Linear Algebra
Week 8

G-07

14 XI 2024

1 Projections

We want to project vectors onto subspaces. It might be intuitive in 1 dimensional case, or 2 dimensional case if you paint a mental picture. However we need a more general way to describe the projection of the vector $\mathbf{b} \in \mathbb{R}^m$ onto a subspace $S \subseteq \mathbb{R}^m$ as follows:

$$proj_S(\mathbf{b}) = \arg\min_{p \in S} \|b - p\|$$

A lot to unpack here:

- $\arg\min_{p \in S}$ means we want the **argument p** that minimizes the expression following arg min and not the minimum value itself. The result is therefore the vector **p**.
- In words, the expression $proj_S(\mathbf{b})$ is the vector in S that is closest to b, which matches our intuition of projection.

1.1 Projecting onto a line

Before we move to the more complicated case, let's assume our subset S, onto which we want to project our vector **b**, has dimension 1. This means we are projecting onto a line. The figure below visualizes this. Let's have the formula first and then look at ways to derive it.

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

$$proj_S(\mathbf{b}) = \frac{aa^\top}{a^\top a} b \tag{1}$$

You can prove this algebraically as it is done in the lecture notes, where you treat the expression $\|b - p\|^2$ as a convex, quadratic, differentiable function in one variable λ where $p = \lambda * a$. Then you should derive by λ and take everything else like a constant. This is the formal definition and there is nothing to add to the notes in lecture document. Yet there is another way to derive these equations if you assume the error vector e is orthogonal to a. **Be careful, in the lecture you don't assume $e \perp a$ and later prove this intuition on page 11. In the following we nevertheless make this assumption. So take this as intuition and not the actual proof.**

How to see the (1) geometrically

First assume $a \perp e$. This means we have $a \perp (b-p)$ since we define our error vector $e = b - p$. We also know that we can write $p = \lambda a$ since we know the projection vector is on the line spanned by a and hence p is a scalar multiple of a.

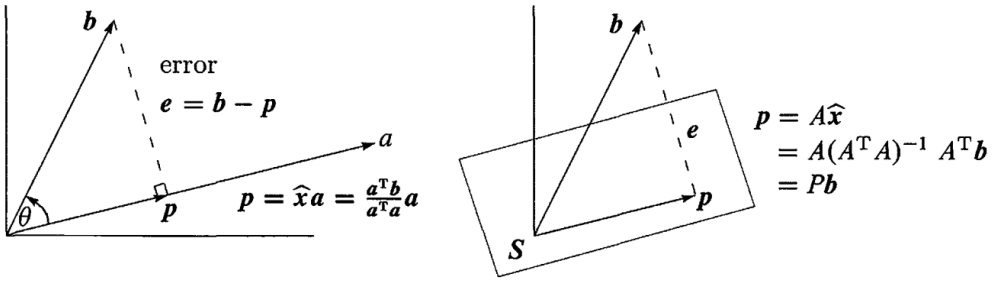
$$\begin{aligned}
a \perp (b-p) &\iff a^\top (b-p) = 0 \\
&\iff a^\top (b - \lambda a) = 0 \\
&\iff a^\top b - a^\top \lambda a = 0 \\
&\iff a^\top b = a^\top \lambda a \\
&\iff a^\top b = \lambda a^\top a \\
&\iff \lambda = \frac{a^\top b}{a^\top a}
\end{aligned}$$

Where we first used $v \perp u \iff v^\top u = 0$, then plugged in λa for p, then used the distributivity of the vector multiplication. In the step before last we used the fact that scalars commute in the expression as they want. Then we divided by $a^\top a$, which we can do since $a^\top a$ is a nonzero real number, as the squared norm of a nonzero vector a.

Now that we have the λ we can plug it into $p = \lambda a$ to get the projection vector:

$$p = \frac{a^\top b}{a^\top a} a = \frac{aa^\top}{a^\top a} b$$

Seeing the equality might be confusing. We can actually pull this off because $a^\top b$ is a scalar, it is a real number. So it makes no difference if we have $(a^\top b)a$ or $a(a^\top b)$. Then we use associativity to get from $a(a^\top b)$ to $(aa^\top)b$.



