Math 310 Linear Algebra

# Projecting a Vector on the Column Space of a Matrix.

Suppose $A$ is an $m \times n$ matrix, and $\mathbf{b} \in \mathbb{R}^m$. If the linear system $A\mathbf{x} = \mathbf{b}$ is inconsistent then there is no exact solution. One way to think of the situation is to imagine looking at $A\mathbf{x}$ for every possible $\mathbf{x} \in \mathbb{R}^n$. Of course, none of those is actually $\mathbf{b}$, since the system is inconsistent. But we could try to choose an $\mathbf{x}$ for which $A\mathbf{x}$ is close as possible to $\mathbf{b}$. Here it is helpful to recognize that all the different $A\mathbf{x}$'s fill up the column space of $A$, which can also be thought of as the range space since it is the range of the function $\mathbf{x} \to A\mathbf{x}$. The inconsistency tells us that col $A$ is not all of $\mathbb{R}^m$, and in fact $\mathbf{b}$ is outside col $A$. The key idea is that the nearest element of col $A$ to $\mathbf{b}$ can be found by projecting $\mathbf{b}$ orthogonally onto col $A$. This handout will describe how this projection can be found, and show that it always exists and exists uniquely.

**Definition of Projection.** Conceptually, we want to decompose $\mathbf{b}$ into two vectors – one that is in col $A$ and one that is orthogonal to col $A$, meaning orthogonal to every vector in col $A$. The first is the part we call the projection, $\mathbf{p}$. The second part must then be $\mathbf{b} - \mathbf{p}$. These two vectors fill our requirements if and only if $\mathbf{p} \in$ col $A$ and $\mathbf{b} - \mathbf{p}$ is orthogonal to col $A$. This will be our definition for the projection.

**Definition:** If $A$ is an $m \times n$ matrix, and $\mathbf{b} \in \mathbb{R}^m$, then the vector $\mathbf{p} \in \mathbb{R}^m$ is called the projection of $\mathbf{b}$ on col $A$ if and only if the following hold:

i. $\mathbf{p} \in$ col $A$

ii. $\mathbf{b} - \mathbf{p}$ is orthogonal to col $A$.

Note that when $\mathbf{b} \in$ col $A$, we may take $\mathbf{p} = \mathbf{b}$ and both of the above conditions will be satisfied. Because we will show that $\mathbf{p}$ is uniquely determined, we can say that any $\mathbf{b} \in$ col $A$ is equal to its own projection.

**An Equation for p.** The projection $\mathbf{p}$ of $\mathbf{b}$ on col $A$ always exists, and is uniquely defined. In fact, it can be obtained with a simple formula:

$$\mathbf{p} = A\hat{\mathbf{x}} \text{ where } \hat{\mathbf{x}} \text{ is any solution of } A^TA\mathbf{x} = A^T\mathbf{b}.$$

The goal now is to justify this statement. We will do so in two steps:

1. Show the equation $A^TA\mathbf{x} = A^T\mathbf{b}$ is always consistent.
2. Show that a vector $\mathbf{p}$ satisfies the definition of projection if and only if it is $A\hat{\mathbf{x}}$ for some solution $\hat{\mathbf{x}}$ of $A^TA\mathbf{x} = A^T\mathbf{b}$

Afterward, we will see that $\mathbf{p}$ is uniquely determined and is the closest element of col $A$ to $\mathbf{b}$.

**The equation $A^TA\mathbf{x} = A^T\mathbf{b}$ is always consistent.** Recall that an equation is consistent when the vector on the right is in the column space of the matrix on the left. That is the same thing as saying that the system is consistent when the vector on the right is expressible as a linear combination of the columns of the matrix on the left. In the equation $A^TA\mathbf{x} = A^T\mathbf{b}$ the vector on the right is $A^T\mathbf{b}$ and the matrix on the left is $A^TA$. So we must show that $A^T\mathbf{b} \in$ col $A^TA$ for every choice of $\mathbf{b}$.

