Chapter 4: Orthogonality

Val Anthony Balagon

January 2019

Abstract

This chapter focuses on the orthogonality of the four subspaces, orthogonal vectors, projections, least squares approximations, and the Gram-Schmidt factorization.

1 Orthogonality of the Four Subspaces

Two vectors are orthogonal when their dot product is zero $\boldsymbol{v} \cdot \boldsymbol{w} = \boldsymbol{v}^T\boldsymbol{w} = 0$. This chapter will revolve around orthogonal subspaces, orthogonal bases, and orthogonal matrices.

Definition 1.1. Orthogonal vectors have the following properties:

i. $\boldsymbol{v}^T\boldsymbol{w} = 0$

ii. $||\boldsymbol{v}||^2 + ||\boldsymbol{w}||^2 = ||\boldsymbol{v} + \boldsymbol{w}||^2 \to \boldsymbol{v}^T\boldsymbol{v} + \boldsymbol{w}^T\boldsymbol{w} = (\boldsymbol{v} + \boldsymbol{w})^T(\boldsymbol{v} + \boldsymbol{w})$

Remark. *The zero vector is orthogonal to any vector.*

Remark. *The subspaces have orthogonal properties.*

1. ***The rowspace*** $C(A^T)$ ***is perpendicular to the nullspace*** $N(A)$. *Every row of* A *is perpendicular to the solution of* $A\boldsymbol{x} = \boldsymbol{0}$.

2. ***The column space*** $C(A)$ ***is perpendicular to the left nullspaces*** $N(A^T)$. *When* $\boldsymbol{b}$ *is outside of the column space when we're trying to solve for* $A\boldsymbol{x} = \boldsymbol{b}$, *then this nullspace of* A^T *comes into its own. It contains the error* $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ *in the least-squares solution.*

Definition 1.2. Two subspaces $\boldsymbol{V}$ and $\boldsymbol{W}$ of a vector space are orthogonal if every vector $\boldsymbol{v}$ in $\boldsymbol{V}$ is perpendicular to every vector $\boldsymbol{w}$ in $\boldsymbol{W}$.

$$\boldsymbol{v}^T\boldsymbol{w} = 0 \text{ for all } \boldsymbol{v} \text{ in } \boldsymbol{V} \text{ and all } \boldsymbol{w} \text{ in } \boldsymbol{W}.$$

Theorem 1. Every vector $\boldsymbol{x}$ in the nullspace is perpendicular to every row of A, because $A\boldsymbol{x} = \mathbf{0}$. The nullspace $N(A)$ and the row space $C(A^T)$ are orthogonal subspaces of $\mathbb{R}^n$.

$$A\boldsymbol{x} = \begin{bmatrix} \text{row 1} \\ \vdots \\ \text{row m} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{gathered} C_1(\text{row}_1^T) = 0 \\ C_2(\text{row}_2^T) = 0 \\ \vdots \\ C_m(\text{row}_m^T) = 0 \end{gathered}$$

(row 1) $\cdot\boldsymbol{x}$ is zero and (row m) $\cdot\, \boldsymbol{x}$ is also zero. Every row has a zero dot product with $\boldsymbol{x}$. Then $\boldsymbol{x}$ is perpendicular to every combination of the rows. **The whole row space $C(A^T)$ is orthogonal to $N(A)$.**

Proof. The vectors in the row space are combinations of $A^T\boldsymbol{y}$ of the rows. We take the dot product of $A^T\boldsymbol{y}$ with any $\boldsymbol{x}$ in the nullspace.

$$\boldsymbol{x} \cdot (A^T\boldsymbol{y}) = \boldsymbol{x}^T(A^T\boldsymbol{y}) = (A\boldsymbol{x})^T\boldsymbol{y} = 0^T\boldsymbol{y} = 0$$

□

Example 1. The rows of A are perpendicular to $\boldsymbol{x} = (1, 1, -1)$ in the nullspace:

$$A\boldsymbol{x} = \begin{bmatrix} 1 & 3 & 4 \\ 5 & 2 & 7 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1+3-4 \\ 5+2-7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In this example, the column space is all of $\mathbb{R}^2$. The nullspace of A^T is the zero vector. The column space of A and the nullspace of A^T are always orthogonal subspaces.

Theorem 2. Every vector $\boldsymbol{y}$ in the nullspace of A^T is perpendicular to every column of A. The left nullspace $N(A^T)$ and the column space $C(A)$ are orthogonal in $\mathbb{R}^m$.

