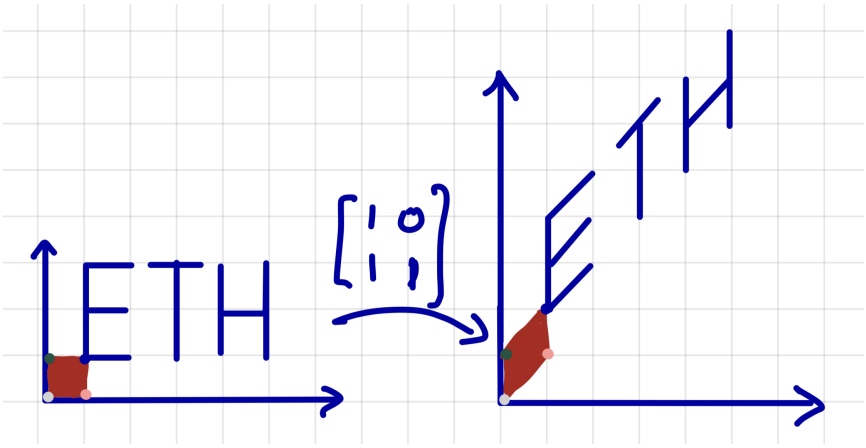
5.0.1. *Linear Transformations as transformations of* $\mathbb{R}^n$.

We will now focus on linear transformations from $\mathbb{R}^n$ to itself, corresponding to square matrices $A \in \mathbb{R}^{n \times n}$. Instead of thinking simply of how $x \to Ax$ maps a vector x to Ax, it is useful to think of this transformation being applied to all of $\mathbb{R}^n$ and to view it as a transformation of the entire $\mathbb{R}^n$. Let us focus on $\mathbb{R}^2$ for better visualization and look at a few examples.

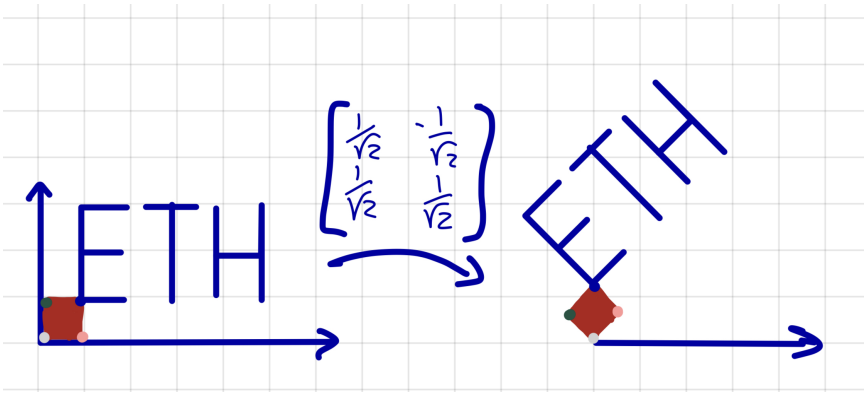
Example 5.0.1 (Stretch)**.** *The matrix* $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ *corresponds to stretching by a factor of 2 in the vertical axis. Notice: the first column of A is the image of* e_1 *and the second the image of* e_2*.*



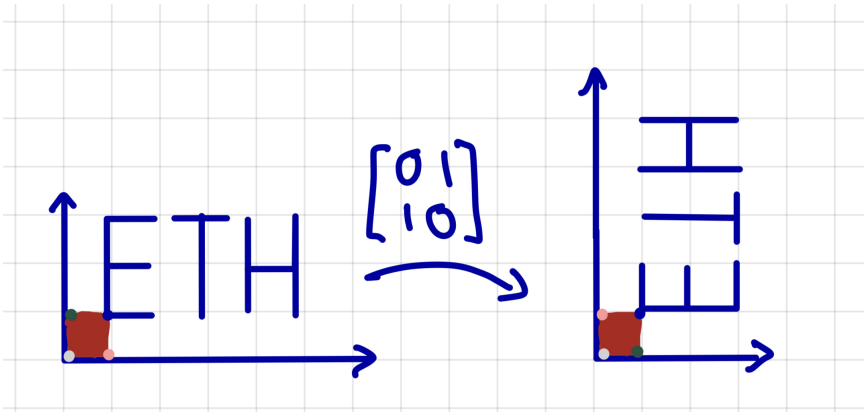
Example 5.0.2 (Shear)**.** *The matrix* $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ *corresponds to a shearing transformation given by* $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 + x_2 \end{bmatrix}$.



Example 5.0.3 (Rotation). *The matrix* $A = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ *corresponds to a counter-clockwise rotation by* $\frac{\pi}{4}$ *(or* 45^o*)*



Example 5.0.4 (Reflection). *The matrix* $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ *corresponds to a reflection by the diagonal line* $x_2 = x_1$.



Challenge 26. Draw a few more linear transformations. Draw also one corresponding to a non-invertible matrix A. Try to draw one in $\mathbb{R}^3$ as well.

Challenge 27. Show that a linear transformation takes a triangle to either a triangle, a line segment connecting two points, or a point. What can you say about the rank or invertibility of the corresponding matrix, depending on which of the three objects is the image of a triangle?

5.0.2. *Definition of Linear Transformations.*

We now treat linear transformations more formally and write the definition of Linear Transformation for any vector space U and V even though in this section we will only treat $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$ (so you can, for now, fully restrict your attention to this setting). Later on, this more general definition will allow us, for example, to discuss linear transformations between subspaces of $\mathbb{R}^n$ and $\mathbb{R}^m$.

Definition 5.0.5 (Linear Transformation)**.** *Given two vectors spaces U and V, a Linear Transformation is a function $T : U \to V$ such that, for all $u_1, u_2 \in U$ and $\alpha \in \mathbb{R}$ we have*

$$T(u_1 + u_2) = T(u_1) + T(u_2)$$

and

$$T(\alpha u_1) = \alpha T(u_1).$$

Before proceeding let us "collect" a few facts about Linear Transformations.

Proposition 5.0.6. *Let $T : U \to V$ be a linear transformation and k a positive integer. For all $u_1, \ldots, u_k \in U$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ we have*

$$T\left(\alpha_1 u_1 + \cdots + \alpha_k u_k\right) = \alpha_1 T(u_1) + \cdots + \alpha_k T(u_k).$$

Challenge 28. Prove Proposition 5.0.6, iteratively (by induction) using the properties of Linear Transformations.

The most central implication of Proposition 5.0.6 is the fact that the value of T in a basis of U fully determines T.

Proposition 5.0.7. *Let $T : U \to V$ and $L : U \to V$ be two linear transformations that take the same value in a basis $u_1, \ldots, u_n$ of U. Then $T = L$.*

Proof. Since $u_1, \ldots, u_n$ is a basis of U, any $u \in U$ can be written as $u = \alpha_1 u_1 + \cdots + \alpha_n u_n$. Using Proposition 5.0.6 we have

$$\begin{aligned} T(u) = T\big(\alpha_1 u_1 + \cdots + \alpha_n u_n\big) &= \alpha_1 T(u_1) + \cdots + \alpha_n T(u_n) = \\ &= \alpha_1 L(u_1) + \cdots + \alpha_n L(u_n) = L\big(\alpha_1 u_1 + \cdots + \alpha_n u_n\big) = L(u) \end{aligned}$$

□

Proposition 5.0.8. *Given a basis $u_1, \ldots, u_n$ of U, and any $v_1, \ldots, v_n \in V$ there is a Linear Transformation $T : U \to V$ such that, for all $1 \leq i \leq n$, $T(u_i) = v_i$.*

Challenge 29. Prove Proposition 5.0.8.

