Chap 3. Orthogonality

3.1 Orthogonal Vectors and Subspaces

For \( x = (x_1, \ldots, x_n) \), 
\[ ||x|| = \sqrt{x^T \cdot x} \] 
the length of \( x \).
\[ ||x||^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \]

\( x, y \in \mathbb{R}^n \): orthogonal \( \iff \) \( x^T \cdot y = x_1y_1 + \cdots + x_ny_n = 0 \)
the inner product of \( x \) and \( y \).

**3A**
The inner product \( x^T \cdot y = 0 \) iff \( x \) and \( y \) are orthogonal.

If \( x^T \cdot y > 0 \), their angle is less than 90°.

If \( x^T \cdot y < 0 \), their angle is greater than 90°.

Useful Fact: If nonzero vectors \( v_1, \ldots, v_k \) are mutually orthogonal, then those vectors are linearly independent.

Suppose \( c_1v_1 + \cdots + c_kv_k = 0 \).
\[ \Rightarrow \quad v_i \cdot (c_1v_1 + \cdots + c_kv_k) = c_i v_i^T \cdot v_i = 0 \]
Since \( v_i \) are nonzero, \( v_i^T \cdot v_i \neq 0 \) and \( c_i = 0 \)

The coordinate vectors \( e_1, \ldots, e_n \in \mathbb{R}^n \) 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 \( \mathbb{R}^2 \), \( v_1 = (\cos \theta, \sin \theta), v_2 = (-\sin \theta, \cos \theta) \) form an orthonormal basis.

Orthogonal Subspaces

**3B** Two subspaces \( V \) and \( W \) of \( \mathbb{R}^n \) are orthogonal if every vector \( v \in V \) is orthogonal to every \( w \in W \):
\[ v^T \cdot w = 0 \] for all \( v \) and \( w \).
Fundamental Theorem of Orthogonality

The row space is orthogonal to the nullspace (in \( \mathbb{R}^n \)).

\[ \text{row space} \perp \text{nullspace} \]

The column space is orthogonal to the left nullspace (in \( \mathbb{R}^m \)).

\[ \text{column space} \perp \text{left nullspace} \]

- For \( x \in N(A) \): \( A^T x = 0 \) \( \Rightarrow \) \( y^T x = z^T A^T x = 0 \)
- For \( v \in C(A^T) \): \( v = A^T z \) for some \( z \)

\[ \therefore N(A) \perp C(A^T) \]

Definition (Orthogonal Complement)

- \( \mathcal{V} \): a subspace of \( \mathbb{R}^n \).
- \( \mathcal{V}^\perp \): the orthogonal complement of \( \mathcal{V} \) is the space of all vectors orthogonal to \( \mathcal{V} \).

Thus, \( \mathcal{V}^\perp \) is denoted as \( \mathcal{V} \perp \).

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 \( \mathbb{R}^m \).

- \( A x = b \) is solvable if and only if \( y^T b = 0 \) whenever \( y^T A = 0 \)
- For \( b \in C(A) \): \( b \in (N(A^T))^\perp \)

Example:

\[ A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \Rightarrow y = (1, 1, 1)^T \in N(A^T) \]

Since \( y^T A = 0 \).

Thus, \( A x = b \) is solvable \( \Leftrightarrow \) \( b_1 + b_2 + b_3 = 0 \).

\( \therefore y^T b \).
The Matrix and the Subspaces

If $W = V^\perp$, then $V = W^\perp$ and $\dim V + \dim W = n$.
In other words $V^{\perp \perp} = V$. 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.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{matrix_subspaces_diagram.png}
\caption{The true action $Ax = A(x_{\text{row}} + x_{\text{null}})$ of any $m \times n$ matrix.}
\end{figure}

\begin{itemize}
  \item Let $b \in \text{C}(A)$, $b = Ax = Ax_r$, where $x = x_r + x_n$ and $Ax_r = 0$.
  \item Let $x^{\prime}$ in the row space and $Ax^{\prime} = b$.
  \item Then $A(x_r - x^{\prime}) = b - b = 0$, and $x_r - x^{\prime} \in N(A)$.
  \item And $x_r - x^{\prime}$ is also in the row space $\Rightarrow x_r = x^{\prime}$.
\end{itemize}

0. $A^T$ goes from $\mathbb{R}^m$ to $\mathbb{R}^n$ and from $\text{C}(A)$ to $\text{C}(A^T)$.
But $A^T \neq A^{-1}$ 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 $A^+$:
$A^+A x = x$ for $x \in \text{C}(A^T)$ 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.