Proof. The nullspace of A^T is orthogonal to the row space of A^T, which is the column space of A.

$$A^T\boldsymbol{y} = \begin{bmatrix} (\text{column 1})^T \\ \vdots \\ (\text{column n})^T \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

□

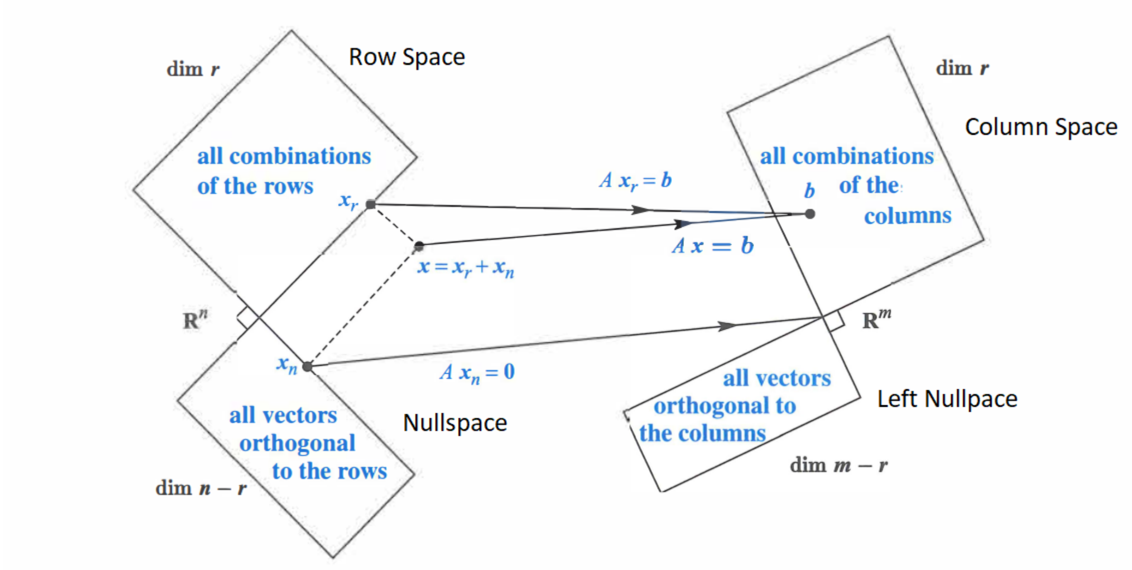

Figure 1: The Four Subspaces. There are two pairs of orthogonal subspaces.

Theorem 3. If a vector $\boldsymbol{v}$ is orthogonal to itself, then $\boldsymbol{v}$ is the zero vector.

Theorem 4. Fundamental Theorem of Linear Algebra, Part 2:
$N(A)$ **is the orthogonal complement of the row space** $C(A^T)$ **in** $\mathbb{R}^n$**.**
$N(A^T)$ **is the orthogonal complement of the column space** $C(A)$ **in** $\mathbb{R}^m$**.**

Things to note from Figure 1

1. When A multiplies to $\boldsymbol{x} = \boldsymbol{x}_r + \boldsymbol{x}_n$, it goes to $\boldsymbol{b}$ which is in the column space.
2. When A multiplies to $\boldsymbol{x}_r$, it goes to $\boldsymbol{b}$ which is also in the column space.
3. When A multiplies to $\boldsymbol{x}_n$, the nullspace component goes to $\mathbf{0}$.

1.1 Combining Bases from Subspaces

Theorem 5. Any independent vectors in $\mathbb{R}^n$ must span $\mathbb{R}^n$. So they are a basis.
Any n vectors that span $\mathbb{R}^n$ must be independent. So they are a basis

Theorem 6. If the n columns of A are independent, they span $\mathbb{R}^n$. So $A\boldsymbol{x} = \boldsymbol{b}$ is solvable.
If the n columns span $\mathbb{R}^n$, they are independent. So $A\boldsymbol{x} = \boldsymbol{b}$ has only one solution.

2 Projections

Let's say we are given an arbitrary vector $\boldsymbol{b}$. When $\boldsymbol{b}$ is projected onto a line, its projection $\boldsymbol{p}$ is the part of $\boldsymbol{b}$ along that line. If $\boldsymbol{b}$ is projected onto a plane, $\boldsymbol{p}$ is a part in that plane. The projection $\boldsymbol{p}$ is $P\boldsymbol{b}$. The projection matrix P multiplies $\boldsymbol{b}$ to give $\boldsymbol{p}$.

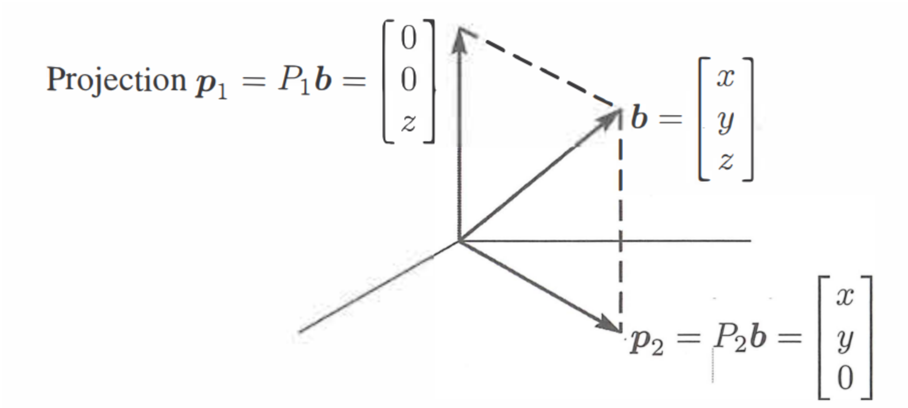

Figure 2: The projections $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ onto the z axis and the xy plane.

