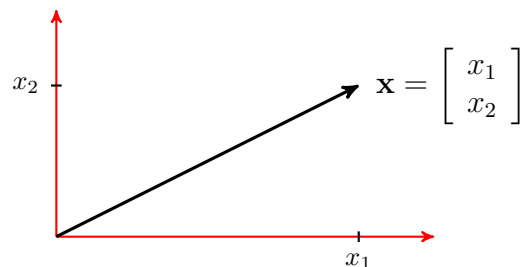


Lecture 15

We now take a closer look at the geometry of vectors and its relationship to the fundamental subspaces of a matrix. The important concepts in this unit include the development of orthogonal bases for the column space of a matrix and their use in data fitting and the very important **least-squares problem**.

Vector Geometry

The length of a vector can be defined in the usual geometric sense in \mathbb{R}^2 or \mathbb{R}^3 .



From the Pythagorean Theorem we see that the length of the vector \mathbf{x} is given by $\sqrt{x_1^2 + x_2^2}$. We denote the length or **norm** of a vector by the following:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$$

The idea of a length of a vector generalizes naturally to vectors in \mathbb{R}^n despite the fact that we can't picture vectors nor lengths in such high dimensions. Regardless, we define for $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Notice that the sum of squares inside the square root can also be written as the dot product of \mathbf{x} with itself. So we have alternatively

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\mathbf{x}^T \mathbf{x}} \quad \text{or} \quad \|\mathbf{x}\|^2 = \mathbf{x}^T \mathbf{x}$$

In many applications it's useful to define a vector \mathbf{u} that points in the same direction as a vector \mathbf{v} but has length one. Clearly we can obtain a vector such as this by scaling the vector \mathbf{v} by the reciprocal of it's length, i.e.

$$\mathbf{u} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

We call such a vector a **unit vector**.

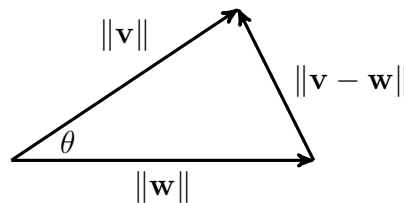
We showed in class previously that two vectors are perpendicular or **orthogonal** if their dot products are zero. We say \mathbf{u} and \mathbf{v} are orthogonal, or $\mathbf{u} \perp \mathbf{v}$ if

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = 0$$

It turns out that the dot product of two vectors can tell us something about their relationship even if they're not orthogonal. In general the dot product of two vectors is equal to the product of the norms of the vectors and the cosine of the angle between them. ””

Fact: $\mathbf{u} \cdot \mathbf{w} = \|\mathbf{u}\| \|\mathbf{w}\| \cos \theta$

Geometric Proof: Consider vectors \mathbf{v} and \mathbf{w} in \mathbb{R}^2 and orient them tail-to-tail along with the difference vector $\mathbf{v} - \mathbf{w}$.



Then, from the Law of Cosines we have

$$\|\mathbf{v} - \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\|\mathbf{v}\|\|\mathbf{w}\| \cos \theta$$

But from the relationship between the norm and the dot product we have

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|^2 &= (\mathbf{v} - \mathbf{w})^T (\mathbf{v} - \mathbf{w}) \\ &= \mathbf{v}^T \mathbf{v} - 2\mathbf{v}^T \mathbf{w} + \mathbf{w}^T \mathbf{w} \\ &= \|\mathbf{v}\|^2 - 2\mathbf{v}^T \mathbf{w} + \|\mathbf{w}\|^2 \end{aligned}$$

Setting the right-hand side of these two identities equal to each other we have

$$\|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\|\mathbf{v}\|\|\mathbf{w}\|\cos\theta = \|\mathbf{v}\|^2 - 2\mathbf{v}^T\mathbf{w} + \|\mathbf{w}\|^2$$

and canceling like terms on either side gives

$$\|\mathbf{v}\|\|\mathbf{w}\|\cos\theta = \mathbf{v}^T\mathbf{w}.$$

□

The cosine formula tells us several things about how the dot product of two vectors relates to their orientation to one another.

