Linear Algebra for Computer Scientists

Lecture Notes Part II

A. S. Bandeira and R. Weismantel
ETH Zürich

Last update on October 29, 2025

“Read me” for Part II

These lecture notes serve as a continuation of Part I, taught by Prof. Bernd Gärtner. Please read the Preface there. Please note there may be some changes in notation.

For Part II we will roughly continue to follow, in structure and content, [Str23], with several deviations, for instance, the detailed description of the scenarios that are possible when one considers a system of equations.

Linear Algebra is a beautiful topic. It connects Algebra with Geometry, as you will likely see later in your academic life. For example, when Joseph Fourier invented Fourier Series to develop a theory of heat transfer he was essentially finding good orthonormal bases for functions. Linear Algebra has countless applications making it a key component/ technology of almost all quantitative pursuits. We sincerely hope you will enjoy the course as much as we enjoy teaching this beautiful subject!

As your mathematical level matures over the semester, the notes will have less illustrations and examples, but more definitions and mathematical statements. We will put a focus on the algebraic derivation of the important results. The geometry and additional examples will be shown in the lecture. Our 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 help you greatly in learning the material.

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

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, we 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, we will try to add them in. This might change the numbering of things downstream.

CONTENTS

5. Orthogonality and Projections

5.1. **Orthogonality of vectors and subspaces.**

Let us begin by describing shortly the rational of studying orthogonality. Our task is to explore orthogonality as a geometric and algebraic tool in order to be able to decompose a space into subspaces. Among many results our knowledge will then put us in position to understand how to solve systems of linear equations.

We begin by introducing orthogonality.

Definition 5.1.1. *Two vectors $v, w \in \mathbb{R}^n$ are called orthogonal if $v^T w = \sum_{i=1}^n v_i w_i = 0$. Two subspaces V and W are orthogonal if for all $v \in V$ and $w \in W$, the vectors v and w are orthogonal.*

As an example, consider a vector $(a,b)^T \in \mathbb{R}^2$ with entries $a,b \in \mathbb{R}$. Then $(-b,a)^T$ is orthogonal to $(a,b)^T$. By direct calculation one can also see that the two subspaces

$$V = \{\lambda(1,2,3)^T \mid \lambda \in \mathbb{R}\} \subseteq \mathbb{R}^3 \text{ and } W = \{\mu(1,-2,1)^T \mid \mu \in \mathbb{R}\}$$

are orthogonal. Indeed, it is enough here to verify that $(1,2,3)^T$ and $(1,-2,1)^T$ are orthogonal. This is a general phenomenon. In order to check whether two subspaces V and W are orthogonal it is enough to verify it for the vectors forming a basis of V and W, respectively.

Lemma 5.1.2. *Let $v_1, \dots, v_k$ be a basis of subspace V. Let $w_1, \dots, w_l$ be a basis of subspace W. V and W are orthogonal if and only if v_i and w_j are orthogonal for all $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, l\}$.*

Proof. Let us begin by proving the statement from left to right: Suppose V and W are orthogonal. Since $v_i \in V$ for all $i \in \{1, \dots, k\}$ and $w_j \in W$ for all $j \in \{1, \dots, l\}$, we have that $v_i^T w_j = 0$ for all $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, l\}$.

For the converse direction, assume that $v_i^T w_j = 0$ for all $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, l\}$. Let $v \in V$ and $w \in W$. Then, there exist real multipliers such that

$$v = \sum_{i=1}^k \lambda_i v_i \text{ and } w = \sum_{j=1}^l \mu_j v_j.$$

Then

$$v^T w = \sum_{i=1}^k \lambda_i v_i^T w = \sum_{i=1}^k \sum_{j=1}^l \mu_j \lambda_i v_i^T w_j = 0.$$

□

One can say more about the structure of orthogonal subspaces. One such central statement is presented next.

Lemma 5.1.3. *Let V and W be two orthogonal subspaces of $\mathbb{R}^n$. Let $v_1, \dots, v_k$ be a basis of subspace V. Let $w_1, \dots, w_l$ be a basis of subspace W. The set of vectors $\{v_1, \dots, v_k, w_1, \dots, w_l\}$ are linearly independent.*

Proof. Consider the linear combination

$$(*) \quad \sum_{i=1}^{k} \lambda_i v_i + \sum_{j=1}^{l} \mu_j w_j = 0.$$

We want to show that $\lambda_i = 0$ for all $i \in \{1, \dots, k\}$ and $\mu_j = 0$ for all $j \in \{1, \dots, l\}$.

Let $v = \sum_{i=1}^{k} \lambda_i v_i$. The relation $(*)$ is equivalent to $v = -\sum_{j=1}^{l} \mu_j w_j$. We obtain

$$v^T v = -\sum_{j=1}^{l} \mu_j v^T w_j = 0.$$

Hence, $v = 0$ and with this relation (*) turns into $\sum_{j=1}^{l} \mu_j w_j = 0$.

This implies that $\lambda_i = 0$ and $\mu_j = 0$ for all $i \in \{1, \dots, k\}$, $j \in \{1, \dots, l\}$ since $v_1, \dots, v_k$ and $w_1, \dots, w_l$ are linearly independent vectors. □

Lemma 5.1.3 allows us to derive an important fact about orthogonal subspaces. Namely we can take bases of the two subspaces V and W and their union gives a basis for the subspace

$$V + W = \{\lambda v + \mu w \mid \lambda, \mu \in \mathbb{R},\ v \in V,\ w \in W\}.$$

Note that if V and W are subspaces of $\mathbb{R}^n$, then $V + W$ is indeed also a subspace of $\mathbb{R}^n$. You should by now be able to prove this result.

The proof of Lemma 5.1.3 showed that if a vector v is in the intersection of two orthogonal subspaces, then $v = 0$. This gives the following statement.

Corollary 5.1.4. *Let V and W be orthogonal subspaces. Then $V \cap W = \{0\}$.*

Moreover, if $\dim(V) = k$ *and* $\dim(W) = l$*, then* $\dim(V + W) = k + l \leq n$.

So far we have explored general subspaces V and W that are orthogonal. Next consider a subspace V. Then there is a special orthogonal subspace attached to V.

Definition 5.1.5. *Let V be a subspace of $\mathbb{R}^n$. We define the orthogonal complement of V as*

$$V^{\perp} = \{w \in \mathbb{R}^n \mid w^T v = 0 \, \textit{for all } v \in V\}.$$

It is important to know that $V^{\perp}$ is a subspace of $\mathbb{R}^n$. It might again be a good exercise to verify this fact.

The concept of orthogonal subspaces allows us to decompose the space $\mathbb{R}^n$. An important example of this idea is to take a matrix $A \in \mathbb{R}^{m\times n}$ and decompose $\mathbb{R}^n$ into two orthogonal subspaces: the nullspace of A and the column space of A^T.

Theorem 5.1.6. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix.*

$$N(A) = C(A^T)^{\perp} = R(A)^{\perp}.$$

Proof. Let us first show that $N(A) \subseteq C(A^T)^{\perp}$.

Let $x \in N(A)$. Take any $b \in C(A^T)$. By definition, $b = A^T y$ for some $y \in \mathbb{R}^m$. Then $b^T x = y^T Ax = 0$. Hence, $x \in C(A^T)^{\perp}$.

Conversely, we want to show that $C(A^T)^{\perp} \subseteq N(A)$. To this end let $x \in C(A^T)^{\perp}$. Then $b^T x = 0$ for all $b \in C(A^T)$. Take $y := Ax \in \mathbb{R}^m$ Then $b := A^T y \in C(A^T)$ and hence, $x^T b = 0$. We obtain

$$0 = x^T b = x^T A^T y = x^T A^T Ax = ||Ax||^2.$$

This implies that $Ax = 0$, i.e., $x \in N(A)$. □

From the lecture in Chapter 3.5 we know already that if $r = \dim(R(A))$, then $n - r = \dim(N(A)$. This fact together with Theorem 5.1.6 allows us to prove a general decomposition theorem of the space $\mathbb{R}^n$.

Theorem 5.1.7. *Let V, W be orthogonal subspaces of $\mathbb{R}^n$.*

The following statements are equivalent.

(i) $W = V^{\perp}$.
(ii) $\dim(V) + \dim(W) = n$.
(iii) *Every $u \in \mathbb{R}^n$ can be written as $u = v + w$ with unique vectors $v \in V$, $w \in W$.*

Proof. Let $v_1, \ldots, v_k$ be a basis of V and $w_1, \ldots, w_l$ a basis of W. From Lemma 5.1.2 V and W are orthogonal if and only if $v_i^T w_j = 0$ for all $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, l\}$.

(i) implies (ii): Define $A \in \mathbb{R}^{k\times n}$ to be the matrix with row vectors $v_1, \ldots, v_k$. Then $V = R(A) = C(A^T)$. Moreover, $W = V^{\perp} = N(A)$ from Theorem 5.1.6. From our remark before $\dim(V) = k$ and hence, $\dim(W) = n - k$.

(ii) implies (iii): From Lemma 5.1.3 the vectors in the set $\{v_1, \ldots, v_k, w_1, \ldots, w_l\}$ are linearly independent. Since by assumption $l = n - k$, this set is a basis of $\mathbb{R}^n$. Hence, every vector $u \in \mathbb{R}^n$ has a unique representation in form of

$$u = \sum_{i=1}^{k} \lambda_i v_i + \sum_{j=1}^{l} \mu_j w_j, \text{ where } \lambda_1, \dots, \lambda_k, \mu_1, \dots, \mu_l \in \mathbb{R}.$$

Define the unique vectors $v := \sum_{i=1}^{k} \lambda_i v_i$, $w := \sum_{j=1}^{l} \mu_j w_j$. This gives the statement.

(iii) implies (i): We need to show that $W = V^{\perp}$. Note that $W \subseteq V^{\perp}$ since W is orthogonal to V. To show the reverse inclusion, take any vector $u \in V^{\perp} \subseteq \mathbb{R}^n$ Hence, from our assumption in (iii) we know that $u = v + w$ where $v \in V$ and $w \in W$. Then

$$0 = u^T v = v^T v + v^T w = v^T v = \|v\|^2.$$

Hence, $v = 0$ and it follows that $u = w \in W$. □

Indeed Theorem 5.1.7 allows us to decompose the space $\mathbb{R}^n$ according to a given subspace $V \subseteq \mathbb{R}^n$. We write

$$\mathbb{R}^n = V + V^{\perp} = \{v + w \mid v \in V,\ w \in V^{\perp}\}.$$

This decomposition is symmetric in the sense that we can also take the subspace $V^{\perp}$ and then write

$$\mathbb{R}^n = V^{\perp} + (V^{\perp})^{\perp} = V^{\perp} + V.$$

From Theorem 5.1.7 it follows that $V = (V^{\perp})^{\perp}$. We emphasize this important fact.

Lemma 5.1.8. *Let V be a subspace of $\mathbb{R}^n$. Then $V = (V^{\perp})^{\perp}$.*

This lemma in combination with Theorem 5.1.6 allows us to conclude that for a matrix A we have that $C(A^T) = N(A)^{\perp}$.

Corollary 5.1.9. *Let $A \in \mathbb{R}^{m \times n}$. $N(A) = C(A^T)^{\perp}$ and $C(A^T) = N(A)^{\perp}$.*

There is one final link between the nullspace of a matrix A and the nullspace of the matrix $A^T A$ that we will need in our analysis of projections later on.

Lemma 5.1.10. *Let $A \in \mathbb{R}^{m \times n}$. Then $N(A) = N(A^T A)$ and $C(A^T) = C(A^T A)$.*

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

If $x \in \mathbb{N}(A^{\top} A)$ then $A^{\top} A x = 0$. This implies

$$0 = x^{\top} 0 = x^{\top} A^{\top} A x = (Ax)^{\top} (Ax) = \|Ax\|^2.$$

This shows that Ax must be a vector of norm zero which implies that $Ax = 0$. Hence, $x \in \mathbb{N}(A)$.

For the second statement we utilize Corollary 5.1.9. We have

$$C(A^T) = N(A)^\perp = N(A^TA)^\perp = C((A^TA)^T) = C(A^TA).$$

□

5.2. **Projections.**

The study of this topic is devoted to develop and answer the following very natural question. Given a system of linear equations that has no solution, how do we find a "solution" that has the smallest error? This question is central in countless applications. It is in a sense what Machine Learning is all about. Let us make precise what we mean by projections of vectors onto a subspace.

Definition 5.2.1 (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*

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

Note that this is only a proper definition if the minimum exists and is unique. This will be clarified below.

Let us build us some intuition by starting with one-dimensional subspaces.

5.2.1. *The one-dimensional case.*

Let S be the subspace corresponding to the line that goes through the vector $a \in \mathbb{R}^m \setminus \{0\}$, i.e. $S = \{\lambda a \mid \lambda \in \mathbb{R}\} = C(a)$. By drawing a two dimensional example one can see that the projection p is the vector in the subspace S such that the "error vector" $e = b - p$ is perpendicular to a (i.e. $b - p \perp a$). This geometric intuition turns out to be correct. We will verify it later algebraically.

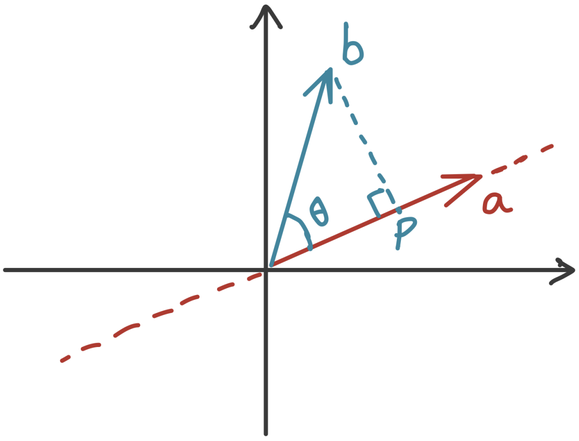


FIGURE 1. Projection on a line.

Lemma 5.2.2. *Let $a \in \mathbb{R}^m \setminus \{0\}$. The projection of $b \in \mathbb{R}^m$ on $S = \{\lambda a \mid \lambda \in \mathbb{R}\} = C(a)$ is given by*

$$\operatorname{proj}_S(b) = \frac{aa^T}{a^T a} b.$$

Proof. Let $p \in S$, $p = \lambda a$ for $\lambda \in \mathbb{R}$. Then

$$\|b-p\|^2 = (b-p)^T(b-p) = b^T b - 2b^T p + p^T p = \|b\|^2 - 2\lambda b^T a + \lambda^2 \|a\|^2 = g(\lambda).$$

g is a convex, quadratic function in one variable λ. Hence, the minimizer is obtained at the point λ^* where the derivative vanishes. We obtain

$$g'(\lambda) = -2b^T a + 2\lambda \|a\|^2 = 0 \iff \lambda^* = \frac{b^T a}{a^T a}.$$

Hence, we have shown that

$$\operatorname{proj}_S(b) = \lambda^* a = a\frac{b^T a}{a^T a} = a\frac{a^T b}{a^T a} = \frac{aa^T}{a^T a} b.$$

□

In particular, the proof of Lemma 5.2.2 shows us that the minimizer $\operatorname{proj}_S(b)$ exists and is unique.

Let us next verify that our initial geometric understanding is indeed correct: the projection p should be the vector in the subspace S such that the "error vector" $e = b - p$ is perpendicular to a, i.e.,

$$(b - \operatorname{proj}_S(b)) \perp \operatorname{proj}_S(b).$$

Indeed by substituting what we just computed we obtain

$$(b - \frac{aa^T}{a^T a} b)^T \frac{aa^T}{a^T a} b = \frac{b^T aa^T b}{a^T a} - b^T \frac{aa^T}{a^T a}\frac{aa^T}{a^T a} b = \frac{(a^T b)^2}{a^T a} - \frac{b^T aa^T b}{a^T a} = 0.$$

The projection of a vector $b = \lambda a$ on the subspace $C(a)$ should be the point b. This follows from our formula:

$$\operatorname{proj}_S(b) = \operatorname{proj}_S(\lambda a) = \frac{1}{a^T a} \lambda aa^T a = \lambda a = b.$$

5.2.2. <u>*The general case.*</u>

For general subspaces the idea is precisely the same as with dimension one. Let S be a subspace in $\mathbb{R}^m$. Let $a_1, \dots, a_n$ be vectors in S that span S, i.e.,

$$S = \operatorname{span}(a_1, \dots, a_n) = C(A) = \{A\lambda \mid \lambda \in \mathbb{R}^n\}, \text{ where}$$

A is the matrix with column vectors $a_1, \dots, a_n$.