Say $\boldsymbol{b} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}^T$, its projections are $\boldsymbol{p}_1 = \begin{bmatrix} 0 & 0 & 4 \end{bmatrix}^T$ and$\boldsymbol{p}_2 = \begin{bmatrix} 2 & 3 & 0 \end{bmatrix}^T$ in Figure 2. $\boldsymbol{p}_1$and$\boldsymbol{p}_2$ are the projections of $\boldsymbol{b}$ onto the z axis and xy axis, respectively. The projection matrices P_1 and P_2 are 3×3 matrices. Projection onto a line comes from a rank one matrix, while a projection onto a planes comes from a rank two matrix.

$$\text{Onto the } z \text{ axis: } P_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{Onto the } xy \text{ plane: } P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{p}_1 = P_1\boldsymbol{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix} \qquad \boldsymbol{p}_2 = P_2\boldsymbol{b} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$$

The projections p_1 and p_2 are perpendicular. The xy plane and the z axis are orthogonal subspaces. The line and plane are also orthogonal complements. Their dimensions add to $1 + 2 = 3$. Every vector $\boldsymbol{b}$ in the whole space is the sum of its parts in the two subspaces. The projections $\boldsymbol{p}_1$and$\boldsymbol{p}_2$ are exactly those two parts of $\boldsymbol{b}$:

$$\text{The vectors give } \boldsymbol{p}_1 + \boldsymbol{p}_2 = \boldsymbol{b} \qquad \text{The matrices give } P_1 + P_2 = I$$

In general, the objective is to find $\boldsymbol{p}$ in each subspace, and the projection matrix P that produces that part $\boldsymbol{p} = P\boldsymbol{b}$. Every subspace of $\mathbb{R}^m$ has its own $m \times m$ projection matrix. To compute P, we need a good description of the subspace that it projects onto. The best description of a subspace is a basis. We put the basis vectors into the columns of A. Now we are projecting onto the column space of A. The problem now is to project any $\boldsymbol{b}$ onto the column space of any $m \times n$ matrix.

2.1 2-D Case

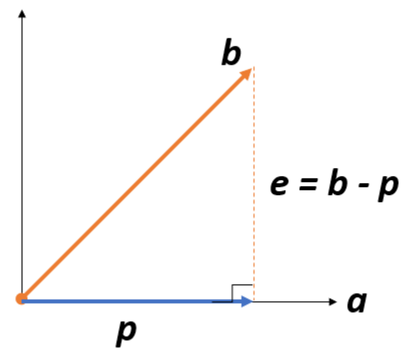

Figure 3: The projection of $\boldsymbol{b}$ onto $\boldsymbol{a}$

A line goes through the origin in the direction $\boldsymbol{a}$. Along that line, we want the point $\boldsymbol{p}$ closest to $\boldsymbol{b}$. The key to projection is orthogonality: The line from $\boldsymbol{b}$ to $\boldsymbol{p}$ is perpendicular to the vector $\boldsymbol{a}$. This is the dotted line marked $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ for the error on the left side of Figure 3. We now comute $\boldsymbol{p}$ by algebra.

The projection $\boldsymbol{p}$ will be a multiple of $\boldsymbol{a}$. We call it $\boldsymbol{p} = \hat{\boldsymbol{x}}\boldsymbol{a}$. Computing this number $\hat{\boldsymbol{x}}$ will give the vector $\boldsymbol{p}$. Then from the formula $\boldsymbol{p}$, we will read off the projection matrix P. These three steps will lead to all projection matrices: **find $\hat{\boldsymbol{x}}$, then find the vector $\boldsymbol{p}$, then fin the matrix $\boldsymbol{P}$**.

Projecting $\boldsymbol{b}$ onto $\boldsymbol{a}$ with error $\boldsymbol{e} = \boldsymbol{b} - \hat{\boldsymbol{x}}\boldsymbol{a}$:

$$\boldsymbol{a} \cdot (\boldsymbol{b} - \hat{\boldsymbol{x}}\boldsymbol{a}) = 0 \quad \text{or} \quad \boldsymbol{a} \cdot \boldsymbol{b} - \hat{\boldsymbol{x}}\boldsymbol{a} \cdot \boldsymbol{a} = 0$$

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\boldsymbol{a} \cdot \boldsymbol{a}} = \boxed{\frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}}$$

Theorem 7. The projection of $\boldsymbol{b}$ onto the line through $\boldsymbol{a}$ is the vector $\boldsymbol{p} = \hat{\boldsymbol{x}}\boldsymbol{a} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a}$

Special Case 1: If $\boldsymbol{b} = \boldsymbol{a}$, then $\hat{\boldsymbol{x}} = 1$. The projection of $\boldsymbol{a}$ onto $\boldsymbol{a}$ is itself. $P\boldsymbol{a} = \boldsymbol{a}$.

Special Case 2: If $\boldsymbol{b}$ is perpendicular to $\boldsymbol{a}$, then $\boldsymbol{a}^T\boldsymbol{b} = 0$. The projection is $\boldsymbol{p} = 0$.