To do so we take a rather indirect approach. We will show that col $A^T$ = col $A^TA$. Then since $A^T\mathbf{b}$ is clearly in col $A^T$, it must also be in col $A^TA$. Notice here that $A^T$ is $n \times m$ and $A^TA$ is $n \times n$, so that their column spaces are both subspaces of $\mathbb{R}^n$. In fact it is easy to see that col $A^TA$ lies inside of col $A^T$. Afterall, an element in col $A^TA$ is anything of the form $A^TA\mathbf{x}$, and regrouping as $A^T(A\mathbf{x})$ we see that this vector is also an element of col $A^T$. This tells us that col $A^TA$ is a subspace of col $A^T$. So if we can show that both of these spaces have the same dimension, then col $A^TA$ fills up col $A^T$, so they must actually be equal.

So now we want to consider the dimensions of col $A^TA$ and col $A^T$. Again we proceed indirectly, this time by looking first at null spaces, by proving the following result.

**Lemma:** For any $m \times n$ matrix $A$, nul $A^TA$ = nul $A$.

**Proof:** Suppose that $\mathbf{x}$ is in nul $A$. That means $A\mathbf{x} = 0$. But then we also have $A^TA\mathbf{x} = 0$, showing that $\mathbf{x}$ is in nul $A^TA$. On the other hand, if $\mathbf{x}$ is in nul $A^TA$, then $A^TA\mathbf{x} = 0$. But then we also have $\mathbf{x}^TA^TA\mathbf{x} = 0$. That can be rewritten $(\mathbf{x}^TA^T)A\mathbf{x} = (A\mathbf{x})^TA\mathbf{x} = 0$. This is the same as the dot product of the vector $A\mathbf{x}$ with itself, and the only way that can be 0 is for the vector itself to be 0. We conclude that $A\mathbf{x} = 0$, and that shows that $\mathbf{x}$ is in nul $A$. Put together, these two arguments show that exactly the same elements are contained in nul $A^TA$ and nul $A$, so they are in fact the same sets, completing the proof.

Next, we use the fact that the dimension of the null space and column space of a matrix are related. In fact we have seen a theorem that says that the dimension of the null space plus the dimension of the column space equals the total number of columns. But it is worth recalling why that is true. The dimension of the column space is the same as the number of pivot columns, because those form a basis for the column space. The dimension of the null space is the same as the number of nonpivot columns, because each of those corresponds to a free variable in the solution of the homogeneous equation. So saying that the dimension of the null space plus the dimension of the column space equals the number of columns is really just saying that you can divide the columns up into pivot and nonpivot columns.

How do we apply that result to the case at hand? We have established that nul $A^TA$ and nul $A$ are equal, and so have the same dimension. Also, $A^TA$ and $A$ each have $n$ columns. So now we can see that col $A^TA$ and col $A$ have equal dimension. Going one further step, we also know that col $A$ and col $A^T$ have the same dimension. This is the rank theorem, and once again it is worth recalling why it is true. The dimension of col $A$ is the number of independent columns in $A$, and hence the number of pivot columns. The dimension of col $A^T$ is the number of independent *rows* in $A$, and that is the same as the number of nonzero rows in the rref of $A$. As we have seen many times, the number of pivot entries is equal to both the number of pivot columns and also the number of nonzero rows. So that shows that col $A$ and col $A^T$ have the same dimension.

Drawing this part of the argument to a close, we now know that col $A^T$ and col $A$ have the same dimension. But we already knew that col $A$ and col $A^TA$ have the same dimension. Therefore col $A^T$ and col $A^TA$ have the same dimension, and as argued above, these subspaces of $\mathbb{R}^n$ must actually be equal. In turn, that shows that for any $\mathbf{b} \in \mathbb{R}^m$, since $A^T\mathbf{b}$ is in col $A^T$, it is also in col $A^TA$, making the equation $A^TA\mathbf{x} = A^T\mathbf{b}$ consistent.

