Chap 3. Orthogonality

3.1 Orthogonal Vectors and Subspaces

For \( \mathbf{x} = (x_1, \ldots, x_n) \), \( \|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}} \): the length of \( \mathbf{x} \),
\[
\|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2
\]
\( \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \): orthogonal \( \iff \) \( \mathbf{x}^T \mathbf{y} = x_1 y_1 + \cdots + x_n y_n = 0 \)
the inner product of \( \mathbf{x} \) and \( \mathbf{y} \).

3A

The inner product \( \mathbf{x}^T \mathbf{y} = 0 \) iff \( \mathbf{x} \) and \( \mathbf{y} \) are orthogonal.
If \( \mathbf{x}^T \mathbf{y} > 0 \), their angle is less than 90°.
If \( \mathbf{x}^T \mathbf{y} < 0 \), their angle is greater than 90°.

Useful Fact: If nonzero vectors \( \mathbf{v}_1, \ldots, \mathbf{v}_k \) are mutually orthogonal, then those vectors are linearly independent.

Suppose \( c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k = \mathbf{0} \).
\[ \Rightarrow \mathbf{v}_i (c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k) = c_i \mathbf{v}_i^T \mathbf{v}_i = 0 \]
Since \( \mathbf{v}_i \) are nonzero, \( \mathbf{v}_i^T \mathbf{v}_i \neq 0 \) and \( c_i = 0 \).

The coordinate vectors \( \mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n \) are mutually orthogonal unit vectors. When they are rotated, the result is a new orthonormal basis: a new system of mutually orthogonal unit vectors in \( \mathbb{R}^n \). In \( \mathbb{R}^3 \), \( \mathbf{v}_1 = (\cos \theta, \sin \theta, 0), \mathbf{v}_2 = (-\sin \theta, \cos \theta, 0) \) form an orthonormal basis.

Orthogonal Subspaces

3B Two subspaces \( \mathbf{V} \) and \( \mathbf{W} \) of \( \mathbb{R}^n \) are orthogonal if every vector \( \mathbf{v} \in \mathbf{V} \) is orthogonal to every \( \mathbf{w} \in \mathbf{W} \):
\[ \mathbf{v}^T \mathbf{w} = 0 \] for all \( \mathbf{v} \) and \( \mathbf{w} \).
3C Fundamental Theorem of Orthogonality
The row space is orthogonal to the nullspace (in \( \mathbb{R}^n \)).
The column space is orthogonal to the left nullspace (in \( \mathbb{R}^m \)).
\[ \begin{align*}
\forall x \in \text{N}(A) & \iff A x = 0 \\
\forall v \in \text{C}(A^T) & \iff v = A^T z \text{ for some } z \\
\Rightarrow v^T x &= z^T A x = 0 \\
\therefore \text{N}(A) & \perp \text{C}(A^T)
\end{align*} \]

Definition (Orthogonal Complement)
\( V \): a subspace of \( \mathbb{R}^n \),
\( V^\perp \): the orthogonal complement of \( V \) is the space of all vectors orthogonal to \( V \).
\( \therefore V^\perp \)

3D Fundamental Theorem of Linear Algebra, Part II
The nullspace is the orthogonal complement of the row space in \( \mathbb{R}^n \).
The left nullspace is the orthogonal complement of the column space in \( \mathbb{R}^m \).

3E \( A x = b \) is solvable iff \( y^T 1 b = 0 \) whenever \( y^T A = 0 \)
\[ \begin{align*}
\forall 0 b \in \text{C}(A) & \iff 1 b \in (\text{N}(A^T))^+ \\
\therefore A x = 1 b \text{ is solvable} & \iff b_1 + b_2 + b_3 = 0 .
\end{align*} \]

Ex:
\[ A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \Rightarrow y = (1, 1, 1)^T \in \text{N}(A^T) \text{ since } y^T A = 0 .
\]
Thus, \( A x = 1 b \) is solvable \( \iff b_1 + b_2 + b_3 = 0 . \)
The Matrix and the Subspaces

If $W = V^\perp$, then $V = W^\perp$ and $\dim V + \dim W = n$. In other words $V^\perp = V$. The whole space $\mathbb{R}^n$ can be decomposed into two perpendicular parts.

**3F** From the row space to the column space, $A$ is actually invertible. Every vector $b$ in the column space comes from exactly one vector $x_r$ in the row space.

