6. Orthogonality and Least-Squares

6.1 Inner product, length, and orthogonality

**Definition**

The inner product of two vector \( \mathbf{u} \) and \( \mathbf{v} \) in \( \mathbb{R}^n \) is written \( \mathbf{u} \cdot \mathbf{v} \).

If \( \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \), \( \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \), then \( \mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n \).

**Theorem 1**

(a) \( \mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u} \)

(b) \( (\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w} \)

(c) \( (c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v}) \)

(d) \( \mathbf{u} \cdot \mathbf{u} \geq 0 \), and \( \mathbf{u} \cdot \mathbf{u} = 0 \iff \mathbf{u} = \mathbf{0} \).
Definition
The length (or norm) of $\mathbf{v}$ is the non-negative scalar $||\mathbf{v}||$ defined by
$$||\mathbf{v}|| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + \cdots + v_n^2} \text{ and } ||\mathbf{v}||^2 = \mathbf{v} \cdot \mathbf{v}.$$ 

Note: $||c\mathbf{v}|| = |c||\mathbf{v}||$.

Definition
The distance between $\mathbf{u}$ and $\mathbf{v}$, written as $\text{dist}(\mathbf{u}, \mathbf{v})$, is the length of the vector $\mathbf{u} - \mathbf{v}$. That is, $\text{dist}(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}||$.

Orthogonal vectors

Definition
Two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$ are orthogonal if $\mathbf{u} \cdot \mathbf{v} = 0$.

Theorem 2 (The Pythagorean Theorem, 畢氏定理)
Two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $||\mathbf{u} + \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2$.

$$(L \cos \theta)^2 + (L \sin \theta)^2 = L^2$$
Orthogonal complements

Definition
The set of all vectors $u$ that are orthogonal to every vector $w$ in $W$, then we say that the set is the orthogonal complement of $W$, and denote by $W^\perp$. (next page)

Note
1. A vector $x$ is in $W^\perp$ if and only if $x$ is orthogonal to every vector in a set that spans subspace $W$.
2. $W^\perp$ is a subspace of $R^n$ for any subset $W$ of $R^n$.
3. Zero vector is orthogonal to any vector.

Theorem 3
Let $A$ be an $m \times n$ matrix. Then

\begin{align*}
(i) \quad & (\text{Row } A)^\perp = \text{Nul } A. \quad \{\mathbf{x} \mid A^T \mathbf{x} = 0\} \equiv \{\mathbf{x} \mid \mathbf{x}^T A = 0\} \equiv \\
(ii) \quad & (\text{Col } A)^\perp = \text{Nul } A^T. \quad \text{left null space of } A)
\end{align*}

\[W^\perp = \{ x \mid x \in R^n, x \perp w_i's \} \]
\( W \perp \) is a subspace of \( R^n \) for any subset \( W \) of \( R^n \).

If \( w \) is an arbitrary vector in \( W \), then
1. \( 0 \in W \perp \)
2. If \( u \in W \perp \),
   then \( cu \cdot w = c (u \cdot w) = 0 \); thus \( cu \in W \perp \)
3. If \( u \) and \( v \in W \perp \),
   then \( (u+v) \cdot w = u \cdot w + v \cdot w = 0 \);
   thus \( (u+v) \in W \perp \)

\( W \) is a subset of \( R^n \), then \( (W \perp) \perp = W \) ?
If \( W \) is a subspace of \( R^n \), then \( (W \perp) \perp = W \) ?

Exercises of Section 6.1.

6.2 Orthogonal sets

A set of vectors \( \{u_1, \ldots, u_n\} \) in \( R^n \) is said to be an orthogonal set if \( u_i \cdot u_j = 0 \) for all \( i \neq j \).

Theorem 4
If \( S = \{u_1, \ldots, u_p\} \) is an orthogonal set of nonzero vectors in \( R^n \), then \( S \) is linearly independent and hence is a basis for the subspace spanned by \( S \).

An orthogonal basis for a subspace \( W \) of \( R^n \) is a basis for \( W \) that is also an orthogonal set.
Theorem 5

Let \{u_1, \ldots, u_p\} be an orthogonal basis for a subspace \(W\) of \(\mathbb{R}^n\). Then each \(y\) in \(W\) has a unique representation as a linear combination of \(u_1, \ldots, u_p\). In fact, if
\[
y = c_1 u_1 + \ldots + c_p u_p
\]
then
\[
c_j = \frac{y \cdot u_j}{u_j \cdot u_j}, \quad j = 1, 2, \ldots, p.
\]

Note

1. \(y_j = c_j u_j = \frac{y \cdot u_j}{u_j \cdot u_j} u_j\) is the orthogonal projection of \(y\) onto \(u_j\).
2. \(y - y_j = y - \frac{y \cdot u_j}{u_j \cdot u_j} u_j\) is component of \(y\) orthogonal to \(u_j\).

