Linear Algebra

Lecture Notes
Part II

(401-0131-00L at ETH Zurich)
Fall 2023

Afonso S. Bandeira
ETH Zurich

Last update on November 9, 2023

"READ ME" FOR PART II

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

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

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

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

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

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

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

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

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

MISCELLANEOUS THOUGHTS

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

In many of my side comments (usually in blue) I do not include specific references, but you can use the technical terms I include to start a search online. If you would like specific references, tell me a bit more about your interests and I would be happy to point you to some references.

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

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

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

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

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

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

CONTENTS

—— Week 7 - Part II - 2023.11.08 ——

4. Orthogonality, Projections, and Least Squares

Guiding Question 1. If we have a system of linear equations that has no solution, how do we find the "solution" that has the smallest error? This question is central in countless applications[3].

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

4.2. Projections.

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

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

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

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

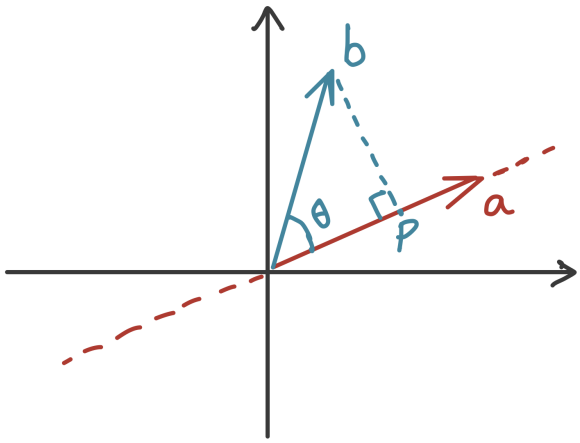

Figure 1. Projection on a line.

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

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

have $a^\top(b-p)=0$. Substituting gives

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

Indeed, we have the following Proposition.

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

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

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

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

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

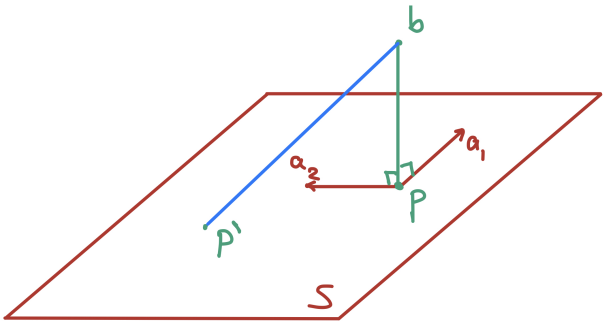

FIGURE 2. Projection on a subspace.

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

$\|p-p'\|^2+\|p-b\|^2$, which implies that $\|p'-b\|^2 \geq \|p-b\|^2$ (with equality holding only when $p=p'$).[4] [5]

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

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

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

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

We just proved the following Proposition.

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

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

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

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

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

Proof. We show this by showing that $A^\top A$ and A have the same nullspace. This is enough because, since $A^\top A$ is a square matrix it is invertible if and only if its nullspace only has the 0 vector, and A has linearly independent columns if and only if its nullspace only has the 0 vector.[6]

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

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

□

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

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

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

Corollary 4.2.5. *If A has linearly independent columns then $A^\top A$ is a square matrix, it is invertible and symmetric.*[7]

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

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

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

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

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

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

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

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

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

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

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

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

Proposition 4.2.106. *Let S be a subspace in $\mathbb{R}^m$ with a basis $a_1, \ldots, a_n$. For $v \in \mathbb{R}^m$, v being orthogonal to all vectors in S is equivalent to being orthogonal to a_k for all $1 \leq k \leq n$.*

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

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

—— Week 8 - 2023.11.10 & 2023.11.15 ——

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

$$Ax = b,$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In matrix-vector notation