![Diagram](image)

**Figure 3.4** The true action $Ax = A(x_{\text{row}} + x_{\text{null}})$ of any $m$ by $n$ matrix.

\[ 0 \leq \|b\| \leq \text{C}(A), \quad \|b\| = Ax = Ax_r, \quad \text{where} \quad x = x_r + x_n \quad \text{and} \quad Ax_n = 0. \]

Let $x_r'$ be in the row space and $Ax_r' = \lambda b$, then $A(x_r - x_r') = \lambda b - \lambda b = 0$, and $x_r - x_r' \in N(A)$ and $x_r - x_r'$ is also in the row space $\Rightarrow x_r = x_r'$.

Exactly one vector in the row space is carried to $\lambda b$.

$A^T$ goes from $\mathbb{R}^m$ to $\mathbb{R}^n$ and from $C(A)$ to $C(A^T)$. But $A^T \neq A^T$ in general. $A^T$ moves the spaces correctly, but not the individual vectors. When $A^T$ fails to exist, the best substitute is the pseudo-inverse $A^+$:

$A^+Ax = x$ for $x \in C(A^T)$ and $A^+y = 0$ for $y \in N(A^T)$. 
3.2 Cosines and Projections onto Lines

**Figure 3.6** The cosine of the angle \( \theta = \beta - \alpha \) using inner products.

\[
\cos \Theta = \cos \beta \cos \alpha + \sin \beta \sin \alpha = \frac{a_1 b_1 + a_2 b_2}{||a|| ||b||}
\]

### Formula

\[
\cos \Theta = \frac{a^T b}{||a|| ||b||}
\]

### Law of Cosines

\[
||b - a||^2 = ||b||^2 + ||a||^2 - 2 ||b|| ||a|| \cos \Theta
\]

\[
\Rightarrow ||b||^2 - 2 \bar{a}^T b + \bar{a}^T a = ||b||^2 + \bar{a}^T a - 2 ||b|| ||a|| \cos \Theta
\]

\[
\Rightarrow \bar{a}^T b = ||a|| ||b|| \cos \Theta
\]

**Figure 3.7** The projection \( p \) of \( b \) onto \( a \), with \( \cos \theta = \frac{Op}{Ob} = \frac{a^T b}{||a|| ||b||} \).

### 3H

The projection of \( b \) onto the line in the direction of \( a \) is

\[
1p = \hat{x}a = \frac{a^T b}{a^T a} a
\]

\[
\Rightarrow (b - \hat{x}a) \perp a
\]

\[
\Rightarrow a^T (b - \hat{x}a) = 0
\]
All vectors $\mathbf{a}$ and $\mathbf{b}$ satisfy the Schwartz inequality:

$$|\mathbf{a}^T \mathbf{b}| \leq ||\mathbf{a}|| \cdot ||\mathbf{b}||,$$

which is $|\cos \theta| \leq 1$ in $\mathbb{R}^n$.

$$\left\| \mathbf{b} - \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \mathbf{a} \right\|^2 = \mathbf{b}^T \mathbf{b} - 2 \frac{(\mathbf{a}^T \mathbf{b})^2}{\mathbf{a}^T \mathbf{a}} + \left( \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \right)^2 \mathbf{a}^T \mathbf{a}$$

$$= \frac{(\mathbf{b}^T \mathbf{b})(\mathbf{a}^T \mathbf{a}) - (\mathbf{a}^T \mathbf{b})^2}{\mathbf{a}^T \mathbf{a}} \geq 0$$

The equality holds iff $\mathbf{b}$ is a multiple of $\mathbf{a}$.

**Ex 1:** Project $\mathbf{b} = (1,2,3)$ onto the line through $\mathbf{a} = (1,1,1)$.

$$\mathbf{x} = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} = \frac{6}{3} = 2 \text{ and } \mathbf{p} = 2 \mathbf{a} = (2,2,2).$$

$$\cos \theta = \frac{||\mathbf{p}||}{||\mathbf{b}||} = \frac{\sqrt{12}}{\sqrt{14}} \text{ and } \cos \theta = \frac{\mathbf{a}^T \mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||} = \frac{6}{\sqrt{14}}.$$

$$\Rightarrow 6 \leq \sqrt{3} \sqrt{14} \text{ and } \sqrt{36} \leq \sqrt{42}.$$ The cosine is less than 1 since $\mathbf{b}$ is not parallel to $\mathbf{a}$.