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

**3.4**  \[ \cos \theta = \frac{a^T b}{\|a\| \|b\|} \]

**Law of Cosines**

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

\[ \Rightarrow \|b\|^2 - 2a^T b + a^T a = \|b\|^2 + a^T a - 2\|b\|\|a\|\cos \theta \]

\[ \Rightarrow a^T b = \|a\| \|b\| \cos \theta \]

Figure 3.7  The projection $p$ of $b$ onto $a$, with $\cos \theta = \frac{O_p}{Ob} = \frac{a^T b}{\|a\| \|b\|}$.

**3H**  The projection of $b$ onto the line in the direction of $a$ is

\[ \hat{p} = \hat{x} a = \frac{a^T b}{a^T a} a \quad \text{[eq. $(b - \hat{x} a) \perp a$] \]

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

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

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

$$\sum_{i=0}^{\infty} ||\mathbf{b} - \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \mathbf{a}||^2 = ||\mathbf{b}||^2 - 2 \frac{(\mathbf{a}^T \mathbf{b})^2}{\mathbf{a}^T \mathbf{a}} + \frac{\left(\mathbf{a}^T \mathbf{b}\right)^2}{\mathbf{a}^T \mathbf{a}} \geq 0$$

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

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

$$\hat{x} = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} = \frac{6}{3} = 2$$

and $\mathbf{p} = \hat{x} \mathbf{a} = (2,2,2)$.

$$\cos \theta = \frac{||\mathbf{p}||}{||\mathbf{b}||} = \frac{\sqrt{12}}{\sqrt{14}}$$

and

$$\cos \theta = \frac{\mathbf{a}^T \mathbf{b}}{||\mathbf{a}|| ||\mathbf{b}||} = \frac{6}{\sqrt{3} \sqrt{14}}.$$  

$$\Rightarrow 6 \leq \sqrt{3} \sqrt{14} \text{ and } \sqrt{36} \leq \sqrt{42}.$$  

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

$$\mathbf{P} = \frac{\mathbf{a} \cdot \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$
Remark on scaling:
The projection matrix is the same if \( \alpha \) is scaled.
\[
\alpha = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{gives} \quad P = \frac{1}{3d^2} \begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} x & y \\ x & y \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}
\]
If \( \alpha \) has a unit length, \( P = \alpha \cdot \alpha^T \) (\( \alpha^T \alpha = 1 \)).

Ex.3: Project onto the \( \theta \)-direction in the xy-plane.
The line goes through \( \alpha = (\cos \theta, \sin \theta) \).
\[
P = \frac{\alpha \cdot \alpha^T}{\alpha^T \alpha} = \frac{1}{c^2 + s^2} \begin{bmatrix} c \\ s \end{bmatrix} \begin{bmatrix} c & s \end{bmatrix} = \begin{bmatrix} \frac{c^2}{c^2 + s^2} & \frac{cs}{c^2 + s^2} \\ \frac{cs}{c^2 + s^2} & \frac{s^2}{c^2 + s^2} \end{bmatrix}
\]

This matrix \( P \) was discovered in Section 2.6.
Now we know \( P \) in any dimensions.

To project \( \mathbf{b} \) onto \( \alpha \), multiply by the projection matrix \( \mathbf{P} \).
\[
P = \mathbf{P} \mathbf{b}.
\]
3.3 Projections and Least Squares

Least squares solution for $Ax = lb$ by minimizing

$$E^2 = ||Ax - lb||^2 = (a_1 x - b_1)^2 + \ldots + (a_m x - b_m)^2.$$ 

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

The least squares solution to a problem $Ax = lb$ in one unknown is $\hat{x} = a^T lb / a^T a$.

Orthogonality: 

$$a^T (lb - \hat{x} a) = a^T lb - \frac{a^T lb \cdot a^T a}{a^T a} = 0$$

$\mathbf{e}$: the error vector

Least-Squares Problem with Several Variables

$Ax = lb$, where $A$ is an $m \times n$ matrix.

$\Rightarrow E = ||Ax - lb||$: error, which is the distance from $lb$ to the point $Ax$ in the column space.

