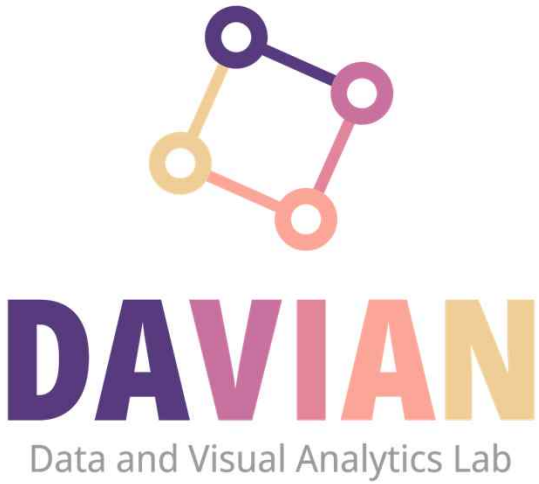
LINEAR ALGEBRA

LECTURE 6: LEAST SQUARES



Lecture Overview

- Elements in linear algebra
- Linear system
- Linear combination, vector equation,
  Four views of matrix multiplication
- Linear independence, span, and subspace
- Linear transformation
- Least squares
- Eigendecomposition
- Advanced eigendecomposition
- Singular value decomposition

Over-determined Linear Systems (#equations ≫ #variables)

- Recall a linear system:

Person ID	Weight	Height	Is_smoking	Life-span
1	60kg	5.5ft	Yes (=1)	66
2	65kg	5.0ft	No (=0)	74
3	55kg	6.0ft	Yes (=1)	78

$$60x_1+5.5x_2+1\cdot x_3=66$$
$$65x_1+5.0x_2+0\cdot x_3=74$$
$$55x_1+6.0x_2+1\cdot x_3=78$$

Over-determined Linear Systems (#equations $\gg$ #variables)

- Recall a linear system:
- What if we have much more data examples?

Person ID	Weight	Height	Is_smoking	Life-span
1	60kg	5.5ft	Yes (=1)	66
2	65kg	5.0ft	No (=0)	74
3	55kg	6.0ft	Yes (=1)	78
⋮	⋮	⋮	⋮	⋮

$$\begin{aligned} 60x_1 + 5.5x_2 + 1 \cdot x_3 &= 66 \\ 65x_1 + 5.0x_2 + 0 \cdot x_3 &= 74 \\ 55x_1 + 6.0x_2 + 1 \cdot x_3 &= 78 \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad & \vdots \end{aligned}$$

- Matrix equation: $\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ \vdots \end{bmatrix}$

$A \qquad \mathbf{x} = \mathbf{b}$

$m \gg n$: more equations than variables

➡ Usually no solution exists

Vector Equation Perspective

- Vector equation form: $\begin{bmatrix}60\\65\\55\\\vdots\end{bmatrix}x_1 + \begin{bmatrix}5.5\\5.0\\6.0\\\vdots\end{bmatrix}x_2 + \begin{bmatrix}1\\0\\1\\\vdots\end{bmatrix}x_3 = \begin{bmatrix}66\\74\\78\\\vdots\end{bmatrix}$

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 = \mathbf{b}$$

- Compared to the original space $\mathbb{R}^n$, where $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{b} \in \mathbb{R}^n$, Span $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ will be a thin hyperplane, so it is likely that $\mathbf{b} \notin$ Span $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$

  ➡ No solution exists.

Motivation for Least Squares

- Even if no solution exists, we want to approximately obtain the solution for an over-determined system.

- Then, how can we define the best approximate solution for our purpose?

Inner Product

- Given $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we can consider $\mathbf{u}$ and $\mathbf{v}$ as $n \times 1$ matrices.
- The transpose $\mathbf{u}^T$ is a $1 \times n$ matrix, and the matrix product $\mathbf{u}^T\mathbf{v}$ is a 1×1 matrix, which we write as a scalar without brackets.
- The number $\mathbf{u}^T\mathbf{v}$ is called the **inner product** or **dot product** of $\mathbf{u}$ and $\mathbf{v}$, and it is written as $\mathbf{u} \cdot \mathbf{v}$.
- For $\mathbf{u} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v} = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} = [14]$

$(1 \times 3)(3 \times 1) = 1 \times 1$

Properties of Inner Product

- **Theorem:** Let $\mathbf{u}, \mathbf{v},$ and $\mathbf{w}$ be vectors in $\mathbb{R}^n$, and let c be a scalar. Then
  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 $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = \mathbf{0}$
- Properties (b) and (c) can be combined to produce the following useful rule:

$$(c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p) \cdot \mathbf{w} = c_1(\mathbf{u}_1 \cdot \mathbf{w}) + \cdots + c_p(\mathbf{u}_p \cdot \mathbf{w})$$

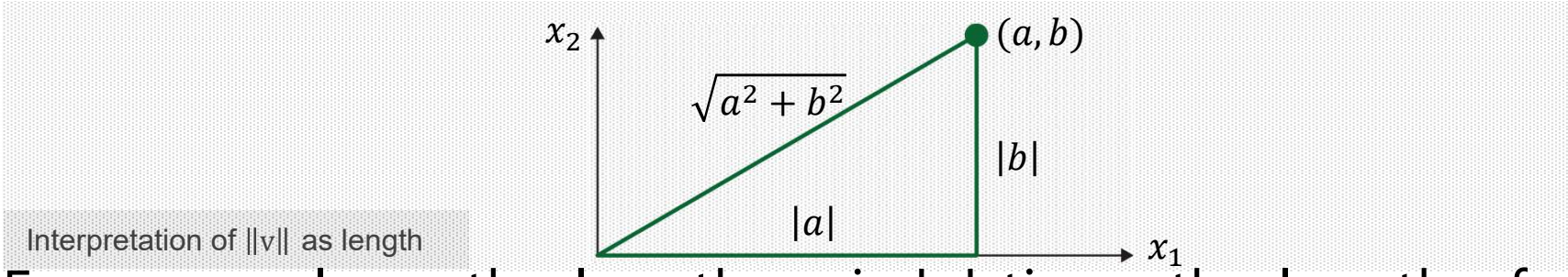
Vector Norm

- For $\mathbf{v} \in \mathbb{R}^n$, with entries $v_1, \ldots, v_n$, the square root of $\mathbf{v} \cdot \mathbf{v}$ is defined because $\mathbf{v} \cdot \mathbf{v}$ is nonnegative.
- **Definition**: The length (or norm) of $\mathbf{v}$ is the non-negative scalar $\|\mathbf{v}\|$ defined as the square root of $\mathbf{v} \cdot \mathbf{v}$:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} \text{ and } \|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$$