A set \(\{u_1, \ldots, u_p\}\) is an orthonormal set if it is an orthogonal set of unit vectors.

Normalize the length of vector \(u = [u_1 \ldots u_n]^T\) to \(u'\)
\[
u' = \frac{1}{\sqrt{u_1^2 + \ldots + u_n^2}} [u_1 \ldots u_n]^T.
\]

Theorem 6

An \(m \times n\) matrix \(U\) has orthonormal columns if and only if \(U^T U = I\).

Proof.
\[
U^T U = \begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix} [u_1 \ldots u_n] = \begin{bmatrix} u_1^T u_1 & \ldots & u_1^T u_n \\ \vdots & \ddots & \vdots \\ u_n^T u_1 & \ldots & u_n^T u_n \end{bmatrix}.
\]
Theorem 7

Let \( U \) be an \( m \times n \) matrix with orthonormal columns, and let \( x \) and \( y \) be in \( \mathbb{R}^n \). Then

(a) \( \|Ux\| = \|x\| \) (preserving length)
(b) \( (Ux) \cdot (Uy) = x \cdot y \) (preserving orthogonality)
(c) \( (Ux) \cdot (Uy) = 0 \) if and only if \( x \cdot y = 0 \)

Proof.

(a) \( Ux = [u_1 \ u_2 \ \ldots \ u_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 u_1 + x_2 u_2 + \ldots + x_n u_n \)

\[
\|Ux\| = \sqrt{(Ux) \cdot (Ux)} = \sqrt{(x_1 u_1 + \ldots + x_n u_n) \cdot (x_1 u_1 + \ldots + x_n u_n)}
\]
\[
= \sqrt{x_1 u_1 \cdot (x_1 u_1 + \ldots + x_n u_n) + \ldots + x_n u_n \cdot (x_1 u_1 + \ldots + x_n u_n)}
\]
\[
= \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{x \cdot x} = \|x\|.
\]

(b) \( (Ux) \cdot (Uy) \)
\[
= (x_1 u_1 + \ldots + x_n u_n) \cdot (y_1 u_1 + \ldots + y_n u_n)
\]
\[
= x_1 y_1 + x_2 y_2 + \ldots + x_n y_n
\]

Exercises of Section 6.2.
6.3 Orthogonal projections

**Purpose**
To find the projection of a vector on a subspace.

**Theorem 8 (The Orthogonal Decomposition Theorem)**
Let $W$ be a subspace of $\mathbb{R}^n$. Then each $y$ in $\mathbb{R}^n$ can be written uniquely in the form $y = \hat{y} + z$, where $\hat{y}$ is in $W$ and $z$ is in $W^\perp$. In fact, if $\{u_1, \ldots, u_p\}$ is any orthogonal basis of $W$, then

$$\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \cdots + \frac{y \cdot u_p}{u_p \cdot u_p} u_p$$

and $z = y - \hat{y}$.

The vector $\hat{y}$ in Eq.(1) is called the orthogonal projection of $y$ onto $W$ and is often written as $\text{proj}_W y$.

**For example,**
The orthogonal projection of $y$ onto $W$.

**Ex.2.**
Let $u_1 = \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix}$, $u_2 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$, and $y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. Observe that $\{u_1, u_2\}$ is an orthogonal basis for $W = \text{Span} \{u_1, u_2\}$. Write $y$ as the sum of vector in $W$ and vector orthogonal to $W$. 
Solution.
The orthogonal projection of $y$ onto $W$ is
\[
\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \frac{y \cdot u_2}{u_2 \cdot u_2} u_2
\]
\[
= \frac{9}{30} \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix} + \frac{3}{6} \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = \frac{9}{30} \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix} + \frac{15}{30} \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{30} \begin{bmatrix} 60 \\ 6 \end{bmatrix} = \begin{bmatrix} -2/5 \\ 2 \\ 1/5 \end{bmatrix}
\]
So
\[
y - \hat{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \begin{bmatrix} -2/5 \\ 2 \\ 1/5 \end{bmatrix} = \begin{bmatrix} 7/5 \\ 0 \\ 14/5 \end{bmatrix}
\]
The desired decomposition of $y$ is
\[
y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -2/5 \\ 2 \\ 1/5 \end{bmatrix} + \begin{bmatrix} 7/5 \\ 0 \\ 14/5 \end{bmatrix}
\]

**Properties of orthogonal projections**

- If $\{u_1, u_2, \ldots, u_p\}$ is an orthogonal basis for $W$ and
  if $y$ is in $W = \text{Span}\{u_1, u_2, \ldots, u_p\}$, then $\text{proj}_W y = y$.