Example 2. Project $\boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ onto $\boldsymbol{a} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$ to find $\boldsymbol{p} = \hat{\boldsymbol{x}}\boldsymbol{a}$.

$$\boldsymbol{p} = \hat{\boldsymbol{x}}\boldsymbol{a} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a} = \frac{\begin{bmatrix} 1 & 2 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 & 2 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}}\boldsymbol{a} = \boxed{\frac{5}{9}\boldsymbol{a}}$$

The error vector between $\boldsymbol{b}$ and $\boldsymbol{p}$ is $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$. Those vectors $\boldsymbol{p}$ and $\boldsymbol{e}$ will add to $\boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

$$\boldsymbol{p} = \left(\frac{5}{9}, \frac{10}{9}, \frac{10}{9}\right) \qquad \text{and} \qquad \boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p} = \left(\frac{4}{9}, -\frac{1}{9}, -\frac{1}{9}\right)$$

To get the magnitude of $\boldsymbol{p}$:

$$||\boldsymbol{p}|| = \frac{||\boldsymbol{a}||||\boldsymbol{b}|| \cos\theta}{||\boldsymbol{a}||^2} ||\boldsymbol{a}|| = ||\boldsymbol{b}|| \cos\theta$$

Now to find the projection matrix,

$$\boldsymbol{p} = \boldsymbol{a}\hat{\boldsymbol{x}} = \boldsymbol{a}\frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}} = P\boldsymbol{b} \text{ then } P = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}$$

$$P = \begin{bmatrix} \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \end{bmatrix}$$

P is a column times a row! Then column is $\boldsymbol{a}$, the row is $\boldsymbol{a}^T$. Then divide by the number $\boldsymbol{a}^T\boldsymbol{a}$. The projection matrix P is an $m \times m$ matrix, but its rank is 1. We are projecting onto a 1-D subspace, a line through $\boldsymbol{a}$. *That line is the column space of* P. Next, we try to project a second time.

$$P^2 = \begin{bmatrix} \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \end{bmatrix} \begin{bmatrix} \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \end{bmatrix} = \begin{bmatrix} \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \\ \frac{2}{9} & \frac{4}{9} & \frac{4}{9} \end{bmatrix} = P$$

$$\boxed{P^2 = P}$$

Notice that projecting a second time does not change anything. Notice also that P is symmetric.

To summarize:

1. $C(P)$ is a line through A
2. $\text{rank}(P) = 1$
3. $P^T = P$, P is symmetric
4. $P^2 = P$, projecting a second time does not change anything

2.2 Projection Onto A Subspace

We now move to n vectors $a_1, \ldots, a_n$ in $\mathbb{R}^m$. We assume that these $\boldsymbol{a}$'s are linearly independent. The problem now becomes: how do we find the combination $\boldsymbol{p} = \hat{\boldsymbol{x}}_1 a_1 + \ldots + \hat{\boldsymbol{x}}_n a_n$ closest to a given vector $\boldsymbol{b}$? We are projecting $\boldsymbol{b}$ in $\mathbb{R}^m$ onto the subspace spanned by the $\boldsymbol{a}$'s.

We compute projections onto n-dimensional subspaces in three steps as before: ***Find a vector $\hat{\mathbf{x}}$, find the projection $\mathbf{p} = A\hat{\mathbf{x}}$, find the projection matrix P.***

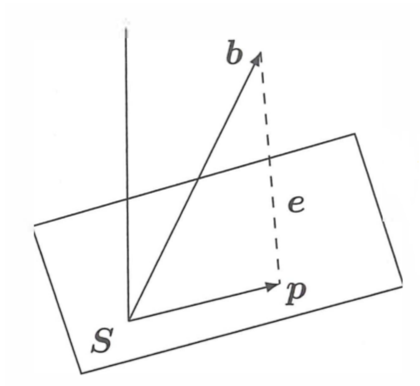

Figure 4: The projection $\boldsymbol{p}$ of $\boldsymbol{b}$ onto $\boldsymbol{S}$ = column space of A

The error vector $\boldsymbol{b} - A\hat{\boldsymbol{x}}$ is perpendicular to the subspace. In n-dimensions, there will be n equations for $\hat{\boldsymbol{x}}$. The error makes a right angle with all the vectors $a_1, \ldots, a_n$ in the base. The n right angles give the n equations for $\hat{\boldsymbol{x}}$.

$$\begin{bmatrix} \text{---} & a_1^T & \text{---} \\ & \vdots & \\ \text{---} & a_n^T & \text{---} \end{bmatrix} \tag{1}$$

The combination $\boldsymbol{p} = \hat{x_1}\boldsymbol{a}_1 + \ldots + \hat{x_n}\boldsymbol{a}_n = A\hat{\boldsymbol{x}}$ that is closest to $\boldsymbol{b}$ comes from $\hat{\boldsymbol{x}}$:

$$\boxed{A^T(\boldsymbol{b} - A\hat{\boldsymbol{x}}) = \mathbf{0}} \qquad \text{or} \qquad \boxed{A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}} \tag{2}$$

This symmetric $A^T A$ is $n \times n$. It is invertivle if $\boldsymbol{a}$'s are independent. The solution $\hat{\boldsymbol{x}} = (A^T A)^{-1} A^T \boldsymbol{b}$. The projection of $\boldsymbol{b}$ onto the subspace is $\boldsymbol{p}$:

$$\boxed{\boldsymbol{p} = A\hat{\boldsymbol{x}} = A(A^T A)^{-1} A^T \boldsymbol{b}} \tag{3}$$