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

where

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

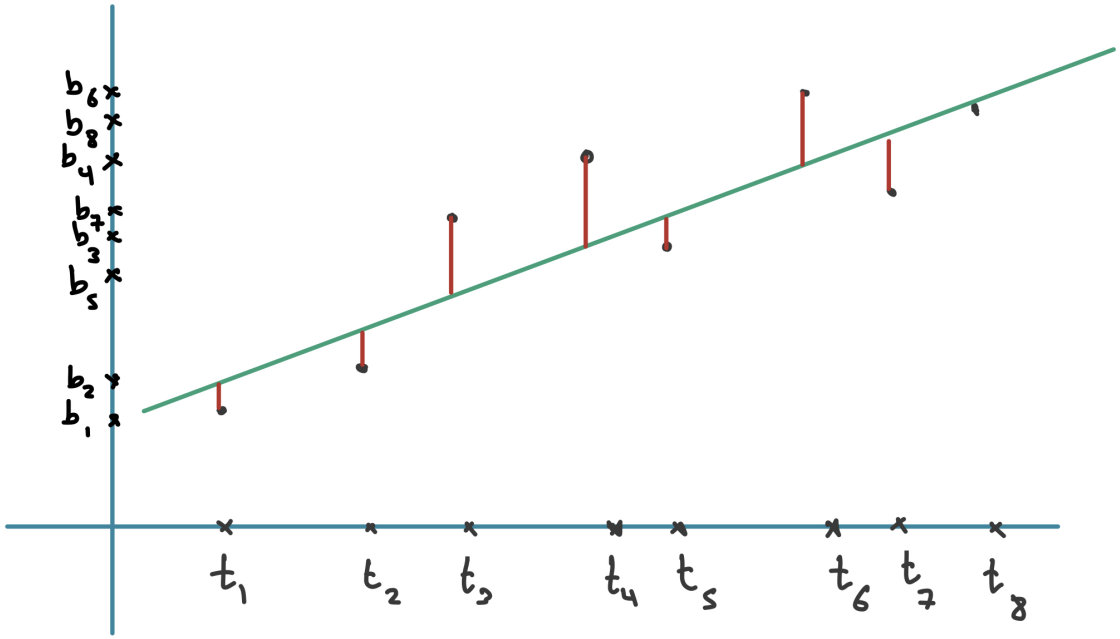

FIGURE 3. Fitting a line to points

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

Further Reading 14. A lot of Machine Learning includes Linear Regression as a key component. The idea is to create, find, or *learn* features of the data points. Given n data points $t_1, \ldots t_n$ (which now can be perhaps pixel images, rather than just timepoints) we might want to do classification (for example, in the case of images, maybe we want a function that is large when the picture has a dog in it and small when it has a cat in it). It is hard to imagine that this can be done with a

linear fit, but if we build good feature vectors $\varphi(t_k) \in \mathbb{R}^p$ for very large p then the function can depend on all coordinated 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, that time-permitting we might do a "CS lenses" on.

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

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

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

where δ_{ij} is the Kronecker delta

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

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

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

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

When Q is a square matrix then $Q^T Q = I$ implies also that $Q^T Q = I$ and so $Q^{-1} = Q^\top$. We call such matrices *orthogonal matrices*. This corresponds to the case when the q_i's are an orthonormal basis for all of $\mathbb{R}^n$.

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

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

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

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

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

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

is an orthogonal matrix.

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

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

Challenge 17. Show that indeed permutation matrices are orthogonal matrices

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

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

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

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

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

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

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

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

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

It is writing x as a linear combination of the orthonormal basis $\{q_i\}_{i=1}^n$ (as we will see later this is sometimes referred to as a change of basis*).*[8]

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

The idea is simple: If we have 2 linearly independent vectors a_1 and a_2 which span a subspace S, it is straightforward to 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:

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

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

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

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

note that the denominator is not zero because a_1 and a_2 are linearly independent.

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

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

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

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

Proof. [9] We prove this by induction.[10] Let S_k be the subspace spanned by $a_1, \ldots, a_k$. Then $S = S_n$. We will show, by induction, that $q_1, \ldots, q_k$ are an orthonormal basis for S_k. It is enough to show that they are orthonormal and are in S_k since orthonormality implies linearly independence and S_k has dimension k.

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

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

- 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 $\|q_k\| = 1$.
- Let $1 \le i \le k-1$. Since $q_1, \ldots, q_{k-1}$ are orthonormal, we have
  $$q_j^\top \left(a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_i\right) = q_j^\top a_k - \sum_{i=1}^{k-1}(a_k^\top q_i)q_j^\top q_i = q_j^\top a_k - (a_k^\top q_j) = 0,$$
  and $q_j^\top q_k = \frac{1}{\|q'_k\|} q_j^\top q'_k = 0$.

□

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

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

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

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

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

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

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

$$A = QR,$$

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

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

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

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

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

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

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

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

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

—— Week 9 - 2023.11.17 & 2023.11.22 ——

Appendix A. Some Important Preliminaries and Remarks on Notation

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

Some further important preliminaries and/or remarks:

(1) The dot product $x \cdot y$ between two vectors is sometimes also called inner product and written as $\langle x, y \rangle$.
(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, and it has an $r \times r$ identity as a submatrix, corresponding to the locations of the first r linearly independent columns of A.

Appendix B. Weekly Schedule

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

Predictions are in red and may be inaccurate.

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

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

(7) 8.11.2023: Introductions; 4.2.
(8) 4.3. Least Squares, 4.3. Fitting a line, 4.4
(9) 4.5. Start Chapter/Section 5 - Determinants
(10) Finish 5. 6.1
(11) 6.2, 6.3. and start Complex Numbers (and 6.4). We will skip 6.5 in [Str23]
(12) continue 6.4 SVD (7.1.), Applications (7.2.)
(13) (maybe) PCA (7.3.), and Linear Transformations (depending on pace)

(14) 22.12.2023: We will have an entire lecture of “CS lenses”, a sort of technical “Ask Me Anything” session that won’t cover core material of the course. If you have any particular topic you would like to me cover, let me know!

References

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