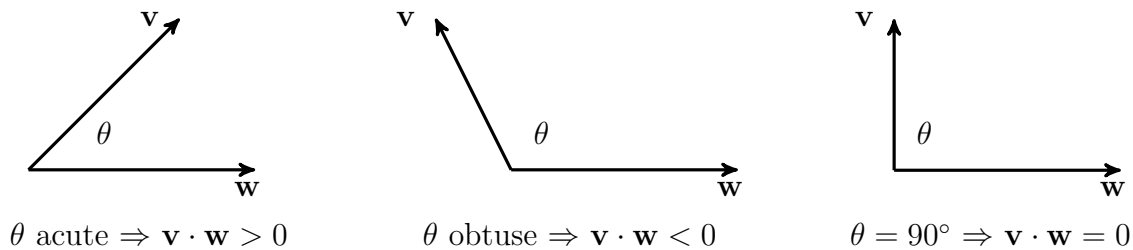
First, if we fix the lengths of \mathbf{v} and \mathbf{w} and consider how the dot product changes as the angle between them changes, we see that the dot product takes on its maximum positive value when the vectors are parallel and point in the same direction. In this case we have $\theta = 0$ and

$$\mathbf{v}^T \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos(0) = \|\mathbf{v}\| \|\mathbf{w}\|$$

Second, the dot product of vectors \mathbf{v} and \mathbf{w} takes on its largest negative value when the vectors are parallel but pointed in opposite directions. In this case we have $\theta = \pi$ and

$$\mathbf{v}^T \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos(\pi) = -\|\mathbf{v}\| \|\mathbf{w}\|$$

Finally, we see that if the angle takes on any other value in between the dot product will be of smaller magnitude. In general the dot product will be positive if the vectors point in generally the same direction (the angle less than perpendicular) and will be negative if the vectors point in generally differing directions (the angle is passed perpendicular).



Rule of Thumb: The dot product of two vectors tell us generally how much two vectors align with one another.

From this definition of the dot product we have two very import inequalities:

Schwarz's Inequality: $|\mathbf{v} \cdot \mathbf{w}| \leq \|\mathbf{v}\| \|\mathbf{w}\|$

Proof: The proof follows immediately from the cosine formula for dot products and the fact that the cosine function is always between minus one and one. Taking the absolute value of the dot product gives

$$|\mathbf{v} \cdot \mathbf{w}| = \|\mathbf{v}\| \|\mathbf{w}\| |\cos \theta| \leq \|\mathbf{v}\| \|\mathbf{w}\|$$

□

The Triangle Inequality: $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$

In class we did a handwavy geometric proof of the Triangle Inequality. Here we give an algebraic proof that relies on Schwarz's Inequality.

Proof: Note that since all of the terms in the inequality are positive it is still true if we square both sides. So instead we'll prove the equivalent statement

$$\|\mathbf{v} + \mathbf{w}\|^2 \leq (\|\mathbf{v}\| + \|\mathbf{w}\|)^2$$

Starting with the left-hand side, we have

$$\begin{aligned}\|\mathbf{v} + \mathbf{w}\|^2 &= (\mathbf{v} + \mathbf{w})^T (\mathbf{v} + \mathbf{w}) \\ &= \|\mathbf{v}\|^2 + 2\mathbf{v}^T \mathbf{w} + \|\mathbf{w}\|^2\end{aligned}$$

Now note that it's possible the dot product of \mathbf{v} and \mathbf{w} is negative. So if we throw absolute values around it it can only make the expression on the right-hand side of the equation larger. Thus we have

$$\begin{aligned}\|\mathbf{v} + \mathbf{w}\|^2 &= (\mathbf{v} + \mathbf{w})^T (\mathbf{v} + \mathbf{w}) \\ &= \|\mathbf{v}\|^2 + 2\mathbf{v}^T \mathbf{w} + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2|\mathbf{v}^T \mathbf{w}| + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2\|\mathbf{v}\|\|\mathbf{w}\| + \|\mathbf{w}\|^2 \\ &= (\|\mathbf{v}\| + \|\mathbf{w}\|)^2\end{aligned}$$

