Four Ways to Solve Least Squares Problems

- Many applications lead to unsolvable linear equations Ax=b → produce a best solution
- The least squares method chooses x to make $||b-Ax||^2$ as small as possible
- A^TA: attractive symmetry, but size, condition number
- Large, ill-posed problems?

1. Solve $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ to minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$

2. Gram-Schmit $\mathbf{A} = \mathbf{Q}\mathbf{R}$ leads to $\mathbf{x} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b}$

3. The pseudoinverse directly multiplies $\mathbf{b}$ to give $\mathbf{x}$

4. The best is the limit of $\left(\mathbf{A}^T\mathbf{A} + \delta\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{b}$ as $\delta \to 0$

A^TA or A^TCA

applications	A^TA (or A^TCA), C: positive diagonal matrix
Mechanical engineering	Stiffness matrix
Circuit theory	Conductance matrix
Graph theory	Graph Laplacian
Mathematics	Gram matrix (inner products of columns of A)

- In large problems, A^TA is expensive and often dangerous to compute
  - How to avoid? $A=QR \rightarrow Rx=Q^Tb$
  - Try not to compute: orthogonal matrices, triangular matrices

Normal Equations (1)

$\mathbf{A}^T\mathbf{A}$ is inverible exactly when $\mathbf{A}$ has independent columns.

If $\mathbf{Ax} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$

Always $\mathbf{A}$ and $\mathbf{A}^T\mathbf{A}$ have the same null space!

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{0} \to \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} = \mathbf{0} \to \|\mathbf{Ax}\|^2 = \mathbf{0} \to \mathbf{Ax} = \mathbf{0}$$

$$N\left(\mathbf{A}^T\mathbf{A}\right) = N(\mathbf{A}),\ C\left(\mathbf{A}\mathbf{A}^T\right) = C(\mathbf{A}), \text{rank}\left(\mathbf{A}^T\mathbf{A}\right) = \text{rank}\left(\mathbf{A}\mathbf{A}^T\right) = \text{rank}(\mathbf{A})$$

$$\text{fit a } \underbrace{\text{straight line}}_{C+D\mathbf{x}} \text{ to } b_1,\ldots,b_m \to \left\{\begin{array}{l}\mathbf{Ax} = \mathbf{b} \to \mathbf{A}^{-1} \text{ when } m = n = r \\ \mathbf{b}:\ \text{vector of measurements}\end{array}\right\} \to \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

$$\text{minimize } \|\mathbf{Ax}-\mathbf{b}\|^2 = (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\underbrace{\mathbf{b}^T\mathbf{Ax}}_{\mathbf{x}^T\mathbf{A}^T\mathbf{b}} + \mathbf{b}^T\mathbf{b}$$

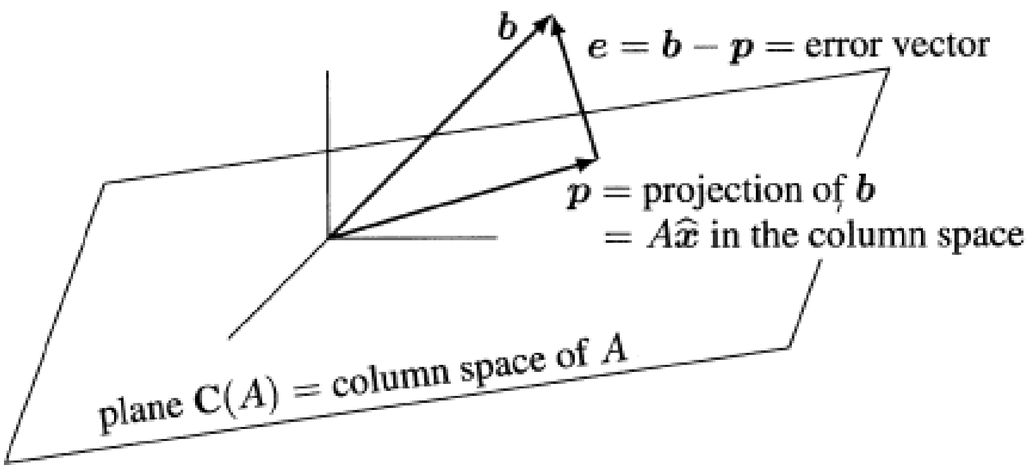
$$\xrightarrow{\text{optimality}} \mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$$

