Linear Algebra

주재걸
고려대학교 컴퓨터학과

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 one more 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

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

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

Errors

$$\underbrace{\begin{bmatrix} 60 & 5.5 & 1 \\ 65 & 5.0 & 0 \\ 55 & 6.0 & 1 \\ 50 & 5.0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} -0.4 \\ 20 \\ -20 \end{bmatrix}}_{\mathbf{x}} = \begin{bmatrix} 66 \\ 74 \\ 78 \\ 60 \end{bmatrix} \neq \underbrace{\begin{bmatrix} 66 \\ 74 \\ 78 \\ 72 \end{bmatrix}}_{\neq\, \mathbf{b}} \quad \Bigg| \quad \begin{array}{c} (\mathbf{b} - A\mathbf{x}) \\ 0 \\ 0 \\ 0 \\ -12 \end{array}$$

Back to Over-Determined System

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

Errors

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

Which One is Better Solution?

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

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

Least Squares: Best Approximation Criterion

- Let's use the squared sum of errors:

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

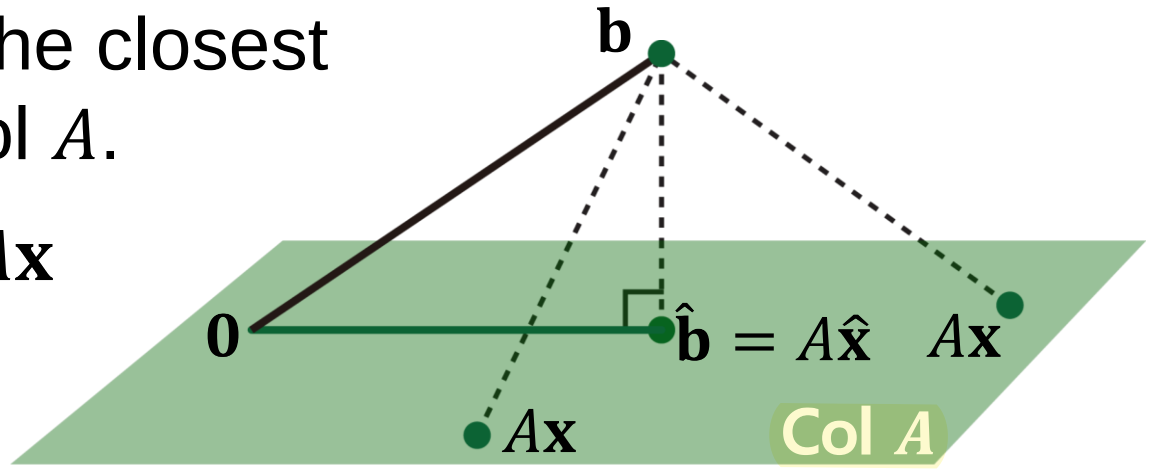
- **Definition**: Given an overdetermined system $A\mathbf{x} \simeq \mathbf{b}$ where $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^{n}$, and $m \gg n$, a least squares solution $\hat{\mathbf{x}}$ is defined as

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

- 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

- Consider $\hat{\mathbf{x}}$ such that $\hat{\mathbf{b}} = A\hat{\mathbf{x}}$ is the closest point to $\mathbf{b}$ among all points in Col A.
- That is, $\mathbf{b}$ is closer to $\hat{\mathbf{b}}$ than to $A\mathbf{x}$ for any other $\mathbf{x}$.
- To satisfy this, the vector $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to Col A.



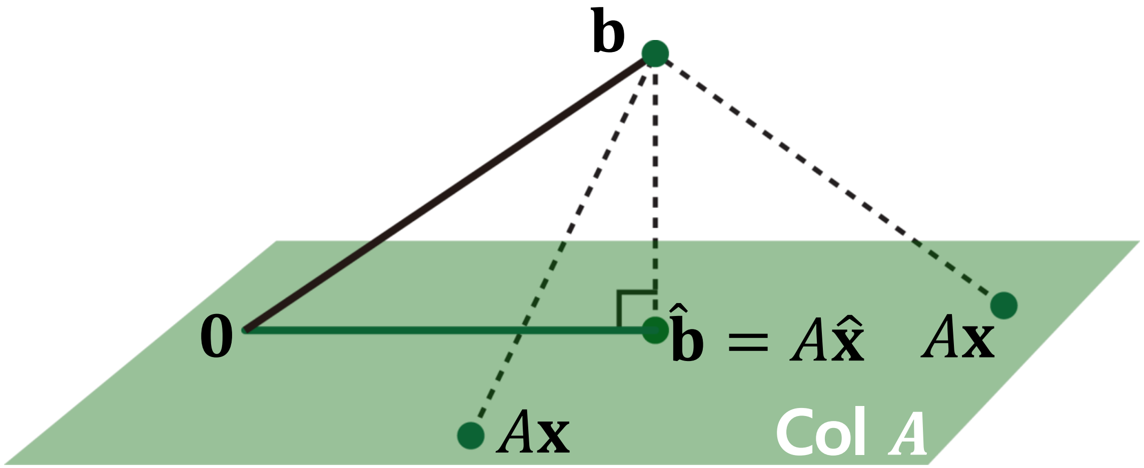
- This means $\mathbf{b} - A\hat{\mathbf{x}}$ should be orthogonal to any vector in Col A:

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

Geometric Interpretation of Least Squares

기저 벡터와 수직이면 된다

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



- Or equivalently,

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

$A^T b - A^T A \hat{x} = 0$

수직은 내적이 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.

normal equation은 일반적으로 해가 존재한다

- 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

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

역행렬이 존재한다는 것은 A 행렬이 독립이라는 말이다

일반적으로 역행렬은 존재한다
이유는 A 행렬의 dimension이 길어질 수록 (data가 많을 때/방정식 수가 많아질때) 독립이 될 경우가 많기 때문임

역행렬이 존재하지 않으면 해가 무수히 많다

Another Derivation of Normal Equation

$\|x\|^2 = x^T x = [3\ 2]\begin{bmatrix}3\\2\end{bmatrix} = 9+4$

$\begin{bmatrix}3\\2\end{bmatrix}^2 = 9+4 = 13$

제곱의 최솟값을
구해도 같은 의미

벡터 미분

$f(x) = a^T x = x a^T$

$\frac{df}{dx} = a = \begin{bmatrix}3\\2\end{bmatrix}$

$[3\ 2]\begin{bmatrix}x_1\\x_2\end{bmatrix} = 3x_1 + 2x_2$

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

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

(컬럼벡터)

$f \cdot g = f'g + f g'$

함수 곱의 미분

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

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

최솟값을 구하기 위해 미분했을때 zero 값을 구한다

- 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

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

- 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. 직각을 이루는 수선 발은 항상 존재
- When $C = A^T A$ is NOT invertible?
  If and only if the columns of A are linearly dependent. Why? feature
- However, $C = A^T A$ is usually invertible. Why?

equation(data)가 늘어나면서 의존이 되는 경우가 거의 사라지기때문

의존 가능성 낮아짐