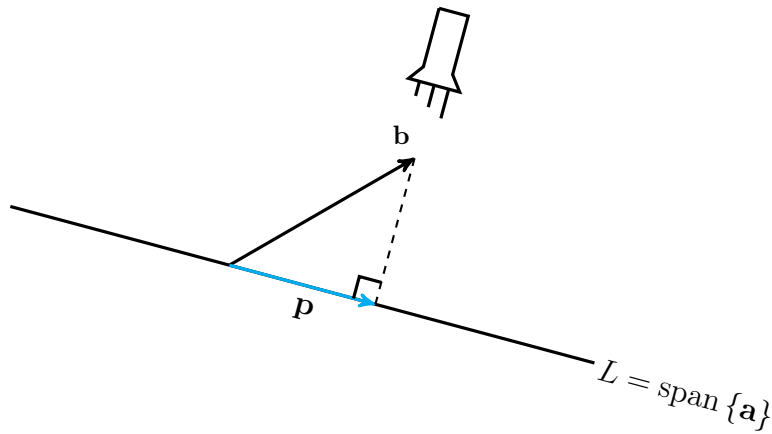
where the second-to-last line follows from Schwartz's Inequality and the last line comes from factoring.

□

Projections

As we start to talk about the least-squares problem, the notion of the *closest vector to a subspace* will become immensely important. Suppose that we have a vector \mathbf{b} and a one-dimensional subspace of \mathbb{R}^3 (a.k.a. a line) spanned by a single vector \mathbf{a} . If the vector \mathbf{b} does not lie along the line then we might ask What is the vector that lies on the line that is in some sense *closest* to the vector \mathbf{b} ?

This closest vector is called the projection of \mathbf{b} onto the subspace representing the line and is denoted as the vector \mathbf{p} . You can think of the projection vector \mathbf{p} as the shadow that the vector \mathbf{b} would make on the line if you shined a flashlight on the vector \mathbf{b} and perpendicular to the line.

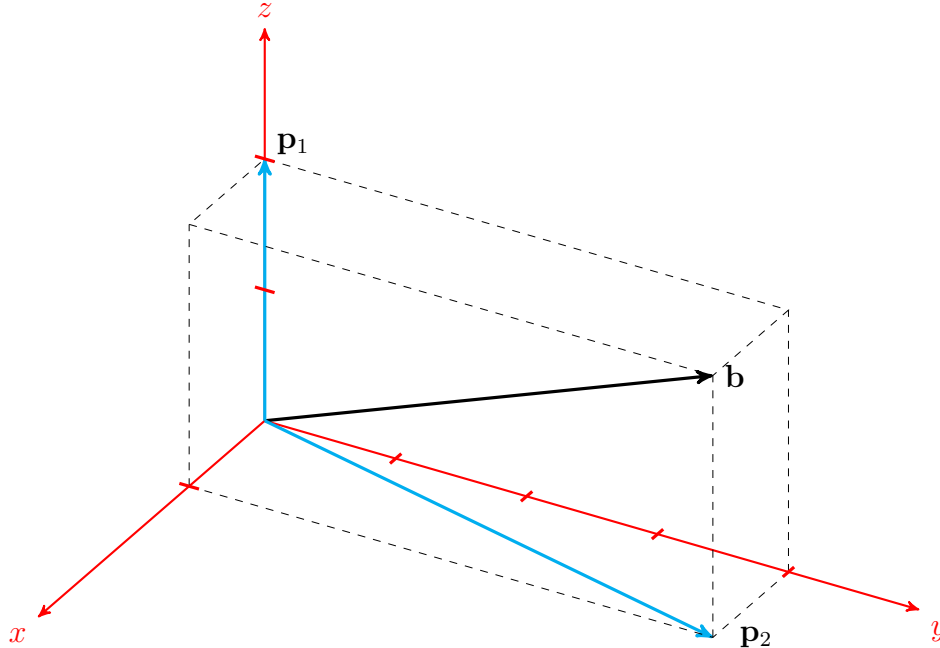


Note that the projection vector \mathbf{p} lies **in** the subspace of the line and points in the same general direction as \mathbf{b} .

We can also talk about projecting vectors into higher dimensional subspaces as well, such as planes. Consider the following example.

Example 1: Find the projection of the vector $\mathbf{b} = [1, 4, 2]^T$ onto (1) the z -axis and (2) the xy -plane.

Using the flashlight and shadow analogy it's easy to see that the projection of \mathbf{b} onto the z -axis is the vector $\mathbf{p}_1 = [0, 0, 2]^T$ and the projection of \mathbf{b} onto the xy -plane is the vector $\mathbf{p}_2 = [1, 4, 0]^T$.



Note that in each case the projection vector \mathbf{p} lies **in** the subspace that \mathbf{b} was projected onto.

