Numerical Problems in Solving the Normal Equation

In general, it is not a good idea to solve the normal eqn:

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

by explicitly forming $A^T A$, and then compute $(A^T A)^{-1}$.

Why?

1) Forming $A^T A \Rightarrow$ loss of info.
2) $\kappa(A^T A) = \kappa(A)^2$, i.e.,

the cond. number of $A^T A$ is much worse than that of $A$ in general.

This example is a bit extreme ... Show previous

Ex. Forming $A^T A$ is bad. MATLAB example

$$A = \begin{bmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}, \text{ say } \varepsilon = 10^{-8}$$

in double precision floating point sys.

Then $A^T A = \begin{bmatrix} 1 + \varepsilon^2 & 1 \\ 1 & 1 + \varepsilon^2 \end{bmatrix}$

$$\approx \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \text{ because } \varepsilon^2 = 10^{-16}$$

How about the condition numbers?

$\kappa(A) \approx 1.4142 \times 10^8$ already bad.

$\kappa(A^T A) \approx +\infty$ in double precision.
If we set $\varepsilon = 10^{-7}$ instead of $10^{-8}$, then $\kappa(A) \approx 1.4142 \times 10^7$ and $\kappa(ATA) \approx 1.9903 \times 10^{14}$.
This is still too bad to get any reliable LS solution for such $A$.

Often such situations occur when some of the column vectors of $A$ are “close to parallel”, i.e., they become almost linearly dependent.

**Def.** Let $A \in \mathbb{R}^{m \times n}$. Then $A$ is called rank deficient if $\text{rank}(A) < \min(m,n)$, i.e., if $A$ is not of full rank.

In general, we should avoid computing a solution for a given LS problem by forming $ATA$ explicitly and computing $(ATA)^{-1}A^Tb$.

$\Rightarrow$ Better to use the methods based on QR decomposition or SVD (we’ll discuss these later in this course.)
Orthogonality

The above discussion should convince you that $A$ is quite "good" if its column vectors are mutually orthogonal.

Suppose $A = [e_1, e_2]$, $\tilde{A} = [e_1, \tilde{e}_2]$ in $\mathbb{R}^2$. You can see that $A$ is much more "well-balanced" and convenient than $\tilde{A}$. For example, suppose we want to represent $x = [1, 1]^T$ in the basis of $\{e_1, e_2\}$ and that of $\{e_1, \tilde{e}_2\}$. Then the coefficient of $x$ w.r.t. $\{e_1, e_2\}$ is the same as $x$ itself since $A^{-1}x = Ax = x$.

But $\tilde{A}^{-1}x$ behaves badly.

Why? Say $c = \tilde{A}^{-1}x$, $c = [c_1, c_2]^T$.

Then $x = \tilde{A}c = [e_1, \tilde{e}_2][c_1^T]$

$= c_1e_1 + c_2\tilde{e}_2$

But $x = e_1 + e_2$, i.e.,

$e_1 + e_2 = c_1e_1 + c_2\tilde{e}_2$
Taking an inner product with $\tilde{e}_2$ on both sides yields

\[
\tilde{e}_2^T(e_1 + \tilde{e}_2) = \tilde{e}_2^T(c_1 e_1 + c_2 \tilde{e}_2) = 0
\]

\[
\implies 1 = c_2 \tilde{e}_2^T \tilde{e}_2
\]

\[
\implies c_2 = \frac{1}{\tilde{e}_2^T \tilde{e}_2}
\]

Could be huge if $\tilde{e}_2$ is close to perpendicular to $e_2$, i.e., close to parallel to $e_1$.

**Orthogonal Vectors**

**Def.** Two vectors $x, y \in \mathbb{R}^m$ are said to be **orthogonal** if

\[
x^T y = 0.
\]

So, the zero vector $0$ is orthogonal to any vector.

- Two **sets** of vectors $X, Y$ are said to be **orthogonal** if

\[
\forall x \in X, \forall y \in Y, x^T y = 0.
\]

- A set of vectors $S$ is said to be **orthogonal** if

\[
\forall x \in S, \forall y \in S, x \neq y \implies x^T y = 0.
\]
A set of vectors $S$ is said to be orthonormal if $S$ is orthogonal and $\forall x \in S$, $\|x\|_2 = 1$.

Even more balanced!

**Theorem.** The vectors in an orthogonal set $S$ are linearly independent.

**Proof.** Let $S = \{v_1, \ldots, v_n\}$.

Suppose they are not lin. indep.