**The Projection Equals $A\hat{\mathbf{x}}$.** Now that we know the equation $A^TA\mathbf{x} = A^T\mathbf{b}$ consistent, consider any solution $\hat{\mathbf{x}}$. Define $\mathbf{p}$ to be $A\hat{\mathbf{x}}$. Clearly $\mathbf{p} \in$ col $A$. We want to show that $\mathbf{b} - \mathbf{p} = \mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to every element $\mathbf{c}$ of col $A$. Such a $\mathbf{c}$ must be $A\mathbf{y}$ for some $\mathbf{y} \in \mathbb{R}^n$, by definition of column space. So we need to show that $A\mathbf{y}$ and $\mathbf{b} - A\hat{\mathbf{x}}$ are orthogonal. Recall that we can write the dot product of vectors $\mathbf{u}$ and $\mathbf{v}$ as the *matrix* product $\mathbf{u}^T\mathbf{v}$. This leads to the following series of equations

$$\begin{aligned}(A\mathbf{y}) \cdot (\mathbf{b} - A\hat{\mathbf{x}}) &= (A\mathbf{y})^T(\mathbf{b} - A\hat{\mathbf{x}}) \\ &= \mathbf{y}^T A^T(\mathbf{b} - A\hat{\mathbf{x}}) \\ &= \mathbf{y}^T(A^T\mathbf{b} - A^T A\hat{\mathbf{x}}) \\ &= \mathbf{y}^T(0)\end{aligned}$$

where the final reduction occurs because $\hat{\mathbf{x}}$ is a solution to the equation $A^T A\mathbf{x} = A^T\mathbf{b}$. As the final equation shows, $A\mathbf{y}$ is orthogonal to $\mathbf{b} - A\hat{\mathbf{x}} = \mathbf{b} - \mathbf{p}$ for every $\mathbf{y}$. That shows that $\mathbf{b} - \mathbf{p}$ is orthogonal to col $A$, and verifies that $\mathbf{p} = A\hat{\mathbf{x}}$ satisfies the definition of projection of $\mathbf{b}$ on col $A$.

We now know that the formula presented above does in fact produce a valid projection. But could there be other projections, obtained by other formulas? The answer is no – the projection is unique. We will see that next.

**Uniqueness of the Projection.** In general, to show that something is unique, one assumes two such things and then proves them to be equal. In this case, we imagine $\mathbf{p}$ and $\mathbf{q}$ each satisfy the definition of projection of $\mathbf{b}$ on col $A$. That means $\mathbf{p}$ and $\mathbf{q}$ are both in col $A$, and since col $A$ is a subspace, $\mathbf{p} - \mathbf{q}$ is also in col $A$. We also know that $\mathbf{b} - \mathbf{q}$ and $\mathbf{b} - \mathbf{p}$ are both orthogonal to col $A$, as is their difference, $(\mathbf{b} - \mathbf{q}) - (\mathbf{b} - \mathbf{p})$ since the orthogonal complement of col $A$ is also a subspace. But what is that difference? It is $(\mathbf{b} - \mathbf{q}) - (\mathbf{b} - \mathbf{p}) = \mathbf{p} - \mathbf{q}$. So we have now shown that $\mathbf{p} - \mathbf{q}$ is simultaneously in col $A$, and orthogonal to every element of col $A$. That makes $\mathbf{p} - \mathbf{q}$ orthogonal to itself, and hence equal to the zero vector. In conclusion, $\mathbf{p} = \mathbf{q}$. This shows that the projection vector $\mathbf{p}$ is uniquely determined.

Note: that is not the same as saying that $\hat{\mathbf{x}}$ is unique. In fact, we can analyze the uniqueness of $\hat{\mathbf{x}}$ using what we know about linear systems, in this case, the system $A^T A\mathbf{x} = A^T\mathbf{b}$. To find solutions, we would find the rref of the augmented matrix $[A^T A | A^T\mathbf{b}]$. We know the system is consistent, so there are two possibilities. If the square matrix $A^T A$ has all pivot columns, there will be no free variables in the solution to the homogeneous equation, and the solution to the system will be unique. As we have seen, this will occur if and only if $A^T A$ is invertible. When that is true, we can write $\hat{\mathbf{x}} = (A^T A)^{-1} A^T\mathbf{b}$, and therefore obtain the formula

$$\mathbf{p} = A(A^T A)^{-1} A^T\mathbf{b}.$$

On the other hand, if $A^T A$ is not invertible, the system $A^T A\mathbf{x} = A^T\mathbf{b}$ will have an infinite number of solutions. But all those solutions lead to the same vector $\mathbf{p} = A\hat{\mathbf{x}}$.