It is often useful to think of the projection operation as the operation of a matrix on the vector \mathbf{b} . In general, we want to write

$$\mathbf{p} = P\mathbf{b}$$

where P is the so-called **projection matrix**. For this example, we'd like to find matrices P_1 and P_2 such that

$$P_1 \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \quad \text{and} \quad P_2 \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$$

For this simple example it's very easy to see that the projection matrices that project onto the z -axis and xy -plane, respectively, are

$$P_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Note that because the two subspaces we've projected onto are orthogonal subspaces, we have the nice result that

$$\mathbf{p}_1 + \mathbf{p}_2 = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \mathbf{b} \quad \text{and} \quad P_1 + P_2 = I$$

□

Referring to the subspaces projected onto as axes and coordinate planes is pretty easy in three dimensions, but with more complicated and higher dimensional subspaces we need a more systematic method to describe the space. Notice that in both of the cases in the previous example, instead of talking about projecting onto a geometric space, we can talk about projecting onto the column space of a matrix. For example, projecting onto the z -axis is equivalent to projecting onto the column space of

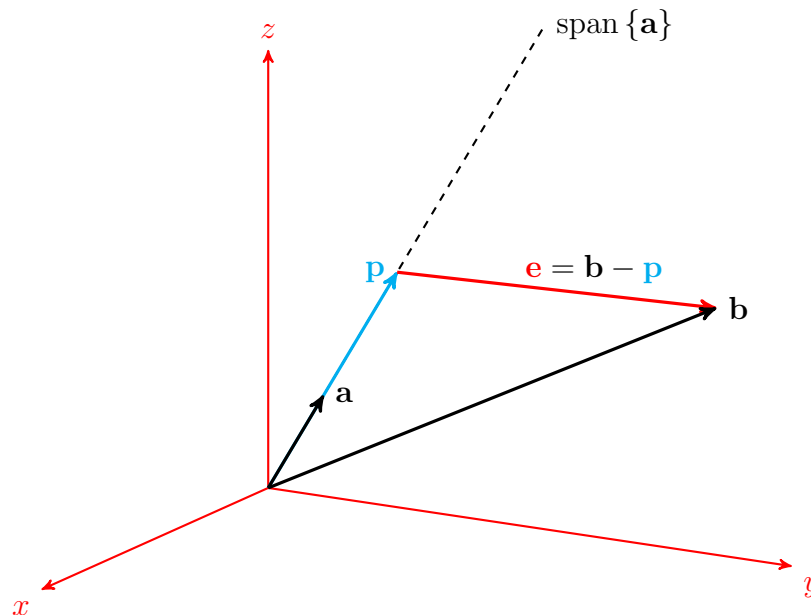
$$A_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{or} \quad A_1 = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

Similarly projecting onto the xy -plane is equivalent to projecting onto the column space of

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{or} \quad A_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{bmatrix}.$$

Projection onto a Line

Now we'll look at the process of projecting onto a line (a one-dimensional subspace) in greater detail. Suppose we want to find the projection \mathbf{p} of the vector \mathbf{b} onto the subspace spanned by the vector \mathbf{a} . The general picture looks as follows



Notice that the vector $\mathbf{e} = \mathbf{b} - \mathbf{p}$ is the difference between the original vector \mathbf{b} and the projection vector \mathbf{p} . Notice also that if the projection vector \mathbf{p} is really the vector on the line that is closest to \mathbf{b} then the difference vector \mathbf{e} will be orthogonal to \mathbf{p} . We label the difference vector \mathbf{e} because it can be thought of as the **error** made if we approximate the vector \mathbf{b} by its projection \mathbf{p} .

OK, so how do we compute the projection vector \mathbf{p} given the original vector \mathbf{b} and the vector \mathbf{a} that describes the line?

Note that the projection \mathbf{p} is in the subspace $\text{span}\{\mathbf{a}\}$ which means it can be written as a scalar multiple of \mathbf{a} . Let the (as yet unknown) multiplier be \hat{x} , then we have

$$\mathbf{p} = \hat{x}\mathbf{a}$$

Remember that we said that \mathbf{p} is the vector in $\text{span}\{\mathbf{a}\}$ that makes the error vector $\mathbf{e} = \mathbf{b} - \mathbf{p}$ orthogonal to \mathbf{a} . We can enforce this condition by dotting \mathbf{e} by \mathbf{a} and setting it to zero. We then find the correct multiplier \hat{x} that makes the equation true.