Lemma 5.2.3. *The projection of a vector $b \in \mathbb{R}^m$ to the subspace $S = C(A)$ is well defined. It can be written as*

$$\operatorname{proj}_S(b) = A\hat{x}, \text{ where } \hat{x} \text{ satisfies the normal equations } A^T A\hat{x} = A^T b.$$

Proof. Let $b \in \mathbb{R}^m$. The vector b can be written as $b = p^* + e$ where $p^* \in S$ and $e \in S^\perp$, i.e., $(p^*)^T e = 0$. Now consider another point $p' \in S$. Then $p^* - p' \in S$ and hence, $e^T(p^* - p') = 0$. This gives

$$\|p' - b\|^2 = \|p' - p^* + p^* - b\|^2 = \|p' - p^* - e\|^2 = \|p' - p^*\|^2 + \|e\|^2 \geq \|e\|^2 = \|p^* - b\|^2.$$

This shows that

$$p^* = \operatorname{proj}_S(b) = \operatorname*{argmin}_{p \in S} \|b - p\| \tag{2}$$

is well defined because in our chain of relations before we have that $\|p' - b\|^2 > \|p^* - b\|^2$ if and only if $p^* \neq p'$. Hence, we have shown that $p^* = \operatorname{proj}_S(b) = A\hat{x} \in S$ where $b = p^* + e$ with $e \in S^\perp$. This shows us that

$$(b - \operatorname{proj}_S(b)) \perp a_i \text{ for all } i = 1, \ldots, n \iff a_i^T(b - \operatorname{proj}_S(b)) = 0 \text{ for all } i = 1, \ldots, n.$$

This is equivalent to saying that

$$A^T(b - \operatorname{proj}_S(b)) = 0 \iff A^T(b - A\hat{x}) = 0 \iff A^T A\hat{x} = A^T b.$$

□

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.

Lemma 5.2.4. *$A^\top A$ is invertible if and only if A has linearly independent columns.*

Proof. This follows essentially from Lemma 5.1.10 where we convinced ourselves that $A^\top A$ and A have the same nullspace. This is enough because $A^\top A$ is a square matrix and it is invertible if and only if its nullspace only has the 0 vector. On the other hand, A has linearly independent columns if and only if its nullspace only has the 0 vector. □

From now on we can apply the fact that if A has linearly independent columns then $A^\top A$ is a square matrix that is symmetric and invertible. This will be a key ingedient in 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 5.2.5. *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*

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

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 any vector $b \in \mathbb{R}^m$ to its projection Pb on a subspace S. For the case of lines, P was given by $P = \frac{aa^\top}{a^\top a} = a\frac{1}{a^\top a}a^\top$.

The matrices A (and $A^\top$) are not necessarily square, and so they do not have inverses in general. 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).

Let us summarize a few facts about the projection matrix P. P can be viewed as a mapping or as an operator: once we constructed a projection matrix P we can apply it for many vectors b to obtain their projection given by $\mathrm{proj}_S(b) = Pb$.

Remark 5.2.6.

- *If $b \in \mathbb{R}^m$, then $\mathrm{proj}_S(\mathrm{proj}_s(b)) = \mathrm{proj}_S(b)$ by definition. This requires us to have that $PPb = Pb$, i.e., we should have $P^2 = P$. Indeed*
$$P^2 = \left(A\left(A^\top A\right)^{-1}A^\top\right)^2 = A\left(A^\top A\right)^{-1}A^\top A\left(A^\top A\right)^{-1}A^\top = A\left(A^\top A\right)^{-1}A^\top = P.$$
- *Let $S^\perp$ be the orthogonal complement of S and P the projection matrix onto the subspace S, i.e., $\mathrm{proj}_S(b) = Pb$. Then $I-P$ is the projection matrix that maps $b \in \mathbb{R}^m$ to $\mathrm{proj}_{S^\perp}(b)$. This follows since $b = e + \mathrm{proj}_S(b) = e + Pb$ where $e \in S^\perp$. Hence,*
$$(I-P)b = b - Pb = e = \mathrm{proj}_{S^\perp}(b).$$
- *Note that – as it should be – we have that $(I-P)^2 = I - 2P + P^2 = I - P$.*

We have so far seen that a single point can be projected to a subspace. This is only a very special case. Indeed, one can project sets onto sets in smaller dimension in more generality. This topic is absolutely fundamental and has numerous applications in many areas of mathematics and beyond. The reason for this is that projections are a way to reduce the dimension of the initial question about feasibility of a set in $\mathbb{R}^n$ to a question about feasibility of a set in smaller dimension. We will not consider this topic any further, but end our discussions with a definition of a projection operation of a set onto a subspace generated by unit vectors.

Let $Q \subseteq \mathbb{R}^n$. Let $S = \{1,\ldots,s\}$. The projection of Q on the subspace $\mathbb{R}^s$ associated with the variables in the subset S is

$$\mathrm{proj}_S(Q) := \{x \in \mathbb{R}^s \mid \exists y \in \mathbb{R}^{n-s} \text{ such that } (x,y) \in Q\}.$$

6. Applications of orthogonality and projections

The concept of orthogonality of vectors or subspaces and related to it the idea of projecting points to subspaces is the basis of understanding many topics in linear algebra and beyond. We start with a direct application of the concept of projections to least squares approximation.

6.1. **Least Squares Approximation.**

In this section we consider a second application of the concept of orthogonality and projections of vectors. 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 $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Suppose we are given a linear system

$$Ax = b,$$

for which no solution x exists, for example, if there are too many equations, which could happen if $m > n$. See also Example 6.2.5 for such an example.

A natural approach is to try to find x for which Ax is as close as possible to b

$$\min_{\hat{x}\in\mathbb{R}^n} \|A\hat{x} - b\|^2. \tag{3}$$

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 will 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 $\mathbb{C}(A)$, $A\hat{x}$ is precisely the projection of b on $\mathbb{C}(A)$. As we discussed before, this means that

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

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

$$A^\top A\hat{x} = A^\top b. \tag{4}$$

Recall that we had shown in Lemma 5.1.10 that for any matrix A, $\mathbb{C}(A^\top) = \mathbb{C}(A^\top A)$. Hence, the system (4) always has a solution. We also know that if A has linearly independent columns, then $A^\top A$ is invertible and so we can write $\hat{x} = (A^\top A)^{-1}A^\top b$.

Fact 6.1.1. *A minimizer of* (3) *is also a solution of* (4)*. When A has independent columns the unique minimizer $\hat{x}$ of* (3) *is given by*

$$\hat{x} = (A^\top A)^{-1}A^\top b \tag{5}$$

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

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

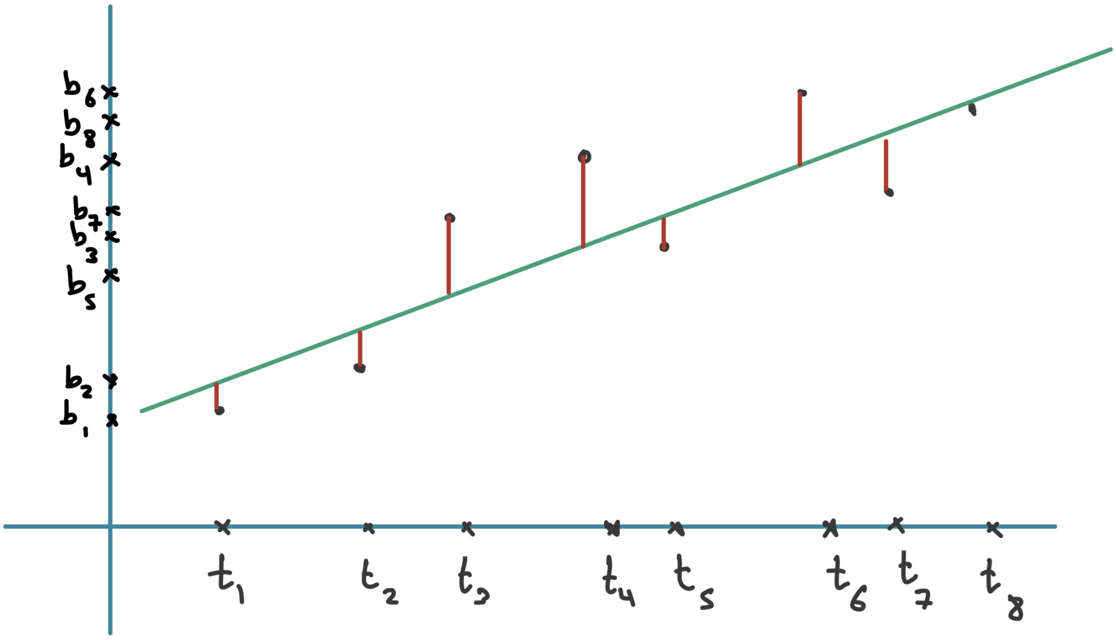


FIGURE 2. Fitting a line to points

See Figure 2. 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

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

where

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

We can assume w.l.o.g. that A has independent columns, see Lemma 6.1.2. Hence, the solution to (6) is given by

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

Lemma 6.1.2. *The columns of the $m \times 2$ matrix A defined before are linearly dependent if and only if $t_i = t_j$ for all $i \neq j$.*

Proof. Suppose that there are two indices $i \neq j$ such that $t_i \neq t_j$. Let $\mathbf{1}$ be the all ones-vector in $\mathbb{R}^m$ and t the vector with components $t_1, \dots, t_m$. Consider the system in variables λ, μ

$$\lambda \mathbf{1} + \mu t = 0.$$

Since $t_i \neq t_j$ we can subtract row j from row i to obtain

$$\lambda 0 + \mu(t_i - t_j) = 0 \iff \mu = 0 \text{ since } t_i - t_j \neq 0.$$

This implies that $\lambda = 0$ and hence A has full column rank.

Conversely, if $t_i = t_j$ for all i and j, then $t = t_1\mathbf{1}$. Then the two columns of A are linearly dependent. □

Remark 6.1.3. *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:*

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

This is an instance where having orthogonal vectors is beneficial. In a later section we will see how to build an orthonormal basis for subspaces and discuss more rigorously some of the many benefits they have.

Example 6.1.4 (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 is not 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

$$\min_{\alpha_0,\alpha_1,\alpha_2} \left\| b - A \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \right\|^2, \tag{7}$$

where

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

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

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

A few final comments are appropriate. 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.*

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 $t_1, \dots t_n$ (which now can be perhaps pixel images, rather than just time points) 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.

We next revisit the topic of systems of linear equations and refine our understanding by using orthogonality.

6.2. **The set of all solutions to a system of linear equations.**

The machinery developed in the previous chapters allows us to understand the set of solutions to a system of linear equations over the reals. This is one fundamental application of the concept of orthogonality.

To make the setting precise, let $A \in \mathbb{R}^{m \times n}$. There are two important subspaces associated with A:

$$N(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \text{ and } R(A) = C(A^T) = \{x \in \mathbb{R}^n \mid \exists y \in \mathbb{R}^m \text{ such that } x = A^T y\}.$$

We have learned that $N(A)$ is the orthogonal complement of $R(A)$. Vice versa, $R(A)$ is the orthogonal complement of $N(A)$. Hence all of $\mathbb{R}^n$ can be written as the sum of two elements: one is from $N(A)$ and the other one from $R(A)$. In other words:

$$\forall x \in \mathbb{R}^n \text{ there exist } x_0 \in N(A) \text{ and } x_1 \in R(A) \text{ such that } x = x_0 + x_1 \text{ and } x_1^T x_0 = 0.$$

Let us first verify that for vectors $x, y \in C(A^T)$ we have that

$$x \neq y \iff Ax \neq Ay.$$

Lemma 6.2.1. *Let $A \in \mathbb{R}^{m\times n}$. Let $x, y \in C(A^T)$. We have that*

$$Ax = Ay \iff x = y.$$

Proof. Let $x, y \in C(A^T)$. Then x,y $\in N(A)^{\perp}$.

$$Ax = Ay \iff x - y \in N(A) \iff x^T(x-y) = 0 = y^T(x-y) \iff (x-y)^T(x-y) = 0.$$

From the latter statement the result follows by noticing that

$$(x-y)^T(x-y) = 0 \iff \|x-y\|^2 = 0 \iff x = y.$$

□

In view of Lemma 6.2.1 we are now in position to describe the set of all solutions to a system of equations in detail. Of course, this requires us to assume that $\{x \in \mathbb{R}^n \mid Ax = b\} \neq \emptyset$. This translated into geometry tell us that b is a linear combination of the columns of A.

Theorem 6.2.2. *Suppose that $\{x \in \mathbb{R}^n \mid Ax = b\} \neq \emptyset$. Then*

$$\{x \in \mathbb{R}^n \mid Ax = b\} = x_1 + N(A) \textit{ where } x_1 \in R(A) \textit{ is unique such that } Ax_1 = b.$$

Proof. Let us first show that

$$\{x \in \mathbb{R}^n \mid Ax = b\} \subseteq x_1 + N(A).$$

From the assumptions of the statement there exists $\hat{x} \in \mathbb{R}^n$ such that $A\hat{x} = b$. Consider $\hat{x}$. Then it follows that $\hat{x} = x_1 + y$ where $y \in N(A)$ and $x_1 \in R(A)$. Since

$$b = A\hat{x} = Ax_1,$$

we have established that there exists $x_1 \in R(A)$ such that $Ax_1 = b$. In view of Lemma 6.2.1 x_1 is the unique vector in $R(A)$ such that $Ax_1 = b$. Hence, every $\hat{x} \in \{x \in \mathbb{R}^n \mid Ax = b\}$ satisfies $\hat{x} = x_1 + y$ with $y \in N(A)$, so

$$\{x \in \mathbb{R}^n \mid Ax = b\} \subseteq x_1 + N(A).$$

The reverse inclusion follows since $x_1 \in R(A)$ satisfies $Ax_1 = b$ and every $y \in N(A)$ satisfies $Ay = 0$. □

Recall that for a matrix $A \in \mathbb{R}^{m\times n}$ we have shown in Lemma 5.1.10 that $N(A) = N(A^TA)$ and $R(A) = C(A^T) = C(A^TA)$. Hence we obtain

Corollary 6.2.3. *Suppose that $\{x \in \mathbb{R}^n \mid Ax = b\} \neq \emptyset$. Then there exists a unique vector $x_1 \in C(A^TA)$ such that $Ax_1 = b$.*

It remains to analyze the case when the given system of linear equations has no solution, i.e.,

$$\{x \in \mathbb{R}^n \mid Ax = b\} = \emptyset.$$

This task seems particularly difficult, because how can one convince someone that a system has no solution? The other person might always claim that it is my fault not to be able to find a solution. Hence, we would like to give an algebraic certificate that such a system has no solution. Ideally such a certificate is a system of equations that then must have a solution. Our knowledge about projections allows us to accomplish this task.

Theorem 6.2.4.

$$\{x \in \mathbb{R}^n \mid Ax = b\} = \emptyset \iff \{z \in \mathbb{R}^m \mid A^Tz = 0,\ b^Tz = 1\} \neq \emptyset.$$

Proof. Set $P = \{x \in \mathbb{R}^n \mid Ax = b\}$ and $D = \{z \in \mathbb{R}^m \mid A^Tz = 0,\ b^Tz = 1\}$. Let us us first verify that the scenario $P \neq \emptyset$ and $D \neq \emptyset$ is impossible. Indeed, if $x \in P$ and $z \in D$ we obtain the following contradiction.

$$0 = 0^Tx = z^TAx = z^Tb = 1.$$

From this we conclude that $P = \emptyset$ or $D = \emptyset$. It remains to show that if $P = \emptyset$, then $D \neq \emptyset$.