We can push this analysis one step further. When is $A^T A$ invertible? Answer: when $A$ has independent columns. To see this, recall that we already saw nul $A$ = nul $A^T A$. We know from the invertible matrix theorem that $A^T A$ is invertible if and only if its null space is $\{0\}$, which is the same as saying if and only if nul $A = \{0\}$. But that is exactly the condition for the columns of $A$ to be independent. Another way to say the same thing is that the rank of $A$ is $n$.

**The Projection is the Closest Thing in col $A$ to b.** Geometrically, we are aware that the hypotenuse of a right triangle is always longer than either of the two sides. This fact motivates a simple proof that the projection of $\mathbf{b}$ on col $A$ is the closest element of col $A$ to $\mathbf{b}$. However, we can formulate a completely algebraic proof, that works in any number of dimensions. As a precursor, recall the following version of the Pythagorean Theorem:

**Vector Pythagorean Theorem:** If $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^n$, and if $\mathbf{u} \cdot \mathbf{v} = 0$, then

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

**Proof:** Using the fact that $\|\mathbf{w}\|^2 = \mathbf{w} \cdot \mathbf{w}$ for all $\mathbf{w}$, we have the following series of equations:

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

where the last result follows because we know $\mathbf{u} \cdot \mathbf{v} = 0$.

Now let's apply this in the context of projections. Let $\mathbf{q}$ be any element of col $A$. How far is it from $\mathbf{b}$? The square of the distance is $\text{DS} = \|\mathbf{b}-\mathbf{q}\|^2$. Next use the projection to replace $\mathbf{b}$ with $\mathbf{p}+(\mathbf{b}-\mathbf{p})$, obtaining

$$\begin{aligned} \text{DS} &= \|\mathbf{p}+(\mathbf{b}-\mathbf{p})-\mathbf{q}\|^2 \\ &= \|(\mathbf{b}-\mathbf{p})+(\mathbf{p}-\mathbf{q})\|^2 \\ &= \|\mathbf{u}+\mathbf{v}\|^2 \end{aligned}$$

where $\mathbf{u} = \mathbf{b}-\mathbf{p}$ and $\mathbf{v} = \mathbf{p}-\mathbf{q}$. Since $\mathbf{p}$ and $\mathbf{q}$ are both in col $A$, so is $\mathbf{v}$. On the other hand, by definition of the projection we know that $\mathbf{b}-\mathbf{p} = \mathbf{u}$ is orthogonal to col $A$. This shows that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, and the Pythagorean Theorem thus gives us

$$\text{DS} = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2.$$

As we vary $\mathbf{q}$ (and hence $\mathbf{v}$), we can see that DS will assume a minimum value of $\|\mathbf{u}\|^2$ exactly when $\mathbf{v} = 0$, that is, when $\mathbf{q} = \mathbf{p}$. This shows that the minimum distance from $\mathbf{b}$ to a $\mathbf{q}$ in col $A$ is obtained when $\mathbf{q} = \mathbf{p}$.

**Least Squares Solutions to Inconsistent Systems.** The discussion of these ideas of projections is motivated by the problem of approximating solutions to inconsistent linear systems. The context is this: $A$ is an $m \times n$ matrix, and $\mathbf{b} \in \mathbb{R}^m$. We assume the linear system $A\mathbf{x} = \mathbf{b}$ is inconsistent – there is no exact solution. That means we cannot make $A\mathbf{x}-\mathbf{b}$ zero. So instead let's minimize it. That is, let's pick an $\mathbf{x}$ that minimizes the length of $A\mathbf{x}-\mathbf{b}$. But that is exactly what the projection analysis gives us. That is, we have seen that the closest $A\mathbf{x}$ to $\mathbf{b}$ is actually $\mathbf{p}$, the projection of $\mathbf{b}$ on col $A$. Our main interest is the solution $\mathbf{x}$, not the projection $A\mathbf{x}$. And we know how to find it: solve the consistent system $A^TA\mathbf{x} = A^T\mathbf{b}$. If $A$ has independent columns, then we actually have a formula

$$\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}.$$

But even if the columns of $A$ are not independent, we can find the solutions $\hat{\mathbf{x}}$ by forming the rref of the augmented matrix $[A^TA|A^T\mathbf{b}]$. In either case, if the original system is actually consistent, the least squares analysis leads to the exact solutions of the system.

