Least Squares Problem

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
⋮	⋮	⋮	⋮	⋮

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

- Matrix equation:

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

$m \gg n$: 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?

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

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

- 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

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

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

$$\underbrace{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}}_{\mathbf{x}} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ 60 \end{bmatrix} \neq \underbrace{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}_{\mathbf{b}}$$

Errors

$$(\mathbf{b} - A\mathbf{x}) = \begin{matrix} 0 \\ 0 \\ 0 \\ 12 \end{matrix}$$

Back to Over-Determined System

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

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

Errors

Which One is Better Solution?

Errors

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

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

Least Squares: Best Approximation Criterion

- Let's use the squared sum of errors:

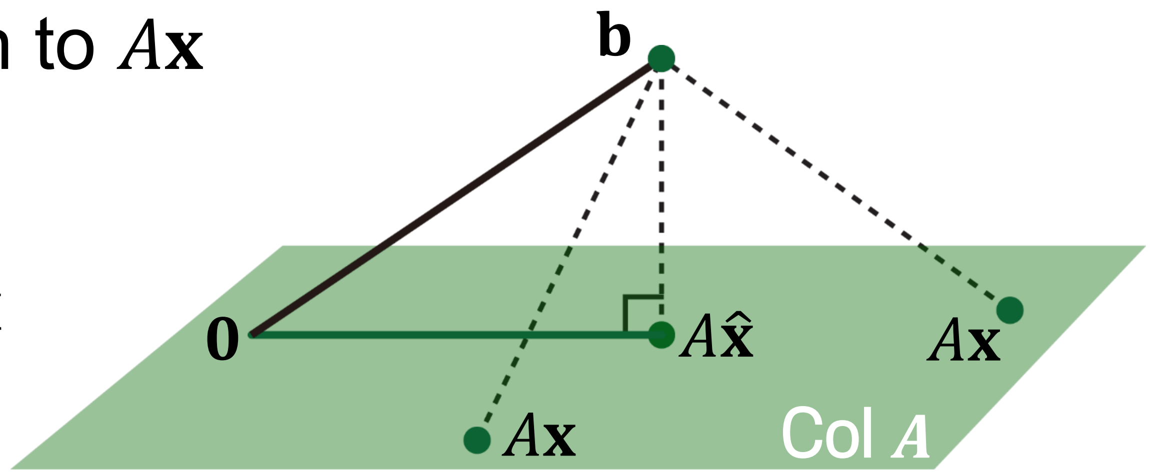
	Errors	Sum of squared errors
$\underset{A}{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}} \underset{\mathbf{x}}{\begin{bmatrix} -0.12 \\ 16 \\ -9.5 \end{bmatrix}} = \begin{bmatrix} 71.3 \\ 69 \\ 79.9 \\ 64.5 \end{bmatrix} \neq \underset{\mathbf{b}}{\begin{bmatrix}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

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



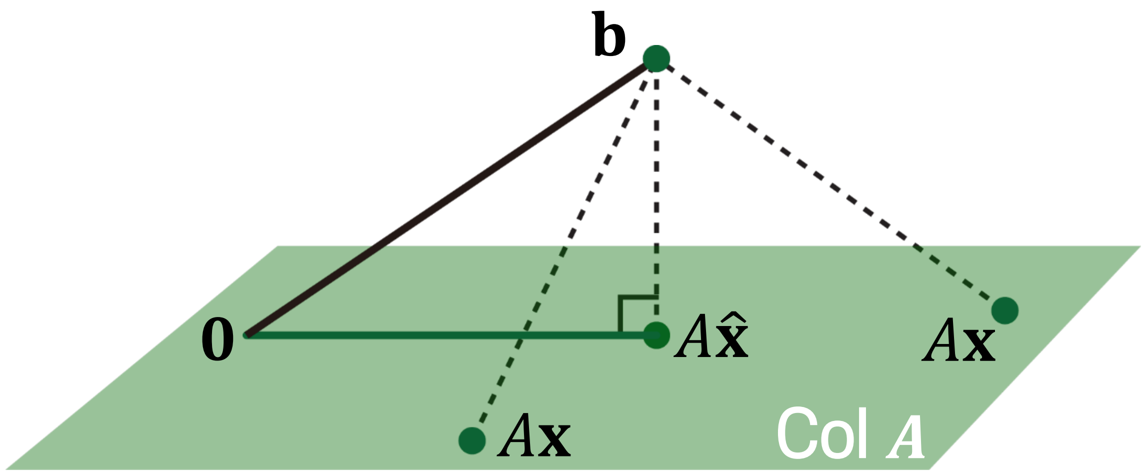
- To satisfy this, the vector $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to Col A.