Normal Equations (2)

$\mathbf{e}$ is perpendicular to the plane (column space of $\mathbf{A}$)

$$\rightarrow (\mathbf{Ax})^T (\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0} \rightarrow \mathbf{x}^T \mathbf{A}^T (\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0} \rightarrow \mathbf{A}^T (\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0}$$

$$\begin{cases} \text{normal equation for } \hat{\mathbf{x}}\text{: } \mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b} \\ \text{least square solution to } \mathbf{Ax} = \mathbf{b}\text{: } \hat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \\ \text{projection to } \mathbf{b} \text{ onto the column space of } \mathbf{A}\text{: } \mathbf{p} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \\ \text{projection matrix that multiplies } \mathbf{b} \text{ to give } \mathbf{p}\text{: } \mathbf{P} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T \rightarrow \mathbf{P}^2 = \mathbf{P} \end{cases}$$



Gram-Schmit

columns of $\mathbf{A}$ are still assumed to be independent, but not to be orthogonal

$\mathbf{A}^T\mathbf{A}$ is a diagonal matrix and solving $\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$ needs work.

orthogonalize the columns of $\mathbf{A} \rightarrow \hat{\mathbf{x}}$ is easy to find

operation counts doubled, but orthogonal vectors provide numerical stability $\left(\mathbf{A}^T\mathbf{A}: \text{ nearly singular?}\right)$

$$\operatorname{cond}\left(\mathbf{A}^T\mathbf{A}\right) = \left\|\mathbf{A}^T\mathbf{A}\right\|\left\|\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\right\| = \frac{\sigma_1^{\,2}}{\sigma_n^{\,2}} \rightarrow \text{large? orthogonalize in advance!}$$

$$\operatorname{cond}(\mathbf{Q}) = \|\mathbf{Q}\|\left\|\mathbf{Q}^{-1}\right\| = 1$$

$$\left.\begin{aligned}
&\mathbf{q}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|} \\
&\mathbf{A}_2 = \mathbf{a}_2 - \left(\mathbf{a}_2^{\,T}\mathbf{q}_1\right)\mathbf{q}_1 \rightarrow \mathbf{q}_2 = \frac{\mathbf{A}_2}{\|\mathbf{A}_2\|} \rightarrow \mathbf{q}_1^{\,T}\mathbf{A}_2 = 0\,? \\
&\mathbf{A}_3 = \mathbf{a}_3 - \left(\mathbf{a}_3^{\,T}\mathbf{q}_1\right)\mathbf{q}_1 - \left(\mathbf{a}_3^{\,T}\mathbf{q}_2\right)\mathbf{q}_2 \rightarrow \mathbf{q}_3 = \frac{\mathbf{A}_3}{\|\mathbf{A}_3\|}
\end{aligned}\right\} \rightarrow \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ & r_{22} & r_{23} \\ & & r_{33} \end{bmatrix}$$

$$\mathbf{A} = \mathbf{QR} \rightarrow \mathbf{R} = \mathbf{Q}^T\mathbf{A} = \text{inner product of } \mathbf{q}\text{'s with } \mathbf{a}\text{'s!} \rightarrow r_{ij} = \mathbf{q}_i^{\,T}\mathbf{a}_j$$

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b} \rightarrow \mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\hat{\mathbf{x}} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b} \rightarrow \mathbf{R}\hat{\mathbf{x}} = \mathbf{Q}^T\mathbf{b}, \text{ safe and fast}$$

Pseudoinverse of A: A^+ (when A is not invertible)

$\mathbf{A}(m\times n)\rightarrow$ pseudoinverse $\mathbf{A}^+(n\times m)$: $\mathbf{A}\mathbf{A}^+\approx\mathbf{I}$

if $\mathbf{A}^{-1}$ exists, $\mathbf{A}\mathbf{A}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{I}$, then $\mathbf{A}^+=\mathbf{A}^{-1}$: square, ful rank matrix

$$\text{in case of }\begin{cases}\text{rectangular}\\ \text{zero eigenvalues}\\ \text{square, but has null space other than 0 vecter = columns are dependent}\end{cases}$$

if $\mathbf{A}$ has independent columns, then $\mathbf{A}^+=\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$ and so $\mathbf{A}^+\mathbf{A}=\mathbf{I}$