Suppose that $P = \emptyset$, i.e., b is not a linear combination of the columns of A. This implies that the orthogonal complement of b minus the projection of b on the subspace $C(A)$ generated by the columns of A is not zero. Let us put this sentence into a formula. Lemma 5.2.3 shows us that

$$\text{proj}_{C(A)}(b) = A\hat{x}, \text{ where } \hat{x} \text{ satisfies the normal equations } A^TA\hat{x} = A^Tb.$$

In particular, $\text{proj}_{C(A)}(b) \in C(A)$ and hence we can write $b = \text{proj}_{C(A)}(b) + p,$ where $p \in C(A)^\perp = N(A^T)$ and $p \neq 0$. Hence, we obtain a feasible solution $z \in D$ by defining

$$z := \frac{1}{p^Tp}p \in \mathbb{R}^m.$$

Since $0 \neq p \in N(A^T)$, we obtain the claim:

$$A^Tz = 0 \text{ and } b^Tz = \frac{1}{p^Tp}p^Tp = 1.$$

□

Let us first see an example of how to apply Theorem 6.2.4.

Example 6.2.5. *Consider the system of two equations in three variables*

$$P = \{x \in \mathbb{R}^3 \mid x_1 + 2x_2 - x_3 = 1,\ 2x_1 + 4x_2 - 2x_3 = 0\}.$$

The system $D = \{z \in \mathbb{R}^m \mid A^T z = 0,\ b^T z = 1\}$ *referenced in Theorem 6.2.4 is*

$$D = \{z \in \mathbb{R}^2 \mid z_1 + 2z_2 = 0,\ 2z_1 + 4z_2 = 0,\ -z_1 - 2z_2 = 0,\ z_1 = 1\}.$$

$P = \emptyset$ *and* $D \neq \emptyset$*, because* $z = (1, -\frac{1}{2})^T \in D$.

There are numerous ways to apply Theorem 6.2.4 and obtain interesting results. For instance, suppose we know that our given matrix $A \in \mathbb{R}^{m \times n}$ has linearly independent rows. Then for every vector $b \in \mathbb{R}^m$ there exists a solution to the system of equations $Ax = b$. A very elegant way to see this is to apply Theorem 6.2.4. Indeed, since the rows of A are linearly independent, the only solution to $z^T A = 0$ is $z = 0$. Hence $z^T b = 0 \neq 1$ for all $b \in \mathbb{R}^m$. This shows that

$$\{z \in \mathbb{R}^m \mid A^T z = 0,\ b^T z = 1\} = \emptyset.$$

Hence $Ax = b$ always has a solution.

In a similar vein, Theorem 6.2.4 can be used to show that a vector $b \in \mathbb{R}^m$ is linearly independent from a set of vectors $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^m$. Define A to be the matrix with column vectors $a_1, \ldots, a_n$. b is linearly independent from $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^m$ if and only if $Ax = b$ has no solution $x \in \mathbb{R}^n$. By Theorem 6.2.4 the latter fact requires us to find a solution $z \in \mathbb{R}^m$ such that $A^T z = 0$, $b^T z = 1$. Hence checking for linear independence can be accomplished by solving a system of linear equations.

The topic of generating a certificate for the non-existence of a given set is fundamental and simply beautiful. You will certainly see several examples of this kind in your further academic live. As an outlook, imagine we are given a matrix $A \in \mathbb{R}^{m \times n}$. Instead of considering a system of equations we associate with every row a of A and a scalar $\alpha \in \mathbb{R}$ the halfspace

$$\{x \in \mathbb{R}^n \mid a^T x \leq \alpha\}.$$

For a matrix $A \in \mathbb{R}^{m \times n}$ and a given vector $b \in \mathbb{R}^m$, we write $Ax \leq b$ to denote the intersection of all halfspaces associated with all rows of A, i.e.,

$$\{x \in \mathbb{R}^n \mid Ax \leq b\} = \cap_{i=1}^m \{x \in \mathbb{R}^n \mid A_{i\cdot}^T x \leq b_i\}.$$

Then the famous Farkas lemma states that

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\} = \emptyset \iff D = \{y \in \mathbb{R}^m \mid y_i \geq 0 \text{ for all } i = 1, \ldots, m,\ y^T A = 0,\ y^T b < 0\} \neq \emptyset.$$

Similar to the first part of the proof of Theorem 6.2.4 it is not difficult to verify that the two sets P and D cannot simultaneously have a solution. For the purpose of deriving a contradiction, assume that $P \neq \emptyset$ and $D \neq \emptyset$. Take any $x \in P$ and $y \in D$ to obtain the following contradiction:

$$0 = 0^T x = y^T Ax = \sum_{i=1}^m y_i A_{i\cdot}^T x \leq \sum_{i=1}^m y_i^T b_i = y^T b < 0.$$

The difficult part is to show that if one set is empty, then the other one is non-empty. We will touch this topic in a CS lenz. The interested reader is refereed to the text book

Alexander Schrijver, Theory of linear and integer programming, Wiley 1986.

There are also other extensions of systems of equations over the reals. For instance, let $A \in \mathbb{Z}^{m\times n}$ be a matrix with integral entries of full row rank. For a given vector $b \in \mathbb{Z}^m$ it is now natural to ask whether b is in the "lattice" generated by the columns of A, i.e., is there a solution to the system of equations over the integers

$$Ax = b,\ x \in \mathbb{Z}^n?$$

If you are interested in understanding such extensions you might want to consider to take the course Discrete Optimization taught by the second author of these notes.

6.3. **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 as we will see later. What can we say about these bases? How to construct them?

Definition 6.3.1 (Orthonormal vectors)**.** *Vectors $q_1,\dots,q_n \in \mathbb{R}^m$ are orthonormal if they are orthogonal and have norm 1. In other words, for all $i, j \in \{1,\dots,n\}$*

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

where δ_{ij} is the Kronecker delta

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \tag{8}$$

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 $Q^\top Q = I$. Note though that Q may not be a square matrix. Hence, it is not necessarily the case that $QQ^\top = I$.

Example 6.3.2. *A classical example of an orthonormal set of vectors is the canonical basis, $e^1,\dots,e^n \in \mathbb{R}^n$ where e^i is the vector with a 1 in the i-th entry and 0 in all other entries, i.e., $(e^i)_j = \delta_{ij}$.*

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 6.3.3 (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 $\mathbb{R}^n$.*

Example 6.3.4. *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. Indeed,

$$R_\theta^T R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = I.$$

Example 6.3.5. *Permutation matrices are another example of orthogonal matrices. A permutation is a map*

$$\pi:\ \{1,\dots,n\} \mapsto \{1,\dots,n\} \text{ such that } \pi(i) \neq \pi(j) \text{ for } i \neq j.$$

A permutation matrix $P \in \mathbb{R}^{n\times n}$ associated with π has entries $P_{ij} = 1$ if $\pi(i) = j$ and $P_{ij} = 0$, otherwise. From this definition one can derive that P^T is the permutation matrix associated with the permutation σ defined as $\sigma(j) = i$ for $\pi(i) = j$. Hence, $P^T P = I$, i.e., P is an orthogonal matrix.

It is also not so difficult to show that for every permutation matrix P there exists a positive integer k such that $P^k = I$. The reason is that $P^2, P^3, \dots, P^k, \dots$ are all permutation matrices. Since the number of different permutations is finite there exist two indices $k < l$ such that a permutation matrix Q say, is visited twice, i.e.,

$$\text{there exist indices } k < l \text{ such that } P^l = Q = P^k \iff I = P^{l-k}.$$

This idea can be turned into an elegant clean proof. We leave it to the interested reader to work this out.

Proposition 6.3.6. *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 equality 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 $\|x\| \geq 0$. Hence, it suffices to show that the squares are equal. Indeed, we obtain $\|Qx\|^2 = (Qx)^\top(Qx) = x^\top x = \|x\|^2$. □

One advantage of having access to an orthonormal basis is that projections become much simpler. The reason is easy to explain. 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 6.3.7. *Let S be a subspace of $\mathbb{R}^m$ and $q_1, \dots, q_n$ 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 $\hat{x} = Q^\top b$.*

What can one say when Q is a square matrix? When Q is square, then the projection $QQ^\top$ is 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

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

It is writing x as a linear combination of the orthonormal basis $\{q_i\}_{i=1}^n$. We will see later that this operation is sometimes referred to as a *change of basis*. 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.

By now we have given some evidence that orthonormal bases are useful. Fortunately, there is a relatively simple process to construct orthonormal bases that will also suggest a new matrix factorization.

The idea is based on computing projections and is simple to describe: 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:

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

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, $(a_2^\top q_1)q_1 = \mathrm{proj}_{\mathrm{Span}(q_1)}(a_2)$.

For more vectors, the idea is to apply 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 6.3.8. *[Gram-Schmidt Process] Given n linearly independent vectors $a_1, \ldots, a_n$ that span a subspace S, the Gram-Schmidt process constructs $q_1, \ldots q_n$ in the following way:*

- $q_1 = \frac{a_1}{\|a_1\|}$.
- *For $k = 2, \ldots, n$ set*

```
    $q'_k = a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_i$
    $q_k = \frac{q'_k}{\|q'_k\|}$.
```

Theorem 6.3.9 (Correctness of Gram-Schmidt)**.** *Given n linearly independent vectors $a_1, \ldots, a_n$, the Gram-Schmidt process returns an orthonormal basis for the span of $a_1, \ldots, a_n$.*

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

- 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 $1 \le j \le k-1$. Since $q_1, \dots, q_{k-1}$ 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 $q_j^\top q_k = \frac{1}{\|q'_k\|} q_j^\top q'_k = 0$.

□

It might be a good exercise to perform the Gram-Schmidt process for the columns of the matrix

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

What do you observe? Is it true that the Gram-Schmidt process applied to the columns of an upper triangular matrix with non-zero diagonal elements always produces a subset of the canonical basis?

The Gram-Schmidt process has numerous applications. One is that it provides us with a new matrix factorization. Let A be an $m \times n$ matrix with linearly independent columns $a_1, \dots, a_n$ and Q the $m \times n$ matrix whose columns are $q_1, \dots, q_n$ as returned by Algorithm 6.3.8. Let $R = Q^\top A$. R is upper triangular because each q_k is orthogonal to every a_i for $i < k$. Note that 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 allows us to conclude that $QQ^\top A = A$, so we have that $QR = QQ^\top A = A$. We call $A = QR$ the QR decomposition.

Definition 6.3.10 (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 6.3.8, on the columns of A) and R is an upper triangular matrix given by $R = Q^\top A$.

It requires us to show that indeed this is a proper definition. In particular, we need to convince ourselves that R is upper triangluar.

Lemma 6.3.11. *The matrix R defined in Definition 6.3.10 is upper triangular and invertible. Moreover, $QQ^T A = A$ and hence, $A = QR$ is well defined.*

Proof. $q_k^T q_i = 0$ for all $i = 1, \dots k-1$. Since $q_1, \dots, q_{k-1}$ and $a_1, \dots, a_{k-1}$ span the same subspace S_{k-1} we have that $q_k^T a_i = 0$ for all $i = 1, \dots, k-1$. Hence $R = Q^T A$ is upper triangular. Moreover, QQ^T is the projection onto the subspace $C(Q) = C(A)$. Hence, for every index i,

$$\text{proj}_{S_n}(a_i) = a_i = QQ^T a_i.$$

This is equivalent to $QQ^T A = QR = A$. Finally, $N(A) = \{0\}$ and since $A = QR$, we must have that $N(R) = \{0\}$. Since R is an $n \times n$ matrix, R is invertible. □

Note that in the proof we showed that R is invertible. This also implies that R^T is invertible. Hence we obtain

Fact 6.3.12. *The QR decomposition greatly simplifies calculations involving Projections and Least Squares.*

- *Since $\mathbb{C}(A) = \mathbb{C}(Q)$ then projections on $\mathbb{C}(A)$ can be done with Q which means they are given by* $\text{proj}_{\mathbb{C}(A)}(b) = QQ^\top b$.
- *The least squares solution to $Ax = b$ denoted by $\hat{x}$ is defined as a solution of the normal equations (recall* (4)*)*

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

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

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

  Since R^T is invertible we can simplify (9) *to*

  $$R\hat{x} = Q^\top b, \tag{10}$$

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

6.4. **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. Such an analogue is called a pseudoinverse, or the Moore-Penrose Inverse, and we will denote it by $A^\dagger$. It is also commonly denoted by A^+. What is the right idea to define a "pseudoinverse" for any matrix A? It should be a matrix that is, in a sense, closest to being an inverse for A? What should "closest to being an inverse" mean?

There are 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 an 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 in Section 6.1. If the columns of A are linearly independent then it might 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 is the following definition.

Definition 6.4.1 (Pseudoinverse for matrices of 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 6.4.2. *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, $A^\dagger A = (A^\top A)^{-1}A^\top A = I$. □

As a next step let us 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 a pseudoinverse is based on the observation that $A^\top$ has 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 6.4.3 (Pseudoinverse for matrices of 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*

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

Lemma 6.4.4. *For $A \in \mathbb{R}^{m\times n}$ with* $\operatorname{rank}(A) = m$*, the pseudoinverse $A^\dagger$ is a right inverse of A, meaning that $AA^\dagger = I$.*

Proof. Since $\operatorname{rank}(A) = m$, $AA^\top$ 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.[1] In other words

[1] 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.

we solve

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

where s.t. stands for “subject to” or “such that”.

Lemma 6.4.5. *For any matrix A and a vector $b \in C(A)$, the (unique) solution to* (11) *is given by the vector $\hat{x} \in \mathbb{C}(A^\top)$ that satisfies the constraint $A\hat{x} = b$.*

Proof. From Theorem 6.2.2 it follows that

$$\{x \in \mathbb{R}^n \mid Ax = b\} = \hat{x} + N(A) \text{ where } \hat{x} \in C(A^T) \text{ is unique such that } A\hat{x} = b.$$

Moreover, for all $y \in N(A)$ we notice that $\hat{x}^T y = 0$ since $\hat{x} \in N(A)^\perp$. This gives

$$\|\hat{x} + y\|^2 = \hat{x}^T\hat{x} + 2\hat{x}^T y + y^T y = \|\hat{x}\|^2 + \|y\|^2 \geq \|\hat{x}\|^2.$$

□

$A^\dagger$ is precisely the matrix that maps b to a point $\hat{x}$ that corresponds to a solution of (11).

Proposition 6.4.6. *For a full row rank matrix A, the (unique) solution to* (11) *is given by the vector $\hat{x} = A^\dagger b$.*

Proof. By using Lemma 6.4.5 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 $\mathbb{C}(A^\top)$. Both these are easy to verify: $A\hat{x} = AA^\dagger b = AA^\top(AA^\top)^{-1}b = b$ and $\hat{x} = A^\dagger b = A^\top\left((AA^\top)^{-1}b\right)$ and so $\hat{x} \in \mathbb{C}(A^\top)$. □

Our next task is to define $A^\dagger$ for all matrices, not just full rank matrices. The idea is to write A as a product of two matrices, one which is of full column rank and one which is of full row rank. Recall that in Part I of the lecture we achieved this task by introducing the CR-decomposition. For $A \in \mathbb{R}^{m\times n}$, with $\operatorname{rank}(A) = r$, 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 has full column rank and R full row rank.

Definition 6.4.7 (Pseudoinverse for all matrices)**.** *For $A \in \mathbb{R}^{m\times n}$ with* $\operatorname{rank}(A) = r$ *and 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 lemma characterizes what the matrix $A^\dagger$ achieves for us.

Lemma 6.4.8. *Given $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$, the (unique) solution to*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|^2 \\ s.t. \quad & A^\top Ax = A^\top b, \end{aligned} \tag{12}$$

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,

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

It remains to show that $\hat{x}$ is the smallest norm solution. To verify this we use Lemma 6.4.5, i.e., we need to verify that $\hat{x} \in \mathbb{C}(A^\top A)$. From Lemma 5.1.10 we conclude that $C(A^TA) = C(A^T)$. Hence it is enough to show that $\hat{x} \in \mathbb{C}(A^\top)$ and since $\mathbb{C}(A^\top) = \mathbb{C}(R^\top)$ we have that $\hat{x} = R^\top \left(C^\top AR^\top\right)^{-1} C^\top b \in \mathbb{C}(R^\top)$ from which the statement follows. □

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

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

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 6.4.9 are indeed full rank (either full column rank or full row rank). Let us finally summarize a few important properties about the matrix A and its pseudoinverse $A^\dagger$.