Geometric Meaning of Vector Norm

- Suppose $\mathbf{v} \in \mathbb{R}^2$, say, $\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}$.
- $\|\mathbf{v}\|$ is the length of the line segment from the origin to $\mathbf{v}$.
- This follows from Pythagorean Theorem applied to a triangle such as the one shown in the following figure:



Interpretation of ||v|| as length

- For any scalar c, the length $c\mathbf{v}$ is $|c|$ times the length of $\mathbf{v}$ That is,

$$\|c\mathbf{v}\| = |c|\|\mathbf{v}\|$$

Unit Vector

- A vector whose length is 1 is called a unit vector.
- Normalizing a vector: Given a nonzero vector $\mathbf{v}$, if we divide it by its length, we obtain a unit vector $\mathbf{u} = \frac{1}{\|\mathbf{v}\|}\mathbf{v}$.
- $\mathbf{u}$ is in the same direction as $\mathbf{v}$, but its length is 1.

Distance between Vectors in $\mathbb{R}^n$

- **Definition:** For $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the **distance between $\mathbf{u}$ and $\mathbf{v}$**, written as dist $(\mathbf{u}, \mathbf{v})$, is the length of the vector $\mathbf{u} - \mathbf{v}$ That is,

$$\text{dist}\,(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|$$

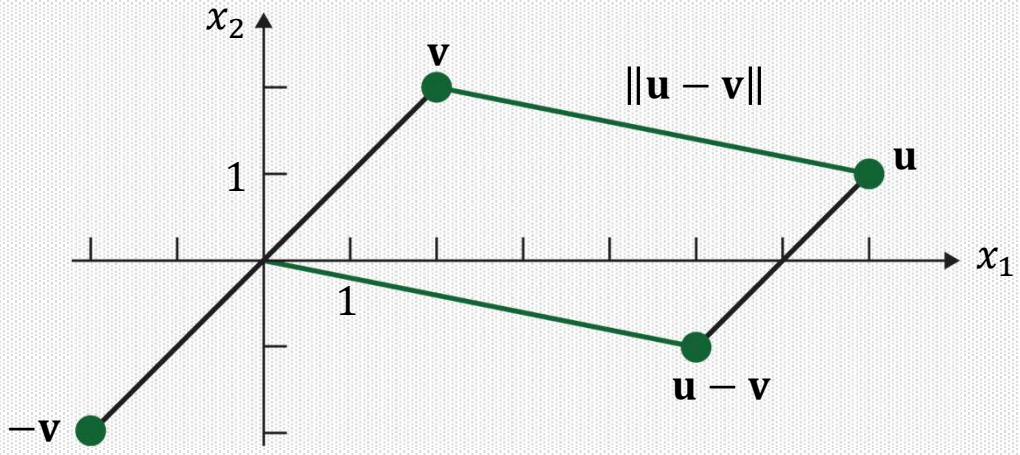
- **Example**: Compute the distance between the vector $\mathbf{u} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$.
- **Solution**: Calculate

$$\mathbf{u} - \mathbf{v} = \begin{bmatrix} 6 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$\|\mathbf{u} - \mathbf{v}\| = \sqrt{3^2 + (-1)^2} = \sqrt{10}$$

Distance between Vectors in $\mathbb{R}^n$

- The distance from $\mathbf{u}$ to $\mathbf{v}$ is the same as the distance from $\mathbf{u} - \mathbf{v}$ to 0.



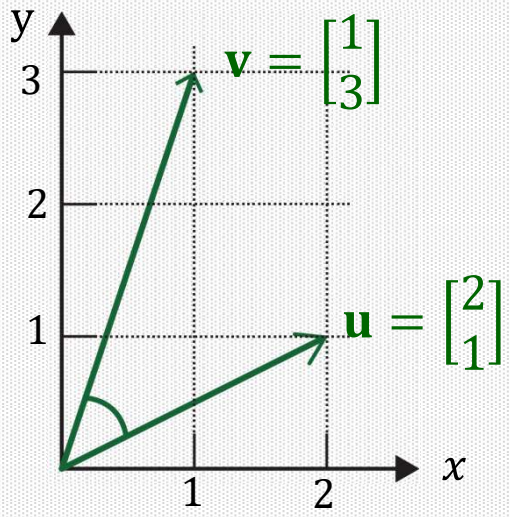
The distance between $\mathbf{u}$ and $\mathbf{v}$ is the length of $\mathbf{u} - \mathbf{v}$

Inner Product and Angle Between Vectors

- Inner product between $\mathbf{u}$ and $\mathbf{v}$ can be rewritten using their norms and angle:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta$$

- **Example**:



$$\mathbf{u} \cdot \mathbf{v} = \begin{bmatrix}2\\1\end{bmatrix} \cdot \begin{bmatrix}1\\3\end{bmatrix} = [2 \quad 1] \begin{bmatrix}1\\3\end{bmatrix} = 5$$

$$\|\mathbf{u}\| = \sqrt{2^2 + 1^2} = \sqrt{5} \quad \|\mathbf{v}\| = \sqrt{1^2 + 3^2} = \sqrt{10}$$

$$\mathbf{u} \cdot \mathbf{v} = 5 = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta = \sqrt{5} \cdot \sqrt{10} \cos\theta$$

$$\Rightarrow \cos\theta = \frac{5}{\sqrt{50}} = \frac{1}{\sqrt{2}}$$

$$\Rightarrow \theta = 45^\circ$$

Orthogonal Vectors

- **Definition:** $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$ are **orthogonal** (to each other) if $\mathbf{u} \cdot \mathbf{v} = 0$