Gil. Strang (2009), Introduction to Linear Algebra, 4th Edition, Page 208

1.2 Projections in the general case

Now we want to project onto any subspace and not only onto a line. The main idea is the same instead now we have more than one vector to project onto. Let $S \subseteq \mathbb{R}^m$ be a subspace of $\mathbb{R}^m$. Additionally $S = span(a_1, a_2, \ldots, a_n)$, where the vectors $a_1, a_2, \ldots, a_n \in \mathbb{R}^m$ are a basis of S. We can pack the basis vectors as the columns of an $m \times n$ matrix $A = [a_1, a_2, \ldots, a_n]$. So we have $S = \{Ax \mid x \in \mathbb{R}^n\}$. In words, this is all linear combinations of the columns of the matrix A, which is the definition of the column space of A. Long story short $S = \mathbf{C}(A)$ and we project onto the column space of A.

Lemma 5.2.3. The projection of a vector $b \in \mathbb{R}^m$ to the subspace $S = \mathbf{C}(A)$ can be written as $proj_S(b) = A\hat{x}$, where $\hat{x}$ satisfies **the normal equations**

$$A^\top A\hat{x} = A^\top b \tag{2}$$

Intuition about (2): -formal proof is on page 12 of the lecture notes-

- We actually still want $proj_S(\mathbf{b}) = \underset{p \in S}{\arg\min} \|b - p\|$. But now $p \in S$ means that $p \in \mathbf{C}(A)$. This is just a fancy way of saying that we can write p as the matrix A multiplied by a vector, in this case: $p = A\hat{x}$.

- In the first part of the proof in lecture notes you show that there can not be more than one projections of b on $\mathbf{C}(A)$ by assuming there are and reaching to the contradiction that the error $\|p' - b\|^2$ is greater than $\|p - b\|^2$. We want the vector p with minimum error, so p must be the unique projection vector.

- In one dimensional case we had $e = (b - proj_S(b)) \perp a$. Now we have this condition for all columns of A. If the error is orthogonal to the basis of the subspace than it is orthogonal to all vectors in the subspace S. So for all $i \in [n]$ we have

$$a_i \perp (b - proj_S(b)) \iff a_i^\top (b - proj_S(b)) = 0$$

- An equation equivalent to this is where we pack a_i's in the matrix A is as follows: -remember $proj_S(b) = p = A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$ -

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

A has a basis as its columns, this is how we constructed it. So its columns are linearly independent. Now since we know $A^\top A$ is square, symmetric, and invertible when A has independent columns (**Lemma 5.2.4.** + **Corollary 5.2.5.**) we can solve the normal equations (2) for $\hat{x}$ by inverting $A^\top A$ as follows: $\hat{x} = (A^\top A)^{-1} A^\top b$. **Caution:** This is not our projection vector p, we have $p = A\hat{x}$. Hence we have to multiply A with $\hat{x}$ to get p.

$$p = A(A^\top A)^{-1} A^\top b$$

is our projection. If you look closely, we have a formula for a matrix - let's call it P - that takes b to its projection p.

Theorem 5.2.6. 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

$$proj_S(b) = Pb$$

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

See that the projection matrix P is **symmetric**:

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

You use the fact that $(A^{-1})^\top = (A^\top)^{-1}$ for invertible A. Better write this one line proof in the exam to be safe before using this fact.