- 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,

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

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

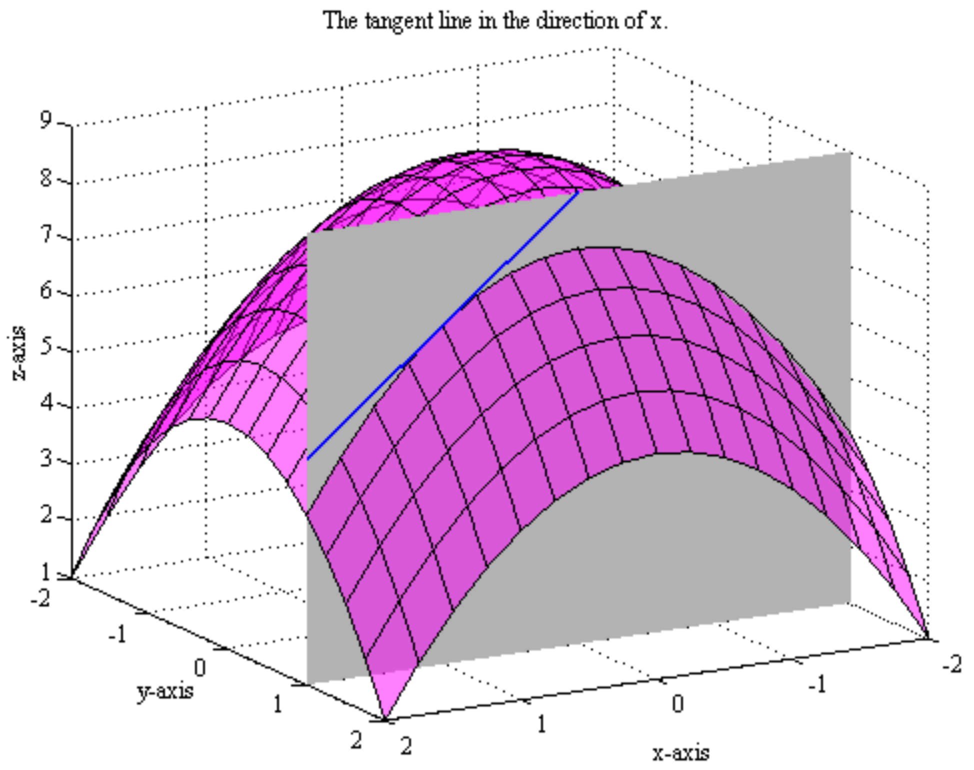
$$-A^T\mathbf{b} - A^T\mathbf{b} + 2A^T A\mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad A^T A\mathbf{x} = A^T\mathbf{b}$$

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

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

Partial Derivative

- For a multi-variate function, e.g., $f(x, y)$, one can consider a univariate function by assigning particular values to all other variables, e.g., $g(x) = f(x, y = 1)$. Then, one can consider a partial derivative $\frac{d}{dx}g(x)$ with respect to x.