To summarize, we have the following statement.

**For an inconsistent system $A\mathbf{x} = \mathbf{b}$, by a least-squares solution we mean any vector $\hat{\mathbf{x}}$ that is a solution to the system $A^TA\mathbf{x} = A^T\mathbf{b}$. A least squares solution produces the minimal possible error $\|A\mathbf{x}-\mathbf{b}\|$ associated with the inconsistent system.**

**Is the Least Squares Solution Really Best? The Error Vector Concept.** The derivation of the concept of least squares solution has at its foundation the concept of distance between two vectors in $n$ dimensional real space, $\mathbb{R}^n$. But remember that for dimensions higher than 3, this distance is entirely a matter

of convention. That is, we defined the length of a vector $\mathbf{v}$ to be $\sqrt{\mathbf{v} \cdot \mathbf{v}}$ in analogy with the geometrically observable formulas in lower dimensions. Since we cannot actually see the geometry of spaces with four or more dimensions, we cannot absolutely verify that our idea of distance is right. So does that leave a question about whether the entire least squares approach might be somehow misconceived?

To put this another way, we can ask what exactly is being minimized in the least squares approach. The answer involves what is called the error vector. Suppose our linear system $A\mathbf{x} = \mathbf{b}$ is inconsistent, and let us introduce the notation

$$\mathbf{y} = A\mathbf{x}.$$

As we vary $\mathbf{x}$, we will never observe $\mathbf{y}$ to exactly equal $\mathbf{b}$. But we can consider $\mathbf{x}$ to be an approximate solution if $\mathbf{y}$ is approximately equal to $\mathbf{b}$. How accurate is this approximation? The answer can be expressed by the vector difference $\mathbf{b} - A\mathbf{x}$, which we call the *error vector*. It represents the the difference between what we want, $\mathbf{b}$ and what our approximate solution actually produces, $A\mathbf{x}$.

In column vector notation, we can express

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Then the error vector is given by

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} = \begin{bmatrix} b_1 - y_1 \\ b_2 - y_2 \\ \vdots \\ b_m - y_m \end{bmatrix}.$$

The numbers $e_1, e_2$, etc. are often called *residuals*.

The least squares solution gives us a vector $\hat{\mathbf{x}}$ that minimizes the length of the error vector, or equivalently, the square of the length of the error vector. As noted, we can question whether our definition of length is geometrically correct. But the derivation of the least squares solution is numerically valid. So we do know that $\hat{\mathbf{x}}$ provides the minimal value of

$$\|\mathbf{b} - \mathbf{y}\|^2 = e_1^2 + e_2^2 + \cdots + e_m^2.$$

Thus, the least squares solution is definitely the one that minimizes the sum of the squares of the residuals. This is where the term *least squares* comes from.

People can (and do) question whether our idea of distance in high dimensional spaces is appropriate for a given setting. For example, if the vector $\mathbf{b}$ arises from measured data, it is possible that some of the values $b_1, b_2, \cdots, b_m$ are more accurately measured than others. In this case we may wish to give greater weight to the most accurate values in trying to obtain a best approximate solution to our inconsistent system. And this in turn can be viewed as adopting a different meaning of distances between vectors in $\mathbb{R}^m$. This leads to modified methods referred to as *weighted least squares*.

Taking everything into consideration leads to the following overview of the least squares methodology. Our extension of geometric ideas from two and three dimensions to higher dimensional settings is based on analogies. This guides an initial concept of the best approximate solution to an inconsistent system. And although the definition of what we wish to find involves reasoning by analogy, the derivation of the solution is algebraically correct. In the end, we have a specific equation for what is being minimized, and a specific method for finding the exact value of the minimum. Finally, by reconsidering whether our initial analogy is appropriate for a specific application, we can adapt the method to alternative formulations of distance in $\mathbb{R}^m$.