Example 5.0.9. *A few examples of linear transformations:*

(1) The identity map $T : \mathbb{R}^n \to \mathbb{R}^n$ given by $T(x) = x$,
(2) For any matrix A, the map $x \to Ax$,
(3) For a vector $v \in \mathbb{R}^n$ the map $T : \mathbb{R}^n \to \mathbb{R}$ given by $T(u) = v^\top x$,
(4) The map $T : \mathbb{R}^n \to \mathbb{R}^n$ given by $T(x) = 0$.

A few examples of functions that are not linear transformations:

(1) For a vector $v \in \mathbb{R}^n$ (such that $v \neq 0$) the map $T : \mathbb{R}^n \to \mathbb{R}$ given by $T(u) = v + x$,
(2) The map $T : \mathbb{R}^n \to \mathbb{R}$ given by $T(x) = \|x\|$,
(3) The map $T : \mathbb{R}^n \to \mathbb{R}$ given by $T(x) = \|x\|^2$,
(4) The map $T : \mathbb{R}^n \to \mathbb{R}^n$ given by $T(x) = \frac{1}{\|x\|}x$.

Challenge 30. Show that the first batch of examples above indeed correspond to linear transformations, and that the second does not.

It is easy to see that given an $m \times n$ matrix A, the function $x \to Ax$ is a Linear Transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, what we will show now that the converse is also true.

Proposition 5.0.10. *For any Linear Transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, there exists an $m \times n$ matrix A such that $T(x) = Ax$ for all $x \in \mathbb{R}^n$.*

Proof. We will prove this proposition constructively. Let $e_1, \ldots, e_n$ be the canonical basis of $\mathbb{R}^n$ (the i-th basis element has a 1 in the i-th entry and zeros elsewhere, or in other words $(e_i)_j = \delta_{ij}$). We write $x = x_1 e_1 + \cdots + x_n e_n$, then by linearity of T,

$$T(x) = x_1 T(e_1) + \cdots + x_n T(e_n) = \begin{bmatrix} | & & | \\ T(e_1) & \cdots & T(e_n) \\ | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = Ax,$$

for $A = \begin{bmatrix} | & & | \\ T(e_1) & \cdots & T(e_n) \\ | & & | \end{bmatrix}$. □

Challenge 31. ($\star$) Can you describe the linear transformation corresponding to $A^\dagger$? (in terms of the linear transformation corresponding to a matrix A)

Exploratory Challenge 32. Can you describe the linear transformation corresponding to $A^\top$? (in terms of the linear transformation corresponding to a matrix A)

Further Remark 33. Since we are restricting ourselves to $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$ we are identifying an element $x \in \mathbb{R}^n$ with its coordinates in the canonical basis $x = x_1 e_1 + \cdots + x_n e_n$. In general, if we use a different basis for U and V we will have a matrix representation for each linear transformation, but it will potentially correspond to a different matrix, it will be the matrix that describes the map from the coordinates in the basis of U to the ones in the basis of V, later in the course we will see some examples, and we will briefly discuss how to go from a matrix representation in one basis to that on another basis, a so-called *change of basis*.

Proposition 5.0.11. *Given two linear transformations $T : \mathbb{R}^n \to \mathbb{R}^m$ and $L : \mathbb{R}^m \to \mathbb{R}^p$, with corresponding matrices (as given by Proposition 5.0.10) $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times m}$ the linear transformation $L \circ T$ (given by $L \circ T(x) = L(T(x))$ corresponds to multiplying by the matrix BA. In other words $L \circ T(x) = BAx$.*

Challenge 34. Prove Proposition 5.0.11.

5.1. **The Determinant.** We will now introduce the notion of determinant $\det(A)$ of a square matrix A. While this has a somewhat involved definition for $n \times n$ matrices, it is useful to first discuss what the determinant geometrically corresponds to, and to focus on small matrices.

In a nutshell, **the determinant of a matrix is a number that corresponds to how much the associated linear transformation inflates space, it corresponds precisely to the volume (or area, in $\mathbb{R}^2$) of the image of the unit cube (the red square in the pictures above in $\mathbb{R}^2$); with a negative sign when the orientation changes (in the pictures above in $\mathbb{R}^2$, when the order of the colored dots, on the red square, changed).** If we think about the determinant this way, then many of the properties we will list below can be intuitively understood (while it is hard to do so from the formula for the $n \times n$ determinant). For this reason, this section will be somewhat less proof-based, and rather focus on the most relevant properties of the determinant.

Remark 5.1.1. *Grant Sanderson has a website* `https://www.3blue1brown.com/` *and Youtube channel* `https://www.youtube.com/3blue1brown` *with excellent animation-heavy explanations of topics in Mathematics, including Linear Algebra. I particularly recommend the video on Determinants, it has also 3 dimensional visualizations that are harder to*

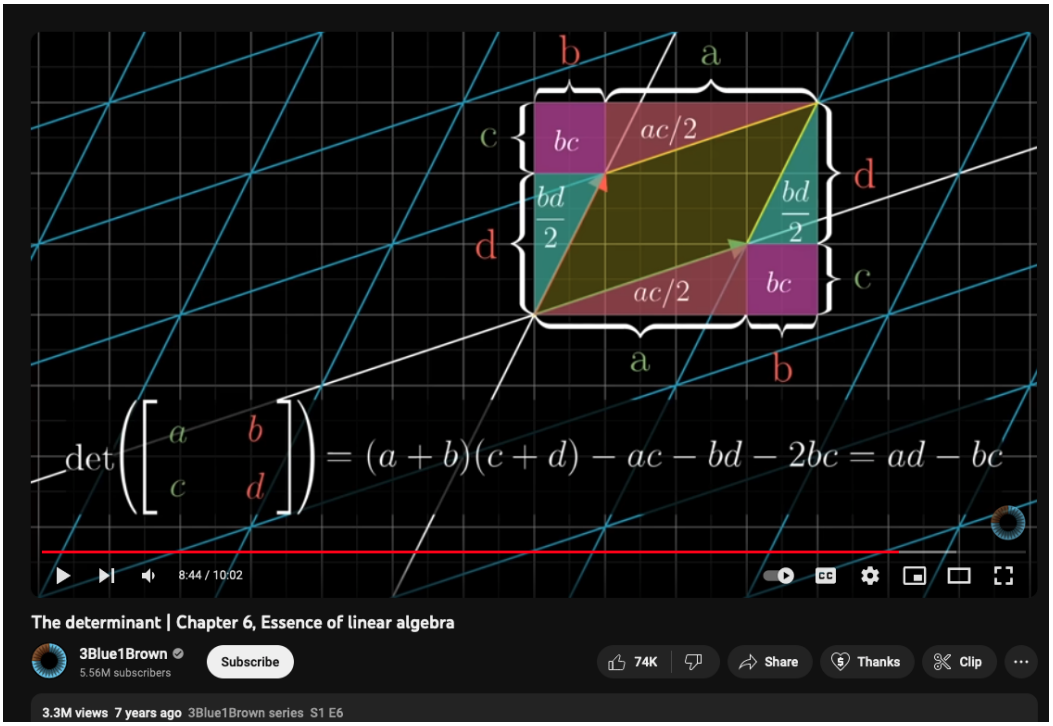


FIGURE 4. Calculation in 3Blue1Brown's video (see Remark 5.1.1) computing the determinant of a 2×2 matrix as the area of the image of the unit square after a linear transformation (that does not change orientation).

do on a static medium. You can find it here `https://youtu.be/Ip3X9LOh2dk` *or here* `https://www.3blue1brown.com/lessons/determinant`. *See also Figure 4.*

A calculation of the area of the image of the unit square by left-multiplication by a 2×2 matrix shows (see Figure 4) that

$$\left| \begin{array}{cc} a & b \\ c & d \end{array} \right| := \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

Before we actually formally define determinant for $n \times n$ matrices we will state some of the most important properties of the determinant, you can find the actual definition in Definition 5.1.6.