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

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

- 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 projection 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

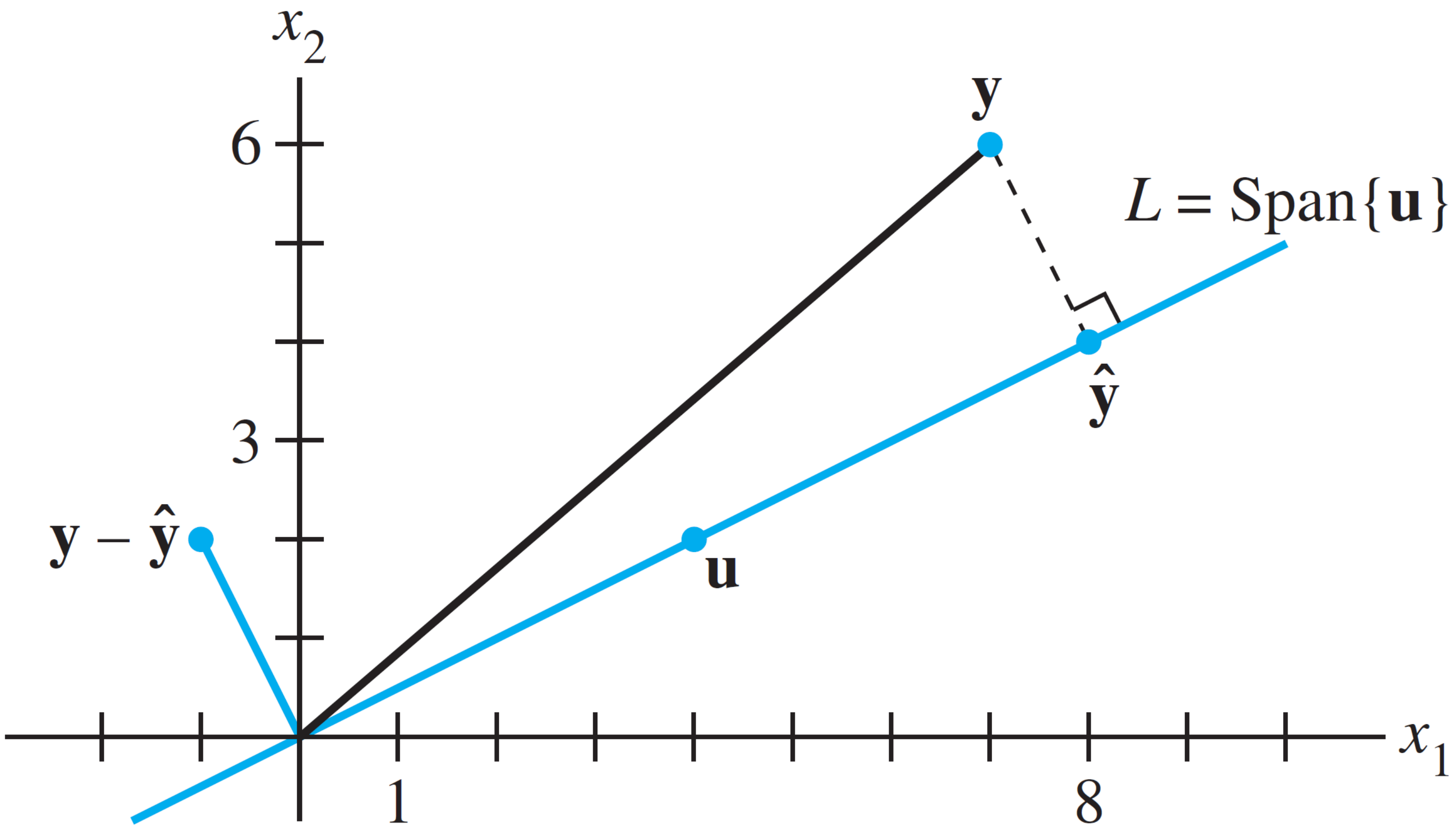
- Consider basis $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ of a p-dimensional subspace W in $\mathbb{R}^n$.
- Can we make it as an orthogonal (or orthonormal) basis?
  - Yes, it can be done by Gram–Schmidt process. → QR factorization.
- Given the orthogonal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$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