**Theorem 9 (The Best approximation theorem)**

Let $W$ be a subspace of $\mathbb{R}^n$, $y$ be any vector in $\mathbb{R}^n$, and $\hat{y}$ be the orthogonal projection of $y$ onto $W$. Then $\hat{y}$ is the closest point in $W$ to $y$, in the sense that
\[
\|y - \hat{y}\| < \|y - v\|
\]
for all $v$ in $W$ distinct from $\hat{y}$. 

![Diagram of orthogonal projection](image)
Ex. 4. The distance from a point $y$ in $\mathbb{R}^n$ to a subspace $W$ is defined as the distance from $y$ to the nearest point in $W$. Find the distance from $y$ to

$$W = \text{Span} \{ u_1, u_2 \}, \quad \text{where} \quad y = \begin{bmatrix} -1 \\ -5 \\ 10 \end{bmatrix}, \quad u_1 = \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}.$$ 

Solution.

By the best approximation theorem, the distance from $y$ to $W$ is $\|y - \hat{y}\|$, where $\hat{y} = \text{proj}_W y$. Since $\{ u_1, u_2 \}$ is orthogonal basis for $W$,

$$\hat{y} = \frac{15}{30} u_1 + \frac{-21}{6} u_2 = \frac{1}{2} \begin{bmatrix} 5 \\ -2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ -8 \\ 4 \end{bmatrix}.$$

$$y - \hat{y} = \begin{bmatrix} -1 \\ -5 \\ 10 \end{bmatrix} - \begin{bmatrix} -1 \\ -8 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 6 \end{bmatrix}$$

$$\|y - \hat{y}\|^2 = 3^2 + 6^2 = 45. \text{ The distance from } y \text{ to } W \text{ is } \sqrt{45} = 3\sqrt{5}.$$ 

Theorem 10

If $\{ u_1, \ldots, u_p \}$ is an orthonormal basis for a subspace $W$ of $\mathbb{R}^n$, then

$$\text{proj}_W y = (y \cdot u_1) u_1 + (y \cdot u_2) u_2 + \ldots + (y \cdot u_p) u_p \quad (4)$$

If $U = [u_1 \ u_2 \ \ldots \ u_p]$, then

$$\text{proj}_W y = U U^T y \quad \text{for all } y \in \mathbb{R}^n. \quad (5)$$

Proof.

$$\text{proj}_W y = (y \cdot u_1) u_1 + (y \cdot u_2) u_2 + \ldots + (y \cdot u_p) u_p = U \begin{bmatrix} y \cdot u_1 \\ y \cdot u_2 \\ \vdots \\ y \cdot u_p \end{bmatrix}$$

Since $y \cdot u_i = u_i \cdot y = u_i^T y$

$$\text{proj}_W y = U \begin{bmatrix} u_1^T y \\ u_2^T y \\ \vdots \\ u_p^T y \end{bmatrix} = U (U^T y) = U U^T y$$

Exercises of Section 6.3.
6.4 The Gram-Schmidt process

- **Purpose**
  The Gram-Schmidt process is a simple algorithm for producing an orthogonal or orthonormal basis for nonzero subspace of $\mathbb{R}^n$.

- **Theorem 11 (The Gram-Schmidt process)**
  Let a basis $\{x_1, \ldots, x_p\}$ for a subspace $W$ of $\mathbb{R}^n$, define $v_1 = x_1$, $v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1$, $v_3 = x_3 - \frac{x_3 \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_3 \cdot v_2}{v_2 \cdot v_2} v_2$, ..., $v_p = x_p - \frac{x_p \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_p \cdot v_{p-1}}{v_{p-1} \cdot v_{p-1}} v_{p-1}$.

  Then $\{v_1, \ldots, v_p\}$ is an orthogonal basis for $W$. In addition, $\text{Span} \{v_1, \ldots, v_k\} = \text{Span} \{x_1, \ldots, x_k\}$ for $1 \leq k \leq p$.

- **The geometric meaning of the Gram-Schmidt process**
  (A three-dimensional case)

  Note that the angle between $x_i$ and $v_i$ is less than 90°.

  $$Q^T A = \begin{bmatrix} -v_1 & -v_2 & -v_3 \\ \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$
Proof.

\[ x_2 = \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1 + \frac{x_2 \cdot v_2}{v_2 \cdot v_2} v_2 \]

\[ \Rightarrow \frac{x_2 \cdot v_2}{v_2 \cdot v_2} v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} v_1. \]

Orthonormal bases

- An orthonormal basis is constructed easily from an orthogonal basis \( \{ v_1, \ldots, v_p \} \); simply normalize all \( v_k \) by \( v_k / \| v_k \| \rightarrow v_k \).