**Projection Matrix of Rank 1**

$$\mathbf{P} = \mathbf{a} \cdot \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \Rightarrow \mathbf{P} = \mathbf{a} \cdot \frac{\mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} : \text{ projection matrix}$$

1. $\mathbf{P}$ is a symmetric matrix
2. Its square is itself: $\mathbf{P}^2 = \mathbf{P}$.

**Ex 2:** $\mathbf{a} = (1,1,1)$

$$\mathbf{P} = \frac{\mathbf{a} \cdot \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$
Remark on scaling:

The projection matrix is the same if \( \mathbf{a} \) is scaled.

\[
\mathbf{a} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{gives} \quad P = \frac{1}{3x^2} \begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{bmatrix}
\]

If \( \mathbf{a} \) has a unit length, \( P = \mathbf{a} \cdot \mathbf{a}^T \) (\( \mathbf{a}^T \cdot \mathbf{a} = 1 \)).

Ex 3: Project onto the \( \theta \)-direction in the \( xy \)-plane.

The line goes through \( \mathbf{a} = (\cos \theta, \sin \theta) \).

\[
P = \frac{\mathbf{a} \cdot \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} = \frac{1}{\cos^2 \theta + \sin^2 \theta} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \begin{bmatrix} \cos \theta & \sin \theta \end{bmatrix} = \begin{bmatrix} \cos^2 \theta & \cos \theta \sin \theta \\ \cos \theta \sin \theta & \sin^2 \theta \end{bmatrix}
\]

This matrix \( P \) was discovered in Section 2.6.

Now we know \( P \) in any dimensions.

To project \( \mathbf{b} \) onto \( \mathbf{a} \), multiply by the projection matrix \( P \).

\( P = Pb \).
3.3 Projections and Least Squares

Least squares solution for $a x = b$ by minimizing

$$E^2 = ||a x - b||^2 = (a_1 x - b_1)^2 + \cdots + (a_m x - b_m)^2.$$

$$\frac{1}{2} \frac{dE^2}{dx} = (a_1 x - b_1) a_1 + \cdots + (a_m x - b_m) a_m = 0$$

The least squares solution to a problem $a x = b$

in one unknown is

$$\hat{x} = a^T b / a^T a.$$

Orthogonality:

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

of $a$ and $\hat{x}$

$$a^T b - \frac{a^T b}{a^T a} a^T a = 0$$

$\hat{x}$: the error vector

Least-Squares Problem with Several Variables

$$A x = b,$$ where $A$ is an $m \times n$ matrix.

$$E = ||Ax - b||: \text{ error, which is the distance from } b$$

to the point $A \hat{x}$ in the column space.

0. Locate the point $p = A \hat{x}$ that is closer to $b$

than any other point in the column space. the column

0. The error $e = b - A \hat{x}$ must be perpendicular to $b$.

Figure 3.8  Projection onto the column space of a 3 by 2 matrix.
(1) All vectors perpendicular to the column space lie in the left nullspace. Thus the error \( \mathbf{e} = \mathbf{lb} - \mathbf{Ax} \) must be in the nullspace of \( \mathbf{A}^T \):
\[
\mathbf{A}^T (\mathbf{lb} - \mathbf{Ax}) = \mathbf{0} \quad \text{or} \quad \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{lb}.
\]

2. The error must be perpendicular to each column \( \mathbf{a}_1, \ldots, \mathbf{a}_n \) of \( \mathbf{A} \):
\[
\begin{align*}
\mathbf{a}_1^T (\mathbf{lb} - \mathbf{Ax}) &= 0 \\
\vdots & \quad \text{or} \quad \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix} \begin{bmatrix} \mathbf{lb} - \mathbf{Ax} \end{bmatrix} &= \mathbf{0} \\
\mathbf{a}_n^T (\mathbf{lb} - \mathbf{Ax}) &= 0
\end{align*}
\]

This is again \( \mathbf{A}^T (\mathbf{lb} - \mathbf{Ax}) = \mathbf{0} \) and \( \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{lb} \).

3. Taking partial derivatives of \( E^2 = (\mathbf{Ax} - \mathbf{lb})^T (\mathbf{Ax} - \mathbf{lb}) \) gives the same \( 2\mathbf{A}^T \mathbf{Ax} - 2\mathbf{A}^T \mathbf{lb} = \mathbf{0} \).