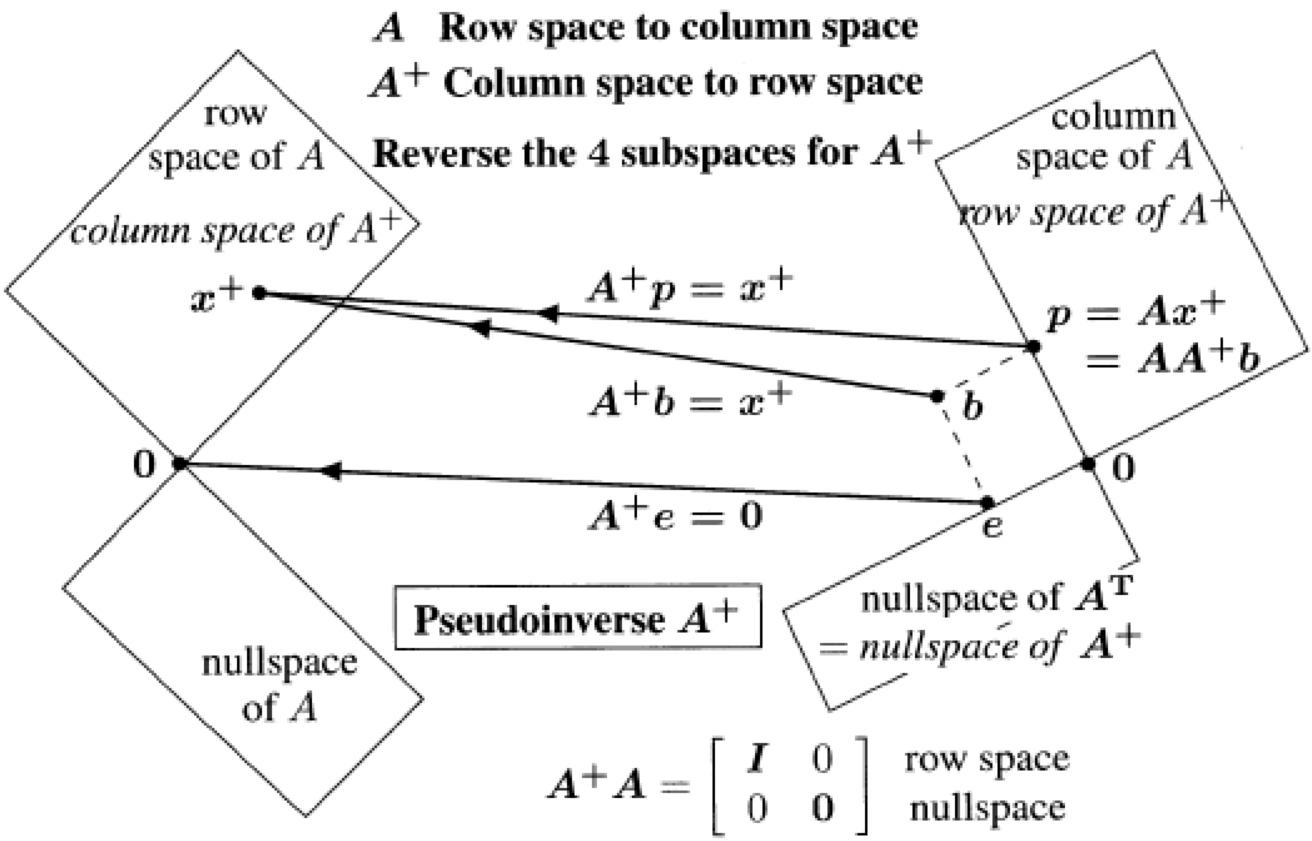
if $\mathbf{A}$ has independent rows, then $\mathbf{A}^+=\mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}$ and so $\mathbf{A}\mathbf{A}^+=\mathbf{I}$

A diagonal matrix $\mathbf{\Sigma}$ is inverted where possible, otherwise $\mathbf{\Sigma}^+$ has zeros

$$\mathbf{\Sigma}=\begin{bmatrix}\sigma_1 & 0 & 0 & 0\\ 0 & \sigma_2 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\rightarrow\mathbf{\Sigma}^+=\begin{bmatrix}1/\sigma_1 & 0 & 0\\ 0 & 1/\sigma_2 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