QR factorization of matrices

- This factorization is widely used in computer algorithm for various computations, such as solving equations and finding eigenvalues. (Exercise 23 of Section 5.2.)

**Theorem 12** *(The QR factorization)*

If \( A \) is an \( m \times n \) matrix with linearly independent columns, then \( A \) can be factored as \( A = QR \), where \( Q \) is an \( m \times n \) matrix whose columns form an orthonormal basis for \( \text{Col} \ A \) and \( R \) is an \( n \times n \) upper triangular invertible matrix with positive entries on its diagonal.

**Ex. 4.**

Factor \( A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \) into \( QR \).

Answer.

(a) Find \( Q \).

\[ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \]

orthogonalize to

\[ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix} \]
(b) To find $R$.

Since $A = QR$ and $Q^TQ = I$ (Theorem 6 of page 10)

$\Rightarrow Q^T A = Q^T Q R \Rightarrow Q^T A = R$.

$\Rightarrow R = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
-\frac{3}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} \\
0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}}
\end{bmatrix} \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 0 & 2
\end{bmatrix} = \begin{bmatrix}
2 & \frac{3}{\sqrt{12}} & 1 \\
0 & \frac{3}{\sqrt{12}} & 2 & \frac{\sqrt{12}}{2} \\
0 & 0 & 2 & \frac{\sqrt{6}}{2}
\end{bmatrix}$.

Exercises of Section 6.4.

Problem \[ A = QR \text{ and } Q^T Q = I \]

$\Rightarrow Q^T A = Q^T Q R \Rightarrow Q^T A = R$

$\Rightarrow Q Q^T A = Q R = A$

$\Rightarrow Q Q^T = I$ ?

No!

$Q = \begin{bmatrix}
\frac{1}{2} & \frac{3}{\sqrt{12}} & 0 \\
\frac{1}{2} & \frac{1}{\sqrt{12}} & -\frac{2}{\sqrt{6}} \\
\frac{1}{2} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{6}} \\
\frac{1}{2} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{6}}
\end{bmatrix} \Rightarrow Q Q^T = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1/2 & 1/2 \\
0 & 0 & 1/2 & 1/2
\end{bmatrix}$

$\Rightarrow Q Q^T A = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1/2 & 1/2 \\
0 & 0 & 1/2 & 1/2
\end{bmatrix} \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 \\
1 & 1 & 0 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{bmatrix} = A.$
6.5 Least-squares problems

For a linear system, $Ax = b$, a solution is demanded but none exists. The best one can do is to find an $x$ that makes $Ax$ as close as possible to $b$. The general least-squares problem is just to find an $x$ that makes $\|b - Ax\|$ as small as possible. The term least-squares arises from the fact that $\|b - Ax\|$ is the square root of a sum of squares.

Definition

If $A$ is an $m \times n$ matrix and $b$ is in $\mathbb{R}^m$, a least-squares solution of $Ax = b$ is an $\hat{x}$ in $\mathbb{R}^n$ such that

$$\|b - A\hat{x}\| \leq \|b - Ax\| \text{ for all } x \text{ in } \mathbb{R}^n.$$ 

Theorem 9 on page 16

Note

No matter that $x$ is selected, the vector $Ax$ will necessarily be in the column space $\text{Col} \ A$. So we seek an $x$ that makes $Ax$ the closest point in $\text{Col} \ A$ to $b$.

If $\hat{b}$ is the orthogonal projection of $b$ on $\text{Col} \ A$, then $A\hat{x} = \hat{b}$.

It means that $Ax = \hat{b}$ is consistent and there is a solution $\hat{x}$ in $\mathbb{R}^n$. By the orthogonal decomposition principle, the projection $\hat{b}$ has the property that $b - \hat{b}$ is orthogonal to $\text{Col} \ A$, so $b - A\hat{x}$ is orthogonal to each column of $A$. If $a_j$ is any column of $A$, then $a_j$ is orthogonal to $(b - A\hat{x})$ and

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

Since each $a_j^T$ is a row of $A^T$,

$$A^T(b - A\hat{x}) = 0 \implies A^T b - A^T A \hat{x} = 0 \implies A^T A \hat{x} = A^T b.$$
1. **Theorem 13**
   The set of least-squares solutions of \( Ax = b \) coincides with the nonempty set of solutions of the normal equations \( A^T A \hat{x} = A^T b \). If \( A^T A \) is invertible, then \( \hat{x} = (A^T A)^{-1} A^T b \).