$$\begin{aligned} 0 &= \mathbf{a}^T(\mathbf{b} - \hat{x}\mathbf{a}) \\ 0 &= \mathbf{a}^T\mathbf{b} - \hat{x}\mathbf{a}^T\mathbf{a} \\ \hat{x}\mathbf{a}^T\mathbf{a} &= \mathbf{a}^T\mathbf{b} \\ \hat{x} &= \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} \end{aligned}$$

So we have

$$\hat{x} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} \quad \text{and so} \quad \mathbf{p} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\mathbf{a}$$

Example 2: Compute the projection of $\mathbf{b} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$ onto the subspace spanned by $\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

We have $\mathbf{a}^T\mathbf{b} = 2$ and $\mathbf{a}^T\mathbf{a} = 1$ which gives $\mathbf{p} = \frac{2}{1}\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$

What if we had chosen another basis vector for the subspace, for instance $\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$?

Then $\mathbf{a}^T\mathbf{b} = 6$ and $\mathbf{a}^T\mathbf{a} = 9$ which gives $\mathbf{p} = \frac{6}{9}\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$

Hopefully this isn't surprising. We made the entries of \mathbf{a} larger, which in turn made $\mathbf{a}^T \mathbf{b}$ larger in the numerator, but also made $\mathbf{a}^T \mathbf{a}$ comparatively larger in the denominator to compensate. In general you can use any nonzero basis vector in the subspace for the line and the formula will work.

□

Example 3: Compute the projection of $\mathbf{b} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$ onto the subspace spanned by $\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

We have $\mathbf{a}^T \mathbf{b} = 7$ and $\mathbf{a}^T \mathbf{a} = 3$ which gives $\mathbf{p} = \frac{7}{3} \mathbf{a} = \begin{bmatrix} 7/3 \\ 7/3 \\ 7/3 \end{bmatrix}$

Let's look at some special cases and check the formula as well as our intuition.

b already in subspace: Suppose that \mathbf{b} is already in $\text{span}\{\mathbf{a}\}$. Then clearly the closest vector to \mathbf{b} in the subspace is \mathbf{b} itself, so the projection shouldn't change anything. If \mathbf{b} is in the subspace we have $\mathbf{b} = \alpha \mathbf{a}$ for some scalar α . Then

$$\mathbf{a}^T \mathbf{b} = \mathbf{a}^T (\alpha \mathbf{a}) = \alpha \mathbf{a}^T \mathbf{a} \quad \text{then} \quad \mathbf{p} = \frac{\alpha \mathbf{a}^T \mathbf{a}}{\mathbf{a}^T \mathbf{a}} \mathbf{a} = \alpha \mathbf{a} = \mathbf{b} \quad \checkmark$$

b orthogonal to subspace: Suppose that \mathbf{b} is orthogonal to $\text{span}\{\mathbf{a}\}$. Then \mathbf{b} is orthogonal to any vector in $\text{span}\{\mathbf{a}\}$ and in particular \mathbf{b} is orthogonal to \mathbf{a} . Then

$$\mathbf{p} = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \mathbf{a} = \frac{0}{\mathbf{a}^T \mathbf{a}} \mathbf{a} = \mathbf{0}$$

This should make intuitive sense. Suppose you projected a vector \mathbf{b} that lies in the xy -plane onto the z -axis. Since the vector \mathbf{b} does not have any height off the plane, its shadow on the z -axis would be nonexistent.

OK, so how do we come up with the projection matrix P such that $\mathbf{p} = P\mathbf{b}$?

Well, we've already done most of the work. Finding the P matrix just takes some creative rearranging of our formula. Switching the order of the multiplication in the formula (which is fine, because \hat{x} is just a scalar) gives

$$\mathbf{p} = \mathbf{a} \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \mathbf{b} = \left(\frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \right) \mathbf{b} = P\mathbf{b} \quad \text{so} \quad P = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}}$$

Note that the numerator in the expression for P is the **outer-product** of \mathbf{a} with itself, which is a matrix.