5.1.1. *Determinant and invertibility.* Since a square matrix is invertible if the image is full-dimensional, which corresponds to the image of the unit square/cube having non-zero area/volume, then $\det(A) \neq 0$ if and only if A is invertible. This is the following proposition.

Proposition 5.1.2. *A matrix $A \in \mathbb{R}^{n \times n}$ is invertible if and only if*

$$\det(A) \neq 0.$$

In fact, let us try to invert the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, just by naive calculations, not using elimination.[13] If $a = b = 0$ or $c = d = 0$ then the matrix is not invertible (it has a 0 row).[14] Let's assume either a or b is non-zero and that either c or d is nonzero. We are looking for a matrix $\begin{bmatrix} w & x \\ y & z \end{bmatrix}$ such that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since either a or b is non-zero and either c or d is nonzero, neither of the rows of the matrix are zero. Also, the second/first column of the inverse needs to be orthogonal to the first/second row of the original matrix, and vice-versa.

We then must have $\begin{bmatrix} x \\ z \end{bmatrix} = \alpha_1 \begin{bmatrix} -b \\ a \end{bmatrix}$ and $\begin{bmatrix} w \\ y \end{bmatrix} = \alpha_2 \begin{bmatrix} d \\ -c \end{bmatrix}$ for some $\alpha_1, \alpha_2 \in \mathbb{R}$.

Since $\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} w \\ y \end{bmatrix} = 1$ we have: $\alpha_1 \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} d \\ -c \end{bmatrix} = 1$, which gives $\alpha_1 = \frac{1}{ad-bc}$, note that the denominator is exactly $\det(A)$ which is non-zero when A is invertible. A similar calculation gives $\alpha_2 = \frac{1}{ad-bc} = \frac{1}{\det(A)}$. This gives a formula for the inverse of 2×2 matrices.

Proposition 5.1.3. *Given a 2×2 matrix A with $\det(A) \neq 0$, the inverse is given by*

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

5.1.2. *Determinant and volumes.*

Proposition 5.1.4. *Given matrices $A, B \in \mathbb{R}^{n \times n}$ we have*

$$\det(AB) = \det(A)\det(B).$$

While proving this from the definition of determinant is nontrivial, it is relatively easy to intuitively see why it is true if we recall that determinant measures areas/volumes: Since the area of the image of a square/cube does not change depending on the location of the initial square/cube, and any (nice enough) region of $\mathbb{R}^n$ can be approximated by the union of small squares/cubes (see Figure 5), then the determinant is also the area/volume of the image any (nice enough) unit area/volume 1 region, and so $\det(AB) = \det(A)\det(B)$ (since the image by AB of a unit square/cube is the image by B of the image by A of the same unit square/cube).

[13] Elimination is a much better way to do it in general, but bear with me as I am trying to illustrate something, not invert the matrix as efficiently as possible.

[14] Note that this is not a necessary condition for non-invertibility, as the all-ones matrix is not invertible while having no zero rows.

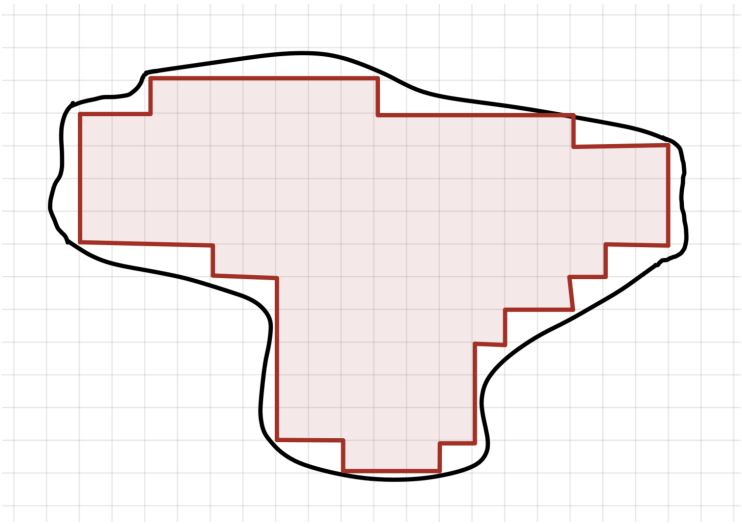


FIGURE 5. Approximation of a region by unit squares

5.1.3. *The Definition.* We now give the definition of determinant for $n \times n$ matrices. Before we define determinant we need first to discuss permutations.

Definition 5.1.5 (Sign of Permutation)**.** *Given a permutation* $\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$ *of* n *elements, its sign* $\operatorname{sgn}(\sigma)$ *can be* 1 *or* -1*. The sign counts the parity of the number of pairs of elements that are out of order (sometimes called inversions) after applying the permutation. In other words,*

$$\operatorname{sgn}(\sigma) = \begin{cases} 1 & \text{if } |(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\} \text{ such that } i < j \text{ and } \sigma(i) > \sigma(j)| \text{ is even,} \\ -1 & \text{if } |(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\} \text{ such that } i < j \text{ and } \sigma(i) > \sigma(j)| \text{ is odd.} \end{cases}$$

Exploratory Challenge 35. The sign of a permutation has many nice properties. Try to prove a couple of them:

(1) The sign of a permutation is multiplicative, i.e.: for two permutations σ, γ we have that $\operatorname{sgn}(\sigma \circ \gamma) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\gamma)$.
(2) For all $n \geq 2$, exactly half of the permutations have sign 1 and exactly half have sign -1.

the identity if 1, the sign of a transposition (a permutation that only swaps two elements) is -1 and for two permutations σ, γ we have that $\operatorname{sgn}(\sigma \circ \gamma) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\gamma)$.

Definition 5.1.6. *Given a square matrix* $A \in \mathbb{R}^{n \times n}$ *the determinant* $\det(A)$ *is defined as*

$$\det(A) = \sum_{\sigma \in \Pi_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)},$$

where Π_n *is the set of all permutations of* n *elements.*

From this Definition one can verify the following propositions.

Proposition 5.1.7. *Given a permutation matrix $P \in \mathbb{R}^{n\times n}$ corresponding to a permutation σ, then* $\det(P) = \operatorname{sgn}(\sigma)$. *We sometimes also write* $\operatorname{sgn}(P)$.

Proposition 5.1.8. *Given a triangular (either upper- or lower-) matrix $T \in \mathbb{R}^{n\times n}$ we have*

$$\det(T) = \prod_{k=1}^{n} T_{kk},$$

in particular, $\det(I) = 1$.

Proposition 5.1.9. *Given a matrix $A \in \mathbb{R}^{n\times n}$ we have*

$$\det(A^\top) = \det(A).$$

The following is a consequence of the propositions above (and the only proof we'll do in this section)

Proposition 5.1.10. *If $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix then*

$$\det(Q) = 1 \quad \textit{or} \quad \det(Q) = -1.$$

Proof. By Propositions 5.1.8 and 5.1.4 we have $1 = \det(I) = \det(Q^\top Q) = \det(Q^\top)\det(Q)$. by Proposition 5.1.9 we have $1 = \det(Q)^2$ and so $\det(Q)$ is 1 or -1. □

Following the same line of argument we also have

Proposition 5.1.11. *Given a matrix $A \in \mathbb{R}^{n\times n}$ such that* $\det(A) \neq 0$*, then A is invertible and*

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

5.1.4. 3×3 *matrices.*

If A is a 1×1 matrix, since there is only one permutation of 1 element (the permutation $\sigma(1) = 1$, which has sign 1), we have $\det(A) = A_{11} = A$.