Then $\exists v_k \in S$ s.t. $v_k \neq 0$ and

$$v_k = \sum_{i=1}^{n} c_i v_i \text{ with } c \neq 0$$

$$c = [c_1, \ldots, c_k, \ldots, c_n]^T$$

Since $S$ is an orthogonal set,

$$v_j^T v_i = 0 \text{ for } \forall j \neq i.$$  

But

$$v_k^T \left( \sum_{i=1}^{n} c_i v_i \right) = \sum_{i=1}^{n} c_i v_k^T v_i = 0$$

$$\iff v_k^T v_k = 0$$

$$\iff \|v_k\|^2 = 0$$

$$\iff v_k = 0$$

# contradiction!

**Components of a vector**

"Inner products can be used to decompose arbitrary vectors into orthogonal components!"
Suppose \( \{ \varphi_1, \ldots, \varphi_n \} \subset \mathbb{R}^m \) is an orthonormal set. \( \varphi_j \in \mathbb{R}^m, 1 \leq j \leq n \).

Let \( \psi \) be an arbitrary vector in \( \mathbb{R}^m \).

\[
\psi = \psi - (\varphi_1^T \psi) \varphi_1 - (\varphi_2^T \psi) \varphi_2 - \cdots - (\varphi_n^T \psi) \varphi_n
\]

This residual vector is \( \perp \) to \( \{ \varphi_1, \ldots, \varphi_n \} \).

Why?

\[
\varphi_j^T \psi = \varphi_j^T \psi - (\varphi_1^T \psi) \varphi_1 - (\varphi_2^T \psi) \varphi_2 - \cdots - (\varphi_j^T \psi) \varphi_j - (\varphi_{j+1}^T \psi) \varphi_{j+1} - \cdots - (\varphi_n^T \psi) \varphi_n
\]

\[
= \varphi_j^T \psi - \varphi_j^T \psi = 0
\]

This is true for any \( j = 1, \ldots, n \).

\[
\Rightarrow \psi = \psi + \sum_{i=1}^n (\varphi_i^T \psi) \varphi_i
\]

Any vector in \( \mathbb{R}^m \) is

\[
= \psi + \sum_{i=1}^n (\varphi_i^T \psi) \varphi_i
\]

where \( Q := [\varphi_1 \ldots \varphi_n] \in \mathbb{R}^{m \times n} \).

If \( \{ \varphi_1, \ldots, \varphi_n \} \) is a basis of \( \mathbb{R}^m \),
then \( n = m \) and \( \psi = 0 \).

i.e.,

\[
\mathbf{v} = \sum_{i=1}^m (\varphi_i^T \psi) \varphi_i = \sum_{i=1}^n (\varphi_i^T \psi) \varphi_i
\]
In fact, $v = QQ^Tv$, i.e.,

$$QQ^T = I$$

**Def.** A square matrix $Q \in \mathbb{R}^{m \times m}$ is said to be **orthogonal** if

$$Q^T = Q^{-1}$$

should be called orthonormal

i.e., $Q^TQ = QQ^T = I$

**Note:** If $Q = [q_1 \ldots q_n] \in \mathbb{R}^{m \times n}$ with $m > n$ and these vectors are orthonormal, then it is always true that $Q^TQ = In_{n \times n}$ but $QQ^T \neq I_{m \times m}$ unless $m = n$

e.g.,

$$Q = \begin{bmatrix}
\sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\
\sqrt{\frac{1}{3}} & 0 \\
\sqrt{\frac{1}{3}} & -\sqrt{\frac{2}{3}}
\end{bmatrix}$$

then $Q^TQ = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2 \times 2}$

But,

$$QQ^T = \begin{bmatrix}
\sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\
\sqrt{\frac{1}{3}} & 0 \\
\sqrt{\frac{1}{3}} & -\sqrt{\frac{2}{3}}
\end{bmatrix}
\begin{bmatrix}
\sqrt{\frac{1}{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{3}} & 0 & -\frac{1}{\sqrt{3}}
\end{bmatrix}
= \begin{bmatrix}
\frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
-\frac{1}{6} & \frac{1}{3} & \frac{5}{6}
\end{bmatrix} \neq I_{3 \times 3}$$

Why? \Rightarrow Next lecture on **Orthogonal Projector**.
Multiplication by an ortho. matrix

\[ y : \text{coef's of expansion in } \{e_1, \ldots, e_m\} \]
\[ Q^T y : \text{coef's of expansion of } y \text{ in } \{e_1, \ldots, e_m\} \]

\[ Q \cdot \]
\[ \cdot Q^T. \]

Note that \( \| y \| = \| Q^T y \| \)
i.e., isometry!

Why?
\[ \| Q^T y \|^2 = (Q^T y)^T (Q^T y) = y^T Q Q^T y = y^T y = \| y \|^2 \]

Compare this with the general situation we discussed before: \( A \in \mathbb{R}^{m \times m} \), nonsingular

\[ y : \text{coef's of expansion in } \{e_1, \ldots, e_m\} \]
\[ A^{-1} y : \text{coef's of expansion of } y \text{ in } \{e_1, \ldots, e_m\} \]
\[ A \cdot \]
\[ \cdot A^{-1}. \]