Example 4: Compute the projection matrix P for the space $\text{span}\{\mathbf{a}\}$ with $\mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

We have $\mathbf{a}^T \mathbf{a} = 1$ and

$$\mathbf{a}\mathbf{a}^T = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \checkmark$$

Example 5: Compute the projection matrix P for the space $\text{span}\{\mathbf{a}\}$ with $\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

We have $\mathbf{a}^T \mathbf{a} = 3$ and

$$\mathbf{a}\mathbf{a}^T = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \Rightarrow P = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

We can check that multiplication by this P gives us the same projection vector we found in Example 3:

$$P\mathbf{b} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 7 \\ 7 \\ 7 \end{bmatrix} = \begin{bmatrix} 7/3 \\ 7/3 \\ 7/3 \end{bmatrix}$$

Lecture 16

Last time we developed a general formula for projecting a vector onto a one-dimensional subspace (i.e. a line). Now we'll look at the general procedure for projecting a vector into higher dimensional subspaces. The projection into the xy -plane that we looked at in the previous lecture is an example of projecting a vector in \mathbb{R}^3 into a two-dimensional subspace. We'll consider general n -dimensional subspaces here, but keep the plane example in the back of your mind.

Suppose we have a basis for an n -dimensional subspace of \mathbb{R}^m and a vector $\mathbf{b} \in \mathbb{R}^m$ that we'd like to project. The projection will have potential contributions from each of the basis vectors, which we'll denote $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \in \mathbb{R}^m$.

Since we know the projection vector \mathbf{p} must lie in the subspace, it must be possible to write it as a linear combination of the basis vectors, i.e.

$$\mathbf{p} = \hat{x}_1 \mathbf{a}_1 + \hat{x}_2 \mathbf{a}_2 + \dots + \hat{x}_n \mathbf{a}_n$$

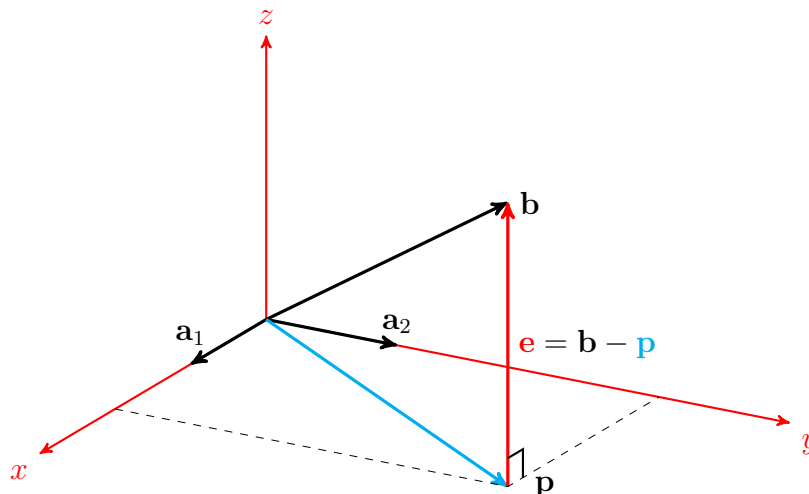
Notice that this linear combination of basis vectors kinda looks like the column-view of a matrix vector product. In fact, the whole process becomes much simpler if we organize the basis vectors into a matrix, $A = [\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_n]$, and the unknown \hat{x} 's into a vector $\hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n]^T$. We then can write the projection vector as

$$\mathbf{p} = A\hat{\mathbf{x}}$$

Note that not much has changed from the one-dimensional case. We still want to

1. Find the unknown vector of coefficients $\hat{\mathbf{x}}$
2. Compute the projection vector $\mathbf{p} = A\hat{\mathbf{x}}$
3. Find a projection matrix P such that $\mathbf{p} = P\mathbf{b}$

Let's go back to the xy -plane projection case to get a better picture of what's going on



The picture is very similar as before. We see that the projection vector \mathbf{p} is closest to \mathbf{b} exactly when the error vector $\mathbf{e} = \mathbf{b} - \mathbf{p} = \mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to the xy -plane. Our strategy will be to enforce the condition that \mathbf{e} is orthogonal to the subspace and then use the resulting equations to solve for the unknown \hat{x} 's.