Theorem 6.4.10. *Let $A \in \mathbb{R}^{m\times n}$.*

$AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$ and $\left(A^\top\right)^\dagger = \left(A^\dagger\right)^\top$.
$AA^\dagger$ is symmetric. It is the projection matrix for projection on $\mathbb{C}(A)$,
$A^\dagger A$ is symmetric. It is the projection matrix for projection on $\mathbb{C}(A^\top)$.

Proof. We calculate

$$AA^{\dagger}A = CRR^T(C^TCRR^T)^{-1}C^TCR = CRR^T(RR^T)^{-1}(C^TC)^{-1}C^TCR = CR = A.$$

To see that $AA^{\dagger}$ is symmetric we calculate

$$AA^{\dagger} = CRR^T(RR^T)^{-1}(C^TC)^{-1}C^T = C(C^TC)^{-1}C^T = \left(C(C^TC)^{-1}C^T\right)^T = (AA^{\dagger})^T.$$

Since the column space of A, $C(A)$ and the column space of C coincide and since C is a basis for $C(A)$ Theorem 5.2.5 applies and shows us that $AA^{\dagger} = C(C^TC)^{-1}C^T$ is the projection matrix for projecting onto $C(A)$. □

The proof of the other statements in Theorem 6.4.10 is left to the reader.

7. 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 7.0.1. *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 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 3.*

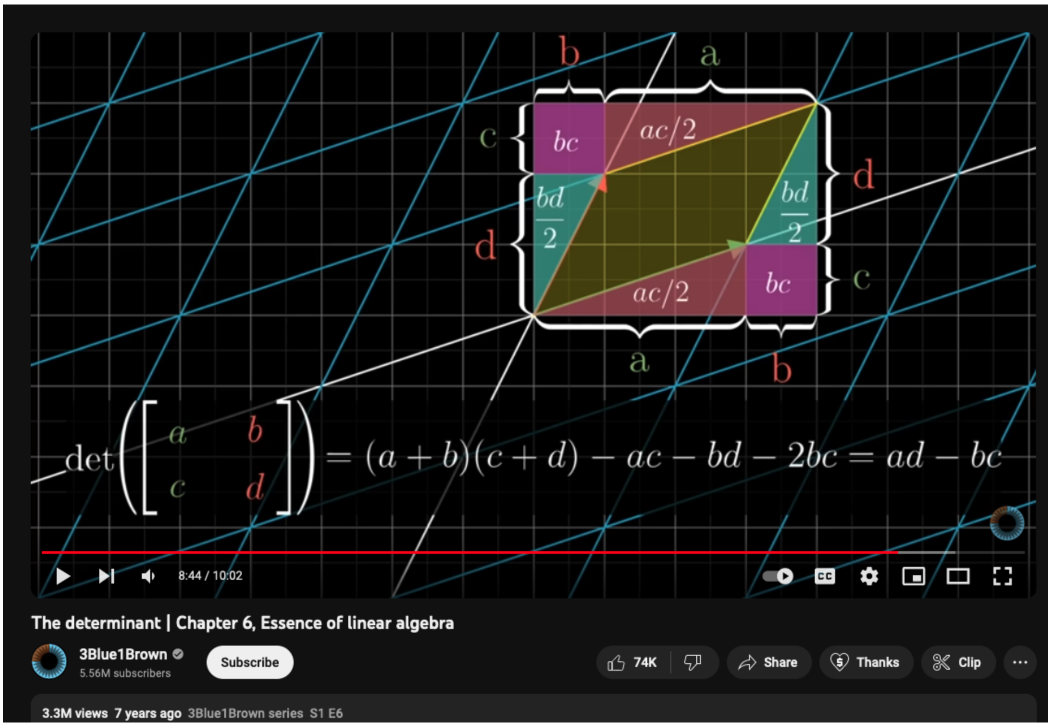
A calculation of the area of the image of the unit square by left-multiplication by a 2 × 2 matrix shows (see Figure 3) that

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

Before we actually formally define the determinant for general $n \times n$ matrices we will first focus on the special case of 2×2- matrices to derive several important properties of the determinant.

7.1. 2×2 - matrices.

This section is based on the following definition for general 2×2 - matrices.



$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = (a+b)(c+d) - ac - bd - 2bc = ad - bc$$

The determinant | Chapter 6, Essence of linear algebra

3Blue1Brown

5.56M subscribers

3.3M views 7 years ago 3Blue1Brown series S1 E6

FIGURE 3. Calculation in 3Blue1Brown's video (see Remark 7.0.1) 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).

Definition 7.1.1. *Let*

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

The determinant of A is $\det(A) = ad - bc$.

Let us first understand how the determinant changes when we multiply matrices. To this end, let

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \text{ and } W = \begin{bmatrix} x & z \\ y & w \end{bmatrix}.$$

Next we multiply these two matrices and obtain an explicit representation of the coefficients

$$AW = \begin{bmatrix} ax+cy & az+cw \\ bx+dy & bz+dw \end{bmatrix}.$$

Using this representation we obtain the following result.

Lemma 7.1.2. *Let* $A, W \in \mathbb{R}^{2\times 2}$. *Then* $\det(AW) = \det(A)\det(W)$.

Proof.

$$\begin{aligned} det(AW) &= (ax+cy)(bz+dw)-(az+cw)(bx+dy) \\ &= axbz+axdw+cybz+cydw-azbx-azdy-cwbx-cwdy \\ &= axdw+cybz-azdy-cwbx \\ &= ad(xw-zy)+cb(zy-xw) = \det(A)\det(W). \end{aligned}$$

□

This computation allows us to derive a characterization of when a 2×2-matrix is invertible.

Lemma 7.1.3. *A matrix $A \in \mathbb{R}^{2\times 2}$ is invertible if and only if* $\det(A) \neq 0$.

Proof. Let

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

If A is invertible, then A^{-1} exists and hence, $AA^{-1} = I$ implies together with the previous lemma that $\det(A)\det(A^{-1}) = 1$. Hence, $\det(A) \neq 0$.

Conversely, if $\det(A) \neq 0$, then $a \neq 0$ or $b \neq 0$. Without loss of generality we can assume that $a \neq 0$. Consider now the system of linear equations $AW = I$.

$$\begin{array}{rcl} ax+cy=1 & \text{implies that} & x = \frac{1-cy}{a} \\ az+cw=0 & \text{implies that} & z = \frac{-cw}{a}. \end{array}$$

By substituting these expressions into the other two equations $bx+dy=0$ and $bz+dw=1$ we obtain

$$\frac{b}{a} - \frac{cyb}{a} + dy = 0 \iff b + y(ad-bc) = 0 \iff y = \frac{-b}{\det(A)}.$$

$$\frac{-bcw}{a} + dw = 1 \iff -bcw + adw = a \iff w = \frac{a}{\det(A)}.$$

This then gives us a formula for the parameters z and x in form of

$$z = \frac{-c}{\det(A)} \text{ and } x = \frac{1+\frac{cd}{\det(A)}}{a} = \frac{d}{\det(A)}.$$

These calculations show that A^{-1} exists whenever $\det(A) \neq 0$. □

Notice that our calculations give us an explicit formula for the inverse of matrix A and its determinant:

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}. \tag{13}$$

7.2. **The general case.**

It turns out that what we have verified in dimension two carries over to general dimensions. It is, however much more involved to verify it algebraically.

We now give the definition of a determinant for $n \times n$ matrices. This requires us to discuss permutations.

Definition 7.2.1 (Sign of a 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,*

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

Example 7.2.2. *Let $n = 4$. Consider the permutation π defined as $\pi(1) = 1$, $\pi(2) = 3$, $\pi(3) = 2$, $\pi(4) = 4$. The pairs (i, j) such that $i < j$ are*

$$(1,2),(1,3),(1,4),(2,3),(2,4),(3,4).$$

For all these listed pairs (i, j) we have that $\pi(i) < \pi(j)$ except for the pair $(2,3)$. Hence, $\operatorname{sgn}(\pi) = -1$.

The sign of a permutation has many nice properties that are not so easy to show.

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

We are now in position to introduce the general notion of a determinant of a square matrix.

Definition 7.2.3. *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.

If A is a 1×1 matrix then there is only one permutation of 1 element: the permutation $\sigma(1) = 1$, which has sign 1. It follows that $\det(A) = A$.

For 2×2 matrices we observe that there exist two permutations. Let us call 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). Hence, for a 2×2 matrix A with entries A_{ij} 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 the definition that we used for the determinant of a 2×2-matrix with entries $A_{11} = a$, $A_{12} = c$, $A_{21} = b$ and $A_{22} = d$, namely

$$\det(A) = ad - bc.$$

By applying Definition 7.2.3 we immediately obtain the following results.

Proposition 7.2.4.

(a) *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* $\operatorname{sgn}(P)$.

(b) *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, $\det(I) = 1$.

(c) *If $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix then*

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

Proof. (a) $P \in \mathbb{R}^{n \times n}$ satisfies that $\prod_{i=1}^{n} P_{i,\sigma(i)} = 1$ and

$$\prod_{i=1}^{n} P_{i,\pi(i)} = 0 \text{ for every permutation } \pi \neq \sigma.$$

This gives the result.

(b) Without loss of generality let $T \in \mathbb{R}^{n \times n}$ satisty $T_{ij} = 0$ for all $j < i$. Let σ denote the identity permutation, i.e., $\sigma(i) = i$ for all $i \in \{1, \ldots, n\}$. Consider any permutation π different from σ, i.e., there exists an index i such that $\pi(i) \neq i = \sigma(i)$. We notice that

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

Hence, if for an index i, $\pi(i) > i$, then there exists another index $l \neq i$ such that $\pi(l) < l$. Hence we observe that there always exists an index k, say such that $\pi(k) < k$. Since $T_{k,\pi(k)} = 0$, we obtain

$$\prod_{i=1}^{n} T_{i,\pi(i)} = 0 \text{ for every permutation } \pi \neq \sigma.$$

This gives the result.

(c) From Part (b) and Theorem 7.2.5 below we obtain

$$1 = \det(I) = \det(Q^\top Q) = \det(Q^\top)\det(Q) = \det(Q)^2.$$

It follows that $1 = \det(Q)^2$ and so $\det(Q)$ is 1 or -1. □

The most important results about determinants are described now. We begin with Theorem 7.2.5 that we already applied.

Theorem 7.2.5. *Given a matrix $A \in \mathbb{R}^{n\times n}$, then*

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

Proof. For a permutation σ let σ^{-1} denote the inverse permutation, i.e.,

$$\sigma(i) = j \iff \sigma^{-1}(j) = i \text{ for all } i, j.$$

We have that for all i $\sigma(\sigma^{-1})(i) = i$ and hence, $\mathrm{sgn}(\sigma)\mathrm{sgn}(\sigma^{-1}) = 1$. It follows that $\mathrm{sgn}(\sigma) = \mathrm{sgn}(\sigma^{-1})$. The conclusion $\det(A^\top) = \det(A)$ follows from observing

$$\begin{aligned}
&\textstyle\sum_{\sigma\in\Pi_n} \mathrm{sgn}(\sigma) \prod_{i=1}^n A_{i,\sigma(i)} = \\
&\textstyle\sum_{\sigma\in\Pi_n} \mathrm{sgn}(\sigma) \prod_{i=1}^n A_{\sigma^{-1}(\sigma(i)),\sigma(i)} = \\
&\textstyle\sum_{\sigma\in\Pi_n} \mathrm{sgn}(\sigma) \prod_{j=1}^n A_{\sigma^{-1}(j),j} = \\
&\textstyle\sum_{\sigma^{-1}\in\Pi_n} \mathrm{sgn}(\sigma^{-1}) \prod_{i=1}^n A_{\sigma^{-1}(i),i} = \\
&\textstyle\sum_{\sigma\in\Pi_n} \mathrm{sgn}(\sigma) \prod_{i=1}^n A_{\sigma(i),i}.
\end{aligned}$$

□

In correspondence with our analysis of the determinant for 2×2-matrices, we can also more generally multiply matrices or derive a formula for the determinant of the inverse of a matrix.

Theorem 7.2.6.

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

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

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

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

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

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

Example 7.2.7. *Let A be a 3×3 matrix. There are $3! = 6$ permutations, so in total there will be 6 terms involved in comnputing the determinant. Notice that we can write its determinant also as a sum of* *determinants of special matrices (where an empty entry corresponds to a zero entry).*

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

7.3. **Cofactors, Cramer's rule and beyond.**

Let us revisit Example 7.2.7 from the previous section. Let A be a 3×3 matrix.

There is another convenient way of writing the determinant of A.

$$\begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} = A_{11} \begin{vmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{vmatrix} - A_{12} \begin{vmatrix} A_{21} & A_{23} \\ A_{31} & A_{33} \end{vmatrix} + A_{13} \begin{vmatrix} A_{21} & A_{22} \\ A_{31} & A_{32} \end{vmatrix}. \tag{14}$$

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

Definition 7.3.1. *Given $A \in \mathbb{R}^{n\times n}$, for each $1 \le i, j \le n$ 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 (14), the determinant can be written in terms of the co-factors.

Proposition 7.3.2. *Let $A \in \mathbb{R}^{n\times n}$, for any $1 \le i \le n$,*

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

The formula we derived above for the inverse of 2×2 matrices (Equation 13), also has an analogue in n dimensions. Verify that Proposition 7.3.3 indeed corresponds to the formula we derived for A^{-1} when $n = 2$.

Proposition 7.3.3. *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.

The formula in Proposition 7.3.3 can be rewritten as

$$AC^{\top} = \det(A)I.$$

Remark 7.3.4. *Computationally speaking, this is not a good way to compute the inverse, as it involves computing many determinants.*

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

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 & 0 & 0 \\ x_2 & 1 & 0 \\ x_3 & 0 & 1 \end{bmatrix} = \begin{bmatrix} b_1 & A_{12} & A_{13} \\ b_2 & A_{22} & A_{23} \\ b_3 & A_{32} & A_{33} \end{bmatrix}.$$

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 from 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 7.3.5 (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 $\mathscr{B}_j$ is the matrix obtained by A by replacing the j-th column of A with the vector b.

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.

The definition we used for the determinant of a square matrix involves a formula with $n!$ terms. This formula 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 $PA = LU$ where P is a permutation matrix and so $\det(P) = \operatorname{sgn}(P)$, U is upper triangular and L is lower triangular with only 1s in the diagonal, so $\det(L) = 1$). This gives us the formula

$$\det(A) = \frac{1}{\det(P)} \det(L) \det(U) = \operatorname{sgn}(P) \det(U). \tag{15}$$

U is a triangular matrix. Hence, its determinant is simply the product of its diagonal entries.

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

Proposition 7.3.6. *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 7.3.7. *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

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

Let us make a final comment on the presentation here. The more mathematical way of presenting this material is to define a determinant as a function that goes from $n \times n$ matrices to $\mathbb{R}$ with the following properties:

(1) it is linear in each column,
(2) $\det(I) = 1$ and
(3) $\det(A) = 0$ 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.

8. Eigenvalues and Eigenvectors

We are now prepared for one very important concept in Linear Algebra, **eigenvalues and eigenvectors**. As we will see below this material requires us to have understood the geometry of subspaces of $\mathbb{R}^n$ and being aware of the notion of the determinant of a square matrix. 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 search for eigenvalues as solutions of $\det(A-\lambda I) = 0$ which is a polynomial in λ. Unfortunately, not all polynomials have real zeros. A zero of a polynomial P is a point x such that $P(x) = 0$. Such a point is also called a root of the polynomial. In German, it is a “Nullstelle”. In fact, a (rather deep) multidimensional version of Theorem 8.1.2, and one of the most important facts in Algebraic Geometry, is called “Hilbert’s Nullstellensatz”.

