Review of linear equations: $Ax = b$, $A \in \mathbb{R}^{m \times n}$.

- $b \in \mathbb{R}(A)$?
  - yes
    - columns of $A$ linearly independent, i.e., $\text{rank}(A) = n$?
      - yes
        - there is a unique way of expressing $b$ as linear combination of columns of $A$, i.e., $Ax = b$ has a unique solution.
      - no
        - what can we do in this case? Use least squares.
  - no
    - there are no solutions

Then $N(A) \neq \{0\}$. So there are infinitely many solutions $x = x_p + w$.

$x_p$ solves $Ax = b$ and $w \in N(A)$.

Typical setup: we'd like to solve $Ax = b$, $A \in \mathbb{R}^{m \times n}$ where $m \gg n$ (m much larger than n). Typically, columns of $A$ will be full rank and $b \notin \mathbb{R}(A)$; so no solution.

Define the residual $r = Ax - b$. 
If we can't make \( r = 0 \) (if \( Ax = b \) has no solutions), instead try to make \( \| r \| \) as small as possible.

Often written as:

\[
\min_{x \in \mathbb{R}^n} \| Ax - b \|^2
\]

This is not necessary, but it's done by convention so that the square-root cancels. i.e. \( \min x^2 + y^2 \) instead of \( \min \sqrt{x^2 + y^2} \).

Geometric intuition. Take the case \( m = 3, \ n = 2 \).

\[
Ax = b \implies x_1 a_{11} + x_2 a_{12} = b.
\]

\[
\implies x_1 \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}
\]

The minimal \( \| r \| \) occurs when \( r \) is perpendicular to \( A\hat{x} \).

i.e. \( (A\hat{x}) \) perpendicular to \( (A\hat{x} - b) \).

\[
S = \text{span}(a_{11}, a_{12}) = \mathbb{R}(A).
\]
Optimal solution to least squares problem:

\[ \hat{x} = \text{arg min}_x \| Ax - b \|^2 \]

happens when \( \hat{r} = Ax - b \) is perpendicular to \( \mathbf{R}(A) \),
( perpendicular to every vector in \( \mathbf{R}(A) \) )

Aside: the set \( S^\perp = \{ w \in \mathbb{R}^n \mid w^T x = 0 \text{ for all } x \in S \} \)
\( "S \text{ perp}" \)

is a subspace! Why? Use the definition:

- \( 0^T x = 0 \) (zero is in \( S^\perp \)).
- if \( w_1^T x = 0 \) for all \( x \in S \) and \( w_2^T x = 0 \) for all \( x \in S \),
  then \( (w_1 + w_2)^T x = \frac{w_1^T x + w_2^T x}{0} = 0 \) whenever \( x \in S \),
- if \( w^T x = 0 \) for all \( x \in S \), then \( (aw)^T x = a(w^T x) = 0. \)

so \( \hat{x} \) is optimal if \( \hat{r} \in \mathbf{R}(A)^\perp \)

\[ w^T (Ax - b) = 0 \text{ for every } w \in \mathbf{R}(A), \text{ i.e. whenever } w = Ax \text{ for some } x. \]

\[ (Ax)^T (Ax - b) = 0 \text{ for all } x. \]

\[ x^T (A^T A x - A^T b) = 0 \text{ for all } x. \]

\[ A^T A \hat{x} = A^T b \]

These are called the normal equations
So if \( \hat{x} = \arg \min_x \|Ax - b\|^2 \)

then \( A^TA\hat{x} = A^Tb \) \( \tag{\text{this is an n x n system of linear equations}} \)

We proved that if \( \hat{x} = \arg \min_x \|Ax - b\|^2 \), then \( A^TA\hat{x} = A^Tb \).

what about the converse? what if we find some \( \hat{x} \) such that \( A^TA\hat{x} = A^Tb \) ... does it follow that \( \hat{x} \) minimizes \( \|Ax - b\|^2 \)?

[Yes, but we didn't prove it yet! To see why, just replace our normal equations by something that's always true, like \( \hat{x} = \hat{x} \).]

proof of the converse: Suppose \( A^TA\hat{x} = A^Tb \). Let's compute the residual for some other candidate point \( x \).

\[
r = Ax - b = A\hat{x} - b + A(x-\hat{x}) = \hat{r} + A(x-\hat{x})
\]

\[
\|r\|^2 = \|\hat{r} + A(x-\hat{x})\|^2
\]

\[
= (\hat{r} + A(x-\hat{x}))^T(\hat{r} + A(x-\hat{x}))
\]

\[
= \hat{r}^T\hat{r} + \hat{r}^T A(x-\hat{x}) + (x-\hat{x})^T A^T \hat{r} + (x-\hat{x})^T A^T A (x-\hat{x})
\]

\[
\|r\|^2 = \|\hat{r}\|^2 + \|A(x-\hat{x})\|^2 \geq 0.
\]

so \( \|r\|^2 \) can't be any smaller than \( \|\hat{r}\|^2 \), i.e. \( \hat{x} \) is optimal.
Example

find \( x \) that minimize \( \| [1\ 0\ 0] [y] - [1\ 1] \| ^2 \)