1. Locate the point $p = A\hat{x}$ that is closer to $lb$ than any other point in the column space.

2. The error $\mathbf{e} = lb - A\hat{x}$ must be perpendicular to the column space.

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 $e=lb-Ax$ must be in the nullspace of $A^T$:
$$A^T(lb-Ax) = 0 \quad \text{or} \quad A^TAx = A^Tlb.$$ 

(2) The error must be perpendicular to each column of $A$:
$$\begin{align*}
a_1^T(lb-Ax) &= 0 \\
\vdots & \quad \vdots \\
a_n^T(lb-Ax) &= 0
\end{align*}$$
This is again $A^T(lb-Ax) = 0$ and $A^TAx = A^Tlb$.

o. Taking partial derivatives of $E = (Ax-lb)^T(Ax-lb)$ gives the same $2A^TAx - 2A^Tlb = 0$.

C. symmetric square matrix

B. When $Ax=lb$ is inconsistent, its least-squares solution minimizes $\|Ax-lb\|^2$:

Normal equations: $A^TAx = A^Tlb$.

$A^TA$ is invertible exactly when the columns of $A$ are linearly independent! Then,

Best estimate $\hat{x}$: $\hat{x} = (A^TA)^{-1}A^Tlb$.

The projection of $lb$ onto the column space is the nearest point $Ax$:

Projection: $\hat{p} = Ax = A(A^TA)^{-1}A^Tlb$.

Ex: $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}$, $lb = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$.

$A^TAx = lb$ has no solution $A^TAx = A^Tlb$ gives the best $x$.

The projection of $lb = (4, 5, 6)$ is $\hat{p} = (4, 5, 0)$.
Solving the normal equations:

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

\[ \hat{x} = (ATA)^{-1}ATlb = \begin{bmatrix} 13 & -5 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \]

Projection \[ iP = AX = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \]

Remarks:

1. \[ lb = AX \text{ (a combination of the columns)} \]
2. \[ lb \in \text{Null}(A) \Rightarrow iP = A(ATA)^{-1}A^TAx = AX = lb \]  
   \[ \therefore \text{The projection of } lb \text{ is still } lb \]  
3. \[ lb \in \text{Null}(A^T) \Rightarrow iP = A(ATA)^{-1}ATlb = A(ATA)^{-1} \cdot 0 = 0 \]
   \[ \therefore lb \text{ projects to the zero vector} \]
4. \[ \text{When } A \text{ is square and invertible, } C(A) = \mathbb{R}^n \text{, the whole space} \]
   \[ A : \text{invertible} \Rightarrow iP = A(ATA)^{-1}ATlb = A(ATA)^{-1}ATlb = lb \]
   \[ \therefore \text{Every vector projects to itself, } iP = lb, \hat{x} = x \]
5. \[ A \text{ has only one column} \Rightarrow ATA = ATA \text{ and } \hat{x} = ATlb/ATA \]

The Cross-Product Matrix \( ATA \)

- \( ATA \) has the same nullspace as \( A \).
- \( AX = 0 \Rightarrow ATAx = 0 \)
- \( ATAx = 0 \Rightarrow x^TATAx = 0, \|Ax\|^2 = 0, A\hat{x} = 0 \).

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

\( P = A(ATA)AT : \) projection matrix

1. \( P = P1b \) is the component of \( 1b \) in the column space, and the error \( \varepsilon = 1b - P1b = (I-P)1b \) is in \( N(AT) \), the orthogonal complement of \( CA \).

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

\[ \therefore P1b \text{ and } (I-P)1b \text{ are two perpendicular components of } 1b. \]

**3N** The projection matrix \( P = A(ATA)AT \) has two basic properties: (i) \( P^2 = P \) and (ii) \( P^T = P \).

Conversely, any symmetric matrix with \( P^2 = P \) represents a projection.

**pf**

1. \( P^2 = A(ATA)^TAT = A(ATA)AT = P \)
\[ P^T = (AT)^T(ATAT)^TAT = A(ATA)^TAT = A(ATA)^TAT = P. \]

2. For the converse, from \( P^2 = P \) and \( P^T = P \), we can show the error vector \( 1b - P1b \) is orthogonal to the column space. For any vector \( PC \in C(P) \),
\[ (1b - P1b)^TPC = 1b^T(I-P)^TPC = 1b^T(P-P^2)C = 0. \]

Thus, \( 1b - P1b \) is orthogonal to the space, and \( P1b \) is the projection onto the column space.