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 $\lambda^2 + 1 = 0$.

8.1. 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 $x^2 = 2$. 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:

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

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

$$\Re(a+ib) := a \quad \text{called the real part of } z = a+ib, \tag{16}$$

$$\Im(a+ib) := b \quad \text{called the imaginary part of } z = a+ib, \tag{17}$$

$$|z| := \sqrt{a^2+b^2} \quad \text{called the modulus of } z = a+ib, \tag{18}$$

$$\overline{a+ib} := a - ib \quad \text{called the complex conjugate of } z = a+ib. \tag{19}$$

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

Remark 8.1.1. *Given $\theta \in \mathbb{R}$, we have*

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{20}$$

This means, in particular, that $e^{i\pi} = -1$. This is usually written as $e^{i\pi} + 1 = 0$ and known as Euler's formula.

A complex number $z \in \mathbb{C}$ can be written as

$$z = re^{i\theta}, \tag{21}$$

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. This is known under the name polar coordinates.

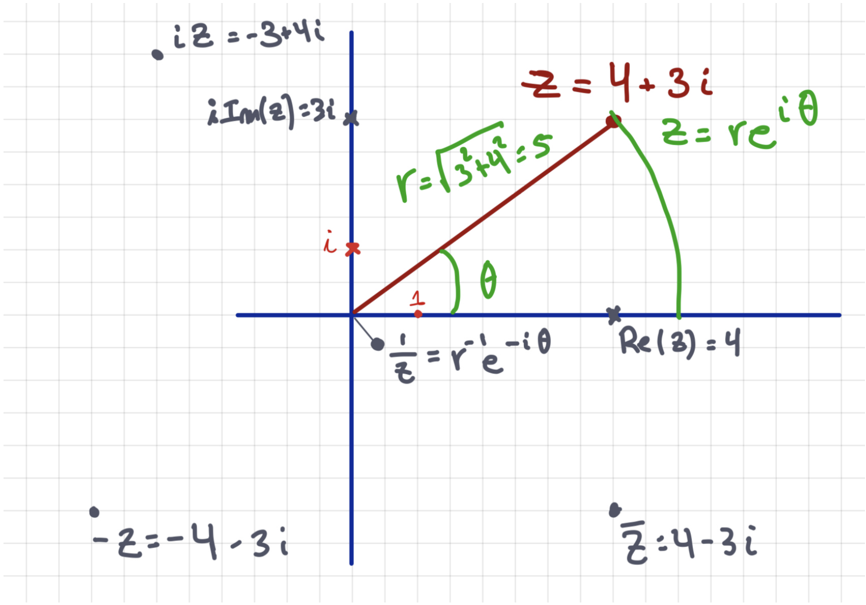


FIGURE 4. A complex number $z = 4+3i$ in the complex plane.

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 do not need to extend numbers further, $\mathbb{C}$ is algebraically closed.

Theorem 8.1.2 (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 $P(\lambda) = 0$.*

As the name suggests, Theorem 8.1.2 is a central result. Proving it is outside the scope of this course. See Appendix B for a relatively elementary proof.

Let us return to the question why Theorem 8.1.2 is so important for us. 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 8.1.3. *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 zeros: $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, perhaps with repetitions, such that*

$$P(z) = \alpha_n(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n). \tag{22}$$

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

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,

$$A^* = \overline{A}^T. \tag{23}$$

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 $\mathbb{R}^n$, we say $v_1, \ldots, v_k \in \mathbb{C}^n$ are linearly independent if there is no (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).

With these definitions one 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 that the first author usually teaches [BM23]. You can also see a discussion of Fourier Transform, circulant matrices, and signal convolutions in [Str23] (end of Section 6.4).

8.2. **Introduction to Eigenvalues and Eigenvectors.**

Even though the theory can be analogously developed for complex valued matrices, we will from now on focus on real valued matrices. We will use a guiding example to illustrate both some of the power and some of the properties of eigenvalues and eigenvectors. The central object of study below is given next.

Definition 8.2.1. *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 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.

Example 8.2.2. *In this example we will derive a formula for the n-th Ficonacci Number. The Fibonacci numbers are defined by the recurrence:*

$$F_0 = 0,\ F_1 = 1,\ \textit{and, for } n \geq 2,\ F_n = F_{n-1} + F_{n-2}. \tag{24}$$

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

$$\begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix}. \tag{25}$$

Defining

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \textit{ and } g_n = \begin{pmatrix} F_{n+1} \\ F_n \end{pmatrix}, \tag{26}$$

the recurrence can be rewritten as

$$g_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \textit{ and } g_n = M g_{n-1}.$$

This leads us to the formula

$$g_n = M^n g_0. \tag{27}$$

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 $\lambda \in \mathbb{R}$ 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).[2] This is equivalent to $\det(M - \lambda I) = 0$ and so we can find the eigenvalues λ with this equation:

$$0 = \det(M - \lambda I) = \begin{vmatrix} 1-\lambda & 1 \\ 1 & 0-\lambda \end{vmatrix} = (1-\lambda)(0-\lambda) - 1 = \lambda^2 - \lambda - 1. \tag{28}$$

By the quadratic formula,[3] the solutions to (28) are given by

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \textit{ and } \lambda_2 = \frac{1-\sqrt{5}}{2}. \tag{29}$$

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.

[2] Normally, we would have to look for $\lambda \in \mathbb{C}$ and $v \in \mathbb{C}^n$ but in this case the eigenvalues, as we will see, are real.

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

> *"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"*

Next 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 $\mathbb{N}\left(A - \frac{1+\sqrt{5}}{2}I\right)$. In other words

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

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 $v_1 = \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{pmatrix}$ is an eigenvector of M associated to the eigenvalue $\lambda_1 = \frac{1+\sqrt{5}}{2}$.

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

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{pmatrix} = \frac{1+\sqrt{5}}{2} \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{pmatrix} \text{ and } \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{pmatrix} = \frac{1-\sqrt{5}}{2} \begin{pmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{pmatrix}. \tag{30}$$

It is left to the reader to carry out the calculations and confirm that we have indeed found two eigenvectors. This requires one to check the two equalities in (30).

The vectors v_1 and v_2 we constructed in the example are not the only possible choices. For example, any non-zero scalar multiple 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 $v \in \mathbb{N}(A - \lambda I) \setminus \{0\}$ 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 $\mathbb{N}(A-\lambda I)$.

Let us first formulate this for real eigenvalues and eigenvectors.

Lemma 8.2.3. *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 $v \in \mathbb{R}^n \setminus \{0\}$ is an eigenvector associated with the eigenvalue λ if (and only if) $v \in \mathbb{N}(A-\lambda I)$.*

Proof. Since $\lambda \in \mathbb{R}$, we have that a solution of $Av=\lambda v$ is real valued. Suppose $v=v_1+iv_2 \in \mathbb{C}^n$. Then

$$Av = Av_1 + iAv_2 = \lambda v_1 + i\lambda v_2 \iff Av_1 = \lambda v_1 \text{ and } Av_2 = \lambda v_2,$$

i.e., v_1 and v_2 are both real eigenvectors corresponding to eigenvalue $\lambda \in \mathbb{R}$. □

A direct inspection of the formula for the determinant (Definition 7.2.3) gives the following.

Proposition 8.2.4. *$\det(A-\lambda I)$ is a polynomial in λ of degree n. The coefficient of the λ^n term is $(-1)^n$.*

The Fundamental Theorem of Algebra (Theorem 8.1.2) immediately implies

Theorem 8.2.5. *Every matrix $A \in \mathbb{R}^{n\times n}$ has an eigenvalue (perhaps complex-valued).*

Let us see an example of a matrix that has no real eigenvalues and only complex-valued ones.

Example 8.2.6. *The eigenvalues of the matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, corresponding to a 90^o counterclockwise rotation, are the solutions to $0=\det(A-\lambda I)=\lambda^2+1$, which are $\lambda_1 = i$ and $\lambda_2=-i$. The eigenvectors are given by $v_1 = \begin{pmatrix} i \\ 1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} -i \\ 1 \end{pmatrix}$.*

In this example the two eigenvalues and eigenvectors are conjugates of each other. Moreover, the modulus of the eigenvalues was one. The latter property is not a surprise. Indeed, the matrix in the example is a particular case of an orthogonal matrix, whose eigenvalues have a special property.

Proposition 8.2.7. *Let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. If $\lambda \in \mathbb{C}$ is an eigenvalue of Q, then $|\lambda|=1$.*

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue of Q and $v \in \mathbb{C}^n$ an associated eigenvector. Then $Qv = \lambda v$. Since Q is an orthogonal matrix we have $\|v\|^2 = \|Qv\|^2 = \|\lambda v\|^2 = |\lambda|^2\|v\|^2$. Since $v \neq 0$ we have $|\lambda| = 1$. □

The other property we observed in this example is generally true. Complex eigenvalue-eigenvector pairs are closed under taking the conjugate-operation.

Lemma 8.2.8. *Let $A \in \mathbb{R}^{n\times n}$. If (λ, v) is an eigenvalue-eigenvector pair, then $(\bar{\lambda}, \bar{v})$ is an eigenvalue-eigenvector pair.*

Proof. Let

$$\lambda = \lambda_1 + i\lambda_2 \text{ and } v = v_1 + iv_2,\ v_1, v_2 \in \mathbb{R}^n,\ \lambda_1, \lambda_2 \in \mathbb{R}.$$

We calculate

$$Av_1 + iAv_2 = Av = \lambda v = \lambda_1 v_1 + i\lambda_1 v_2 + i\lambda_2 v_1 + i^2\lambda_2 v_2 = (\lambda_1 v_1 - \lambda_2 v_2) + i(\lambda_1 v_2 + \lambda_2 v_1)$$

and obtain that $Av_1 = \lambda_1 v_1 - \lambda_2 v_2$ and $Av_2 = \lambda_1 v_2 + \lambda_2 v_1$. This allows us to verify that

$$A\bar{v} = Av_1 - iAv_2 = \lambda_1 v_1 - \lambda_2 v_2 - i(\lambda_1 v_2 + \lambda_2 v_1) = (\lambda_1 - i\lambda_2)(v_1 - iv_2) = \bar{\lambda}\bar{v}.$$

□

Let us return to Example 8.2.2.

Example 8.2.9. *Notice that v_1 and v_2 are linearly independent, and so they are a basis for $\mathbb{R}^2$. We can write $g_0 = \alpha_1 v_1 + \alpha_2 v_2$.*

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

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 – a formal proof would use induction – gives $A^n v_1 = \lambda_1^n v_1$. This means that

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

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

$$F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n. \tag{31}$$

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 true in general as we will show in the next section.

8.3. **Properties of eigenvalues and eigenvectors.**

Eigenvalues and eigenvectors of a matrix tell us many secrets about a matrix. This topic is illuminated in this section.

Proposition 8.3.1.

(a) *If λ and v are an eigenvalue-eigenvector pair of a matrix A, then, for $k \geq 1$, λ^k and v are an eigenvalue-eigenvector pair of the matrix A^k.*
(b) *Let A be an invertible matrix. If λ and v are an eigenvalue-eigenvector pair of the matrix A, then, $\frac{1}{\lambda}$ and v are an eigenvalue-eigenvector pair of the matrix A^{-1}.*

Proof. (a) Proof by Induction. The base case $k = 1$ is obvious.

For the inductive step, assume that λ^{k-1} and v are an eigenvalue-eigenvector pair for A^{k-1}. Then the following calculation shows us that λ^k and v are an eigenvalue-eigenvector pair for A^k:

$$A^k v = A\left(A^{k-1}v\right) = A\left(\lambda^{k-1}v\right) = \lambda^k v.$$

(b) A is invertible and hence, in the statement $\lambda \neq 0$. Since $Av = \lambda v$ we have $A^{-1}(\lambda v) = v$ and so $\lambda A^{-1}v = v$, which (since $\lambda \neq 0$) is equivalent to $A^{-1}v = \frac{1}{\lambda}v$. □

Another important property that we used in Example 8.2.9 was that we were able to write a vector as a linear combination of eigenvectors. This was possible because the eigenvectors were linearly independent.

Lemma 8.3.2. *Let $A \in \mathbb{R}^{n\times n}$ and let $v_1,\dots,v_k \in \mathbb{R}^n$ be eigenvectors corresponding to eigenvalues $\lambda_1,\dots,\lambda_k \in \mathbb{R}$. If $\lambda_1,\dots,\lambda_k$ are all distinct, the eigenvectors $v_1,\dots,v_k$ are linearly independent.*

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

$$v_j = \alpha_1 v_1 + \cdots \alpha_{j-1} v_{j-1}. \tag{32}$$

If we multiply by A both sides we get

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

Replacing the v_j in the left hand side with the right hand side of (32) we get

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

which we can rearrange as

$$\alpha_1\left(\lambda_j - \lambda_1\right) v_1 + \alpha_2\left(\lambda_j - \lambda_2\right) v_2 + \cdots + \alpha_{j-1}\left(\lambda_j - \lambda_{j-1}\right) v_{j-1} = 0. \tag{33}$$

Since $\lambda_j - \lambda_i \neq 0$ for all $i \leq j-1$ and not all α_i's are zero, this is a non-zero linear combination of $v_1, \ldots, v_{j-1}$ adding to zero, which would 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 8.3.3. *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 8.1.3, are all distinct) then there is a basis of $\mathbb{R}^n$, $v_1, \ldots, v_n$, made up of eigenvectors of A.*

Let us next discuss several further properties of the eigenvalues.

Definition 8.3.4. *Let $A \in \mathbb{R}^{n \times n}$.*

$$P(z) = (-1)^n \det(A - zI) = \det(zI - A) = (z - \lambda_1)(z - \lambda_2) \cdots (z - \lambda_n). \tag{34}$$

The polynomial $P(z)$ in (34) *is called the characteristic polynomial of the matrix A. The eigenvalues $\lambda_1, \ldots, \lambda_n$ as they show up in* (34) *are not all distinct in general. The number of times an eigenvalue shows up is called the algebraic multiplicity of the eigenvalue.*

The trace of A is defined as $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$.

As a side remark, there is a converse to this in the sense that any monic polynomial can be written as a characteristic polynomial of a matrix. There is a particularly elegant way to build the matrix, if you are interested in learning more, look-up “companion matrix”.

We continue with exploring further properties of the eigenvalues.

Lemma 8.3.5. *The eigenvalues of $A \in \mathbb{R}^{n \times n}$ are the same as the ones of $A^\top$.*

Proof. This follows from (34), and the fact that, for $\det(A-zI) = \det((A-zI)^\top) = \det(A^\top - zI)$. □

Even though the eigenvalues of A and $A^\top$ are the same, the eigenvectors are in general not! Try to verify this in an example.

A link between eigenvalues, the determinant and the trace of a matrix is established below.

Lemma 8.3.6. *Let $A \in \mathbb{R}^{n\times n}$ and $\lambda_1,\ldots,\lambda_n$ its n eigenvalues as they show up in* (34). *Then*

$$\det(A) = \prod_{i=1}^{n} \lambda_i \text{ and } \operatorname{Tr}(A) = \sum_{i=1}^{n} \lambda_i.$$

Proof. Let us consider the characteristic polynomial.

$$\begin{aligned}(-1)^n \det(A-zI) &= (z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n)\\ &= z^n + (-\textstyle\sum_{i=1}^n \lambda_i)z^{n-1} + \sum_{k=1}^{n-2} b_k z^k + (-1)^n \prod_{i=1}^n \lambda_i,\end{aligned}$$

where $b_k \in \mathbb{C}$.

Set $z=0$ in the expression above. It gives $(-1)^n \det(A) = (-1)^n \prod_{i=1}^n \lambda_i$ as claimed.