For 2×2 matrices: There are two permutations σ_1 the identity permutation (that doesn't move any element, which has sign 1) and σ_2 the permutation that swaps the two elements (which has sign -1). So, for A a 2×2 matrix, we have

$$\det(A) = \sum_{\sigma\in\Pi_2} \operatorname{sgn}(\sigma)\prod_{i=1}^{2} A_{i,\sigma(i)} = (+1)\prod_{i=1}^{2} A_{i,\sigma_1(i)} + (-1)\prod_{i=1}^{2} A_{i,\sigma_2(i)} = A_{11}A_{22} - A_{12}A_{21}.$$

This corresponds precisely to,

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

For 3×3 matrices there are $3! = 6$ permutations, so there will be 6 terms. For A a 3×3 matrix, we can write its determinant as (where an empty entry corresponds to a zero entry)

$$\begin{aligned}
\det(A) &= \begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} \\
&= \begin{vmatrix} A_{11} & & \\ & A_{22} & \\ & & A_{33} \end{vmatrix} + \begin{vmatrix} & A_{12} & \\ A_{21} & & \\ & & A_{33} \end{vmatrix} + \begin{vmatrix} & A_{12} & \\ & & A_{23} \\ A_{31} & & \end{vmatrix} \\
&\quad + \begin{vmatrix} & & A_{13} \\ & A_{22} & \\ A_{31} & & \end{vmatrix} + \begin{vmatrix} & & A_{13} \\ A_{21} & & \\ & A_{32} & \end{vmatrix} + \begin{vmatrix} A_{11} & & \\ & & A_{23} \\ & A_{32} & \end{vmatrix} \\
&= A_{11}A_{22}A_{33} - A_{12}A_{21}A_{33} + A_{12}A_{23}A_{31} - A_{13}A_{22}A_{31} + A_{13}A_{21}A_{32} - A_{11}A_{23}A_{32}.
\end{aligned}$$

There is another convenient way of writing this determinant

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = A_{11} \begin{vmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{vmatrix} - A_{12} \begin{vmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{vmatrix} + A_{13} \begin{vmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{vmatrix}. \tag{12}$$

In general, these terms are called the co-factors of A.

Definition 5.1.12. *Given $A \in \mathbb{R}^{n \times n}$, for each $1 \leq i, j \leq n$ let $\mathscr{A}_{ij}$ denote the $(n-1) \times (n-1)$ matrix obtained by removing row i and column j from A. Then we define the co-factors of A as*

$$C_{ij} = (-1)^{i+j} \det(\mathscr{A}_{ij}).$$

Just as in (12), the determinant can be written in terms of the co-factors.

Proposition 5.1.13. *Let $A \in \mathbb{R}^{n \times n}$, for any $1 \leq i \leq n$,*

$$\det(A) = \sum_{j=1}^{n} A_{ij} C_{ij}.$$

The formula we derived above for the inverse of 2×2 matrices (Proposition 5.1.3), also has an analogue in n dimensions.

Proposition 5.1.14. *Given $A \in \mathbb{R}^{n\times n}$ with $\det(A) \neq 0$ we have*

$$A^{-1} = \frac{1}{\det(A)} C^\top,$$

where C is the $n \times n$ matrix with the co-factors of A as entries.

One good way to think of this proposition is as the identity $AC^\top = \det(A)I$.

Remark 5.1.15. *Computationally speaking, this is not a good way to compute the inverse, as it involves computing many determinants.*

Challenge 36. Verify that Proposition 5.1.14 indeed corresponds to Proposition 5.1.3 when $n = 2$.

Exploratory Challenge 37. Try to prove Proposition 5.1.14 by showing that $AC^\top = \det(A)I$. Perhaps start with $n = 3$. You can also use Cramer's Rule (below) to prove this.

5.1.5. *Cramer's Rule.* The determinant also allows us to write a formula for the solution of the linear system of the type $Ax = b$ when $A \in \mathbb{R}^{n\times n}$ and $\det(A) \neq 0$. The idea is simple, we will illustrate it here for $n = 3$.

If $\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$, then we have

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 & 0 & 0 \\ x_1 & 1 & 0 \\ x_3 & 0 & 1 \end{bmatrix} = \begin{bmatrix} b_1 & A_{12} & A_{13} \\ b_2 & A_{22} & A_{23} \\ b_3 & A_{32} & A_{33} \end{bmatrix}.$$

Since the determinant is multiplicative, and the determinant of the second matrix in the expression is x_1, we have

$$\det(A)x_1 = \det(\mathscr{B}_1),$$

where $\mathscr{B}_1$ is the matrix obtained by A by replacing the first column of A with the vector b.

Since we can do this for any of the columns, we have $x_j = \det(\mathscr{B}_j)/\det(A)$. In general

Proposition 5.1.16 (Cramer's Rule)**.** *Let $A \in \mathbb{R}^{n\times n}$ such that $\det(A) \neq 0$ and $b \in \mathbb{R}^n$ then the solution $x \in \mathbb{R}^n$ of $Ax = b$ is given by*

$$x_j = \frac{\det(\mathscr{B}_j)}{\det(A)},$$

where $\mathscr{B}_|$ is the matrix obtained by A by replacing the j-th column of A with the vector b.

Remark 5.1.17. *As with the formula for the inverse: computationally speaking, this is not a good way to solve linear systems, as it involves computing many determinants.*

5.1.6. *Elimination and the Determinant.* The definition we used for Determinant involves a formula with $n!$ terms, it is computational infeasible for even moderate levels of n (it is faster than exponential! For example, 100! has almost 160 digits!), in practice the determinant of a matrix A is computed by Gaussian Elimination and the matrix decomposition $PA = LU$ (P permutation and so $\det(P) = \operatorname{sgn}(P)$, U is upper triangular and L is lower triangular with only 1s in the diagonal, and so $\det(L) = 1$) and so we would have

$$\det(A) = \frac{1}{\det(P)} \det(L)\det(U) = \operatorname{sgn}(P)\det(U), \tag{13}$$

and since U is a triangular matrix its determinants can be easily computed by Proposition 5.1.8.

Alternatively, one can also think of Gaussian Elimination as directly computing the determinant via the following two propositions

Proposition 5.1.18. *If A is an $n \times n$ matrix and P is a permutation that swaps two elements, meaning that PA corresponds to swapping two rows of A then* $\det(PA) = -\det(A)$.

Proposition 5.1.19. *The determinant is linear in each row (or each column). In other words, for any $a_0, a_1, a_2 \ldots, a_n \in \mathbb{R}^n$ and $\alpha_0, \alpha_1 \in \mathbb{R}$ we have*

$$\begin{vmatrix} — & \alpha_0 a_0^\top + \alpha_1 a_1^\top & — \\ — & a_2^\top & — \\ & \vdots & \\ — & a_n^\top & — \end{vmatrix} = \alpha_0 \begin{vmatrix} — & a_0^\top & — \\ — & a_2^\top & — \\ & \vdots & \\ — & a_n^\top & — \end{vmatrix} + \alpha_1 \begin{vmatrix} — & a_1^\top & — \\ — & a_2^\top & — \\ & \vdots & \\ — & a_n^\top & — \end{vmatrix},$$

and

$$\begin{vmatrix} | & | & & | \\ \alpha_0 a_0 + \alpha_1 a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{vmatrix} = \alpha_0 \begin{vmatrix} | & | & & | \\ a_0 & a_2 & \cdots & a_n \\ | & | & & | \end{vmatrix} + \alpha_1 \begin{vmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{vmatrix}.$$

Exploratory Challenge 38. The more mathematical way of presenting this material is to define Determinant as a function that goes from $n \times n$ matrices to $\mathbb{R}$ that (i) is linear in each column, (ii) $\det(I) = 1$ and (iii) $\det(A) = 0$ whenever A has two identical columns. It is then possible to prove that the only function satisfying these three properties is the determinant as we defined it. Try to prove it!