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

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



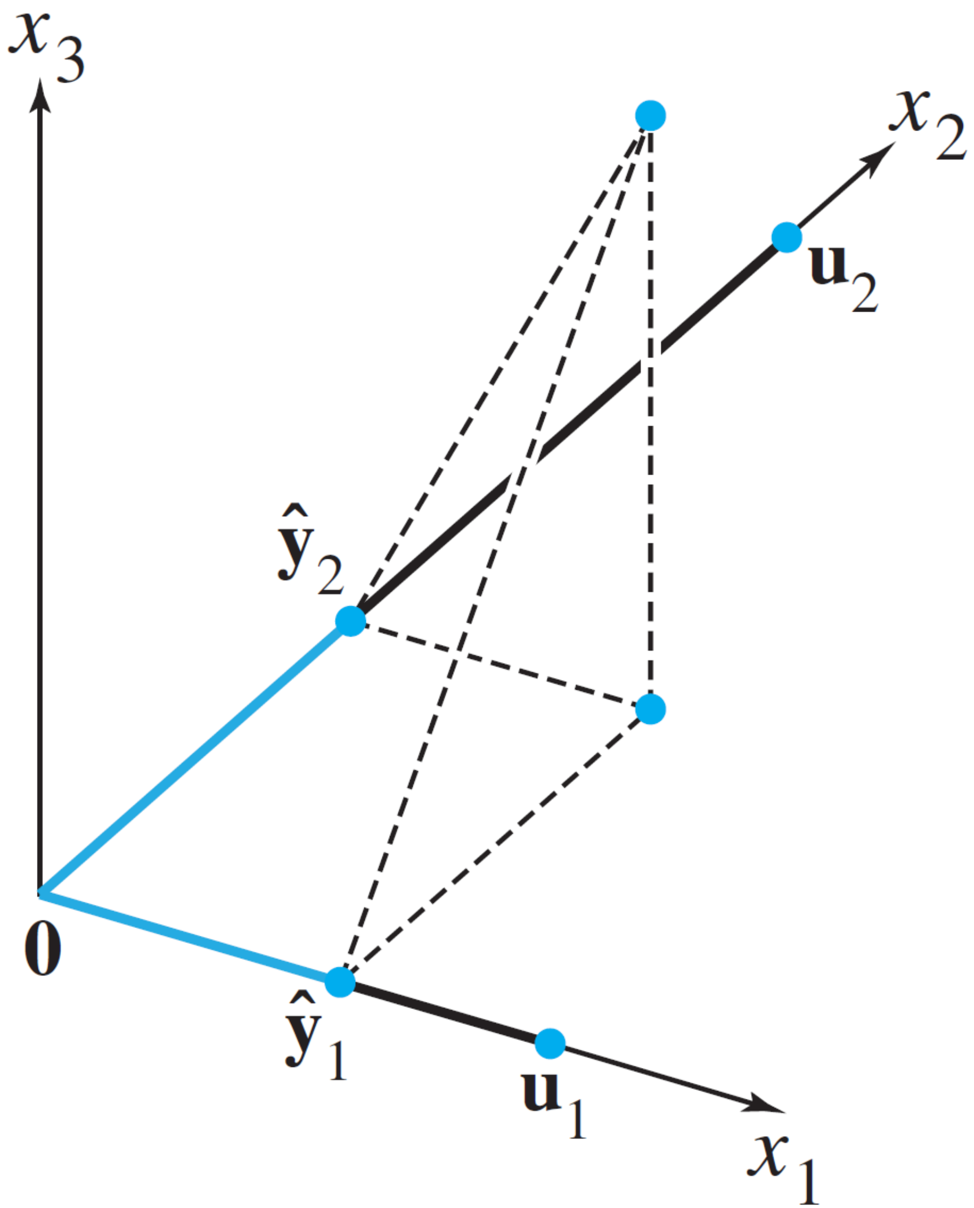
Orthogonal Projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto Plane

- 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.