For the second claim, note that the coefficient of z^{n-1} in the characteristic polynomial (34) is given in the right hand side by $(-\sum_{i=1}^n \lambda_i)$. On the left hand side the coefficient of z^{n-1} can only come from the permutation that takes all diagonal elements in the matrix $zI - A$. Hence it is the coefficient of z^{n-1} of $\prod_{i=1}^n (z - A_{ii})$ which is $-\sum_{i=1}^n A_i i = -\operatorname{Tr}(A)$. □

Lemma 8.3.6 is quite surprising. $A \in \mathbb{R}^{n\times n}$ and hence, $\det(A) \in \mathbb{R}$ and $\operatorname{Tr}(A) \in \mathbb{R}$. The eigenvalues $\lambda_1,\ldots,\lambda_n \in \mathbb{C}$ are in general not real. Still the sum of all eigenvalues is always a real number and their product as well. Can you see why this is true? When calculating eigenvalues, Lemma 8.3.6 can also be useful. For instance, imagine we are given a real 3×3-matrix A and established that $\lambda_1 = 1$ is an eigenvalue of A. Suppose that we also know that $\det(A) = 4$ and $\operatorname{Tr}(A) = 5$. Then A must have eigenvalues $\lambda_2 = \lambda_3 = 2$. It is also elementary to verify the following calculations involving the trace-operator.

Lemma 8.3.7. *For matrices $A,B,C \in \mathbb{R}^{n\times n}$ one has*

$$\operatorname{Tr}(AB) = \operatorname{Tr}(BA) \text{ and } \operatorname{Tr}(ABC) = \operatorname{Tr}(BCA) = \operatorname{Tr}(CAB).$$

Proof.

$$\operatorname{Tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} B_{ji}A_{ij} = \operatorname{Tr}(BA).$$

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(A(BC)) = \mathrm{Tr}((BC)A) = \mathrm{Tr}(B(CA)) = \mathrm{Tr}(CAB).$$

□

Let us end this chapter with a few additional comments. The eigenvalues of $A+B$ are not easily computed from the eigenvalues of A and the ones of B, in particular they are not their sum! Also, the eigenvalues of AB or BA are not easily computed from the eigenvalues of A and the ones of B. They might not be their product!

More importantly, Gaussian Elimination does not preserve eigenvalues and eigenvectors. In this respect, eigenvalues heavily rely on the representation of the given matrix and do less reflect properties of corresponding subspaces etc.

Last, but not least let us return to Example 8.2.2. This example provided 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 fact, Google's celebrated PageRank algorithm is based on the idea of how eigenvectors can be used for ranking. You can read more about it here [BSS]. You might also want to take a look at "Landau on Chess Tournaments and Google's PageRank" by Rainer Sinn and Günter M. Ziegler (`https://arxiv.org/pdf/2210.17300.pdf`).

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 instability/chaos. If it represents the evolution of an economical system over time (or the finances of a company) it can be the difference between growth or ruin.

9. Diagonizable matrices and the singular value decomposition

In this chapter we will explore a deep connection between a matrix defining a linear transformation from the column space to the row space and the representation of these subspaces. Our guiding question here is when can we find a basis of the subspace of the rows and columns such that viewed in this basis our matrix has the simplest possible representation in form of a diagonal matrix. On a first glance, this topic might have little to do with eigenvalues and eigenvectors. This is, however not true as we will show below.

9.1. **From eigenvalues to diagonizable matrices.**

An important part of the success of the strategy we took in Example 8.2.2 was the fact that we were able to build a basis of $\mathbb{R}^2$ with eigenvectors of the matrix M. In Theorem 8.3.3 we showed that we can always build a basis of $\mathbb{R}^n$ with eigenvectors of an $n \times n$ matrix A if A has n distinct real eigenvalues.

Let us show next that whenever we have a basis of $\mathbb{R}^n$ that consists of real eigenvectors of a matrix A, then we can change the basis of the row and column space such that A can be viewed as a diagonal matrix in this new representation. This is a remarkable fact that explains why real eigenvectors of a matrix are important.

Theorem 9.1.1. *Let $A \in \mathbb{R}^{n\times n}$. Suppose that A has eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$ that form a basis of $\mathbb{R}^n$. For $i \in \{1, \ldots, n\}$ let λ_i be the eigenvalue associated to v_i. Let V be the matrix whose columns correspond to the eigenvectors $v_1, \ldots, v_n$. Then,*

$$A = V\Lambda V^{-1}, \tag{35}$$

where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i$ (and $\Lambda_{ij} = 0$ for all $i \neq j$).

Proof. Since $v_1, \ldots, v_n$ is a basis of $\mathbb{R}^n$, the matrix V is invertible. Hence, it suffices to verify that

$$V^{-1}AV = \Lambda. \tag{36}$$

This can be done by direct calculation: For any $1 \leq j \leq n$, the j-th column of the matrix $V^{-1}AV$ is given by

$$\left(V^{-1}AV\right)_{\cdot j} := \left(V^{-1}AV\right)e_j = V^{-1}Av_j = V^{-1}\lambda_j v_j = \lambda_j V^{-1}v_j = \lambda_j e_j,$$

since $V^{-1}v_j = V^{-1}Ve_j = e_j$. Recall that e_j is the vector in $\mathbb{R}^n$ with a 1 in j-th entry and zero elsewhere. Since for any $1 \leq j \leq n$, $\lambda_j e_j$ is also the j-th column of Λ, we have that $V^{-1}AV = \Lambda$. □

Many properties of the eigenvalues that we discussed in the previous chapter are easy to prove by using Theorem 9.1.1. Try it!

Theorem 9.1.1 justifies the following two definitions.

Definition 9.1.2 (Diagonalizable Matrix)**.** *A matrix $A \in \mathbb{R}^{n\times n}$ is called a diagonalizable matrix if there exists an invertible matrix V such that $V^{-1}AV = \Lambda$, where Λ is a diagonal matrix.*

Definition 9.1.3. *If, given a matrix $A \in \mathbb{R}^{n\times n}$, we can build a basis of $\mathbb{R}^n$ with eigenvectors of A we say that A has a complete set of real eigenvectors.*

If we have a matrix $A \in \mathbb{R}^{n\times n}$ with a complete set of real eigenvectors then Theorem 9.1.1 tells us that the corresponding linear transformation, when viewed in the basis $v_1, \ldots, v_n$ is simply a diagonal matrix. This is a remarkable fact: since many matrices have a complete set of real eigenvectors this says that all the corresponding linear combinations, regardless of how complicated they might seem, are actually just a diagonal operation when viewed in the basis $v_1, \dots, v_n$. Let us now study classes of matrices for which there exists a complete set of real eigenvectors.

Example 9.1.4. *For $D \in \mathbb{R}^{n \times n}$ a diagonal matrix, the eigenvalues of D are the diagonal entries of D. The canonical basis $e_1, \dots, e_n$ is a complete set of real eigenvectors of D.*

Another obvious fact is

Example 9.1.5. *The eigenvalues of an $n \times n$ triangular matrix are the n values in the diagonal. However, triangular matrices may not have a complete set of real eigenvectors. Try to find an example!*

The following class of matrices has been studied in the context of projections.

Proposition 9.1.6 (Eigenvalues and Eigenvectors of a Projection Matrix)**.** *Let P be the projection matrix on the subspace $U \subseteq \mathbb{R}^n$. Then P has two eigenvalues, 0 and 1, and a complete set of real eigenvectors.*

Proof. Let m be the dimension of U. Let $u_1, \dots, u_m$ be an orthonormal basis of U, and $w_1, \dots, w_{n-m}$ an orthonormal basis of $U^\perp$. It is easy to see that $Pu_k = 1u_k$ for any $1 \leq k \leq m$ and $Pw_k = 0w_k$ for any $1 \leq k \leq n-m$, so all n vectors are eigenvectors of P (with eigenvalues either 1 or 0). By construction of $U^\perp$, they form an orthonormal basis. □

We can also apply operations on matrices that preserve the property of having a complete set of real eigenvectors.

Definition 9.1.7 (Similar matrices)**.** *We say that $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ are similar matrices if there exists an invertible matrix S such that $B = S^{-1}AS$.*

Proposition 9.1.8. *Similar matrices $A \in \mathbb{R}^{n \times n}$ and $B = S^{-1}AS \in \mathbb{R}^{n \times n}$ have the same eigenvalues. The matrix A has a complete set of real eigenvectors if and only if B does.*

Proof. $v \in \mathbb{R}^n$ is eigenvector associated with eigenvalue $\lambda \in \mathbb{R}$ for matrix A iff

$$Av = \lambda v \iff \lambda S^{-1}v = S^{-1}Av = S^{-1}ASS^{-1}v = B(S^{-1}v).$$

Hence, the eigenvalues of A and B coincide. Moreover, v is eigenvector associated with eigenvalue $\lambda \in \mathbb{R}$ for matrix A iff S^-1v is eigenvector associated with eigenvalue $\lambda \in \mathbb{R}$ for matrix B. All eigenvectors $v_1, \dots, v_n$ of A form a basis of $\mathbb{R}^n$ iff all eigenvectors $S^{-1}v_1, \dots, S^{-1}v_n$ of B form a basis of $\mathbb{R}^n$. This completes the proof.

□

We have by now seen various classes of matrices that are diagonazible. But not all matrices allow us to work with this beautiful and powerful property.

Example 9.1.9.

(a) *The matrix* $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ *does not have two linearly independent eigenvectors. Indeed,* $\det(A - \lambda I) = \lambda^2$ *which means that* $\lambda = 0$ *is the only eigenvalue and it has algebraic multiplicity* 2. *However,* $N(A - 0I) = N(A)$ *only has dimension* 1. *Hence, there is only one eigenvector (and multiples of it) corresponding to eigenvalue* 0.

(b) *The zero matrix* $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ *does have two linearly independent eigenvectors. Indeed,* $\det(A - \lambda I) = \lambda^2$ *which means that* $\lambda = 0$ *is the only eigenvalue and has algebraic multiplicity* 2. *But, unlike in (a), we now notice that* $N(A - 0I) = N(A)$ *has dimension* 2. *Hence there is a basis made up of two eigenvectors (in fact any two linearly independent vectors will be such a basis).*

For the matrix A considered in Example 9.1.9 (a), we obtain that $N\left(A^2\right)$ has dimension 2. When there exist a positive integer k such that $A^k = 0$ we call A nilpotent. There is a rather deep Theorem that essentially says nilpotency is the only obstacle to get a complete set of eigenvectors. It roughly says that when there are "missing" eigenvectors they can be found in the nullspace of powers of $A - \lambda I$. This gives rise to something called "Jordan Normal Form", a topic beyond the scope of this course.

In general when there is an eigenvalue $\lambda \in \mathbb{R}$ with algebraic multiplicity larger than 1, it can be that $\mathbb{N}(A - \lambda I)$ is of large enough dimension to find enough linearly independent eigenvectors. This applies, for instance to projection matrices as we observed above. It does, however, not apply to the nilpotent example.

Definition 9.1.10. *Given a matrix* $A \in \mathbb{R}^{n \times n}$ *and an eigenvalue* λ *of* A *we call the dimension of* $\mathbb{N}(A - \lambda I)$ *the geometric multiplicity of* λ.

A link between the property that a matrix has a complete set of eigenvectors and the algebraic and geometric multiplicities of its eigenvalues is given below.

Lemma 9.1.11. *A matrix has a complete set of real eigenvectors if all its eigenvalues are real and the geometric multiplicities are the same as the algebraic multiplicites of all its eigenvalues.*

Proof. Let $A \in \mathbb{R}^{n \times n}$ be a matrix with eigenvalues $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$. Let m_i be the algebraic multiplicity of eigenvalue λ_i, i.e., $\dim(N(A - \lambda_i I)) = m_i$. For $i \in \{1, \ldots, k\}$ take a basis $v_{i_1}, \ldots, v_{i_{m_i}}$ of the subspace $N(A - \lambda_i I)$. From Lemma 8.3.2 it follows that for $i \neq j$ we have that the vectors v_{i_s} and v_{j_l} are linearly independent. Since $m_1 + m_2 + \ldots + m_k = n$, the set

$$\bigcup_{i=1}^{k} \{v_{i_1}, \ldots, v_{i_{m_i}}\}$$

defines a basis of $\mathbb{R}^n$.

□

So far we established a connection between diagonalize matrices and their eigenvectors, provided that they form a basis of the underlying space. What can we do if this is not true? Or even more generally, the concept of eigenvectors and eigenvalues only applies to square matrices. Is there something more general that we can discover for $m \times n$ matrices? In order to prepare for this task let us take a detour in studying a linear transformation written in different bases.

Let $A \in \mathbb{R}^{m\times n}$ be a matrix representing a linear transformation

$$L : \mathbb{R}^n \to \mathbb{R}^m$$

given by $x \in \mathbb{R}^n \to Ax \in \mathbb{R}^m$, with both input and output written in the canonical bases as

$$x = \sum_{j=1}^{n} x_j e_j \text{ and } Ax = \sum_{i=1}^{m} (Ax)_i e_i.$$

Recall (Example 6.3.2) that $(e_i)_j = \delta_{ij}$, and that $(Ax)_i$ is the i-th entry of the vector Ax.

Suppose that we are given a basis $u_1, \ldots, u_n$ of $\mathbb{R}^n$ and a basis $v_1, \ldots, v_m$ of $\mathbb{R}^m$. Our task is to understand the linear transformation L written in these bases. Then L takes a vector $x = \sum_{j=1}^{n} \alpha_j u_j$ and outputs $L(x) = \sum_{j=1}^{n} \beta_i v_i$. We want to compute the matrix B that takes

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \text{ to } B\alpha = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}.$$

To achieve this, let $U \in \mathbb{R}^{n\times n}$ be the matrix whose columns are the basis elements $u_1, \ldots, u_n$ and $V \in \mathbb{R}^{m\times m}$ the matrix whose columns are the basis elements $v_1, \ldots, v_m$. Then,

$$x = U\alpha \text{ and } L(x) = V\beta. \text{ Hence, } \beta = V^{-1}AU\alpha.$$

The matrix B, corresponding to the linear transformation L represented in the new basis is $B = V^{-1}AU$. Note that we can make such a change of bases between any pair of bases. It is not required that one basis is the canonical basis. In this general case, the role of U and V would be played by the change of the basis matrix (the matrix that maps the coefficients of a vector represented in the old basis to its coefficients when written in the new basis).

$$\begin{aligned} L:\mathbb{R}^n &\to \mathbb{R}^m && \text{linear transformation} \\ L\left(\sum_{j=1}^n x_j e_j\right) &= \sum_{i=1}^n (Ax)_i e_i && x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ L\left(\sum_{j=1}^n \alpha_j u_j\right) &= \sum_{i=1}^n (B\alpha)_i v_i && \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \end{aligned} \tag{37}$$

$$\text{where } B = V^{-1}AU \in \mathbb{R}^{m\times n}, \quad U = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad V = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \in \mathbb{R}^{m\times m}.$$

Let us connect this topic with with our discussions on eigenvectors of a square matrix $A \in \mathbb{R}^{n\times n}$. In particular, let A be a matrix with a complete set of real eigenvectors (in the sense of Definition 9.1.3) and let $v_1,\ldots,v_n \in \mathbb{R}^{n\times n}$ be a basis formed with eigenvectors of A. If we write a vector $x \in \mathbb{R}^n$ as $x = \sum_{i=1}^n \alpha_i v_i$ then $Ax = \sum_{i=1}^n \lambda_i \alpha_i v_i$. One way to think about this is that the linear transformation corresponding to the matrix A, when written in the basis V is simply a diagonal matrix/transformation. This is the key idea behind matrix diagonalization and one of the most important facts in Linear Algebra.

9.2. **Symmetric Matrices and the Spectral Theorem.**

We continue our discussion about classes of matrices that allow us to perform a change of basis so that the corresponding linear transformation becomes diagonal. In fact, this section deals with an important class, namely real symmetric matrices meaning matrices $A \in \mathbb{R}^{n\times n}$ for which $A^\top = A$. The main goal of this section is to prove the Spectral Theorem.

Theorem 9.2.1 (Spectral Theorem)**.** *Any symmetric matrix $A \in \mathbb{R}^{n\times n}$ has n real eigenvalues and an orthonormal basis of $\mathbb{R}^n$ consisting of its eigenvectors.*

Together with Theorem 9.1.1 this implies the following corollary.