1) draw a picture:

so closest point is \([0\ 1]^

which is \([1\ 0\ 0] [1] \)

minimum residual has length 1

2) solve normal equations:

\( A^T A \hat{x} = A^T b \)

\[
\Rightarrow (\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}) (\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}) (\begin{bmatrix}
x \\
y \\
z
\end{bmatrix}) = (\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}) (\begin{bmatrix}
1 \\
t \\
z
\end{bmatrix})
\]

\[
\Rightarrow (\begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0
\end{bmatrix}) (\begin{bmatrix}
x \\
y \\
z
\end{bmatrix}) = (\begin{bmatrix}
1 \\
t \\
z
\end{bmatrix})
\]

\[
\Rightarrow (x = 1) \\
\Rightarrow (y = 1)
\]
So far we have seen that:

\[
\begin{aligned}
\{ \text{x is a solution to} \quad & \min_x \|Ax-b\|^2 \} \\
\leftrightarrow \quad & \{ \text{x is a solution to} \quad A^TAx = Ab \}
\end{aligned}
\]

\textbf{Fact: } \quad A^TAx = A^Tb \text{ always has a solution.}

So \( A^Tb \in \text{R}(A^TA) \) no matter what \( A \) looks like.

We will prove this later on.

\textbf{Fact: } \quad A^TA \text{ is invertible (i.e. } A^TAx = A^Tb \text{ has a unique solution) if and only if } A \text{ has linearly independent columns.} \quad \text{(i.e. } \text{rank}(A) = n, \text{ or } N(A) = \{0\} \).

\textbf{Proof: } \quad \text{observe that if } Ax = 0 \text{ for some } x, \text{ then } A^TAx = 0 \text{ also.}

Similarly, if \( A^TAx = 0 \), then \( x^TA^TAx = 0 \), which is the same as \( \|Ax\|^2 = 0 \), which implies \( Ax = 0 \) (property of norm!).

So \( Ax = 0 \iff A^TAx = 0 \). This implies in particular that \( N(A) = N(A^TA) \), so if \( N(A) = \{0\} \), then \( N(A^TA) = \{0\} \) and vice versa.

\textbf{Conclusion: } \quad \text{if } A \text{ has linearly independent columns, solution to LS problem } \min \|Ax-b\|^2 \text{ is } \hat{x} = (A^TA)^{-1}A^Tb.

Otherwise, \( \hat{x} = \tilde{x} + W \), any element \( \tilde{x} \in N(A) \), any solution to \( A^TAx = Ab \).
Makes sense that a LS problem could have multiple solutions. Even if $Ax = b$ has no solutions, if $w \in \text{N}(A)$ then $\|A(\hat{x} + w) - b\|^2 = \|A\hat{x} + Aw - b\|^2 = \|A\hat{x} - b\|^2$, so there may be multiple $\hat{x}$ that solve $\min \|Ax - b\|^2$, and all achieve the same residual.

Our previous example:

\[
\begin{pmatrix}
1 & 2 \\
2 & 3
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
= 
\begin{pmatrix}
3 \\
5
\end{pmatrix}
\]

add a new column: \[
\begin{pmatrix}
2 \\
1
\end{pmatrix}
\]

Now there are multiple ways of producing \[
\begin{pmatrix}
1 \\
1
\end{pmatrix}
\]
which is still the closest point to \[
\begin{pmatrix}
0 \\
0
\end{pmatrix}
\]
because we haven't changed $R(A)$. So this LS problem is degenerate, (has many solutions).
Example: Linear regression.

Suppose we have data from an experiment to characterize the age of a tree as a function of its height (for a particular species). Collect data: \((X_i, Y_i)\)

\[
\begin{align*}
X_i &= \text{age} & i &= 1, 2, \ldots, N \\
Y_i &= \text{height}
\end{align*}
\]

1) Write equations:

\[
\begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_N
\end{bmatrix}
\begin{bmatrix}
p \\
q
\end{bmatrix}
= 
\begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_N
\end{bmatrix}
\]

Note: \( A \) will always have full rank unless all the \( X_i \) are identical.

2) Solve least squares problem, \( \min_x \| A x - b \|^2 \),

which gives us \( \hat{x} = \begin{bmatrix} \hat{p} \\ \hat{q} \end{bmatrix} \).

3) Predict future heights based on age:

\[
\hat{Y}_{\text{est}} = \hat{p} \times \text{age of a new tree} + \hat{q}
\]