Using a symmetric square matrix

When \( \mathbf{Ax} = \mathbf{lb} \) is inconsistent, its least-squares solution minimizes \( \| \mathbf{Ax} - \mathbf{lb} \|^2 \):

Normal equations: \( \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{lb} \).

\( \mathbf{A}^T \mathbf{A} \) is invertible exactly when the columns of \( \mathbf{A} \) are linearly independent! Then,

Best estimate \( \mathbf{x} \): \( \mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{lb} \).

The projection of \( \mathbf{lb} \) onto the column space is the nearest point \( \mathbf{Ax} \):

Projection: \( \mathbf{\hat{p}} = \mathbf{Ax} = \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{lb} \).

\[\begin{align*}
\text{Ex:} & \quad \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{lb} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \quad \mathbf{Ax} = \mathbf{lb} \text{ has no solution} \\
& \quad \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{lb} \text{ gives the best } \mathbf{x}. \\
& \quad \text{The projection of } \mathbf{lb} = (4,5,6) \text{ is } \mathbf{\hat{p}} = (4,5,0).
\end{align*}\]
Solving the normal equations:

\[ A^T A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 5 \\ 5 & 13 \end{bmatrix} \]

\[ \hat{x} = (A^T A)^{-1} A^T b = \begin{bmatrix} 13 & -5 \\ -5 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 11 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \]

Projection \[ P = A \hat{x} = \begin{bmatrix} 1 & 2 \\ 0 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}. \]

Remarks:

1. \( lb \in \text{row}(A) \Rightarrow P = A (A^T A)^{-1} A^T lb = A \hat{x} = lb. \)
   \[ \therefore \text{The projection of } lb \text{ is still } lb. \]

2. \( lb \in \text{N}(A^T) \Rightarrow P = A (A^T A)^{-1} A^T lb = A (A^T A)^{-1} 0 = 0 \)
   \[ \text{left null space} \]
   \[ \therefore \text{lb projects to the zero vector} \]

3. When \( A \) is square and invertible, \( \text{C}(A) = \mathbb{R}^n \). \( A \) invertible \( \Rightarrow P = A (A^T A)^{-1} A^T lb = A A^T (A^T A)^{-1} A^T lb = lb. \)
   \[ \therefore \text{Every vector projects to itself, } P = lb, \hat{x} = x. \]

4. \( A \) has only one column \( \Rightarrow A^T A = \alpha A^T \) and \( \hat{x} = \alpha lb / A^T lb. \)

The Cross-Product Matrix \( A^T A \)

- \( A^T A \) has the same nullspace as \( A \).

- \( A x = 0 \Rightarrow A^T A x = 0 \)

- \( A^T A x = 0 \Rightarrow x^T A^T A x = 0, \| A x \|^2 = 0, A x = 0. \)

- If \( A \) has independent columns, then \( A^T A \) is square, symmetric, and invertible.
Projection Matrices

We assume the independence of the columns of $A$ in what follows.

$P = A(ATA)^{-1}A^T$; projection matrix

1. $p = Plb$ is the component of $lb$ in the column space, and the error $e = lb - Plb = (I-P)lb$ is in $N(AT)$, the orthogonal complement of $C(A)$.

$(I-P)$ is also a projection matrix!

$	herefore$ $Plb$ and $(I-P)lb$ are two perpendicular components of $lb$.

3N: The projection matrix $P = A(ATA)^{-1}A^T$ has two basic properties: (i) $P^2 = P$ and (ii) $P^T = P$.

Conversely, any symmetric matrix with $P^2 = P$ represents a projection.

\[ P^2 = A(ATA)^{-1}A^T A(ATA)^{-1}A^T = A(ATA)^{-1}A^T = P \]

\[ P^T = (AT)^T (ATA^{-1})^T A^T = A (ATA^{-1})^T A^T = P \]

1. For the converse, from $P^2 = P$ and $P^T = P$, we can show the error vector $lb - Plb$ is orthogonal to the column space. For any vector $Pc \in C(P)$,

\[ (lb - Plb)^T Pc = lb^T (I-P)^T Pc = lb^T (P-P^2)c = 0. \]

Thus, $lb - Plb$ is orthogonal to the space, and $Plb$ is the projection onto the column space.

**Example:**

$A$: an invertible $4 \times 4$ matrix $\Rightarrow C(A) = \mathbb{R}^4$.