6. Eigenvalues and Eigenvectors

We are (almost) ready for one of the most important concepts (if not the most important one) in Linear Algebra, **eigenvalues and eigenvectors**. In a sense, *it has all been building up to this!*

Guiding Strategy 39. Given a square matrix A, as we will see below, an eigenvalue λ and eigenvector v will be, respectively, a scalar and a non-zero vector satisfying $Av = \lambda v$. This means that $(A-\lambda I)v = 0$ and so $(A-\lambda I)$ is not invertible, or equivalently $\det(A-\lambda I) = 0$. We can look for eigenvalues as solutions of $\det(A-\lambda I) = 0$ which is a polynomial[15] in λ but unfortunately, not all polynomials have real zeros. For example if $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $\det(A-\lambda I) = 0$ corresponds to $\lambda^2 + 1 = 0$ which only has solutions in $\mathbb{C}$, the Complex Numbers. For this reason we will start this Chapter with a brief introduction to Complex Numbers. It all starts with asking for a number λ such that $\lambda^2 + 1 = 0$.

Further Reading 40. Complex Analysis is a beautiful topic in Mathematics, what we will cover here is just a tiny peak at it, there is a all bookshelf of excellent books in this topic in our library. I have personally taught a course at ETH on Complex Analysis, and since it was during the COVID pandemic I made videos available online, which are still available at `https://www.youtube.com/playlist?list=PLiud-28tsatLRRGqO_Eg_x0S4LVyxuV5p` (In particular, the first lecture covers roughly the content here).

[15] This is one of the main reasons we had to cover determinants.

6.0. **Complex Numbers.** If we start with the natural numbers $\mathbb{N}$ and want to solve equations like $x+10=1$, we need negative numbers. This motivates considering the integers $\mathbb{Z}$. Similarly, rational numbers $\mathbb{Q}$ are needed to solve equations like $10x=1$ and real numbers $\mathbb{R}$ are needed to solve $x^2=2$.[16] Similarly, the Complex Numbers are needed to solve equations such as $x^2+1=0$. It starts with the introduction of an imaginary number $i\in\mathbb{C}$ such that $i^2=1$. You can think of i as $\sqrt{-1}$.

The complex numbers are numbers of the form $z=a+ib$ for $a\in\mathbb{R}$ and $b\in\mathbb{R}$. $\mathbb{C}=\{a+ib : a,b\in\mathbb{R}\}$. Keeping in mind that $i^2=1$ we can do operations with complex numbers:

- $(a+ib)+(x+iy)=(a+x)+i(b+y)$,
- $(a+ib)(x+iy)=ax+i(ay+bx)+i^2by=ax+i(ay+bx)-by=(ax-by)+i(ay+bx)$,
- $(a+ib)(a-ib)=a^2+b^2$,
- $\frac{a+ib}{x+iy}=\frac{(x-iy)(a+ib)}{(x-iy)(x+iy)}=\frac{(ax+by)+i(bx-ay)}{x^2+y^2}=\left(\frac{ax+by}{x^2+y^2}\right)+i\left(\frac{bx-ay}{x^2+y^2}\right)$.

Given $z\in\mathbb{C}$ with $z=a+ib$ we have the following notation

$$\Re(a+ib) := a \quad \text{called the real part of } z=a+ib, \tag{14}$$

$$\Im(a+ib) := b \quad \text{called the imaginary part of } z=a+ib, \tag{15}$$

$$|z| := \sqrt{a^2+b^2} \quad \text{called the modulus of } z=a+ib, \tag{16}$$

$$\overline{a+ib} := a-ib \quad \text{called the complex conjugate of } z=a+ib. \tag{17}$$

Note that for $z_1,z_2\in\mathbb{C}$, we have $|z|^2=z\bar{z}$, $z_1z_2=z_2z_1$, $\overline{z_1+z_2}=\overline{z_1}+\overline{z_2}$, and $\frac{1}{z}=\frac{\bar{z}}{|z|^2}$.

Fact 6.0.1 (Euler's Formula)**.** *Given $\theta\in\mathbb{R}$, we have*

$$e^{i\theta}=\cos\theta+i\sin\theta. \tag{18}$$

This mean, in particular, that $e^{i\pi}=-1$, or as it is more usually written $e^{i\pi}+1=0$.

Further Reading 41. In order to prove Euler's Formula, we need to first define what we mean by $e^{i\theta}$, this can be done, for example, by the Taylor series of the exponential, but this is outside the scope of this course (see Further Reading 40).

Fact 6.0.2 (Polar Coordinates)**.** *A complex number $z\in\mathbb{C}$ can be written as*

$$z=re^{i\theta}, \tag{19}$$

[16] If you have never seen the proof that there exists no $x\in\mathbb{Q}$ such that $x^2=2$ I highly recommend trying to do it: set $x=a/b$ for $a,b\in\mathbb{Z}$ and try to count how many times 2 divides both a and b and find a contradiction.

where $r \geq 0$ is the modulus of z and $\theta \in \mathbb{R}$ (we can restrict to $\theta \in [0, 2\pi[$) is an angle, also called the argument of z.

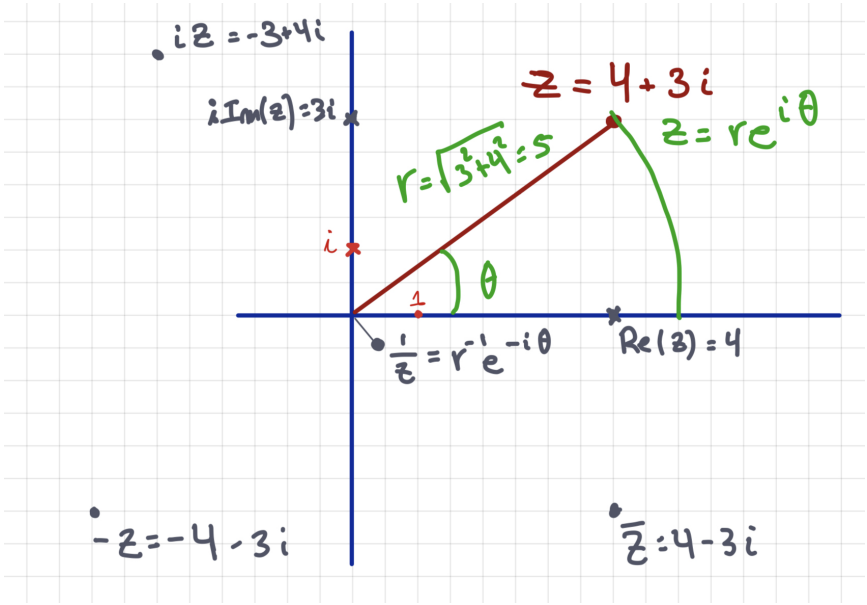


FIGURE 6. A complex number $z = 4 + 3i$ in the Complex plane.

The most important property of Complex Numbers, and what makes them a very natural mathematical object, is that any univariate polynomial equation with complex number coefficients has a (complex) solution, in a certain sense we don't need to extend numbers further, $\mathbb{C}$ is **Algebraically closed**.