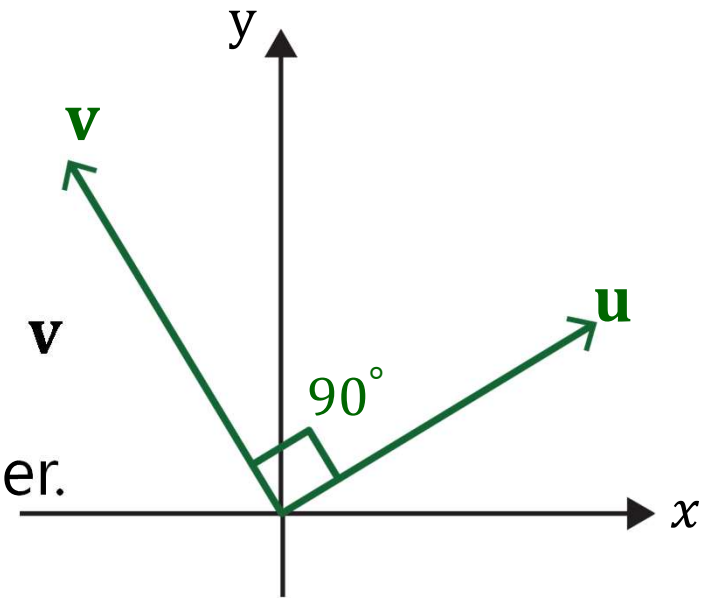
  That is,

  $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \cos\theta = 0.$

  ➡ $\cos\theta = 0$ for nonzero vectors $\mathbf{u}$ and $\mathbf{v}$

  ➡ $\theta = 90°$ ($\mathbf{u} \perp \mathbf{v}$).

  ➡ $\mathbf{u}$ and $\mathbf{v}$ are perpendicular each other.



Back to Over-Determined System

- Let's start with the original problem:

Person ID	Weight	Height	Is_smoking	Life-span
1	60kg	5.5ft	Yes (=1)	66
2	65kg	5.0ft	No (=0)	74
3	55kg	6.0ft	Yes (=1)	78

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \end{bmatrix}}$$

- Using the inverse matrix, the solution is $\mathbf{x} = \begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}$.

Back to Over-Determined System

- Let's add an additional example:

Person ID	Weight	Height	Is_smoking	Life-span
1	60kg	5.5ft	Yes (=1)	66
2	65kg	5.0ft	No (=0)	74
3	55kg	6.0ft	Yes (=1)	78
4	50kg	5.0ft	Yes (=1)	72

$$\overset{A}{\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 0\end{bmatrix}}\overset{\mathbf{x}}{\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}} = \overset{\mathbf{b}}{\begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix}}$$

- Now, let's use the previous solution $\mathbf{x} = \begin{bmatrix}-0.4\\ 20\\ -20\end{bmatrix}$

$$\overset{A}{\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 1\end{bmatrix}}\overset{\mathbf{x}}{\begin{bmatrix}-0.4\\ 20\\ -20\end{bmatrix}} = \begin{bmatrix}66\\ 74\\ 78\\ 60\end{bmatrix} \neq \overset{\mathbf{b}}{\begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix}} \quad \overset{\text{Errors}}{\begin{vmatrix}(\mathbf{b}-A\mathbf{x})\\ 0\\ 0\\ 0\\ 12\end{vmatrix}}$$

Back to Over-Determined System

- How about using slightly different solution $\mathbf{x} = \begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix}$?

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix}} = \begin{bmatrix} 71.3 \\ 72.2 \\ 79.9 \\ 64.5 \end{bmatrix} \neq \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}$$

Errors

$$(\mathbf{b} - A\mathbf{x}) = \begin{bmatrix} -5.3 \\ 1.8 \\ -1.9 \\ 7.5 \end{bmatrix}$$

Which One is Better Solution?

$$\underset{A}{\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 1\end{bmatrix}}\underset{\mathbf{x}}{\begin{bmatrix}-0.12\\ 16\\ -9.5\end{bmatrix}} = \begin{bmatrix}71.3\\ 72.2\\ 79.9\\ 64.5\end{bmatrix} \neq \underset{\mathbf{b}}{\begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix}} \quad \text{Errors: } (\mathbf{b}-A\mathbf{x}) = \begin{bmatrix}-5.3\\ 1.8\\ -1.9\\ 7.5\end{bmatrix}$$

$$\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 1\end{bmatrix}\begin{bmatrix}-0.4\\ 20\\ -20\end{bmatrix} = \begin{bmatrix}66\\ 74\\ 78\\ 60\end{bmatrix} \neq \begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix} \quad \begin{bmatrix}0\\ 0\\ 0\\ 12\end{bmatrix}$$

Least Squares: Best Approximation Criterion

- Let's use the squared sum of errors:

	A	$\mathbf{x}$		$\mathbf{b}$	Errors	Sum of squared errors
$\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 1\end{bmatrix}$		$\begin{bmatrix}-0.12\\ 16\\ -9.5\end{bmatrix}$	$=\begin{bmatrix}71.3\\ 69\\ 79.9\\ 64.5\end{bmatrix}\neq$	$\begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix}$	$\begin{bmatrix}(\mathbf{b}-A\mathbf{x})\\ -5.3\\ 1.8\\ -1.9\\ 7.5\end{bmatrix}$	$\left((-5.3)^2+1.8^2+(-1.9)^2+7.5^2\right)^{0.5}$ $=9.55$ *Better solution*
$\begin{bmatrix}60 & 5.5 & 1\\ 65 & 5.0 & 0\\ 55 & 6.0 & 1\\ 50 & 5.0 & 1\end{bmatrix}$		$\begin{bmatrix}-0.4\\ 20\\ -20\end{bmatrix}$	$=\begin{bmatrix}66\\ 74\\ 78\\ 60\end{bmatrix}\neq$	$\begin{bmatrix}66\\ 74\\ 78\\ 72\end{bmatrix}$	$\begin{bmatrix}0\\ 0\\ 0\\ 12\end{bmatrix}$	$(0^2+0^2+0^2+12^2)^{0.5}$ $=12$

Least Squares Problem

- Now, the sum of squared errors can be represented as $\|\mathbf{b} - A\mathbf{x}\|$.
- **Definition**: Given an overdetermined system $A\mathbf{x} \simeq \mathbf{b}$ where $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^{n}$, and $m \gg n$, a least squares solution $\hat{\mathbf{x}}$ is defined a s