$P = A(ATA)^{-1}A^T = AA^T (A^T)^{-1}A^T = I$.

$I$: symmetric, $I^2 = I$, $lb - Plb = 0$.

The projection onto the whole space is the identity matrix.
Least-Squares Fitting of Data

In a series of experiments, we expect the output $b$ to be a linear function of the input $t$. We look for a straight line: $b = C + Dt$, with two unknowns $C, D$.

\[
\begin{align*}
C + Dt_1 &= b_1 \\
C + Dt_2 &= b_2 \\
&\vdots \\
C + Dt_m &= b_m
\end{align*}
\]

\[
\begin{bmatrix}
1 & t_1 \\
1 & t_2 \\
& \ddots \\
1 & t_m
\end{bmatrix}
\begin{bmatrix}
C \\
D
\end{bmatrix}
= 
\begin{bmatrix}
b_1 \\
b_2 \\
& \ddots \\
b_m
\end{bmatrix}
\]

The best solution $(\hat{C}, \hat{D})$ is the $x$ that minimizes $E^2 = \| lb - A x \|^2 = (b_1 - C - Dt_1)^2 + \ldots + (b_m - C - Dt_m)^2$.

The vector $p = A x$ is as close as possible to $lb$. Of all straight lines $b = C + Dt$, we choose the one that best fits the data. On the graph, the errors are the vertical distances $b - C - Dt$ to the straight line (not perpendicular distances!)

![Figure 3.9](image)

**Figure 3.9** Straight-line approximation matches the projection $p$ of $b$. 
**Example 2:** Three measurements:

\[ b_1 = 1 \text{ at } t_1 = -1; \quad b_2 = 1 \text{ at } t_2 = 1; \quad b_3 = 3 \text{ at } t_3 = 2 \]

\[ A \hat{x} = lb \text{ is } \begin{cases} C - D = 1 \\ C + D = 1 \\ C + 2D = 3 \end{cases} \]

\[ A^T A \hat{x} = A^T lb \text{ is } \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix} \]

\[ \therefore \hat{C} = \frac{9}{7} \text{ and } \hat{B} = \frac{4}{7} \text{ and the best line is } \frac{9}{7} + \frac{4}{7}t \]

**Example 30** The measurements \( b_1, \ldots, b_m \) are given at distinct points \( t_1, \ldots, t_m \). Then the straight line \( \hat{C} + \hat{B}t \) which minimizes \( E^2 \) comes from least squares:

\[ A^T A \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} = A^T lb \text{ or } \begin{bmatrix} m & \sum t_i \\ \sum t_i & \sum t_i^2 \end{bmatrix} \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} = \begin{bmatrix} \sum b_i \\ \sum t_i b_i \end{bmatrix} \]

**Remark:** Given a mixture of two radioactive chemicals with known half-lives \( \lambda \) and \( \mu \), we want to know their unknown amounts \( C \) and \( D \): \( b = CE^{\lambda t} + DE^{-\mu t} \).

\[ A \hat{x} = lb \text{ is } \begin{cases} C e^{\lambda t_1} + D e^{-\mu t_1} \approx b_1 \\ C e^{\lambda t_m} + D e^{-\mu t_m} \approx b_m \end{cases} \]

The least-squares principle will give optimal \( \hat{C} \) and \( \hat{D} \).

But, if we knew the amounts \( C \) and \( D \), and we're trying to discover the decay rates \( \lambda \) and \( \mu \), this is a problem in nonlinear least squares. Setting the derivatives of \( E^2 \) to zero will give nonlinear equations for the optimal \( \lambda \) and \( \mu \).
Weighted Least Squares

The estimate \( \hat{x} \) of weight from two observations \( x = b_1 \) and \( x = b_2 \):

\[
\begin{bmatrix}
1 \\
1
\end{bmatrix}
\begin{bmatrix}
x \\
b
\end{bmatrix}
\Rightarrow
A^TA \hat{x} = A^Tb,
\quad x \hat{x} = b_1 + b_2
\]

Weighted error: 
\[ E^2 = w_1(x - b_1)^2 + w_2(x - b_2)^2 \]

\[
\frac{dE^2}{dx} = 2 \left[ w_1^2(x - b_1) + w_2^2(x - b_2) \right] = 0 \quad \text{at} \quad \hat{x}_w = \frac{w_1^2 b_1 + w_2^2 b_2}{w_1^2 + w_2^2}
\]

The least squares solution to \( WAx = Wlb \) is \( \hat{x}_w \)

\[
\text{Weighted Normal Equation:} \quad (A^TW^TW) \hat{x}_w = A^TW^TWlb.
\]

0. The projection \( A\hat{x}_w \) is still the point on the column space that is closest to \( lb \) under a new meaning of closeness. The perpendicularity test involves \( (Wy)^T(Wx) = 0 \) instead of \( y^Tx = 0 \). The matrix \( W^TW \) appears in the middle. In this new sense, the projection \( A\hat{x}_w \) and the error \( lb - A\hat{x}_w \) are again perpendicular.