To find the projection matrix P:

$$\boxed{P = A(A^T A)^{-1} A^T} \tag{4}$$

The equations above are equivalent to their $n = 1$ counterparts. We use the inverse $(A^T A)^{-1}$ instead of $\frac{1}{\boldsymbol{a}^T \boldsymbol{a}}$. Matrix inversion is a guarantee because the columns of A are linearly independent.
Here are interesting properties:

1. Our subspace is the column space of A
2. The error vector $\boldsymbol{b} - A\hat{\boldsymbol{x}}$ is perpendicular to that column space
3. Therefore $\boldsymbol{b} - A\hat{\boldsymbol{x}}$ is in the nullspace of A^T. This means that $A^T(\boldsymbol{b} - A\hat{\boldsymbol{x}}) = \mathbf{0}$

Example 3. If $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$ find $\hat{\boldsymbol{x}}$ and $\boldsymbol{p}$.

Computing for $A^T A$ and $A^T \boldsymbol{b}$

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

$$A^T \boldsymbol{b} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

Computing for $\hat{\boldsymbol{x}}$

$$\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix} \rightarrow \hat{\boldsymbol{x}} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

Finding $\boldsymbol{p}$:

$$\boldsymbol{p} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ -3 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix}$$

Finding the error:

$$\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

Finding P:

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

$$P = A(A^T A)^{-1} A^T = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} * \frac{1}{6} \begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$

Projecting it twice,

$$P^2 = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} * \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$$
$$\boxed{P^2 = P}$$

If A is a rectangular matrix, then we won't be able to do $(A^TA)^{-1} = A^{-1}(A^T)^{-1}$. This is because A does not have an inverse.

Theorem 8. A^TA is invertible if and only if A has linearly independent columns.

Proof. A^TA and A share the same nullspace in $\mathbb{R}^n$. A is invertible when its nullspace is the zero vector. The same goes for A^TA. Let A be any matrix. If $\boldsymbol{x}$ is in the nullspace of A, then $A\boldsymbol{x} = \mathbf{0}$. Multiplying both sides by A^T gives:

$$A^TA\boldsymbol{x} = \mathbf{0}$$

We prove $A\boldsymbol{x} = \mathbf{0}$ from here. We multiply $\boldsymbol{x}^T$ to both sides.

$$\begin{gathered} (x^T)A^TA\boldsymbol{x} = \mathbf{0} \\ (A\boldsymbol{x})^T(A\boldsymbol{x}) = \mathbf{0} \\ A\boldsymbol{x} \cdot A\boldsymbol{x} = \mathbf{0} \\ ||A\boldsymbol{x}|| = \mathbf{0} \end{gathered}$$

If $A^TA\boldsymbol{x} = \mathbf{0}$, then $A\boldsymbol{x}$ has length zero. Therefore $A\boldsymbol{x} = \mathbf{0}$. Every vector $\boldsymbol{x}$ in the nullspace of A is also in the nullspace of A^TA. If A has dependent columns, then A^TA also has dependent columns. If A has independent columns, then A^TA also has independent columns. In this case, A^TA is invertible. □

Theorem 9. When A has independent columns, A^TA is square and invertible.

Proof.

Take the transpose of A^TA:

$$(A^TA)^T = A^T(A^T)^T = A^TA$$
$$\boxed{(A^TA)^T = A^TA}$$

Hence, A^TA is symmetric. Since A has independent columns, this means that A^T is also invertible. Hence, A^TA is invertible. □

3 Least Squares Approximations

Sometimes $A\boldsymbol{x} = \boldsymbol{b}$ has no solution. This is seen in the case of overdetermined systems where there are more equations than unknowns which means that $\boldsymbol{b}$ is outside the column space of A.

We cannot always get $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$ to zero (with zero error, we can solve the system exactly with $\boldsymbol{x}$). When the length of $\boldsymbol{e}$ is as small as possible, $\hat{\boldsymbol{x}}$ is a least squares solution.

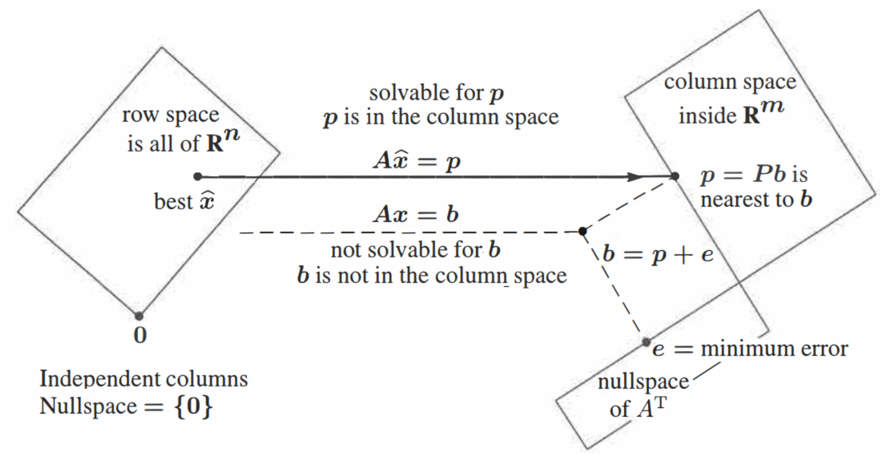

Figure 5: The big picture of least squares. The projection $\boldsymbol{p} = A\hat{\boldsymbol{x}}$ is closest to $\boldsymbol{b}$, so $\hat{\boldsymbol{x}}$ minimizes $E = ||\boldsymbol{b} - A\boldsymbol{x}||^2$

Remark. *When $A\boldsymbol{x} = \boldsymbol{b}$ has no solution, multiply by A^T and solve $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$.*

Example 4. An important application of least squares is fitting a straight line to m points. This problem is called "linear regression" in statistics. We have to find the closest line to the points $(0, 6), (1, 0),$ and $(2, 0)$.