The pseudoinverse of $\mathbf{A}=\mathbf{U\Sigma V}^T$ is $\mathbf{A}^+=\mathbf{V\Sigma}^+\mathbf{U}^T$

Pseudoinverse of A: A^+



Pseudoinverse of A: A$^+$

$$\begin{cases} \mathbf{x} = \mathbf{x}^+ = \mathbf{A}^+\mathbf{b} \text{ makes } \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 \text{ as small as possible} \rightarrow \text{least squares solution} \\ \text{if another } \hat{\mathbf{x}} \text{ achieves that minimum then } \|\mathbf{x}^+\| < \|\hat{\mathbf{x}}\| \rightarrow \text{minimum norm solution} \end{cases}$$

The shortest least squares solution to $\begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}$ is $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b} = \begin{bmatrix} 1/3 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 6 \\ 8 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

all vectors $\begin{bmatrix} 0 \\ x_2 \end{bmatrix}$ are in the nullspace of $\mathbf{A}$

all vectors $\hat{\mathbf{x}} = \begin{bmatrix} 2 \\ x_2 \end{bmatrix}$ minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 = 64$, but $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ is the shortest

$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \rightarrow \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 \rightarrow$ the best $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$

SDV solved the least square problems in one step $\mathbf{A}^+\mathbf{b} \rightarrow$ computational cost?

Least Squares with a Penalty Term (Ridge Regression)

if $\mathbf{A}$ has dependent columns and $\mathbf{Ax} = \mathbf{0}$ has nonzero solutions,

then $\mathbf{A}^T\mathbf{A}$ cannot be invertible $\rightarrow$ we need $\mathbf{A}^+$

regularize least squares: minimize $\|\mathbf{Ax} - \mathbf{b}\|^2 + \delta^2 \|\mathbf{x}\|^2$

$\rightarrow$ solve $\left(\mathbf{A}^T\mathbf{A} + \delta^2\mathbf{I}\right)\mathbf{x}_\delta = \mathbf{A}^T\mathbf{b},\ \mathbf{x}_\delta \rightarrow \hat{\mathbf{x}}$ as $\delta \rightarrow 0$

$$\mathbf{A}(1 \text{ by } 1): \left(\mathbf{A}^T\mathbf{A} + \delta^2\mathbf{I}\right)^{-1}\mathbf{A}^T = \left(\frac{\sigma}{\sigma^2 + \delta^2}\right) \text{ if } \delta \rightarrow 0, \text{ then the limit is } \begin{cases} 0 \text{ if } \sigma = 0 \\ \dfrac{1}{\sigma} \text{ if } \sigma \neq 0 \end{cases} \leftrightarrow \mathbf{A}^+ = 0 \text{ or } \frac{1}{\sigma}$$

$$\mathbf{A} = \mathbf{U\Sigma V}^T \rightarrow \mathbf{A}^T\mathbf{A} + \delta^2\mathbf{I} = \mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U\Sigma V}^T + \delta^2\mathbf{I} = \mathbf{V}\left(\mathbf{\Sigma}^T\mathbf{\Sigma} + \delta^2\mathbf{I}\right)\mathbf{V}^T$$

$$\left(\mathbf{A}^T\mathbf{A} + \delta^2\mathbf{I}\right)^{-1}\mathbf{A}^T = \mathbf{V}\left(\mathbf{\Sigma}^T\mathbf{\Sigma} + \delta^2\mathbf{I}\right)^{-1}\mathbf{V}^T\left(\mathbf{V\Sigma}^T\mathbf{U}^T\right) = \mathbf{V}\left(\mathbf{\Sigma}^T\mathbf{\Sigma} + \delta^2\mathbf{I}\right)^{-1}\mathbf{\Sigma}^T\mathbf{U}^T$$

$$\lim_{\delta \rightarrow 0} \mathbf{V}\left[\left(\mathbf{\Sigma}^T\mathbf{\Sigma} + \delta^2\mathbf{I}\right)^{-1}\mathbf{\Sigma}^T\right]\mathbf{U}^T = \mathbf{V\Sigma}^+\mathbf{U}^T = \mathbf{A}^+$$