0. The inner product of \( x \) and \( y \) is generalized to
\[ y^TCx, \text{ where } C = W^TW \text{ is a symmetric matrix.} \]

For an orthogonal matrix \( W = Q, \ C = Q^TQ = I \) and the inner product is not new.

For any invertible matrix \( W \), these rules define a new inner product and length:

\[
\text{Weighted by } W : \quad (x, y)_W = (Wy)^T(Wx) \quad \text{and} \quad ||x||_W = ||Wx||.
\]
If the errors in the $b_i$ are independent of each other and their variances are $\sigma_i^2$, then the right weights are $w_i = 1/\sigma_i$. A more accurate measurement, with a smaller variance, gets a heavier weight.

The observations may not be independent. Then $W$ has off-diagonal terms. The best unbiased matrix $C = W^TW$ is the inverse of the covariance matrix—whose $i,j$ entry is the expected value of (error in $b_i$) times (error in $b_j$). The main diagonal of $C^{-1}$ contains the variances $\sigma_i^2$; the average of (error in $b_i$)².

Ex3:

Two bridge partners both guess (after the bidding) the total number of spades they hold. For each guess, the errors $-1, 0, 1$ might have equal probability $\frac{1}{3}$. Then the expected error is zero and the variance is $\frac{2}{3}$.

\[
E(e) = \frac{1}{3}(-1) + \frac{1}{3}(0) + \frac{1}{3}(1) = 0
\]

\[
E(e^2) = \frac{1}{3}(-1)^2 + \frac{1}{3}(0)^2 + \frac{1}{3}(1)^2 = \frac{2}{3}
\]

The two guesses are dependent but not identical. Say the chance that they are both too high or both too low is zero, but the chance of opposite errors is $\frac{1}{3}$. Then $E(e_1e_2) = \frac{1}{3}(-1)$, and the inverse of the covariance matrix is $W^TW$:

\[
\begin{bmatrix}
E(e_1^2) & E(e_1e_2)
\end{bmatrix}
\begin{bmatrix}
E(e_2^2) & E(e_1e_2)
\end{bmatrix}^{-1} = \begin{bmatrix}
\frac{2}{3} & -\frac{1}{3}
\end{bmatrix}^{-1} = C = W^TW.
\]

This matrix goes into the middle of the weighted normal equations.
3.4 Orthogonal Bases and Gram-Schmidt

The vectors \( q_1, ..., q_n \) are orthonormal if

\[
q_i^T q_j = \begin{cases} 
0 & \text{if } i \neq j \\
1 & \text{if } i = j
\end{cases}
\]

A matrix with orthonormal columns will be called \( Q \).

The standard basis \( e_1, ..., e_n \) consists of the columns of \( I \). We can rotate these vectors to form other orthonormal bases. A subspace of \( \mathbb{R}^n \) may not contain the standard vectors \( e_k \), but an orthonormal basis can be constructed for the space. This construction is known as Gram-Schmidt orthogonalization.

**Orthogonal Matrices**

**3Q** If \( Q \) has orthonormal columns, then \( Q^T Q = I \).

An orthogonal matrix is a square matrix with orthonormal columns. Then \( Q^T = Q^{-1} \). (Note that \( Q^T Q = I \) even if \( Q \) is rectangular. But then \( Q^T \) is only a left-inverse.)