The general equation for a line is:

$$C + Dx = y$$

From those points, we can generate our matrix A.

$$C + D \cdot (0) = 6$$
$$C + D \cdot (1) = 0$$
$$C + D \cdot (2) = 0$$

We are left with the following system of equations:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \hat{\boldsymbol{x}} = \begin{bmatrix} C \\ D \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$$

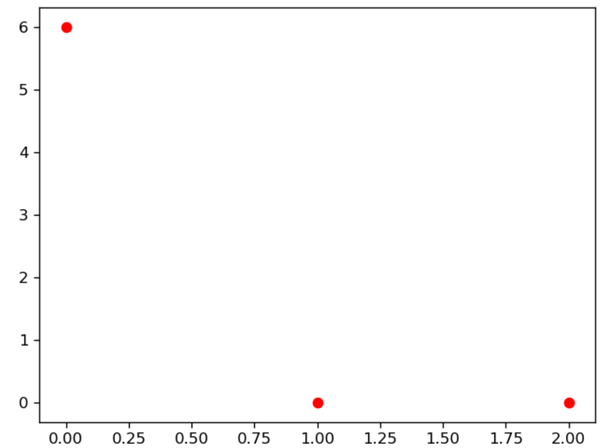

Figure 6: Points that need to be fitted with a straight line.

We want to solve $A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$:

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

$$A^T\boldsymbol{b} = \begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}6\\0\\0\end{bmatrix} = \begin{bmatrix}6\\0\end{bmatrix}$$

Solving for $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} = (A^T A)^{-1}A^T\boldsymbol{b} = \begin{bmatrix}\frac{5}{6} & -\frac{1}{2}\\-\frac{1}{2} & \frac{1}{2}\end{bmatrix}\begin{bmatrix}1 & 1 & 1\\0 & 1 & 2\end{bmatrix}\begin{bmatrix}6\\0\\0\end{bmatrix}$$

$$\hat{\boldsymbol{x}} = \begin{bmatrix}5\\-3\end{bmatrix}$$

Hence, the equation of the fitted line is:

$$\boxed{y = 5 - 3x}$$

This problem can also be solved through calculus. We want to minimize $||A\boldsymbol{x} - \boldsymbol{b}||^2 = ||\boldsymbol{e}||^2$

$$||A\boldsymbol{x} - \boldsymbol{b}|| = ||\boldsymbol{e}||^2 = \boldsymbol{e}_1^2 + \boldsymbol{e}_2^2 + Ve_3^2$$
$$\boldsymbol{e}_1^2 + \boldsymbol{e}_2^2 + \boldsymbol{e}_3^2 = (C + D\cdot(0) - 6)^2 + (C + D\cdot(1))^2 + (C + D\cdot(2))^2$$

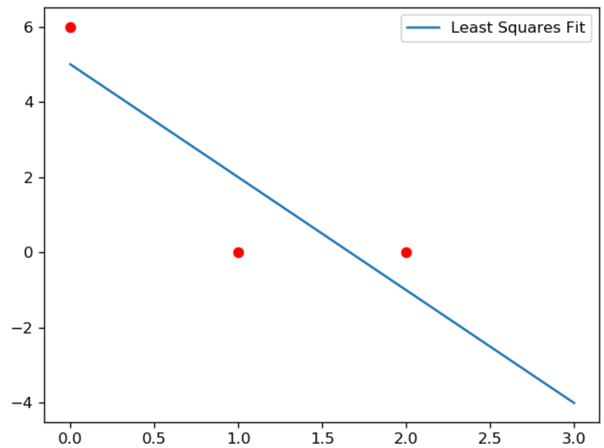

Figure 7: The points and the linear fit.

Setting $||A\boldsymbol{x} - \boldsymbol{b}||^2 = 0$ and taking its partial derivative with respect to $C$and D

$$\frac{\partial ||\boldsymbol{e}||^2}{\partial C} = 2(C + D \cdot (0) - 6) + 2(C + D \cdot (1)) + 2(C + D \cdot (2)) = 0$$
$$\frac{\partial ||\boldsymbol{e}||^2}{\partial D} = 2(C + D \cdot (1)) = 0$$

Reduction will yield $A^T A$:

$$3C + 3D = 6$$
$$3C + 5D = 0$$

Solution:

$$\hat{\boldsymbol{x}} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

Remark. *The least squares solution* $\hat{\boldsymbol{x}}$ *makes* $E = ||A\boldsymbol{x} - \boldsymbol{b}||^2$ *as small as possible.*

Remark. *The partial derivatives of* $||A\boldsymbol{x} - \boldsymbol{b}||^2$ *are zero when* $A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b}$.