Theorem 6.0.3 (Fundamental Theorem of Algebra)**.** *Any degree n non-constant ($n \geq 1$) polynomial $P(z) = \alpha_n z^n + \alpha_{n-1} z^{n-1} + \cdots + \alpha_1 z + \alpha_0$ (with $\alpha_n \neq 0$) has a zero: $\lambda \in \mathbb{C}$ such that $P(\lambda) = 0$.*

Further Reading 42. As the name suggests, Theorem 6.0.3 is a central result in Complex Analysis. Proving it is outside the scope of this course, but the development of complex analysis needed to prove this is a beautiful example of interaction between analysis, algebra, and geometry. In a nutshell the idea is that differentiable functions in the complex plane $f : \mathbb{C} \to \mathbb{C}$ are very special and, in a sense, need to behave like polynomials (this is a deep statement that needs a significant amount of background to properly state and prove). If a polynomial $P(z)$ doesn't have a zero then $1/P(z)$ is a differentiable function that cannot behave like a non-constant polynomial because it does not grow sufficiently far away from zero, and so it must be a constant function which means that $P(z)$ had to be constant, so any non-constant polynomial has a zero. For more on Complex Analysis see Further Reading 40.

Further Remark 43. Once we have λ a zero of $P(z)$, we can divide $P(z)$ by $(z-\lambda)$ to get $P(z) = (z-\lambda)P_1(z)$, then use a zero of P_1 to reiterate, and so on. This argument (carried out carefully) gives the following corollary.

Corollary 6.0.4. *Any degree n non-constant ($n \geq 1$) polynomial $P(z) = \alpha_n z^n + \alpha_{n-1} z^{n-1} + \cdots + \alpha_1 z + \alpha_0$ (with $\alpha_n \neq 0$) has n zero: $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, perhaps with repititions, such that*

$$P(z) = \alpha_n (z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n). \tag{20}$$

The number of times $\lambda \in \mathbb{C}$ appears in this expansion is called the algebraic multiplicity *of the zero.*

6.0.1. *Complex-valued Matrices and Vectors.* Analogously to $\mathbb{R}^n$ we also define $\mathbb{C}^n$ as the set of n-dimentional complex valued vectors. We can have complex valued vectors $v \in \mathbb{C}^n$ and matrices $A \in \mathbb{C}^{m\times n}$. The natural operation of "transposing" for complex vectors and matrices is that of "conjugate transpose" or "hermitian transpose" denoted by A^*, or sometimes A^H,

$$A^* = \overline{A}^T. \tag{21}$$

Given $v \in \mathbb{C}^n$ we have

$$\|v\|^2 = v^* v = \overline{v}^T v = \sum_{i=1}^n \overline{v_i} v_i = \sum_{i=1}^n |v_i|^2.$$

The inner-product (or dot-product) in $\mathbb{C}^n$ is given by $\langle v, w\rangle = w^* v$.

Similarly to the situation in $\mathbb{R}^n$, if say $v_1, \ldots, v_k \in \mathbb{C}^n$ are linearly independent if there is not (complex valued) non-zero linear combination giving zero, meaning that if $\alpha_1 v_1 + \cdots + \alpha_k v_k = 0$ for $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$ we must have $\alpha_1 = \cdots = \alpha_k = 0$. Also, the span of $v_1, \ldots, v_k \in \mathbb{C}^n$ is the set of possible linear combinations $\alpha_1 v_1 + \cdots + \alpha_k v_k$ for $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$. If $v_1, \ldots, v_k$ is a spanning set of a subspace and linearly independent we say it is a basis of that subspace. As with $\mathbb{R}^n$ if we have $v_1, \ldots, v_n \in \mathbb{C}^n$ that are either a spanning set of $\mathbb{C}^n$ or linearly independent then they must actually be both (and so are a basis).

Further Reading 44. With these definitions you can already understand the Discrete Fourier Transform (which is the linear transformation corresponding to the DFT matrix, one of the most important complex valued matrices). This is the key object behind signal processing, you can read more about it on the lecture notes of another course I usually teach [BM23]. You can also see a discussion of Fourier Transform, circulant matrices, and signal convolutions in [Str23] (end of Section 6.4).

6.1. **Introduction to Eigenvalues and Eigenvectors.** Even though the theory can be analogously developed for complex valued matrices, we will focus on real valued matrices.

Guiding Example 45. We will use a guiding example to illustrate both some of the power, and some of the properties, of eigenvalues and eigenvectors. In Guiding Example numbers 45 through 51 we will derive a formula for the n-th Ficonacci Number. The Fibonacci numbers are defined by the recurrence:

$$F_0 = F_1 = 1 \text{ and, for } n \geq 2,\ F_n = F_{n-1} + F_{n-2}. \tag{22}$$

The recurrence can be rewritten in linear algebraic notation as, for $n \geq 2$,

$$\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix}. \tag{23}$$

Definiting

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } g_n = \begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix}, \tag{24}$$

the recurrence can be rewritten as

$$g_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } g_n = M g_{n-1},$$

meaning that

$$g_n = M^n g_0. \tag{25}$$

Definition 6.1.1. *Given $A \in \mathbb{R}^{n\times n}$, we say $\lambda \in \mathbb{C}$ is an eigenvalue of A and $v \in \mathbb{C}^n \setminus \{0\}$ is an eigenvector of A, associated with with the eigenvalue λ, when the following holds:*

$$Av = \lambda v.$$

We call them an eigenvalue-eigenvector pair. If $\lambda \in \mathbb{R}$ then we will call λ a real eigenvalue, and the associated eigenvalue-eigenvector pair a real eigenvalue-eigenvector pair.

Guiding Example 46. Let us try to find eigenvalues (and later the eigenvectors) of $M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$. We are looking for $v \in \mathbb{R}^2 \setminus \{0\}$ and $\lambda \in \mathbb{C}$ such that $Mv = \lambda v$, but this can be rewritten as $(M - \lambda I)v = 0$ and since $v \neq 0$ it means that $M - \lambda I$ is non-invertible (also called singular). This is equivalent to $\det(M - \lambda I) = 0$ and so we can find the eigenvalues λ with this equation:

$$0 = \det(M - \lambda I) = \begin{vmatrix} 1-\lambda & 1 \\ 1 & 0-\lambda \end{vmatrix} = (1-\lambda)(0-\lambda) - 1 = \lambda^2 - \lambda - 1. \tag{26}$$

By the quadratic formula,[17] the solutions to (26) are given by

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \text{ and } \lambda_2 = \frac{1-\sqrt{5}}{2}. \tag{27}$$

Further Reading 47 (Golden Ratio)**.** The number $\varphi = \frac{1+\sqrt{5}}{2}$ is the celebrated Golden Ratio; believed, since the ancient Greeks, to be the ideal aspect ratio for a rectangle.

> *"Some of the greatest mathematical minds of all ages, from Pythagoras and Euclid in ancient Greece, through the medieval Italian mathematician Leonardo of Pisa and the Renaissance astronomer Johannes Kepler, to present-day scientific figures such as Oxford physicist Roger Penrose, have spent endless hours over this simple ratio and its properties. [...] Biologists, artists, musicians, historians, architects, psychologists, and even mystics have pondered and debated the basis of its ubiquity and appeal. In fact, it is probably fair to say that the Golden Ratio has inspired thinkers of all disciplines like no other number in the history of mathematics."*
>
> — The Golden Ratio: The Story of Phi, the World's Most Astonishing Number

The following is the original definition which dates back to Euclid around 2300 years ago (they called the number "extreme and mean ratio" back then)

> *"A straight line is said to have been cut in extreme and mean ratio when, as the whole line is to the greater segment, so is the greater to the lesser"*

Guiding Example 48. Now we can try to find the eigenvectors v_1 and v_2 such that $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_2 v_2$.