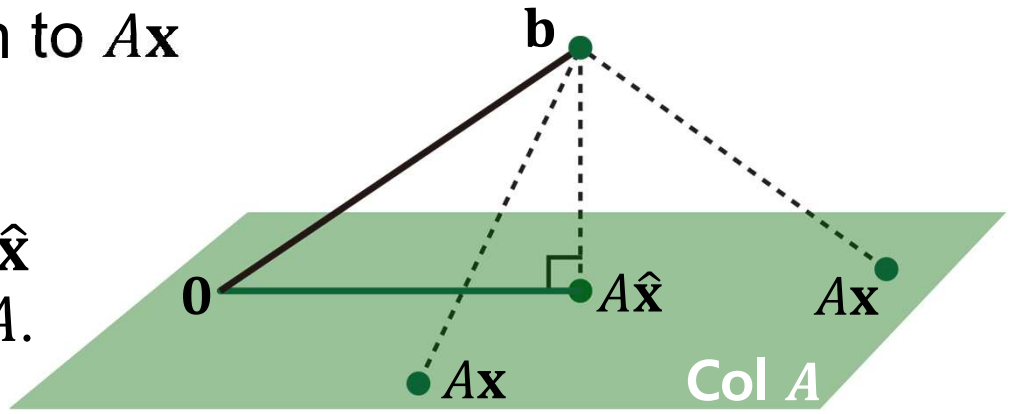
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|$$

- The most important aspect of the least-squares problem is that no matter what $\mathbf{x}$ we select, the vector $A\mathbf{x}$ will necessarily be in the column space Col A.
- Thus, we seek for $\mathbf{x}$ that makes $A\mathbf{x}$ as the closest point in Col A to $\mathbf{b}$.

Geometric Interpretation of Least Squares

- The vector $\mathbf{b}$ is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other $\mathbf{x}$.

- To satisfy this, the vector $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to Col A.

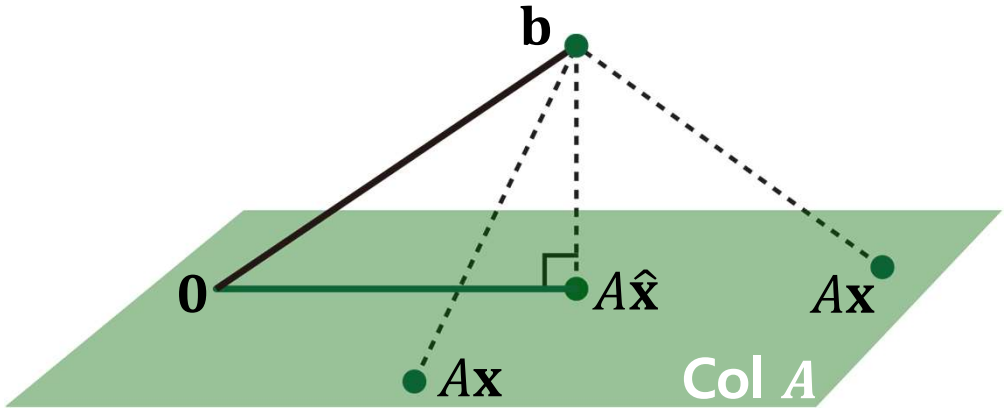


- This means $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to any vector in Col A:

$$\mathbf{b} - A\hat{\mathbf{x}} \perp (x_1\mathbf{a}_1 + x_2\mathbf{a}_2 \cdots + x_n\mathbf{a}_n) \text{ for any vector } \mathbf{x}$$

Geometric Interpretation of Least Squares

- $\mathbf{b} - A\hat{\mathbf{x}} \perp (x_1\mathbf{a}_1 + x_2\mathbf{a}_2 \cdots + x_n\mathbf{a}_n)$ for any vector $\mathbf{x}$



- Or equivalently,

$$(\mathbf{b} - A\hat{\mathbf{x}}) \perp \mathbf{a}_1 \qquad \mathbf{a}_1^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0$$
$$(\mathbf{b} - A\hat{\mathbf{x}}) \perp \mathbf{a}_2 \Rightarrow \quad \mathbf{a}_2^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0 \Rightarrow \quad A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}$$
$$\vdots \qquad\qquad \vdots$$
$$(\mathbf{b} - A\hat{\mathbf{x}}) \perp \mathbf{a}_n \qquad \mathbf{a}_n^T(\mathbf{b} - A\hat{\mathbf{x}}) = 0$$

Normal Equation

- Finally, given a least squares problem, $A\mathbf{x} \simeq \mathbf{b}$, we obtain

$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b},$$

  which is called a normal equation.

- This can be viewed as a new linear system, $C\mathbf{x} = \mathbf{d}$,
  where a square matrix $C = A^T A \in \mathbb{R}^{n\times n}$, and $\mathbf{d} = A^T\mathbf{b} \in \mathbb{R}^n$.

- If $C = A^T A$ is invertible, then the solution is computed as

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

Another Derivation of Normal Equation

- $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\| = \arg\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|^2$

  $= \arg\min_{\mathbf{x}} (\mathbf{b} - A\mathbf{x})^T (\mathbf{b} - A\mathbf{x}) = \mathbf{b}^T\mathbf{b} - \mathbf{x}^T A^T\mathbf{b} - \mathbf{b}^T A\mathbf{x} + \mathbf{x}^T A^T A\mathbf{x}$

- Computing derivatives w.r.t. $\mathbf{x}$, we obtain

$$-A^T\mathbf{b} - A^T\mathbf{b} + 2A^T A\mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad A^T A\mathbf{x} = A^T\mathbf{b}$$

- Thus, if $C = A^T A$ is invertible, then the solution is computed as

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

Life-Span Example

Person ID	Weight	Height	Is_smoking	Life-span
1	60kg	5.5ft	Yes (=1)	66
2	65kg	5.0ft	No (=0)	74
3	55kg	6.0ft	Yes (=1)	78
4	50kg	5.0ft	Yes (=1)	72

$$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} \simeq \underset{\mathbf{b}}{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}$$

- The normal equation $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$ is

$$\begin{bmatrix} 60 & 65 & 55 & 50 \\ 5.5 & 5.0 & 6.0 & 5.0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 60 & 65 & 55 & 50 \\ 5.5 & 5.0 & 6.0 & 5.0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}$$

$$\begin{bmatrix} 13350 & 1235 & 165 \\ 1235 & 116.25 & 16.5 \\ 165 & 16.5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 16600 \\ 1561 \\ 216 \end{bmatrix}$$

What If $C = A^T A$ is NOT Invertible?