Example 5. A has orthogonal columns when the measurement times t_i add to zero. Suppose $\boldsymbol{b} = (1, 2, 4)$ and $t = (-2, 0, 2)$. We want to fit these points into a line.

$$C + D \cdot (-2) = 1$$
$$C + D \cdot (0) = 2$$
$$C + D \cdot (2) = 4$$

Notice that the two columns of A are orthogonal via the dot product test ($\boldsymbol{col}_1 \cdot \boldsymbol{col}_2 = 0$). A and $\boldsymbol{b}$ becomes:

$$A\boldsymbol{x} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

When the columns of A are orthogonal, $A^T A$ will be a diagonal matrix.

$$A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{b} \rightarrow \begin{bmatrix} 3 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

We see that $A^T A$ is a diagonal matrix. This makes solving for $\hat{\boldsymbol{x}}$ a lot easier.

4 Orthonormal Bases and Gram-Schmidt

Orthogonality is good. Dot products are zero, so $A^T A$ will be diagonal. It becomes easier to find $\hat{\boldsymbol{x}}$ and $\boldsymbol{p} = A\hat{x}$. The goal of this subsection is to construct orthogonal vectors. The Gram-Schmidt process chooses original basis vectors to produce right angles. Those original vectors are the columns of a non-orthogonal matrix A. The orthonormal basis vectors will be the columns of a new matrix Q.

The vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are orthogonal if their dot products $\boldsymbol{q}_i \cdot \boldsymbol{q}_j$ are zero whenever $i \neq j$. When we divide each vector by its length, the vectors become orthogonal unit vectors. Their lengths are all 1. Then the basis is called ***orthonormal***.

Definition 4.1. The vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are orthonormal if

$$\boldsymbol{q}_i^T \boldsymbol{q} = \begin{cases} 0 & \text{when } i \neq j \text{ (orthogonal vectors)} \\ 1 & \text{when } i = j \text{ (unit vectors: } ||\boldsymbol{q}_i|| = 1) \end{cases}$$

A matrix with orthonormal columns is assigned the special letter Q.

Remark. *A matrix Q with orthonormal columns satisfies $Q^T Q = I$.*

$$Q^T Q = \begin{bmatrix} — & \boldsymbol{q}_1^T & — \\ & \vdots & \\ — & \boldsymbol{q}_n^T & — \end{bmatrix} \begin{bmatrix} | & \cdots & | \\ \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ | & \cdots & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Theorem 10. When Q is a square matrix, $Q^T Q = I$ means that $Q^T = Q^{-1}$.

Proof.

$$Q^T Q = I$$

Multiplying from the right by Q^{-1}

$$Q^T Q Q^{-1} = I Q^{-1}$$

By definition, $QQ^{-1} =$ I. Therefore,

$$Q^T = Q^{-1}$$

$\square$

Example 6. A rotation matrix is an orthogonal matrix. Q rotates each vector in the plane by the angle θ.

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

These columns give an orthonormal basis for $\mathbb{R}^2$.

Example 7. A permutation matrix is also an orthogonal matrix.

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} y \\ z \\ x \end{bmatrix}$$

Theorem 11. Rotations, reflections, and permutations preserve the length of every vector. So does multiplication by any orthogonal matrix Q - **lengths and angles do not change!**.

Proof.

$$(Q\boldsymbol{x})^T (Q\boldsymbol{x}) = \boldsymbol{x}^T Q^T Q \boldsymbol{x} = \boldsymbol{x}^T I \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{x}$$
$$||Q\boldsymbol{x}||^2 = ||\boldsymbol{x}||^2$$

$\square$

Theorem 12. If Q has orthonormal columns ($Q^T Q = I$), it leaves lengths unchanged.

$$||Q\boldsymbol{x}|| = ||\boldsymbol{x}|| \text{ for every vector } \boldsymbol{x}$$

Q also preserves dot products: $(Q\boldsymbol{x})^T (Q\boldsymbol{y}) = \boldsymbol{x}^T Q^T Q \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y}$

4.1 Projections Using Orthonormal Bases: Q Replaces A

Orthonormal matrices simplify many projection problems. There are no matrices to invert because we can simply use the transpose.

Non-orthonormal	*Orthonormal*
$A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$	$\cancel{Q^TQ}\hat{\boldsymbol{x}} = Q^T\boldsymbol{b} \rightarrow \hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$
$\boldsymbol{p} = A^T\hat{\boldsymbol{x}}$	$\boldsymbol{p} = Q^T\hat{\boldsymbol{x}} = QQ^T\boldsymbol{b}$
$P = A(A^TA)^{-1}A^T$	$P = Q(\cancel{Q^TQ})^{-1}Q^T = QQ^T$

Projection onto $\boldsymbol{q}$'s:

$$\boldsymbol{p} = \begin{bmatrix} | & & | \\ \boldsymbol{q}_1 & \dots & \boldsymbol{q}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1\boldsymbol{b} \\ \vdots \\ \boldsymbol{q}_n\boldsymbol{b} \end{bmatrix} = \boldsymbol{q}_1(\boldsymbol{q}_1^T\boldsymbol{b}) + \ldots + \boldsymbol{q}_n(\boldsymbol{q}_n^T\boldsymbol{b})$$

Important Case: When Q is square, the subspace is the whole space. Then $Q^T = Q^{-1}$ and $\hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$ is the same as $\hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$. The solution is exact. The projection of $\boldsymbol{b}$ onto the whole space is $\boldsymbol{b}$ itself. In this case, $\boldsymbol{p} = \boldsymbol{b}$ and $P = QQ^T = I$.