Corollary 9.2.2. *For any symmetric matrix $A \in \mathbb{R}^{n\times n}$ there exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ (whose columns are eigenvectors of A) such that*

$$A = V\Lambda V^\top,$$

where $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the eigenvalues of A in its diagonal (and $V^\top V = I$).

Remark 9.2.3 (Eigendecomposition)**.** *The decompositions in Corollary 9.2.2 and Theorem 9.1.1 are called Eigendecompositions.*

The following follows easily from the Spectral Theorem.

Corollary 9.2.4. *The rank of a real symmetric matrix A is the number of non-zero eigenvalues (counting repetitions).*

Remark 9.2.5. *For general $n \times n$ (non-symmetric) matrices, the rank is n minus the dimension of the nullspace, so it is n minus the geometric multiplicity of $\lambda = 0$. Since symmetric matrices always have a complete set of eigenvalues and eigenvectors, the geometric multiplicities are always the same as the algebraic multiplicities.*

Proposition 9.2.6. *Let A be a real $n \times n$ symmetric matrix and let $v_1, \dots, v_n$ be an orthonormal basis of eigenvectors of A (the columns of the matrix V in Corollary 9.2.2) and $\lambda_1, \dots, \lambda_n$ the associated eigenvalues. Then*

$$A = \sum_{k=1}^{n} \lambda_i v_i v_i^\top$$

Proof. Follows directly from Corollary 9.2.2. □

We "build up" to the proof of Theorem 9.2.1 with a few observations. The first one shows us that symmetric matrices have a very strong property. Whereas for all square matrices, eigenvectors corresponding to distinct eigenvalues are linearly independent, the symmetry of a matrix implies a much stronger property.

Lemma 9.2.7. *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $\lambda_1 \neq \lambda_2 \in \mathbb{R}$ be two distinct eigenvalues of A with corresponding eigenvectors v_1, v_2. Then v_1 and v_2 are orthogonal.*

Proof. Suppose that $\lambda_1 \neq \lambda_2$ are eigenvalues of a real symmetric matrix A and $v_1, v_2 \in \mathbb{R}^n \setminus \{0\}$ corresponding eigenvectors. Then

$$\lambda_1 v_1^\top v_2 = (Av_1)^\top v_2 = v_1^\top A^\top v_2 = v_1^\top A v_2 = v_1^\top (Av_2) = \lambda_2 v_1^\top v_2.$$

The fact that $\lambda_1 \neq \lambda_2$ implies that $v_1^\top v_2 = 0$ □

A second observation is also important to know.

Lemma 9.2.8. *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $\lambda \in \mathbb{C}$ an eigenvalue of A, then $\lambda \in \mathbb{R}$.*

Proof. Let $v \in \mathbb{C}^n$ be an eigenvector associated with the eigenvalue λ. We have $Av = \lambda v$. Recall that, for a matrix (or vector) M, its Hermitian conjugate is given by $M^* = \overline{M}^\top$. Since A is real symmetric we have $A^* = A$. Thus, we have

$$\overline{\lambda}\|v\|^2 = \overline{\lambda}v^*v = (\lambda v)^*v = (Av)^*v = v^*A^*v = v^*Av = v^*\lambda v = \lambda\|v\|^2.$$

Since $v \neq 0$, then $\|v\| \neq 0$ and so $\lambda = \overline{\lambda}$. This implies that $\lambda \in \mathbb{R}$. □

This, together with Theorem 8.2.5, immediately implies the following.

Corollary 9.2.9. *Every symmetric matrix $A \in \mathbb{R}^{n\times n}$ has a real eigenvalue λ.*

Corollary 9.2.9 is a great example of the usefulness of complex numbers. Even though it is a statement just about real matrices and real eigenvalues we proved it by going through the complex numbers and using results in Complex Analysis.

We are now prepared to prove the spectral theorem. The fact that two eigenvectors of a real symmetric matrix corresponding to distinct eigenvalues are orthogonal follows from Lemma 9.2.7. The main difficulty is proving that the matrix indeed has a complete set of eigenvectors.

Proof. [of Theorem 9.2.1]

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. We will prove the following by induction, which for $k = n$ implies the theorem we want to show:

for any $1 \leq k \leq n$ there are k orthogonal eigenvectors of A.

The base case $k = 1$ follows from Corollary 9.2.9 as we can normalize an eigenvector to have norm one.

We now assume that the statement is true for k and show it for $k+1$. We will show that if a real symmetric matrix A has k (with $1 \leq k < n$) orthonormal eigenvectors then we can build an extra one, orthogonal to the others. Note that in order to achieve norm 1 we simply need to normalize the vector.

Let $v_1, \ldots, v_k$ denote k orthonormal eigenvectors of A and $\lambda_1, \ldots, \lambda_k$ the respective eigenvalues. Let $u_{k+1}, \ldots, u_n$ be an orthonormal basis of the orthogonal complement of the span of $v_1, \ldots, v_k$. Let V_k be the $n \times n$ matrix whose i-th column is v_i if $i \leq k$ and u_i if $i > k$. V_k is an orthogonal matrix. Moreover, let us define $B \in \mathbb{R}^{n\times n}$ as $B = V^\top AV$, then:

$$\begin{aligned}
B = V^\top AV \quad &= \begin{bmatrix} — & v_1^\top & — \\ & \vdots & \\ — & v_k^\top & — \\ — & u_{k+1}^\top & — \\ & \vdots & \\ — & u_n^\top & — \end{bmatrix} \begin{bmatrix} | & & | & | & & | \\ Av_1 & \cdots & Av_k & Au_{k+1} & \cdots & Au_n \\ | & & | & | & & | \end{bmatrix} \\
&= \begin{bmatrix} — & v_1^\top & — \\ & \vdots & \\ — & v_k^\top & — \\ — & u_{k+1}^\top & — \\ & \vdots & \\ — & u_n^\top & — \end{bmatrix} \begin{bmatrix} | & & | & | & & | \\ \lambda v_1 & \cdots & \lambda v_k & Au_{k+1} & \cdots & Au_n \\ | & & | & | & & | \end{bmatrix} \\
&= \begin{bmatrix} \Lambda_k & 0_{k\times(n-k)} \\ 0_{(n-k)\times k} & C \end{bmatrix},
\end{aligned}$$

where Λ_k is a diagonal matrix with $\lambda_1,\dots,\lambda_k$ in the diagonal, $0_{(n-k)\times k}$ and $0_{k\times(n-k)}$ are zero matrices of size respectively $(n-k)\times k$ and $k\times(n-k)$. C is a $(n-k)\times(n-k)$ symmetric matrix. Note that we obtain here zero matrices of size $0_{(n-k)\times k}$ and $0_{k\times(n-k)}$, respectively because

for all i and j we have $u_j^T v_i = 0$ and since A is symmetric $v_i^T Au_j = v_i^T A^T u_j = (Av_i)^T u_j = \lambda_i v_i^T u_j = 0$.

Since C is a $(n-k)\times(n-k)$ symmetric matrix, Theorem 9.2.9 implies it has a real eigenvalue λ_{k+1} and a real eigenvector $y \in \mathbb{R}^{n-k}$. Let $w \in \mathbb{R}^n$ be the vector with 0 in the first k coordinates and y in the remaining $n-k$, in other words

$$w_i = \begin{cases} 0 & \text{if } \; i \le k \\ y_{i-k} & \text{if } \; i > k. \end{cases}$$

We have

$$Bw = \begin{bmatrix} \Lambda_k & 0_{k\times(n-k)} \\ 0_{(n-k)\times k} & C \end{bmatrix} \begin{pmatrix} 0_{k\times 1} \\ y \end{pmatrix} = \begin{pmatrix} 0_{k\times 1} \\ Cy \end{pmatrix} = \begin{pmatrix} 0_{k\times 1} \\ \lambda_{k+1} y \end{pmatrix} = \lambda_{k+1} w.$$

Let $v_{k+1} := Vw$. Since V is orthogonal we have that $A = VBV^\top$. Thus,

$$Av_{k+1} = VBV^\top v_{k+1} = VBw = V\lambda_{k+1} w = \lambda_{k+1} v_{k+1},$$

so v_{k+1} is an eigenvector of A. To see that it is orthogonal to $v_1,\dots,v_k$ note that the inner products $v_i^\top v_{k+1}$ for $i \le k$ appear in the first k entries of $V^\top v_{k+1} = w$ and that w has its first k coordinates 0 by construction. By normalizing the vector we can ensure that it attains unit norm. □

A first application of the Spectral Theorem is presented next.

Proposition 9.2.10 (Rayleigh Quotient)**.** *Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ the Rayleigh Quotient, defined for $x \in \mathbb{R}^n \setminus \{0\}$, as*

$$R(x) = \frac{x^\top Ax}{x^\top x}$$

attains its maximum at $R(v_{\max}) = \lambda_{\max}$ and its minimum at $R(v_{\min}) = \lambda_{\min}$ where $\lambda_{\max}$ and $\lambda_{\min}$ are, respectively, the largest and smallest eigenvalues of A and $v_{\max}$, $v_{\min}$ their associated eigenvectors.

Proof. It is easy to see that $R(v_{\max}) = \lambda_{\max}$ and $R(v_{\min}) = \lambda_{\min}$. Thus it suffices to show that, for all $x \in \mathbb{R}^n \setminus \{0\}$ we have $\lambda_{\min} \le R(x) \le \lambda_{\max}$. Using Proposition 9.2.6 we can write, for $x \in \mathbb{R}^n \setminus \{0\}$,

$$R(x) = \frac{x^\top \left(\sum_{i=1}^n \lambda_i v_i v_i^\top\right) x}{\|x\|^2} = \frac{\sum_{i=1}^n \lambda_i \left(x^\top v_i\right)^2}{\|x\|^2},$$

where $v_1, \dots, v_n$ form an orthonormal basis of eigenvectors of A and $\lambda_1, \dots, \lambda_n$ are the associated eigenvalues. Since $\left(x^\top v_i\right)^2 \ge 0$ for all $1 \le i \le n$ we have that, for all $1 \le i \le n$,

$$\lambda_{\min} \left(x^\top v_i\right)^2 \le \lambda_i \left(x^\top v_i\right)^2 \le \lambda_{\max} \left(x^\top v_i\right)^2.$$

Collecting all these inequalities we get

$$\lambda_{\min} \frac{\sum_{i=1}^n \left(x^\top v_i\right)^2}{\|x\|^2} \le \frac{\sum_{i=1}^n \lambda_i \left(x^\top v_i\right)^2}{\|x\|^2} \le \lambda_{\max} \frac{\sum_{i=1}^n \left(x^\top v_i\right)^2}{\|x\|^2}.$$

To conclude the proof note that, since the v_i's are orthonormal, the matrix V with the v_i's as columns is orthogonal and $\sum_{i=1}^n \left(x^\top v_i\right)^2 = \|Vx\|^2 = \|x\|^2$ and so $\frac{\sum_{i=1}^n \left(x^\top v_i\right)^2}{\|x\|^2} = 1$. □

Proposition 9.2.10 allows us to give a characterization of an important class of matrices. This class is called Positive Definite and Positive Semidefinite matrices and defined below.

Definition 9.2.11. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is said to be Positive Semidefinite (PSD) if all its eigenvalues are non-negative. If all the eigenvalues of A are strictly positive then we say A is Positive Definite (PD).*

The following characterization can be derived from Proposition 9.2.10.

Proposition 9.2.12. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is Positive Semidefinite if and only if $x^\top Ax \ge 0$ for all $x \in \mathbb{R}^n$. Analogously, a symmetric matrix $A \in \mathbb{R}^{n\times n}$ is Positive Definite if and only if $x^\top Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.*

Note that PSD and PD matrices are closed under taking additions: if A and B are PSD (or PD), then so is the sum. This can be verified by applying Proposition 9.2.10. Try it!

Definition 9.2.13 (Gram Matrix)**.** *Given n vectors, $v_1, \dots, v_n$ in $\mathbb{R}^m$ we call their Gram Matrix the $n \times n$ matrix of inner products*

$$G_{ij} = v_i^\top v_j.$$

If $V \in \mathbb{R}^{m\times n}$ is the matrix with columns $v_1, \dots, v_n$ in $\mathbb{R}^m$, then $G = V^\top V$ is the Gram matrix of V.

Remark 9.2.14. *Given a matrix $A \in \mathbb{R}^{m\times n}$, as an abuse of notation, we sometimes also call $AA^\top$ a Gram matrix of A. Notice that, if $a_1, \dots, a_n \in \mathbb{R}^m$ are the columns of A then $AA^\top$ is $m \times m$ and*

$$AA^\top = \sum_{i=1}^{n} a_i a_i^\top. \tag{38}$$

Proposition 9.2.15. *Given a real matrix $A \in \mathbb{R}^{m\times n}$, the non-zero eigenvalues of $A^\top A \in \mathbb{R}^{n\times n}$ are the same as the ones of $AA^\top \in \mathbb{R}^{m\times m}$. Both matrices are symmetric and positive semidefinite.*

Proof. Let r be the rank of A. We know

$$\operatorname{rank}(A) = \operatorname{rank}(A^\top) = \operatorname{rank}(A^\top A) = \operatorname{rank}(AA^\top).$$

Both matrices $A^\top A$ and $AA^\top$ are obviously symmetric. Let us check that they are positive semidefinite. To see this, we have $x^\top A^\top Ax = ||Ax||^2 \geq 0$ for all x which implies $A^\top A$ is PSD, and the same argument can be used for $AA^\top$.

Now, both $AA^\top$ and $A^\top A$ have a complete set of real eigenvalues and orthogonal eigenvectors. Let $\lambda_1, \dots, \lambda_r$ be the r non-zero eigenvalues of $A^\top A$ and $v_1 \dots, v_r$ be the corresponding eigenvectors. We have, for $1 \leq k \leq r$, $A^\top A v_k = \lambda_k v_k$, multiplying by A both sides we get $AA^\top A v_k = \lambda_k A v_k$ and so λ_k is an eigenvalue of $AA^\top$ with eigenvector Av_k (note that $Av_k \neq 0$). Furthermore, For $j \neq k$ we have $(Av_j)^\top (Av_k) = v_j^\top A^\top A v_k = v_j^\top \lambda_k v_k = \lambda_k v_j^\top v_k = 0$ and so the r eigenvectors of $AA^\top$ built this way are orthogonal, and so $\lambda_1, \dots, \lambda_r$ are the nonzero eigenvectors of $AA^\top$. □

Proposition 9.2.16. *[Cholesky decomposition] Every symmetric positive semidefinite matrix M is a Gram matrix of an upper triangular matrix C. $M = C^\top C$ is known as the Cholesky Decomposition.*[4]

Proof. Let M be a symmetric positive semidefinite matrix. Corollary 9.2.2 gives us a decomposition $M = V\Lambda V^\top$ with Λ a diagonal matrix with the eigenvalues of M in the diagonal. Since M is PSD, the diagonal entries of Λ are non-negative and so we can build $\Lambda^{1/2}$ by taking the square root of each diagonal entry of Λ. Then $M = \left(V\Lambda^{1/2}\right)\left(V\Lambda^{1/2}\right)^\top$. To make the matrices in the decomposition be upper triangular, simply take the QR decomposition (recall Definition 6.3.10) $\left(V\Lambda^{1/2}\right)^\top = QR$ with Q such that $Q^\top Q = I$ and R upper triangular. We have $M = \left(V\Lambda^{1/2}\right)\left(V\Lambda^{1/2}\right)^\top = (QR)^\top (QR) = R^\top Q^\top QR = R^\top R$. Taking $C = R$ establishes the Proposition. □

At first glance, symmetric matrices look very special (since we must have $A^\top = A$), but they actually appear very often in both applications and pure mathematics. There are (at least) two reasons for this:

(i) For any matrix B we can form a symmetric matrix $B^\top B$ from which we can study B. This is going to be the key idea behind the Singular Value Decomposition that we discuss next.

(ii) In many instances, matrices represent relationship between objects — for example, A_{ij} can represent a friendship connection (or a similarity measure) between person (or data point) i and j and in many cases such relationships are symmetric.

9.3. **Singular Value Decomposition.**

We are now reaching "the ultimate theorem of our class", the Singular Value Decomposition (SVD). The SVD is a way to generalize the eigendecomposition to non-symmetric, and even non-square, matrices. Instead of eigenvalues we will have singular values and instead of eigenvectors we will have (right and left) singular vectors.