- Given $A^{\mathrm{T}}A\mathbf{x} = A^{\mathrm{T}}\mathbf{b}$, what if $C = A^T A$ is NOT invertible?
- Remember that in this case, the system has either no solution or infinitely many solutions.
- However, the solution always exist for this "normal" equation, and thus infinitely many solutions exist.
- When $C = A^T A$ is NOT invertible?
  If and only if the columns of A are linearly dependent. Why?
- However, $C = A^T A$ is usually invertible. Why?

Orthogonal Projection Perspective

- Back to the case of invertible $C = A^T A$, consider the orthogonal pr ojection of $\mathbf{b}$ onto Col A as

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

  where $C = A(A^T A)^{-1} A^T$.

- One can see that the orthogonal projection is actually a **linear transformation** $f(\mathbf{b}) = C\mathbf{b}$ where the standard matrix is defined as $C = A(A^T A)^{-1} A^T$.

- What if A has orthonormal columns? (More in the next slides.)

Orthogonal and Orthonormal Sets

- **Definition**: A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ in $\mathbb{R}^n$ is an **orthogonal set** if each pair of distinct vectors from the set is orthogonal That is, if $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ whenever $i \neq j$.

- **Definition**: A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ in $\mathbb{R}^n$ is an **orthonormal set** if it is an orthogonal set of unit vectors.

- Is an orthogonal (or orthonormal) set also a linearly independent set? What about its converse?

Orthogonal and Orthonormal Basis

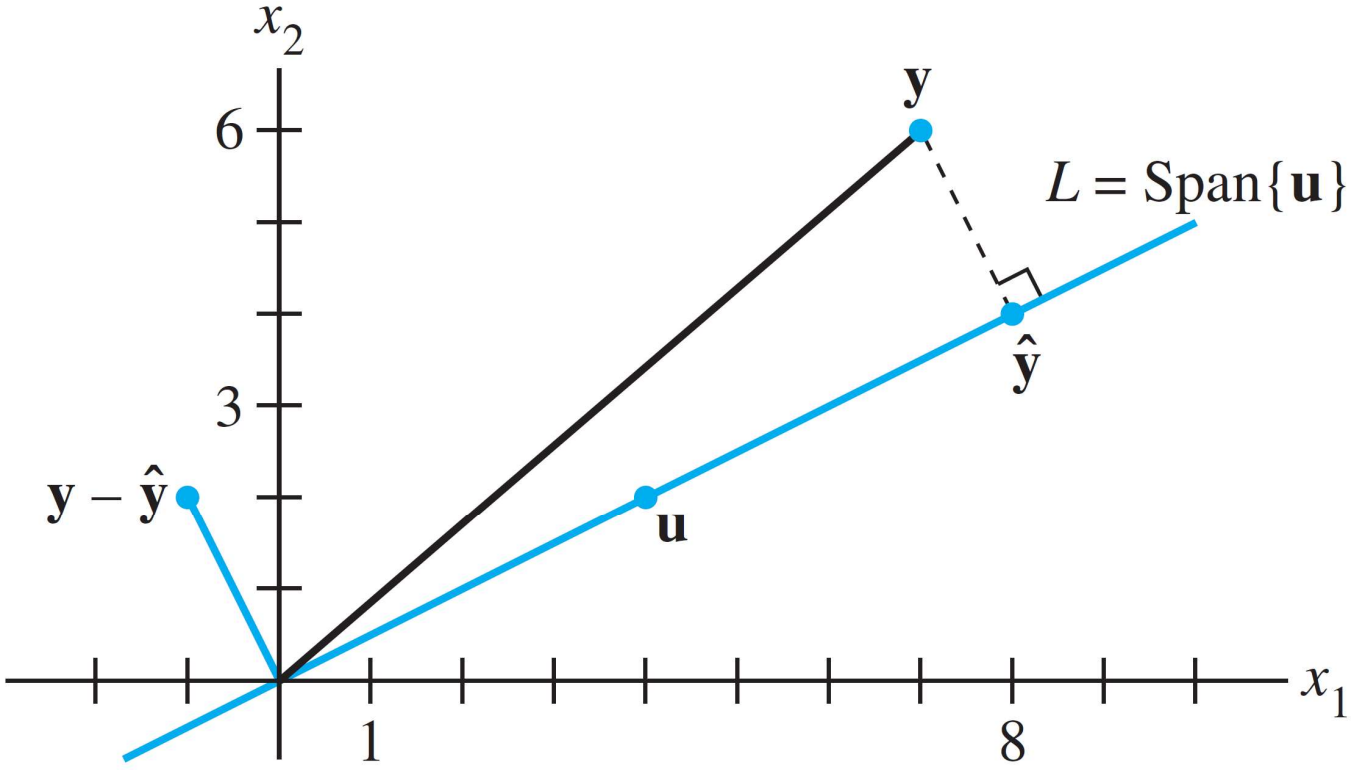
- Consider basis $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ of a p-dimensional subspace W in $\mathbb{R}^n$.
- Can we make it as an orthogonal (or orthonormal) basis?
  - Yes, it can be done by Gram–Schmidt process. → QR factorization.
- Given the orthogonal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$ of W,
  let's compute the orthogonal projection of $\mathbf{y} \in \mathbb{R}^n$ onto W.