Remark 5.2.7:

- $P^2 = P$ projecting twice on the same subset shouldn't change anything.
- If P is the projection matrix for S than $(I-P)$ is the projection matrix for projection onto orthogonal complement of S $(I-P)b = b - Pb = e = proj_{S^\perp}(b)$
- $(I-P)^2 = (I-P)$

2 Least Squares Approximation

2.1 The Approximation

In the first half we tried to solve equations of type $Ax = b$ and calculated **x**. But what do we do when the equation doesn't have any solution? For example when we have too many equations and $A \in \mathbb{R}^{m\times n}$ with $m > n$? Then we choose the best possible x we can choose. We want Ax to be as close to b as it can be. In the lecture notes you have this expression:

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

This expression is equal to the minimal error between $A\hat{x}$ and b. This is not directly what we want. We want the $\hat{x}$ that minimizes this error. Keep it in mind that you are trying to find ***the best solution*** when you are dealing with a least squares problem. You want a solution $\hat{x}$ that you can plug into the equation $Ax = b$ and get away with the minimum error.

That being said, since you want the vector $A\hat{x}$ which is in $\mathbf{C}(A)$ to be as close as possible to the vector b, $A\hat{x}$ is exactly the projection of b on $\mathbf{C}(A)$. This implies:

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

Which brings us back to the *normal equations*:

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

We know that for any matrix A we have $\mathbf{C}(A^\top A) = \mathbf{C}(A^\top)$ by Lemma 5.1.11 (see previous week). This tells us the normal equations always have a solution. We also know that if A has independent columns we can invert $A^\top A$ and solve the normal equations for $\hat{x}$:

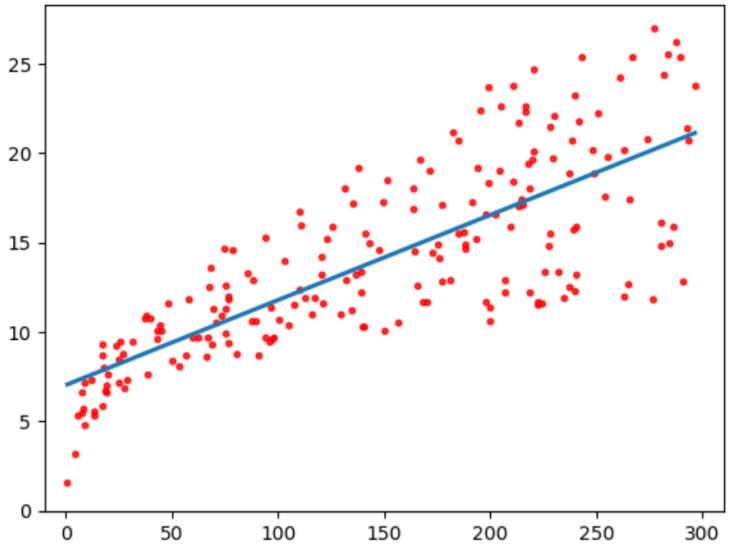
Fact 5.3.1. *("Least Squares Solution")* A minimizer of (3) is also a solution of the normal equations. When A has independent columns the unique minimizer $\hat{x}$ of (3) is given by

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

Notice how similar this is to a projection matrix. But there is a difference: a projection matrix is $P = A(A^\top A)^{-1}A^\top$ whereas here we multiply b with $(A^\top A)^{-1}A^\top$, without the A on the left. This is because we want our best solution $\hat{x}$ that yields the projection of b when multiplied with A but not the projection itself.

2.2 Linear Regression

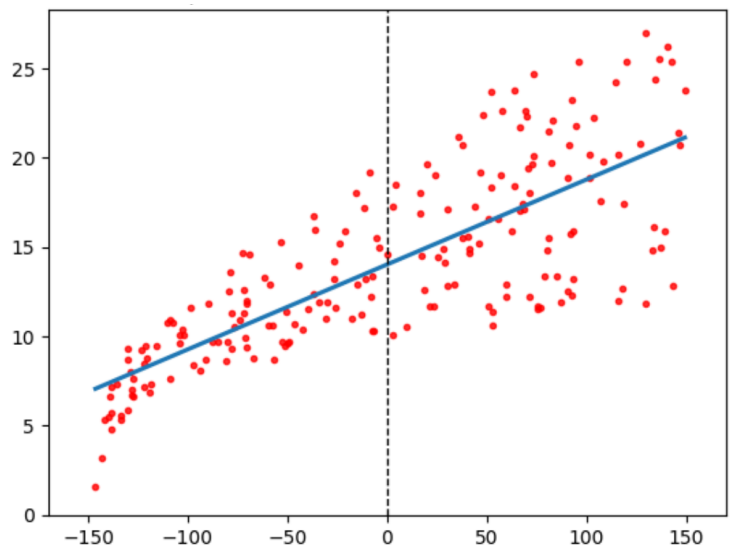
Linear regression can be seen as one of the babysteps of today's powerful AI. In this context you use least squares approximation to fit a function to a given set of data points. You have an example plot in the lecture notes. Here is another one:



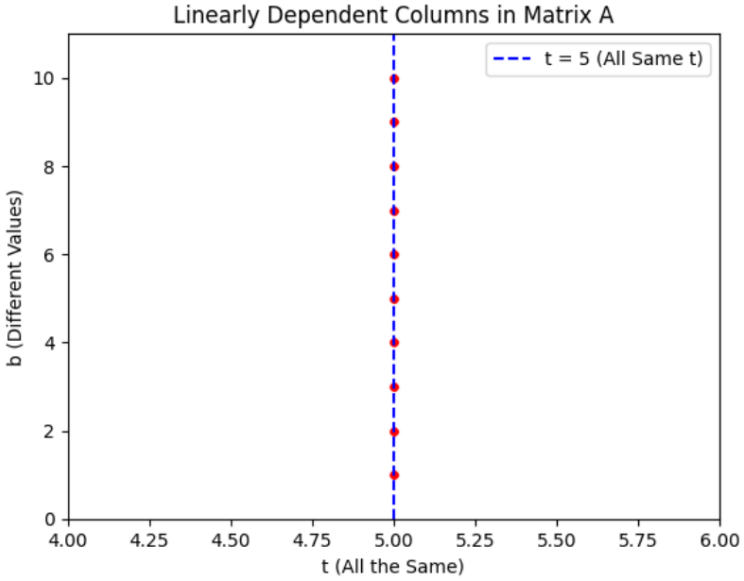
There is typically one task in the exam, where you have to use least squares and/or linear regression. You might also have to apply this to fit a parabola or any polynomial of some other degree (maybe not too large). The best way to get better at this is to apply the approximation by yourself. I have nothing to add to the notes on linear regression in the lecture notes. I definitely recommend reading them. For the matrices where you have scary summations as elements, just multiply A and b and also calculate $A^\top A$ you can get those results by direct computation.

Here I am going to provide two more plots to visualize two cases from the lecture notes:

If you do a change of variables as described in case of **Remark 5.3.3.** you get your data points equally distributed around 0. This is applied to the data points from the plot above. Don't get confused by the axes and actually ignore the axis on the left. You have to shift the fitted line to get the line in the original problem. See Assignment 8 Exercise 4.



And if just as in **Lemma 5.3.2.** the 2 columns of your A are linearly dependent and you have the same value for all t_i's this visualizes to the following plot, which in practice doesn't make sense since this corresponds to the case where we get different results for multiple measurements at the same time point.



The source code can be found on my website under additional material for week 8.

3 Hints

1. Solved in class. It helps to write the equations stacked on each other.

2. Create a new matrix $A' := \Lambda^{\frac{1}{2}} A$ Now you have your usual normal equations with A'.

3. The formula for projection on a line is $\frac{vv^\top}{\|v\|^2}$ What happens to $\|v\|^2$ when v is a unit vector?

4. See above for a small spoiler for a). For b) you compute the scalar product and check if it's zero. You can solve c) by direct computation set α to the given new value and try to play around with $\|A'\alpha' - \mathbf{b}\|^2$ to get $\|A\alpha - \mathbf{b}\|^2$.

5. Try writing $\mathbf{0}$ as a linear combination of the vectors of two bases.

6. See Theorem 5.1.7. and Corollary 5.1.9. $\mathbf{x}$ can be decomposed into components from $\mathbf{C}(A^\top)$ and $\mathbf{N}(A)$. After proving existence you should also prove uniqueness.

mkilic

♠