Recall that to enforce that a vector is orthogonal to a subspace it is sufficient to enforce that the vector is orthogonal to each of the basis vectors for the subspace. In the general n -dimensional case we have the basis $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$. Enforcing that $\mathbf{e} = \mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to each of these gives the following equation

$$\begin{aligned} \mathbf{a}_1^T (\mathbf{b} - A\hat{\mathbf{x}}) &= 0 \\ \mathbf{a}_2^T (\mathbf{b} - A\hat{\mathbf{x}}) &= 0 \\ &\vdots \\ \mathbf{a}_n^T (\mathbf{b} - A\hat{\mathbf{x}}) &= 0 \end{aligned}$$

Now notice that each of the expressions is a dot product of a column of A with the error vector $(\mathbf{b} - A\hat{\mathbf{x}})$. Since the columns of A are the rows of A^T we can also say that each expression is the dot product of a row of A^T with the error vector $(\mathbf{b} - A\hat{\mathbf{x}})$. This is exactly the matrix-vector product of A^T with $\hat{\mathbf{x}}$. So an equivalent set of equations is

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

So the vector of coefficients $\hat{\mathbf{x}}$ that give us the projection vector is the vector $\hat{\mathbf{x}}$ that solves this system of equations. Rearranging the system we have

$$A^T (\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0} \quad \Rightarrow \quad A^T \mathbf{b} - A^T A \hat{\mathbf{x}} = \mathbf{0} \quad \Rightarrow \quad A^T A \hat{\mathbf{x}} = A^T \mathbf{b}$$

Take-Away: To find the vector of unknown coefficients that give the projection vector we can solve the system of equations $A^T A \hat{\mathbf{x}} = A^T \mathbf{b}$ for $\hat{\mathbf{x}}$. Then to find the projection vector we form $\mathbf{p} = A \hat{\mathbf{x}}$.

The system of equations $A^T A \hat{\mathbf{x}} = A^T \mathbf{b}$ are called the **normal equations** because they arise from enforcing the conditions that the error vector is orthogonal to the column space of A and another word for orthogonal is **normal**. Note that if A is an $m \times n$ matrix then the matrix $A^T A$ that must be solved when solving the normal equations is $n \times n$. It turns out that because the columns of A are a basis for the subspace (and hence linearly independent) that the matrix $A^T A$ is always nonsingular, and thus the normal equations always have a unique solution $\hat{\mathbf{x}}$ (see a proof of this at the end of this lecture).

OK, since $A^T A$ is nonsingular, its inverse $(A^T A)^{-1}$ exists, and we can write the projection vector \mathbf{p} as

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} \quad \Rightarrow \quad \mathbf{p} = A \hat{\mathbf{x}} \quad \Rightarrow \quad \mathbf{p} = A (A^T A)^{-1} A^T \mathbf{b}$$

From this expression it's easy to see that the form of the projection matrix is

$$P = A (A^T A)^{-1} A^T$$

It's instructive to check that the formula for the harder n -dimensional case reduces to the formula we worked out for projecting a vector onto a one-dimensional subspace. In that case we have just $A = \mathbf{a}$, and the formula gives

$$P = \mathbf{a} (\mathbf{a}^T \mathbf{a})^{-1} \mathbf{a}^T = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \quad \checkmark$$

Example 1: Find the projection matrix P that projects a vector in \mathbb{R}^3 into the xy -plane.

We need a matrix A whose columns form a basis for the xy -plane. We'll start with the most obvious choice:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \Rightarrow \quad A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

Plugging into the formula we have

$$P = A (A^T A)^{-1} A^T = A I^{-1} A^T = A A^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \checkmark$$

Alternatively we could use a more interesting basis for the xy -plane with $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}$.

$$A^T A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \Rightarrow (A^T A)^{-1} = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}$$

$$P = A(A^T A)^{-1}A^T = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Example 2: Find the projection of $\mathbf{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$ onto the column space of $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

Let's do this one in two ways. First we'll solve the normal equations for $\hat{\mathbf{x}}$ and for $\mathbf{p} = A\hat{\mathbf{x}}$. Then we'll build the projection matrix P and use it to compute the projection. We have

$$A^T A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \quad A^T \mathbf{b} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

Solving the augmented system for $\hat{\mathbf{x}}$ gives

$$\left[\begin{array}{cc|c} 3 & 3 & 6 \\ 3 & 5 & 0 \end{array} \right] \sim \left[\begin{array}{cc|c} 3 & 3 & 6 \\ 0 & 2 & -6 \end{array} \right] \sim \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & -3 \end{array} \right] \Rightarrow \hat{\mathbf{x}} = \begin{bmatrix} 2 - 1(-3) \\ -3 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

Then the projection vector is

$$\mathbf{p} = A\hat{\mathbf{x}} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ -3 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix}$$

Forming the projection matrix directly, we have

$$(A^T A)^{-1} = \frac{1}{6} \begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix} \Rightarrow P = A(A^T A)^{-1}A^T = \frac{1}{6} \begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$

And finally

$$\mathbf{p} = P\mathbf{b} = \frac{1}{6} \begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 30 \\ 12 \\ -6 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix} \quad \checkmark$$

Fun Facts about Projection Matrices

Fact: P is a projection matrix if and only if P is square and satisfies $P^2 = P$.