Orthogonal Projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto Line

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto one-dimensional subspace

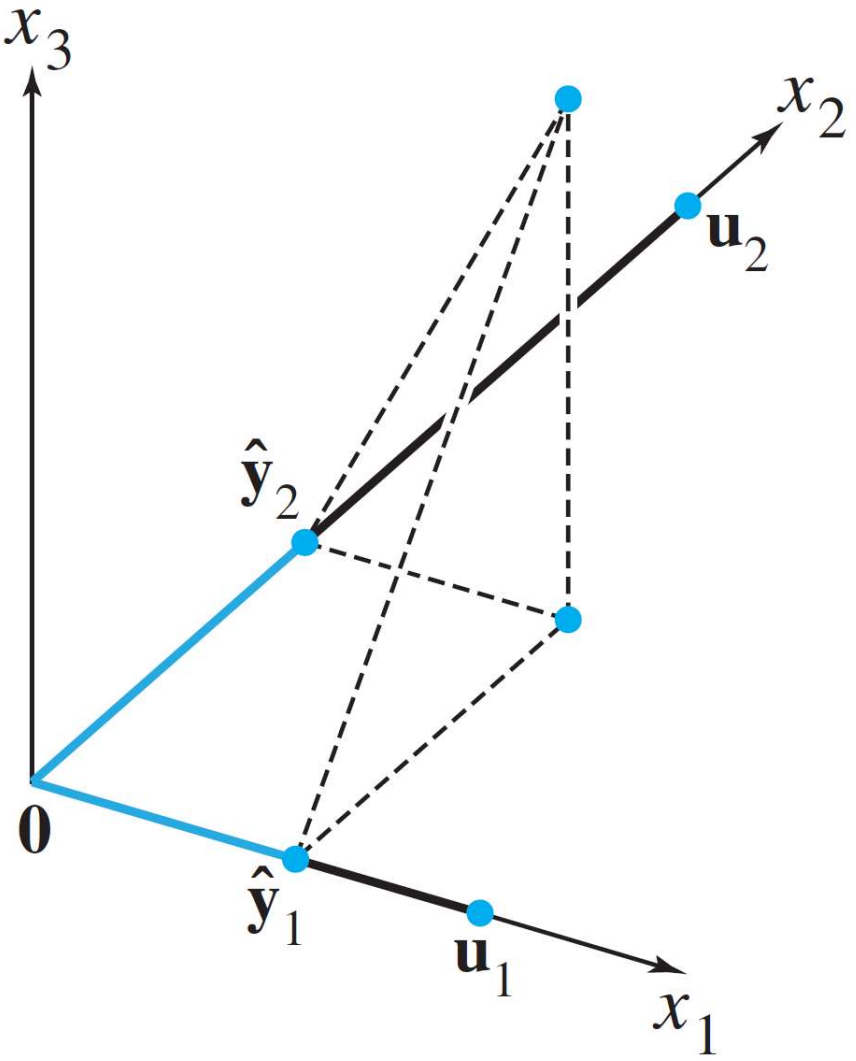
- $\hat{\mathbf{y}} = \text{proj}_L \mathbf{y} = \frac{\mathbf{y} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}$

- If $\mathbf{u}$ is a unit vector,
  $\hat{\mathbf{y}} = \text{proj}_L \mathbf{y} = (\mathbf{y} \cdot \mathbf{u})\mathbf{u}$



Orthogonal Projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto Plane

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto two-dimensional subspace W
- $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = \frac{\mathbf{y}\cdot\mathbf{u}_1}{\mathbf{u}_1\cdot\mathbf{u}_1}\mathbf{u}_1 + \frac{\mathbf{y}\cdot\mathbf{u}_2}{\mathbf{u}_2\cdot\mathbf{u}_2}\mathbf{u}_2$
- If $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors,
  $\hat{\mathbf{y}} = \text{proj}_L\mathbf{y} = (\mathbf{y}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{y}\cdot\mathbf{u}_2)\mathbf{u}_2$
- Projection is done independently on each orthogonal basis vector.



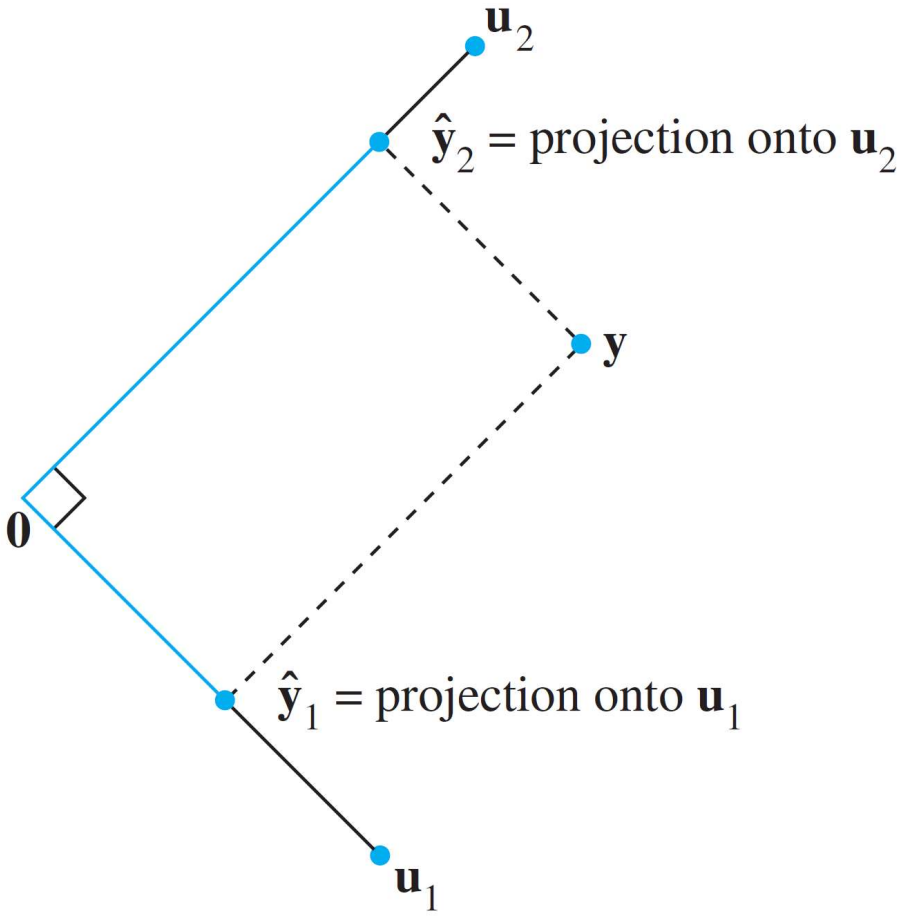
Orthogonal Projection when $\mathbf{y} \in W$

- Consider the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto two-dimensional subspace W, where $\mathbf{y} \in W$

- $\hat{\mathbf{y}} = \text{proj}_L \mathbf{y} = \mathbf{y} = \frac{\mathbf{y}\cdot\mathbf{u}_1}{\mathbf{u}_1\cdot\mathbf{u}_1}\mathbf{u}_1 + \frac{\mathbf{y}\cdot\mathbf{u}_2}{\mathbf{u}_2\cdot\mathbf{u}_2}\mathbf{u}_2$

- If $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors, $\hat{\mathbf{y}} = \mathbf{y} = (\mathbf{y}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{y}\cdot\mathbf{u}_2)\mathbf{u}_2$

- The solution is the same as before. Why?