2. **Ex. 1.**
   Find a least-squares solution of the inconsistent system
   \[
   Ax = b \quad \text{for} \quad A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}.
   \]

   Solution. \( \hat{x} = \left( \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 4 & 0 & 2 \\ 0 & 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 5 & -1 & 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} \]
   \[
   = \frac{1}{84} \begin{bmatrix} 84 \\ 168 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.
   \]

3. **Ex. 2.** Find a least-squares solution of \( Ax = b \) for

   \[
   A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix},
   \]
   where \( A^T A \) is not invertible.

   Solution. \( A^T A \hat{x} = A^T b \) \( \Rightarrow \)

   \[
   \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \\ 2 \\ 6 \end{bmatrix}
   \]

   The augmented matrix
   \[
   \begin{bmatrix} 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \end{bmatrix} \rightarrow ... \rightarrow \begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}
   \]

   \( \Rightarrow \)

   \[
   \begin{bmatrix} x_1 \\ x_2 - x_4 = 3 \\ x_3 - x_4 = -5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ -5 + x_4 \\ -2 + x_4 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ -5 \\ -2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.
   \]
Theorem 14
The matrix $A^TA$ is invertible if and only if the columns of $A$ are linearly independent. In this case, the equation $Ax = b$ has only one least-squares solution $\hat{x}$, and it is given by $\hat{x} = (A^TA)^{-1}A^Tb$.

Note
The distance from $b$ to $A\hat{x}$ is called the least-squares error.

Alternative calculations of least-square solutions

Meaning
In some cases, the normal equations for a least-squares problem can be ill-conditioned; that is, small errors in the calculations of the entries of $A^TA$ can sometimes cause relatively large errors in the solution $\hat{x}$. If the columns of $A$ are linearly independent, the least-squares solution can often be computed more reliably through a $QR$ factorization of $A$.

ill-conditioned
The condition number associated with the linear equation $Ax = b$ gives a bound on how inaccurate the solution $x$ will be after approximate solution.

Conditioning is a property of the matrix, not the algorithm or floating point accuracy of the computer used to solve the corresponding system.

The rate at which the solution $x$ will change with respect to a change in $b$. Thus, if the condition number is large, even a small error in $b$ may cause a large error in $x$. 
The condition number is the maximum ratio of the relative error in $x$ divided by the relative error in $b$,

$$\frac{\|A^{-1}e\|}{\|A^{-1}b\|} = \|A\|\|A^{-1}\|,$$

where $\|x\|$ is 2 norm.

* For example,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 2.001 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

where condition number is

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix} \begin{bmatrix} 1001 & -1000 \\ -1000 & 1000 \end{bmatrix} = \sqrt{4.002001^2 + 4002001} = 4002$$

* The larger condition number is, the more ill conditioned the coefficient matrix is.

---

**Theorem 15**

Given an $m \times n$ matrix $A$ with linearly independent columns, let $A = QR$ be a $QR$ factorization of $A$. Then for each $b$ in $\mathbb{R}^m$, the equation $Ax = b$ has a unique least-squares solution,

$$\hat{x} = R^{-1}Q^Tb.$$

**Proof.**

$$\hat{x} = (A^TA)^{-1}A^Tb \quad \text{since} \quad A = QR$$

$$= (R^TQ^TQR)^{-1}R^TQ^Tb$$

$$= (R^TIR)^{-1}R^TQ^Tb$$

$$= (R^TR)^{-1}R^TQ^Tb$$

$$= R^{-1}(R^T)^{-1}R^TQ^Tb$$

$$= R^{-1}Q^Tb.$$
Note

Since $R$ is an upper triangular matrix, $\hat{X}$ should be calculated from the equation

$$R \hat{X} = Q^T b.$$

It is much faster to solve this equation by back-substitution or row operations than to compute $R^{-1}$ and use the preceding equation.

Ex.5.

Find the least-squares solution of $Ax = b$ for

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \\ 1 & 3 & 3 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 3 \\ 5 \\ 7 \\ -3 \end{bmatrix}.$$

Solution.

$$A = QR = \begin{bmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & -1/2 \\ 1/2 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix}.$$

$$R\hat{X} = Q^T b = \begin{bmatrix} 2 & 4 & 5 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -6 \\ 4 \end{bmatrix}.$$

$$\Rightarrow \hat{X} = \begin{bmatrix} 10 \\ -6 \\ 2 \end{bmatrix}.$$

Exercises for Section 6.5.