When $\boldsymbol{p} = \boldsymbol{b}$, our formula assembles $\boldsymbol{b}$ out of its 1-dimensional projections. If $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ is an orthonormal basis for the whole space, then Q is square. Every $\boldsymbol{b} = QQ^T\boldsymbol{b}$ is the sum of its components along the $\boldsymbol{q}$'s.

$$\boldsymbol{b} = \boldsymbol{q}_1(\boldsymbol{q}_1^T\boldsymbol{b}) + \ldots + \boldsymbol{q}_n(\boldsymbol{q}_n^T\boldsymbol{b}) \tag{5}$$

Remark. *$QQ^T = I$ is the foundation of Fourier series and all great "transforms" of applied mathematics. They break vectors $\boldsymbol{b}$ or functions $f(x)$ into perpendicular pieces. Then by adding the pieces in Equation 6, the inverse transform puts $\boldsymbol{b}$ and $f(x)$ back together.*

4.2 Gram-Schmidt Process

Example 8. Start with three independent vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$. Our goal is to make three orthogonal vectors $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. At the end, we divide each orthogonal vector by their lengths. The result is $\boldsymbol{q}_1 = \boldsymbol{A}/||A||$, $\boldsymbol{q}_2 = \boldsymbol{B}/||B||$, $\boldsymbol{q}_3 = \boldsymbol{C}/||C||$.

Begin by choosing the first orthogonal vector as:

$$\boldsymbol{A} = \boldsymbol{a}$$

The next direction $\boldsymbol{B}$ must be perpendicular to $\boldsymbol{A}$. Start with $\boldsymbol{b}$ and subtract its projection along $\boldsymbol{A}$:

$$\boldsymbol{B} = \boldsymbol{b} - \frac{\boldsymbol{A}^T\boldsymbol{b}}{\boldsymbol{A}^T\boldsymbol{A}}\boldsymbol{A}$$

$\boldsymbol{B}$ is the same as the error vector $\boldsymbol{e}$ perpendicular to $\boldsymbol{A}$. For the next Gram-Schmidt step, we subtract off C's components in the two directions of A and $\boldsymbol{B}$ to get a perpendicular direction $\boldsymbol{C}$.

$$\boldsymbol{C} = \boldsymbol{c} - \frac{\boldsymbol{A}^T\boldsymbol{c}}{\boldsymbol{A}^T\boldsymbol{A}}\boldsymbol{A} - \frac{\boldsymbol{B}^T\boldsymbol{c}}{\boldsymbol{B}^T\boldsymbol{B}}\boldsymbol{B}$$

The goal of the Gram-Schmidt process is to subtract from every new vector its projections in the directions already set. Finally, the next step is to normalize the orthogonal vectors $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

4.3 Factorization of $A = QR$

The goal of QR factorization is to relate A and Q using $A = QR$, where R is a triangular matrix. QR decomposition is Gram-Schmidt in a nutshell.

$$A = QR$$
$$\begin{bmatrix} | & | & | \\ \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \\ | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | \\ \boldsymbol{q}_1 & \boldsymbol{q}_2 & \boldsymbol{q}_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1^T\boldsymbol{a} & \boldsymbol{q}_1^T\boldsymbol{b} & \boldsymbol{q}_1^T\boldsymbol{c} \\ & \boldsymbol{q}_2^T\boldsymbol{b} & \boldsymbol{q}_2^T\boldsymbol{c} \\ & & \boldsymbol{q}_3^T\boldsymbol{c} \end{bmatrix}$$

To get R, we simply multiply both sides by Q^T then we get $R = Q^TA$.

Definition 4.2. From independent vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, Gram-Schmidt constructs orthonormal vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$. The matrices with these columns satisfy $A = QR$. Then $R = Q^TA$ is upper triangular because later $\boldsymbol{q}$'s are orthogonal to earlier $\boldsymbol{a}$'s

Any $m \times n$ matrix A with independent column vectors can be factored into $A = QR$. The $m \times n$ matrix Q has orthonormal columns, and the square matrix R is upper triangular with a positive diagonal. QR factorization is important to least squares because:

$$A^TA = (QR)^T(QR) = R^TQ^TQR = R^TR \tag{6}$$

Then the least squares equation $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ simplifies to $R^TR\hat{\boldsymbol{x}} = R^TQ^T\boldsymbol{b}$. Then we reach:

$$R\hat{\boldsymbol{x}} = Q^T\boldsymbol{b} \tag{7}$$

Remark. *Least squares becomes* $R^TR\hat{\boldsymbol{x}} = R^TQ^T\boldsymbol{b}$ *or* $R\hat{\boldsymbol{x}} = Q^T\boldsymbol{b}$ *or* $\boxed{\hat{\boldsymbol{x}} = R^{-1}Q^T\boldsymbol{b}}$.

5 Problems

Problem 1.1. asd

Solution. soln □