Transformation: Orthogonal Projection

- Consider a transformation of orthogonal projection $\hat{\mathbf{b}}$ of $\mathbf{b}$, given orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2\}$ of a subspace W:

$$\begin{aligned}\hat{\mathbf{b}} &= f(\mathbf{b}) = (\mathbf{b}\cdot\mathbf{u}_1)\mathbf{u}_1 + (\mathbf{b}\cdot\mathbf{u}_2)\mathbf{u}_2 \\ &= (\mathbf{u}_1^T\mathbf{b})\mathbf{u}_1 + (\mathbf{u}_2^T\mathbf{b})\mathbf{u}_2 \\ &= \mathbf{u}_1(\mathbf{u}_1^T\mathbf{b}) + \mathbf{u}_2(\mathbf{u}_2^T\mathbf{b}) \\ &= (\mathbf{u}_1\mathbf{u}_1^T)\mathbf{b} + (\mathbf{u}_2\mathbf{u}_2^T)\mathbf{b} \\ &= (\mathbf{u}_1\mathbf{u}_1^T + \mathbf{u}_2\mathbf{u}_2^T)\mathbf{b} \\ &= [\mathbf{u}_1 \quad \mathbf{u}_2]\begin{bmatrix}\mathbf{u}_1^T \\ \mathbf{u}_2^T\end{bmatrix}\mathbf{b} = UU^T\mathbf{b} = C\mathbf{b} \Rightarrow \text{linear transformation!}\end{aligned}$$

Orthogonal Projection Perspective

- Let's verify the following, when $A = U = [\mathbf{u}_1 \quad \mathbf{u}_2]$ has orthonormal columns:
  Back to the case of invertible $C = A^T A$, consider the orthogon al projection of $\mathbf{b}$ onto Col A as

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

- $C = A^T A = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \end{bmatrix} [\mathbf{u}_1 \quad \mathbf{u}_2] = I$. Thus,

$$\hat{\mathbf{b}} = A\hat{\mathbf{x}} = A(A^T A)^{-1} A^T \mathbf{b} = A(I)^{-1} A^T \mathbf{b} = AA^T \mathbf{b} = UU^T \mathbf{b}$$

Further Study

- Least-squares derivation from maximum likelihood perspectiv e
  (via Gaussian distribution)
  - Kevin Murphy, “Machine Learning: A Probabilistic Perspective,” Ch7.2
- Orthogonal projection and QR decomposition
  - Lay Ch6.2, Ch.6.3, Ch6.4

Gram-Schmidt Orthogonalization

- **Example 1**: Let $W = \text{Span}\{\mathbf{x}_1, \mathbf{x}_2\}$, where $\mathbf{x}_1 = \begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$.
  Construct an orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ for W.

- **Solution:** Let $\mathbf{v}_1 = \mathbf{x}_1$. Next, Let $\mathbf{v}_2$ the component of $\mathbf{x}_2$ orthogonal to $\mathbf{x}_1$, i.e.,

$$\mathbf{v}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} - \frac{15}{45}\begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

- The set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthogonal basis for W.

Gram-Schmidt Orthogonalization

- **Example 2**: Let $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, and $\mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$. Then

$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is clearly linearly independent and thus a basis for a subspace W of $\mathbb{R}^4$. Construct an orthogonal basis for W.

Gram-Schmidt Orthogonalization

- **Solution:**
- **Step 1.** Let $\mathbf{v}_1 = \mathbf{x}_1$ and $W_1 = \text{Span}\{\mathbf{x}_1\} = \text{Span}\{\mathbf{v}_1\}$.
- **Step 2.** Let $\mathbf{v}_2$ be the vector produced by subtracting from $\mathbf{x}_2$ its projection onto the subspace W_1. That is, let

$$\mathbf{v}_2 = \mathbf{x}_2 - \text{proj}_{W_1}\mathbf{x}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 = \begin{bmatrix} -3/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

- $\mathbf{v}_2$ is the component of $\mathbf{x}_2$ orthogonal to $\mathbf{x}_1$, and $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthogonal basis for the subspace W_2 spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

Gram-Schmidt Orthogonalization

- **Step 2′ (optional).** If appropriate, scale $\mathbf{v}_2$ to simplify later computations, e.g.,

$$\mathbf{v}_2 = \begin{bmatrix} -3/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \longrightarrow \mathbf{v}_2' = \begin{bmatrix} -3 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Gram-Schmidt Orthogonalization

- **Step 3.** Let $\mathbf{v}_3$ be the vector produced by subtracting from $\mathbf{x}_3$ its projection onto the subspace W_2. Use the orthogonal basis $\{\mathbf{v}_1, \mathbf{v}_2'\}$ to compute this projection onto W_2:

$$\text{proj}_{W_2}\mathbf{x}_3 = \frac{\mathbf{x}_3 \cdot \mathbf{v}_1}{\mathbf{v}_3 \cdot \mathbf{v}_1}\mathbf{v}_1 + \frac{\mathbf{x}_3 \cdot \mathbf{v}_2'}{\mathbf{v}_3 \cdot \mathbf{v}_2'}\mathbf{v}_2' = \begin{bmatrix} 0 \\ 2/3 \\ 2/3 \\ 2/3 \end{bmatrix}$$

- Then $\mathbf{v}_3$ is the component of $\mathbf{x}_3$ orthogonal to W_2, namely,

$$\mathbf{v}_3 = \mathbf{x}_3 - \text{proj}_{W_2}\mathbf{x}_3 = \begin{bmatrix} 0 \\ -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Gram-Schmidt Orthogonalization