**Ex1:** \( A: \) an invertible \( 4 \times 4 \) matrix \( \Rightarrow CA = 1R^4 \).
\[ P = A(ATA)^TAT = A(ATA)^TAT = I, \]
\[ I: \) symmetric, \( I^2 = I, 1b - II1b = 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: $b = C + D t$, with two unknowns $C, D$.

\[
\begin{align*}
C + D t_1 &= b_1 \\
C + D t_2 &= b_2 \\
&\vdots \\
C + D t_m &= b_m
\end{align*}
\]

or

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

The best solution $(\hat{C}, \hat{D})$ is the $x$ that minimizes

\[
E^2 = \| b - A x \|^2 = (b_1 - C - D t_1)^2 + \cdots + (b_m - C - D t_m)^2.
\]

The vector $p = A x$ is as close as possible to $b$.

Of all straight lines $b = C + D t$, we choose the one that best fits the data. On the graph, the errors are the vertical distances $b - C - D t$ to the straight line (not perpendicular distances!)

![Graph](a)

![Diagram](b)

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

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

Ax = lb is:

\[ \begin{bmatrix} C - D = 1 \\ C + D = 1 \quad \text{or} \quad \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \]

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

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

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

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

Remark: Given a mixture of two radioactive chemicals with known half-lives \( \lambda \) and \( \mu \), we want to know their unknown amounts \( C \) and \( D \):

\[ b = C e^{-\lambda t} + D e^{-\mu t} \]

Ax = lb is:

\[ \begin{bmatrix} C e^{-\lambda t_1} + D e^{-\mu t_1} \approx b_1 \\
C e^{-\lambda t_m} + D e^{-\mu t_m} \approx b_m \end{bmatrix} \]

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

But, if we knew the amounts \( C \) and \( D \), and were 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 \( x = b_2 \):

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

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

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

The least squares solution to \( W A x = W b \) is \( \hat{x}_w \)

**Weighted Normal Equation**

\[
(A^T W^T W A) \hat{x}_w = A^T W^T W b
\]

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

3. The inner product of \( x \) and \( y \) is generalized to \( y^T C x \), where \( C = W^T W \) is a symmetric matrix. For an orthogonal matrix \( W = Q \), \( C = Q^T Q = I \) and the inner product is not new.

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

**Weighted** by \( W \) : \( (x, y)_W = (Wy)^T (Wx) \) and \( \|x\|_W = \|Wx\|_2 \)
0. 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.

0. 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 $\Sigma$ whose $i,j$ entry is the expected value of $(\text{error in } b_i) \times (\text{error in } b_j)$. The main diagonal of $C^{-1}$ contains the variances $\sigma_i^2$, to the average of $(\text{error in } b_i)^2$.

**Ex 3:**

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 $W^TW$:

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

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

The vectors $q_1, \ldots, q_n$ are orthonormal if

$$q_i^\top q_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

A matrix with orthonormal columns will be called $Q$.

The standard basis $e_1, \ldots, e_n$ 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 $e_i$, but an orthonormal basis can be constructed for the space. This construction is known as Gram-Schmidt orthogonalization.

Orthogonal Matrices

If $Q$ has orthonormal columns, then $Q^\top Q = I$.

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

\[ Q = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \Rightarrow Q^\top = Q^{-1} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \]

\[ P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow P^\top = P^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \]

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:
\[\|Qx\| = \|x\|\] 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 \(lb = x_1q_1 + x_2q_2 + \ldots + x_nq_n\).
\[q_i^T lb = x_i.q_i^T q_i \Rightarrow x_i = q_i^T lb\]
\[\therefore lb = (q_1^T lb) q_1 + (q_2^T lb) q_2 + \ldots + (q_n^T lb) q_n\]
0. \(x_1q_1 + \ldots + x_nq_n = lb\) \(\Rightarrow Qx = lb \Rightarrow x = Q^T lb = Q^T l\)
0. \(\|lb\|^2 = (q_1^T lb)^2 + (q_2^T lb)^2 + \ldots + (q_n^T lb)^2 = \|Q^T lb\|^2\).

Remark: Since \(Q^T = Q^1\), we have \(Q^T Q = I\). The rows of a square matrix are orthonormal whenever the columns are.

Rectangular Matrices with Orthonormal Columns.