Definition 9.3.1 (SVD — Singular Value Decomposition)**.** *Let $A \in \mathbb{R}^{m\times n}$. There exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$A = U\Sigma V^\top, \tag{39}$$

where $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix, in the sense that $\Sigma_{ij} = 0$ when $i \neq j$, and the diagonal elements are non-negative and ordered in descending order. $U^\top U = I$ and $V^\top V = I$.

The columns $u_1, \dots u_m$ of U are called the left singular vectors of A and are orthonormal. The columns $v_1, \dots v_n$ of V are called the right singular vectors of A and are orthonormal. The diagonal elements of Σ, $\sigma_i = \Sigma_{ii}$ are called the singular values of A and are ordered as

$$\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}.$$

Remark 9.3.2. *If A has rank r we can write the SVD in a more compact form:*

$$A = U_r\Sigma_r V_r^\top, \tag{40}$$

where $U_r \in \mathbb{R}^{m\times r}$ contains the first r left singular vectors, $V_r \in \mathbb{R}^{n\times r}$ contains the first r right singular vectors and $\Sigma_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the first r singular values. Notice that storing such a decomposition in the computer requires storing $r \times (m+n+1)$ real numbers rather than $m \times n$ real numbers which would be required to store A naively. When a matrix has small rank these are crucial savings. Taking this one step forward, when a matrix is well approximated by a low rank matrix, oftentimes one stores only a small rank approximation of a matrix A. This is a crucial idea in tasks ranging from Image Compressions, Numerical Analysis, and Machine Learning.

Oftentimes the subscript is omitted and the compact SVD is simply written as $U\Sigma V^\top$ while specifying the dimensions of the matrices involved to specify which form of the SVD is being considered.

Let $A \in \mathbb{R}^{m\times n}$ and $A = U\Sigma V^\top$ be its SVD (as in (39)) then

$$AA^\top = U\left(\Sigma\Sigma^\top\right)U^\top,$$

and so the left singular vectors of A, the columns of U, are the eigenvectors of $AA^\top$ and the singular values of A are the square-root of the eigenvalues of $AA^\top$ (note that $\Sigma\Sigma^\top$ is $m\times m$ diagonal). If $m > n$, A has n singular values and $AA^\top$ has m eigenvalues (which is larger than n), but the "missing" ones are 0.

Analogously,

$$A^\top A = V\left(\Sigma^\top\Sigma\right)V^\top,$$

and so the right singular vectors of A, the columns of V, are the eigenvectors of $A^\top A$ and the singular values of A are the square-root of the eigenvalues of $A^\top A$ (note that $\Sigma^\top\Sigma$ is $n\times n$ diagonal). If $n > m$, A has m singular values and $A^\top A$ has n eigenvalues (which is larger than m), but the "missing" ones are 0.

This observation makes it easier to write the singular values and singular vectors of A in terms of eigenvalues and eigenvectors of $AA^\top$ and $A^\top A$, which are symmetric matrices (and directly implies, e.g., uniqueness of singular values; and the fact that the rank of a matrix is the number of nonzero singular values). In fact, the proof of the existence of a SVD will heavily rely on the Spectral Theorem.

Theorem 9.3.3. *Every matrix $A \in \mathbb{R}^{m\times n}$ has an SVD decomposition of the form* (39). *In other words:*

Every linear transformation is diagonal when viewed in the bases of the singular vectors.

Proof. Let $A \in \mathbb{R}^{m\times n}$. Let r be the rank of A. We will build a compact SVD as in (40). It is easy to see that we can get an SVD in the sense of (39) from a compact one by adding singular values that are zero and extending the singular vectors in both U_r and V_r to orthonormal bases.

By Theorem 9.2.1 and Corollary 9.2.2 the matrix $AA^\top$ has a complete set of orthonormal eigenvectors and can be written as

$$AA^\top = U\Lambda U^\top, \tag{41}$$

where $U \in \mathbb{R}^{m\times m}$ is orthogonal and Λ is diagonal. Let us write (41) by ordering the diagonal entries of Λ in decreasing order. Furthermore, let us write (41) also in a compact form, by keeping only the r non-zero eigenvalues (and corresponding eigenvectors), i.e.,

$$AA^\top = U_r\Lambda_r U_r^\top$$

for $U_r \in \mathbb{R}^{m\times r}$ such that $U_r^\top U_r = I$ and Λ_r is $r\times r$ diagonal with the non-zero eigenvalues of $AA^\top$. By Proposition 9.2.15 the eigenvalues of $AA^\top$ are non-negative and so the diagonal entries of Λ_r are positive. Let $\Sigma_r \in \mathbb{R}^{r\times r}$ be the diagonal matrix with diagonal entries $\sigma_i := (\Sigma_r)_{ii} = \sqrt{\Lambda_{ii}}$. Our goal is to show that there is a $n\times r$ matrix V_r, with orthonormal columns, such that $A = U_r\Sigma_r V_r^\top$. We would have $\Sigma_r^{-1}U_r^\top A = \Sigma_r^{-1}U_r^\top U_r\Sigma_r V_r^\top = V_r^\top$, or equivalently $V_r = A^\top U_r\Sigma_r^{-1}$. Motivated by this, let us set

$$V_r := A^\top U_r\Sigma_r^{-1}.$$

This corresponds to a matrix with columns $v_1,\dots,v_r$ given by $v_k = \frac{1}{\sigma_k}A^\top u_k$. To conclude we need to show that this construction indeed gives a compact SVD. This requires us to show two statements:

(1) $V_r^\top V_r = I$. This can be verified by direct computation, while recalling that $AA^\top = U_r\Lambda_r U_r^\top$:

$$V_r^\top V_r = \left(A^\top U_r\Sigma_r^{-1}\right)^\top A^\top U_r\Sigma_r^{-1} = \Sigma_r^{-1}U_r^\top AA^\top U_r\Sigma_r^{-1} = \Sigma_r^{-1}U_r^\top U_r\Lambda_r U_r^\top U_r\Sigma_r^{-1} = \Sigma_r^{-1}\Lambda_r\Sigma_r^{-1} = I$$

(2) $A = U_r\Sigma_r V_r^\top$. Note that

$$U_r\Sigma_r V_r^\top = U_r\Sigma_r\left(A^\top U_r\Sigma_r^{-1}\right)^\top = U_r U_r^\top A.$$

Let us simply verify that $A = U_r U_r^T A$.

- Let $x \in N(A)$. Then $Ax = 0 = U_r U_r^T 0 = U_r U_r^T Ax$.
- Let $x \in C(A^T)$. It follows that $x = A^T y$ for $y \in \mathbb{R}^m$ and hence,

$$Ax = AA^T y = U_r\Lambda_r U_r^T y = U_r I\Lambda_r U_r^T y = U_r U_r^T U_r\Lambda_r U_r^T y = U_r U_r^T AA^T y = U_r U_r^T Ax.$$

Hence, for all $x \in \mathbb{R}^n$ we have verified that $Ax = U_r U_r^T Ax$. Then $A = U_r U_r^T A$ follows.

□

An important direct consequence of the SVD, and in particular of (40) is that we can write any rank-r matrix $A \in \mathbb{R}^{m\times n}$ as a sum of r rank-1 matrices:

Proposition 9.3.4. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix with rank r. Let $\sigma_1,\dots,\sigma_r$ be the non-zero singular values of A, $u_1,\dots,u_r$ the corresponding left singular vectors and $v_1,\dots,v_r$ the corresponding right singular vectors. Then*

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^\top. \tag{42}$$

The SVD is a powerful tool. Many of the results that we derived in this course become significantly simpler to be shown with the SVD. Now that you have the SVD, you might want to read these notes again and try to interpret the results we derived in terms of the SVD. For example, the Moore-Penrose Pseudoinverse has a very simple description of the SVD, it corresponds to swapping U and V and replacing the non-zero singular values by their inverses, while keeping the zero ones zero. Try to derive this!

10. Beyond the topics we discussed

Linear Algebra is absolutely fundamental as a tool to understand mathematical questions of all kinds. It comes as no surprise that we could only touch some of the most important notions and tools in this field. Let us mention below few topics and conjectures for interested readers of the course. This material is not relevant for the exam.

So far, the norm of a vector $x \in \mathbb{R}^n$ was simply given by $\|x\| = \sqrt{x^\top x}$ but there are instances where it makes sense to measure the "lenght" of vectors in other ways. One popular way is called the "Manhattan distance" since when traveling in Manhattan one cannot take advantage of Pythagoras Theorem because that would involve cutting through buildings, that norm is given by $\|x\|_1 = \sum_{i=1}^n |x_i|$. In general, for $1 \leq p \leq \infty$ the ℓ_p norm is given by

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \tag{43}$$

for $p < \infty$, and $\|x\|_\infty = \max_i |x_i|$. Notice that $\|\cdot\|_2$ corresponds to the Euclidean norm that we have used in this course. The ℓ_1 norm is notable for promoting sparsity when one attempts to minimize it to solve underdetermined linear systems. This is the key idea behind "Compressed Sensing", and plays a crucial role in many imaging/sensing technologies. You can read more about it in Section 12 of [BM23] or Chapter 10 of [BSS] and references therein. The following relation can be shown.

Proposition 10.0.1. *For all $x \in \mathbb{R}^n$, we have $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.*

In several situations one also needs to "measure" the size of matrices (for example, when talking about a matrix being close to another one, we need a notion of distance, or norm of the difference).

Definition 10.0.2 (Two matrix norms)**.** *Given a matrix $A \in \mathbb{R}^{m\times n}$ we define two matrix norms:*

- *$\|A\|_F$, known as the Frobenius norm, is defined as*
$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2},$$
- *$\|A\|_{op}$, known as operator or specral norm, is defined as*
$$\|A\|_{op} = \max_{\substack{x\in\mathbb{R}^n \\ s.t. \|x\|=1}} \|Ax\|.$$

One can relate the two matrix norms as follows.

Proposition 10.0.3. *Given $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$. We have*

(1) $\|A\|_F^2 = \mathrm{Tr}\left(A^T A\right)$

(2) $\|A\|_F^2 = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2$

(3) $\|A\|_{op} = \sigma_1$

(4) $\|A\|_{op} \leq \|A\|_F \leq \sqrt{\min\{m,n\}}\|A\|_{op}$.

In the course we have covered the notion of eigenvalues and eigenvectors. There are a few fascinating open questions we can state. These are questions (or conjectures), that we currently do now know the answer to (or that we are not sure they are true).

The first author of this manuscript has a list of 42 open problems in some lecture notes [Ban16]: some have been solved in the meantime, but many remain open.

A few open problems for which you have all the necessary background to understand are listed below

Open Problem 10.0.4 (Hadamard conjecture). *For any n multiple of* 4 *there exists an Hadamard matrix H that is $n \times n$.*

A Hadamard matrix $H \in \mathbb{R}^{n \times n}$ is a matrix with only entries 1 *or* -1 *that is a multiple of an orthogonal matrix. In other words $H_{i,j} = \pm 1$ for all i, j and $H^\top H = nI$. Yet in other words: the columns of H are an orthogonal basis for $\mathbb{R}^n$ formed with only vectors with entries ± 1.*

Open Problem 10.0.5 (Mutually Unbiased Bases). *See Open Problem 6.2 in* [Ban16].

Open Problem 10.0.6 (Zauner's Conjecture). *See Open Problem 6.3 in* [Ban16].

Open Problem 10.0.7 (Komlos Conjecture). *See Open Problem 0.1 in* [Ban16].

Open Problem 10.0.8 (Matrix Spencer Conjecture). *See Open Problem 4.3 in* [Ban16].

Open Problem 10.0.9 (Rank of the Matrix Multiplication Tensor). *What is the rank of the Matrix Multiplication Tensor corresponding to multiplication of* 3×3 *matrices.*

A $d_1 \times d_2 \times d_3$ tensor T is what we can think of as a cubic matrix. It has $d_1 d_2 d_3$ entries given by T_{ijl}. We say T has rank r if r is the smallest integer such that we can write

$$T = \sum_{k=1}^{r} a_k \otimes b_k \otimes c_k,$$

for $a_k \in \mathbb{R}^{d_1}, b_k \in \mathbb{R}^{d_2}, c_k \in \mathbb{R}^{d_3}$, for $k = 1, \ldots, r$. In other words

$$T_{ijl} = \sum_{k=1}^{r} (a_k)_i (b_k)_j (c_k)_l.$$

Recall Proposition 9.3.4 to see why for matrices this corresponds to the notion of rank we have been using. While computing the rank of a matrix is computationally easy, doing so for tensors is notoriously difficult (because they lack a spectral theory of eigenvalues and eigenvectors).

There is a way to think of Strassen's algorithm (that you saw in a CS Lens in Part I of the course) in terms of a decomposition of a certain Tensor in terms of rank-1 tensors. In this description we focus on $n \times n$ matrices, but the same thing can be done for rectangular matrices.

The $n \times n$ matrix multiplication tensor is a $n^2 \times n^2 \times n^2$ tensor, where each dimension is indexed by pairs $(i_1, i_2), (j_1, j_2), (l_1, l_2)$ and T is given by

$$T_{(i_1,i_2),(j_1,j_2),(l_1,l_2)} = \begin{cases} 1 & \text{if} \quad i_1 = j_1, j_2 = l_1, l_2 = i_2 \\ 0 & \text{o.w.} \end{cases}.$$

Strassen's algorithm can be viewed as the fact that the rank of the 2×2 *matrix multiplication tensor (a* $4 \times 4 \times 4$ *tensor) is* ≤ 7*. The rank of the* 3×3 *matrix multiplication tensor (a* $9 \times 9 \times 9$ *tensor) remains unknown.*

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.

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. A "Simple proof" of the Fundamental Theorem of Algebra

In this appendix we present a brief sketch of a (relatively) simple proof of the Fundamental Theorem of Algebra that we learned from Alessio Figalli.

Let $P(z)$ be a polynomial of degree n. Without loss of generality we can assume it is monic $P(z) = z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_0$. Suppose $P(z)$ has no zeros/roots. There is a $r \in \mathbb{R}$ large enough such that the infimum of $|P(z)|$ inside the close disc D_r of radius r centered at zero is smaller than that outside the disc D_r (because far from the origin the term z^n dominates and forces $|P(z)|$ to be large outside of D_r. Since D_r is a compact set and $|P(z)|$ is continuous it needs to attain its minimum at a point $z_0 \in D_r$. It requires some extra analysis/topology background to see that a continuous function defined on a compact (think closed and bounded) set needs to attain a minimum. Let us accept this fact and continue with the proof. Note that $P(z_0) \neq 0$. Write $Q(z) = P(z - z_0)$, it is also a polynomial of degree n, $Q(z) = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + z^n$. Notice that $\beta_0 = P(z_0)$. let k be the first coefficient of $Q(z)$ (not including β_0) that is nonzero (meaning that $\beta_k \neq 0$ but $\beta_i = 0$ for all $0 < i < k$. Then $Q(z) = \beta_0 + \beta_k z^k + \beta_{k+1} z^k + \cdots$. Take $\varepsilon > 0$ arbitrarily small and consider $Q\left(\varepsilon\left(-\frac{\beta_0}{\beta_k}\right)^{\frac{1}{k}}\right)$. It is not difficult to see that for ε small enough the higher order terms are negligible and the term $\beta_0 + \beta_k z^k$ has smaller modulus and so one can pick ε such that $\left| Q\left(\varepsilon \left(-\frac{\beta_0}{\beta_k} \right)^{\frac{1}{k}} \right) \right| < |Q(0)|$ which is a contradiction with the fact that $|P(z_0)|$ was minimum.

References

[Ban16] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. *Available online at:* `https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf`*. See also* `https://ocw.mit.edu/courses/18-s096-topics-in-mathematics-of-data-science-fall-2015/`, 2016.

[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. (Table of contents available at* `https://math.mit.edu/~gs/linearalgebra/ila6/indexila6.html`*).* Wellesley - Cambridge Press., sixth edition, 2023.