Let us start with v_1. We are looking for a non-zero element of $\mathrm{N}\left(A - \frac{1+\sqrt{5}}{2} I\right)$. In other words

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 - \frac{1+\sqrt{5}}{2} & 1 \\ 1 & -\frac{1+\sqrt{5}}{2} \end{bmatrix} \begin{bmatrix} (v_1)_1 \\ (v_1)_2 \end{bmatrix}.$$

This is an under-determined system and we are looking for a non-zero solution, so let us start by setting $(v_1)_2 = 1$. The second equation gives us $(v_1)_1 = \frac{1+\sqrt{5}}{2}$. Indeed $v_1 = \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix}$ is an eigenvector of M associated to the eigenvalue $\lambda_1 = \frac{1+\sqrt{5}}{2}$.

[17] Recall that the quadratic formula says that the zeros of $ax^2 + b + c$ are given by $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

A similar calculation for $\lambda_2 = \frac{1-\sqrt{5}}{2}$ gives that $v_2 = \begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix}$. Indeed

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix} = \frac{1+\sqrt{5}}{2} \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix} = \frac{1-\sqrt{5}}{2} \begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix} \tag{28}$$

Challenge 49. Carry out the calculations in Guiding Example 48 and confirm that we have indeed found two eigenvectors (check the two equalities in (28)).

Further Remark 50. The v_1 and v_2 we constructed in 48 are not the only possible choices, for example any non-zero scalar multiples of these would have also been a possible choice. Normally one picks a unit-norm representative, but in this case we picked vectors that make the calculations the cleanest.

What we carried out in the example above is very general and we now develop the theory for general matrices.

Let λ and v be an eigenvalue-eigenvector pair of a matrix A. Since $v \neq 0$ and $(A-\lambda I)v = Av - \lambda v = 0$ we have that $\det(A-\lambda I) = 0$. Conversely, if $\det(A-\lambda I) = 0$ for some λ, then there exists $v \in \mathrm{N}(A-\lambda I) \setminus \{0\}$ and so λ is an eigenvalue. This gives a procedure to find eigenvalues and eigenvectors: (i) eigenvalues are the solution of $\det(A-\lambda I) = 0$, which is a polynomial equation, and (ii) an associated eigenvector is a non-zero element of $\mathrm{N}(A-\lambda I)$.

Let us first formulate this for real eigenvalues and eigenvectors.

Proposition 6.1.2. *Let $A \in \mathbb{R}^{n\times n}$. $\lambda \in \mathbb{R}$ is a (real) eigenvalue of A if and only if $\det(A-\lambda I) = 0$. A vector v is an eigenvector associated with the eigenvalue λ if (and only if) it is a non-zero element of $\mathrm{N}(A-\lambda I)$.*

A direct inspection of the formula for the determinant (Definition 5.1.6) gives the following.

Proposition 6.1.3. *$\det(A-\lambda I)$ is a polynomial, in λ, of degree n. The coefficient of the λ^n term is $(-1)^n$.*

The Fundamental Theorem of Algebra (Theorem 6.0.3) immediately implies

Theorem 6.1.4. *Every matrix $A \in \mathbb{R}^{n\times n}$ has an eigenvalue (perhaps complex-valued).*

Remark 6.1.5. *For now we will focus on real eigenvalues, and address complex valued ones later on. Essentially all the properties we will describe below also hold for complex valued eigenvalues (just by replacing $\mathbb{R}$ by $\mathbb{C}$ and doing the appropriate adjustments). For example, Proposition 6.1.2 also holds for complex-valued eigenvalues, one just needs to think of $\mathrm{N}(A-\lambda I)$ as a subspace of $\mathbb{C}^n$, meaning the vectors $v \in \mathbb{C}^n$ such that $(A-\lambda I)v = 0$.*

Guiding Example 51. Let us return to our guiding example. Notice that v_1 and v_2 are linearly independent, and so they are a basis for $\mathbb{R}^2$. We can write $g_0 = \alpha_1 v_1 + \alpha_2 v_2$.

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = g_0 = \alpha_1 v_1 + \alpha_2 v_2 = \begin{bmatrix} \alpha_1 \frac{1+\sqrt{5}}{2} + \alpha_2 \frac{1-\sqrt{5}}{2} \\ \alpha_1 + \alpha_2 \end{bmatrix} = \begin{bmatrix} (\alpha_1 + \alpha_2)\frac{1}{2} + (\alpha_1 - \alpha_2)\frac{\sqrt{5}}{2} \\ \alpha_1 + \alpha_2 \end{bmatrix},$$

and so $\alpha_1 = \frac{1}{\sqrt{5}}$ and $\alpha_2 = -\frac{1}{\sqrt{5}}$.

Recall that $g_n = A^n g_0$ and so

$$g_n = A^n \left(\frac{1}{\sqrt{5}} v_1 - \frac{1}{\sqrt{5}} v_2 \right) = \frac{1}{\sqrt{5}} A^n v_1 - \frac{1}{\sqrt{5}} A^n v_2 = \frac{1}{\sqrt{5}} \left(A^n v_1 - A^n v_2 \right).$$

Since $Av_1 = \lambda_1 v_1$ we have that $A^2 v_1 = A(\lambda_1 v_1) = \lambda_1^2 v_1$ and iterating this procedure[18] gives $A^n v_1 = \lambda_1^n v_1$. This means that

$$g_n = \frac{A^n v_1 - A^n v_2}{\sqrt{5}} = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n v_1 - \left(\frac{1-\sqrt{5}}{2}\right)^n v_2}{\sqrt{5}} = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n}{\sqrt{5}} \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix} - \frac{\left(\frac{1-\sqrt{5}}{2}\right)^n}{\sqrt{5}} \begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix}.$$

Since F_n is the second coordinate of g_n, we derived a closed formula for the n-th terms of the Fibonacci sequence:

$$F_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n. \tag{29}$$

An important property that allowed us to do the calculation above was that applying a power of a matrix to an eigenvector was a simple operation, this is the next proposition.

Proposition 6.1.6. *If λ and v are an eigenvalue-eigenvector pair of a matrix A, then, for $n \geq 1$, λ^n and v are an eigenvalue-eigenvector pair of the matrix A^n.*

Proof. Proof by Induction: The base case $n = 1$ is trivial. For the induction step, since λ^n and v are an eigenvalue-eigenvector pair then $A^n v = A\left(A^{n-1} v\right) = A\left(\lambda^{n-1} v\right) = \lambda^n v$. □

Another important property, was that we were able to write a vector as a linear combination of eigenvectors, which was possible because the eigenvectors were linearly independent.

Proposition 6.1.7. *Let $A^{n \times n}$ and let $v_1, \ldots, v_k \in \mathbb{R}^n$ be eigenvectors corresponding to eigenvalues $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$. If $\lambda_1, \ldots, \lambda_k$ are all distinct, the eigenvectors $v_1, \ldots, v_k$ are linearly independent.*

Proof. We will prove this by contradiction. Assume that $v_1, \ldots, v_k$ are linearly dependent. For $i = 1, \ldots, k$, let d_i denote the dimension of the span of $v_1, \ldots, v_i$. Since $v_1 \neq 0$ we have $d_1 = 1$.