**Ex1:** \( Q = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \Rightarrow Q^T = Q^{-1} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \)

**Ex2:** \( P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow P^{-1} = P^T = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \)

0. A reflection also forms an orthogonal matrix. Geometrically, an orthogonal \( Q \) is either a rotation or the product of a rotation and a reflection.
Multiplication by any \( Q \) preserves lengths: 
\[ \| Qx \| = \| x \| \text{ for every vector } x. \]

It also preserves inner products and angles, since 
\[ (Qx)^T(Qy) = x^TQ^TQy = x^Ty. \]

0. Write \( lb \) as a combination \( lb = x_1q_1 + x_2q_2 + \ldots + x_nq_n \).
\[ q_i^Tlb = x_i q_i^Tq_i \Rightarrow x_i = q_i^Tlb \]
\[ : \quad lb = (q_1^Tlb)q_1 + (q_2^Tlb)q_2 + \ldots + (q_n^Tlb)q_n \]
0. \( x_1q_1 + \ldots + x_nq_n = lb \Rightarrow Qx = lb \Rightarrow x = Q^Tlb = Q^Tlb. \)
0. \( \| lb \|^2 = (q_1^Tlb)^2 + (q_2^Tlb)^2 + \ldots + (q_n^Tlb)^2 = \| Q^Tlb \|^2. \)

Remarks: Since \( Q^T = Q^T \), we have \( Q^TQ = I \). The rows of a square matrix are orthonormal whenever the columns are.

Rectangular Matrices with Orthonormal Columns

3C If \( Q \) has orthonormal columns, the least-square problem is easy.

\[ Qx = lb : \text{rectangular system with no solution for most } lb. \]
\[ QTQx = Q^Tlb : \text{normal equation for the best } \hat{x} \text{ -- for which } QTQ = I. \]
\[ \hat{x} = Q^Tlb \]
\[ \hat{x}_i = q_i^Tlb \]
\[ p = Q\hat{x} : \text{the projection of } lb \text{ is } (q_1^Tlb)q_1 + \ldots + (q_n^Tlb)q_n \]
\[ p = QQ^Tlb : \text{the projection matrix is } P = QQ^T \]

Ex 3: Projection of \( lb = (x, y, z) \) onto the xy-plane is \( p = (x, y, 0) \).
\[ q_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } (q_1^Tlb)q_1 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}; \quad q_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } (q_1^Tlb)q_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \]
\[ \Rightarrow p = q_1q_1^T + q_2q_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } P \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}. \]
Ex 4: When the measurement times average to zero, fitting a straight line leads to orthogonal columns.

Take \( t_1 = -3, t_2 = 0, t_3 = 3 \). The attempt to fit \( y = C + Dt \) leads to

\[
\begin{align*}
C + Dt_1 &= y_1 \\
C + Dt_2 &= y_2 \\
C + Dt_3 &= y_3
\end{align*}
\]

We can project \( y \) separately onto each column:

\[
\hat{C} = \frac{[1 \ 1 \ 1] [y_1 \ y_2 \ y_3]^T}{1^2 + 1^2 + 1^2}, \quad \hat{D} = \frac{[-3 \ 0 \ 3] [y_1 \ y_2 \ y_3]^T}{(-3)^2 + 0^2 + 3^2}
\]

\( \hat{C} = (y_1 + y_2 + y_3)/3 \) is the best fit by a horizontal line, whereas \( \hat{D} \) is the best fit by a straight line through the origin. The sum \( \hat{C} + \hat{D} \) is the best fit by any straight line.

If the average \( \bar{E} = (t_1 + \cdots + t_m)/m \) is not zero, then the time can be shifted by \( \bar{E} \). Let \( y = C + D(t - \bar{E}) \).

\[
\hat{C} = \frac{[1 \ \cdots \ 1] [y_1 \ y_2 \ \cdots \ y_m]^T}{1^2 + 1^2 + \cdots + 1^2} = \frac{y_1 + \cdots + y_m}{m}
\]

\[
\hat{D} = \frac{\begin{bmatrix} (t_1 - \bar{E}) & \cdots & (t_m - \bar{E}) \end{bmatrix} [y_1 \ y_2 \ \cdots \ y_m]^T}{(t_1 - \bar{E})^2 + \cdots + (t_m - \bar{E})^2} = \frac{\sum(t_i - \bar{E})y_i}{\sum(t_i - \bar{E})^2}
\]

The best \( \hat{C} \) is the mean and we also get a convenient formula for \( \hat{D} \). The earlier \( A^T A \) had the off-diagonal entries \( \Sigma t_i c_i \) and shifting the time by \( \bar{E} \) made these entries zero. The shift is an example of the Gram-Schmidt process.
The Gram-Schmidt Process

Given three independent vectors \( a, b, c \), let \( q_1 = a / \| a \| \).
Second vector \( B = b - (q_1^T b) q_1 \) and \( q_2 = B / \| B \| \).
Third vector \( C = c - (q_1^T c) q_1 - (q_2^T c) q_2 \) and \( q_3 = C / \| C \| \).