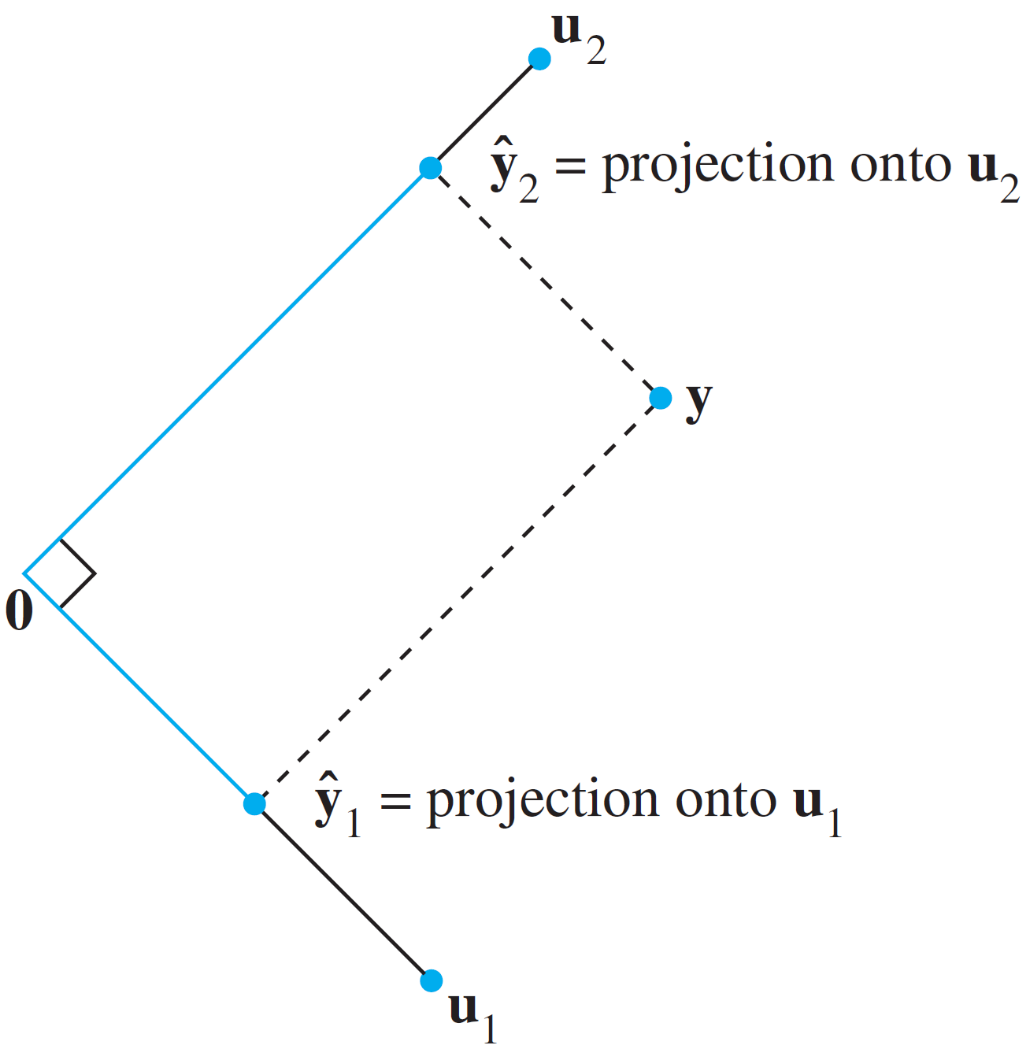
Orthogonal Projection when $\mathbf{y} \in W$

- 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:

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

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 orthogonal 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,

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