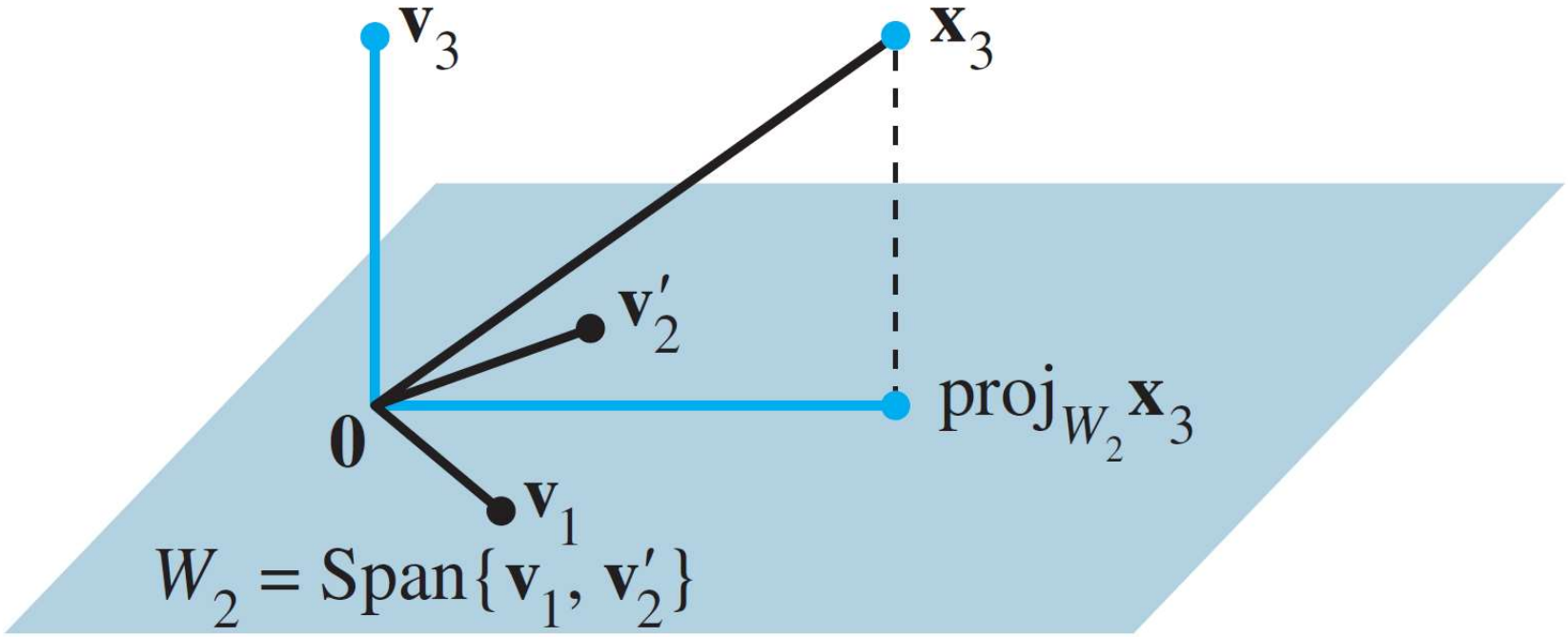


FIGURE 2 The construction of $\mathbf{v}_3$ from $\mathbf{x}_3$ and W_2.

Figure from Lay Ch6.4

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.

Computing QR Factorization

- **Step 1 (Construction of Q):** The columns of A form a basis for $\mathrm{Col}\, A$ since they are linearly independent. Let these columns be $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Then, we can construct the orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ for $\mathrm{Col}\, A$ by the Gram-Schmidt process described by Theorem 11. Using this basis, we can construct Q as

$$Q = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n]$$

Computing QR Factorization

- **Step 2 (Construction of R):** From (1) in Theorem 11, for $k = 1, \dots, n$, $\mathbf{x}_k$ is in $\text{Span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} = \text{Span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$. Therefore, there exist constants $r_{1k}, \dots, r_{kk}$ such that

$$\mathbf{x}_k = r_{1k}\mathbf{u}_1 + \cdots + r_{kk}\mathbf{u}_k + 0 \cdot \mathbf{u}_{k+1} + \cdots + 0 \cdot \mathbf{u}_n$$

- We can always make $r_{kk} \geq 0$ because if $r_{kk} < 0$, then we can multiply both r_{kk} and $\mathbf{u}_k$ by -1. Using this linear combination representation, we can construct $\mathbf{r}_k$, the k-th column of R, as

$$\mathbf{r}_k = \begin{bmatrix} r_{1k} \\ \vdots \\ r_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Computing QR Factorization

- That is, $\mathbf{x}_k = Q\mathbf{r}_k$ for $k = 1, \ldots, n$. Let $R = [\mathbf{r}_1 \quad \cdots \quad \mathbf{r}_n]$. Then,
$$A = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = [Q\mathbf{r}_1 \quad \cdots \quad Q\mathbf{r}_n] = QR$$

- The fact that R is invertible follows easily from the fact that the columns of A are linearly independent (Exercise 19). Since R is clearly upper triangular (from the previous slide) and invertible, the diagonal entries r_{kk}'s should be nonzero. By combining this with the fact that $r_{kk} \geq 0$, r_{kk}'s must be positive.

Example: QR Factorization

- **Example 4:** Find a QR factorization of $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.
- **Solution:** Let $A = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]$. We first obtain $\mathbf{v}_1 = \mathbf{x}_1$ and its normalized vector is $\mathbf{u}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$.
- Thus, $\mathbf{x}_1 = 2\mathbf{u}_1$, which gives us $\mathbf{r}_{11} = 2$, i.e., $\mathbf{r}_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$.

Example: QR Factorization

- Next, we obtain $\mathbf{v}_3$ as $\mathbf{v}_3 = \mathbf{x}_3 - \text{proj}_{W_2}\mathbf{x}_3 = \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\mathbf{u}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2}\mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} - 1\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} - \frac{2}{\sqrt{12}}\begin{bmatrix} -3/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \\ 1/\sqrt{12} \end{bmatrix} = \begin{bmatrix} 0 \\ -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$ and its normalized vector $\mathbf{u}_2$ as $\mathbf{u}_2 = \begin{bmatrix} 0 \\ -2/\sqrt{6} \\ 1/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$.
- Thus, $\mathbf{x}_3 = 1\mathbf{u}_1 + \frac{2}{\sqrt{12}}\mathbf{u}_2 + \frac{2}{\sqrt{6}}\mathbf{u}_3$, i.e., $\mathbf{r}_3 = \begin{bmatrix} 1 \\ 2/\sqrt{12} \\ 2/\sqrt{6} \end{bmatrix}$.

Example: QR Factorization

- In conclusion, $Q = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \mathbf{u_3}] = \begin{bmatrix} 1/2 & -3/\sqrt{12} & 0 \\ 1/2 & 1/\sqrt{12} & -2/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \end{bmatrix}$

and $R = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{r}_3] = \begin{bmatrix} 2 & -3/2 & 1 \\ 0 & -3/\sqrt{12} & 2/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} \end{bmatrix}$.