Ex 5: Gram-Schmidt

\[
\begin{align*}
a &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \Rightarrow q_1 &= a / \| a \| = \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} \\
B &= b - (q_1^T b) q_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \Rightarrow q_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \\
C &= c - (q_1^T c) q_1 - (q_2^T c) q_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \frac{2}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\end{align*}
\]

Orthonormal basis \( Q = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \)

The Gram-Schmidt process starts with independent vectors \( a_1, \ldots, a_n \), and ends with orthonormal vectors \( q_1, \ldots, q_n \). At step \( j \), it subtracts from \( a_j \) its components in the directions \( q_1, \ldots, q_{j-1} \):

\[
A_j = a_j - (q_1^T a_j) q_1 - \cdots - (q_{j-1}^T a_j) q_{j-1}.
\]

Then \( q_j \) is the unit vector \( A_j / \| A_j \| \).

Remark on the calculation

It is easier to compute the orthogonal \( a, b, c \) without forcing their lengths to equal one. Then square roots enter only at the end, when dividing by those lengths.
The Factorization $A = QR$

$$a = (q^T a) \cdot q$$
$$b = (q^T b) \cdot q_1 + (q^T b) q_2$$
$$c = (q^T c) \cdot q_1 + (q^T c) q_2 + (q^T c) q_3$$

$QR$ factors:

$$A = \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} \begin{bmatrix} q^T a & q^T b & q^T c \end{bmatrix} = QR$$

Example:

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

The length of $a, b, c$

34) Every $m \times n$ matrix with independent columns can be factored into $A = QR$. The columns of $Q$ are orthonormal, and $R$ is upper triangular and invertible. When $m = n$ and matrices are square, $Q$ becomes an orthogonal matrix.

The orthogonalization simplifies the least-squares problem $Ax = b$. The normal equations become easier since $A^T A = R^T Q^T Q R = R^T R$.

The fundamental equation $A^T A x = A^T b$ simplifies to a triangular system $R^T R x = R^T Q^T b$ or $R x = Q^T b$.

Instead of solving $QR x = b$, we solve $R x = Q^T b$ by back-substitution. The real cost is the $m n^2$ operations of Gram-Schmidt, which are needed to find $Q$ and $R$. 
Function Spaces

1) Lengths and Inner Products.
\[ ||f||^2 = \int_0^{2\pi} (f(x))^2 \, dx \quad (f,g) = \int_0^{2\pi} f(x)g(x) \, dx \]
The Schwartz inequality is still satisfied:
\[ |(f,g)| \leq ||f|| \cdot ||g|| \]

2) Gram-Schmidt for Functions.
There is no interval \([a,b]\) on which \((1,x) = \int_a^b x^2 \, dx = 0\). Therefore the closest parabola to \(f(x)\) is not the sum of its projections onto \(1, x, x^2\). On the interval \([0,1]\),
\[ A^T A = \begin{bmatrix} (1,1) & (1,x) & (1,x^2) \\ (x,1) & (x,x) & (x,x^2) \\ (x^2,1) & (x^2,x) & (x^2,x^2) \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \]

This is the ill-conditioned Hilbert matrix with a large inverse. The situation becomes impossible if we add a few more axes. It is virtually hopeless to solve \(A^T A x = A^T b\) for the closest polynomial of degree 10. More precisely, it is hopeless to solve this by Gauss elimination. Every roundoff error would be amplified by more than \(10^{13}\).

The right idea is to switch to orthogonal axes (by Gram-Schmidt). On the interval \([-1,1]\), we have \((1,x) = \int_{-1}^{1} x \, dx = 0\) and \((x,x^2) = \int_{-1}^{1} x^3 \, dx = 0\).

Starting with \(v_1 = 1, \quad v_2 = x\),

Orthogonalize \[ v_3 = x^2 - \frac{(1,x^2)}{(1,1)} \cdot 1 - \frac{(x,x^2)}{(x,x)} \cdot x = x^2 - \frac{1}{3} \]

\(1, x, x^2 - \frac{1}{3}\) : Legendre polynomials

The closest polynomial is now computable by projecting onto each of the Legendre polynomials.