If \(Q\) has orthonormal columns, the least-square problem is easy.
\(Qx = lb\): rectangular system with no solution for most \(lb\).
\(QTQx = QTlb\): normal equation for the best \(x\) - in which \(QTQ = I\).
\(\hat{x} = QTlb\)\(\hat{x}_i = q_i^T lb\)
\(p = Q\hat{x}\) - the projection of \(lb\) is \((q_1^T lb) q_1 + \ldots + (q_n^T lb) q_n\)
\(p = QTlb\): the projection matrix is \(P = QT\).

Example: Projection of \(lb = (x,y,z)\) onto the xy-plane is \(p = (x,y,0)\)

\[q_1 = \begin{bmatrix}1 \\ 0 \end{bmatrix}\] and \((q_1^T lb) q_1 = \begin{bmatrix}3 \\ 0 \end{bmatrix}\)

\[q_2 = \begin{bmatrix}0 \\ 1 \end{bmatrix}\] and \((q_2^T lb) q_2 = \begin{bmatrix}2 \\ 0 \end{bmatrix}\)

\[P = q_1q_1^T + q_2q_2^T = \begin{bmatrix}1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\] and \(P \begin{bmatrix}x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix}x \\ y \\ 0 \end{bmatrix}\).
Ex4: 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

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

or

\[
\begin{bmatrix}
1 & -3 \\
1 & 0 \\
1 & 3
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix}
= \begin{bmatrix}
y_1 \\
y_2 \\
y_3
\end{bmatrix}
\]

We can project $y$ separately onto each column:

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

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

If the average $\bar{t} = (t_1 + \ldots + t_m) / m$ is not zero, then the time can be shifted by $\bar{t}$. Let $y = c + dt = c + d(t - \bar{t})$.

\[
\hat{C} = \left[ \frac{1}{1^2 + 1^2 + \ldots + 1^2} \right] \begin{bmatrix} y_1 & \ldots & y_m \end{bmatrix}^T = \frac{y_1 + \ldots + y_m}{m} = \bar{y}
\]

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

The best $\bar{C}$ is the mean and we also get a convenient formula for $\hat{D}$. The earlier $ATA$ had the off-diagonal entries $\Sigma t_i$, and shifting the time by $\bar{t}$ made these entries zero. The shift is an example of the Gram-Schmidt process.
The Gram-Schmidt Process

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

Ex 5: Gram-Schmidt

\[
\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \Rightarrow q_1 = \mathbf{a}/\|\mathbf{a}\| = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}
\]

\[
B = \mathbf{b} - (q_1^T \mathbf{b}) q_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \Rightarrow q_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}
\]

\[
C = \mathbf{c} - (q_1^T \mathbf{c}) q_1 - (q_2^T \mathbf{c}) q_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}
\]

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

3T The Gram-Schmidt Process starts with independent vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$, and ends with orthonormal vectors $q_1, \ldots, q_n$.  At step $j$, it subtracts from $\mathbf{a}_j$ its components in the directions $q_1, \ldots, q_{j-1}$:

\[
\mathbf{A}_j = \mathbf{a}_j - (q_1^T \mathbf{a}_j) q_1 - \cdots - (q_{j-1}^T \mathbf{a}_j) q_{j-1}
\]

Then $q_j$ is the unit vector $\mathbf{A}_j/\|\mathbf{A}_j\|$.  

Remark on the calculation

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

$a = (q_1^T a) \cdot q_1$

$b = (q_1^T b) \cdot q_1 + (q_2^T b) q_2$

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

$QR$ factors:

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

Example:

$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}$

The length of $q_1, b, c$

3. 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 all 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 $A^T A \hat{x} = A^T b$ simplifies to a triangular system $R^T R \hat{x} = R^T Q^T b$ or $R \hat{x} = Q^T b$.

Instead of solving $QR \hat{x} = b$, we solve $R \hat{x} = Q^T b$ by back-substitution. The real cost is the $mn^2$ 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:

\[ |(f, g)| \leq \|f\| \|g\| \]

2. Gram-Schmidt for Functions.

There is no interval \([a, b]\) on which \((1, x^2) = \int_a^b \varphi^2 \, dx = 0\).
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 round-off 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

\[
(1, x) = \int_{-1}^{1} x \, dx = 0, \quad (x, x^2) = \int_{-1}^{1} x^3 \, dx = 0.
\]

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

Orthogonalize \(v_3 = x^2 - \frac{(1, x^2)}{(1, 1)} v_1 - \frac{(x, x^2)}{(x, x)} v_2 = 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.