[18] A formal proof would use induction

By the hypothesis $d_k < k$. Let j be the smallest positive integer for which $d_j < j$. Note that, by construction, $d_{j-1} = d_j = j-1$, this means that $v_1, \ldots, v_{j-1}$ are linearly independent but that v_j is in the span of $v_1, \ldots, v_{j-1}$. We can then write

$$v_j = \alpha_1 v_1 + \cdots \alpha_{j-1} v_{j-1}. \tag{30}$$

If we multiply by A both sides we get

$$\lambda_j v_j = A v_j = A\left(\alpha_1 v_1 + \cdots \alpha_{j-1} v_{j-1}\right) = \alpha_1 \lambda_1 v_1 + \cdots \lambda_{j-1} \alpha_{j-1} v_{j-1}.$$

If $\lambda_j = 0$ this gives a non-zero linear combination of $v_1, \ldots, v_{j-1}$ being zero, which would be a contradiction with $d_{j-1} = j-1$. If $\lambda_j \neq 0$ then we can rewrite it as

$$v_j = \alpha_1 \frac{\lambda_1}{\lambda_j} v_1 + \cdots + \alpha_{j-1} \frac{\lambda_{j-1}}{\lambda_j} v_{j-1},$$

which together with (30) gives

$$0 = \alpha_1 \left(\frac{\lambda_1}{\lambda_j} - 1\right) v_1 + \cdots + \alpha_{j-1} \left(\frac{\lambda_{j-1}}{\lambda_j} - 1\right) v_{j-1},$$

which would give a non-zero linear combination of $v_1, \ldots, v_{j-1}$ being zero, which would also be a contradiction with $d_{j-1} = j-1$. □

A very important consequence of this is that if a matrix has n distinct real eigenvalues then the eigenvectors form a basis for $\mathbb{R}^n$.

Theorem 6.1.8. *Let $A \in \mathbb{R}^{n \times n}$ with n distinct real eigenvalues (meaning that the n zeros of $\det(A - \lambda I)$, as described in Corollary 6.0.4, are all distinct) then there is a basis of $\mathbb{R}^n$, $v_1, \ldots, v_n$, made up of eigenvectors of A.*

Guiding Example 52. Guiding Example 45 is yet to spot providing us with insight into properties of eigenvalues and eigenvectors! Here are a couple of observations, which although outside of the core scope of this course, have significant impact in several areas:

- Notice that since $|\lambda_2| < |\lambda_1|$, the contribution of $\lambda_2^n \alpha_2 v_2$ becomes negligible (when compared to $\lambda_1^n \alpha_1 v_1$) as $n \to \infty$. This observation can be used in a clever way: we can approximate the eigenvector v_1 by $A^n g_0$ and so if we have a fast way to do matrix-vector multiply, we can approximate eigenvalues and eigenvectors. This is often referred to as the *Power Method*. In a CS Lens I plan to show you how Google's celebrated PageRank algorithm is based on the idea of how eigenvectors can be used for ranking (you can also

read more about it here [BSS]), calculating the eigenvector using a version of the Power Method is a crucial part of the algorithm.[19]

- The vector g_n gets larger and larger as $n \to \infty$ because $|\lambda_1| > 1$. If both eigenvalues satisfied $|\lambda| < 1$ then $g_n \to 0$ as $n \to \infty$. This illustrates the importance of the largest absolute values of the eigenvalues of a matrix in understanding the long term behaviour of systems of the form $A^n g_0$ for some A. If it represents a dynamical system it is related to stability or unstability/chaos, if it represents e.g. the evolution of an economical system over time (or the finances of a company) it can be the different between growth or ruin.[20]

[19] An important advantage is that if we already have a good approximation of v_1, e.g. the page ranks from last week, we can compute a better approximation of v_1 (of this week's rankings) with very few matrix multiplies, you can read more about it here [BSS] and in the references therein.

[20] Try to modify the Fibonacci recurrence rule so that the new numbers go to zero as $n \to \infty$. Can you pick a recurrence such that they stabilize as $n \to \infty$ (without going to ∞ or 0)? Maybe linear algebra students in 2823 years will be studying your sequence!

APPENDIX A. SOME IMPORTANT PRELIMINARIES AND REMARKS ON NOTATION

To follow these notes the reader needs to be familiar with basics of vector and matrix operations and manipulations; understand what is dimension of a subspace, and in particular that is well-defined (that every basis of a subspace has the same size); and understand what is the rank of a matrix (and in particular that the dimension of the column space and the row space are the same). Even though Gaussian Elimination is not a core ingredient of this part of the course, we still assume that the reader is familiar with it. The students of 401-0131-00L are familiar with all this via Part I of this course.

Some further important preliminaries and/or remarks:

(1) The dot product $x \cdot y$ between two real valued vectors is sometimes also called inner product and written as $\langle x, y \rangle$ (it is equal to $x^\top y$). For $\mathbb{C}^n$ the inner product is given by $\langle x, y \rangle = y^* x$.

(2) Matrix Factorization for A an $m \times n$ matrix with rank r:
$A = CR$,
C is $m \times r$ with linearly independent columns (they are the first r linearly independent columns of A). R is $r \times n$, it is upper triangular (i.e. $R_{ij} = 0$ if $i > j$), and it has an $r \times r$ identity as a submatrix, corresponding to the locations of the first r linearly independent columns of A.

(3) For V a subspace (or a vector space) with dimension n the following holds:
- Any basis of V has size n.
- Any spanning set of V has size $\geq n$.
- Any spanning set of V with size n is also a basis.
- Any set of linearly independent vectors in V has size $\leq n$.
- Any set of linearly independent vectors in V with size n is also a basis.

APPENDIX B. WEEKLY SCHEDULE

Numbers represent Fr-Wed weeks of 4×45min lectures (except first and last).

Predictions are in red and may be inaccurate.

CS lenses are not in the script (nor do they appear in this schedule).

For Sections not yet available in the script, the numbering corresponds to [Str23], the table of contents of [Str23] is available at `https://math.mit.edu/~gs/linearalgebra/ila6/indexila6.html`.[21]

(7) 8.11.2023: Introductions; 4.2. Projections
(8) 4.3. Least Squares & Fitting a line, 4.4 Orthonormal bases
(9) QR decomposition, 4.5. and Intro to Linear Transformations (subset of 8 in [Str23])
(10) Determinants. Finish 5. (maybe start 6.1, unlikely)
(11) First half of Eigenvalues / first half of 6. Change of basis (small subset of 8 in [Str23])
(12) Second half of Eigenvalues / second half of 6. We will skip 6.5 in [Str23]
(13) SVD and PCA (7)
(14) 22.12.2023: We will have an entire lecture of "CS lenses", a sort of technical "Ask Me Anything" session that won't cover core material of the course. If you have any particular topic you would like to me cover, let me know!

Appendix C. CS Lens Lectures

(1) Kernel Methods: `https://ti.inf.ethz.ch/ew/courses/LA23/slides/CS_Lens_kernels.pdf`
(2) Graphs, Networks, and Linear Algebra `https://ti.inf.ethz.ch/ew/courses/LA23/slides/CS_Lens_Graphs1.pdf`. I also cover this in some classes I have taught, you can see manuscripts here: [BSS, BM23].
(3) ...

References

[BM23] Afonso S. Bandeira and Antoine Maillard. Mathematics of signals, networks, and learning. *Available online at: https://anmaillard.github.io/teaching/msnl_spring_2023.pdf. Videos from an earlier version of the course available at https://youtube.com/playlist?list=PLiud-28tsatL0MbfJFQQS7MYkrFrujCYp*, 2023.

[BSS] A. S. Bandeira, A. Singer, and T. Strohmer. Mathematics of data science. *Book draft available at https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf. Videos available at: https://www.youtube.com/playlist?list=PLiud-28tsatIKUitdoH3EEUZL-9i516IL*.

[Str23] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley - Cambridge Press, sixth edition, 2023.

[21] In some PDF viewers the ~ in the url above does not show as the correct character, if the link appears broken delete the ˜ and write a new ~ on the url.