This should make some intuitive sense. Applying P^2 to a vector \mathbf{b} is the same as applying P to \mathbf{b} and then applying P again. The first application, $P\mathbf{b}$, projects \mathbf{b} into the subspace. The second application, $P(P\mathbf{b})$, does nothing because $P\mathbf{b}$ is already in the subspace that P projects onto. Thus for any vector \mathbf{b} , $P^2\mathbf{b} = P\mathbf{b}$, and so $P^2 = P$.

We can also prove this one algebraically from the formula

$$\begin{aligned} P^2 &= \left[A (A^T A)^{-1} A^T \right] \left[A (A^T A)^{-1} A^T \right] \\ &= A \textcolor{red}{(A^T A)^{-1} (A^T A)} (A^T A)^{-1} A^T \\ &= A I (A^T A)^{-1} A^T \\ &= A (A^T A)^{-1} A^T \\ &= P \end{aligned}$$

Fact: If P is a projection matrix then so is $(I - P)$.

We'll show this by showing that it satisfies $(I - P)^2 = I - P$.

$$\begin{aligned} (I - P)^2 &= (I - P)(I - P) \\ &= I - 2P + P^2 \\ &= I - 2P + P \quad (\text{b/c } P \text{ is a projection}) \\ &= I - P \quad \checkmark \end{aligned}$$

So what does the projection $(I - P)$ do?

Let's look at it for the projection matrix that projects onto the xy -plane:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad I - P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that $(I - P)$ is the projection matrix that projects onto the z -axis. In general, if P is a projection matrix that projects onto a given subspace, $(I - P)$ is the projection matrix that projects onto the orthogonal subspace.

OK, so how do we know that the matrix $A^T A$ that appears in the normal equations is always nonsingular? Recall that A had linearly independent columns because we assumed that the columns of A formed a basis for the subspace that we were projecting onto. In general, the following theorem is true

Theorem: $A^T A$ is nonsingular if and only if A has linearly independent columns.

Before we can prove the main theorem we have to prove a little Lemma.

Lemma: A has the same nullspace as $A^T A$.

Lemma Proof: To show that two subspaces are the same we have to prove two things: If \mathbf{x} is in $N(A)$ then \mathbf{x} is in $N(A^T A)$, and if \mathbf{x} is in $N(A^T A)$ then \mathbf{x} is in $N(A)$. This shows that $N(A)$ is contained in $N(A^T A)$ and conversely that $N(A^T A)$ is contained in $N(A)$. The only way that two spaces can each contain each other is if they're the same space.

$N(A)$ is in $N(A^T A)$: Suppose that $\mathbf{x} \in N(A)$. Then

$$A\mathbf{x} = \mathbf{0} \quad \Rightarrow \quad A^T A\mathbf{x} = A^T \mathbf{0} = \mathbf{0} \quad \Rightarrow \quad \mathbf{x} \text{ is in } N(A^T A)$$

$N(A^T A)$ is in $N(A)$: Suppose that $\mathbf{x} \in N(A^T A)$. Then

$$A^T A\mathbf{x} = \mathbf{0} \quad \Rightarrow \quad \mathbf{x}^T A^T A\mathbf{x} = \mathbf{x}^T \mathbf{0} = 0 \quad \Rightarrow \quad (A\mathbf{x})^T (A\mathbf{x}) = 0 \quad \Rightarrow \quad \|A\mathbf{x}\|^2 = 0$$

The only way that the vector $A\mathbf{x}$ can have length 0 is if it is the zero vector. So $A\mathbf{x} = \mathbf{0}$ and \mathbf{x} is in $N(A)$

□

Theorem Proof: The columns of A are linearly independent if and only if the nullspace of A contains only the zero vector. From the lemma we know this is true if and only if the nullspace of $A^T A$ contains only the zero vector. Finally, the nullspace of $A^T A$ contains only the zero vector if and only if $A^T A$ is nonsingular.

□