The Method of Least Squares

Setup: you have a matrix equation $Ax=b$ which is (generally) inconsistent. What is the best approximate solution?

What do we mean by "best approximate solution"?

Def: $\hat{x}$ is a least squares solution of $Ax=b$ if $\|b-A\hat{x}\|$ is minimized over all vectors $\hat{x}$.

This means $A\hat{x}$ is as close as possible to b.

NB: $\mathrm{Col}(A) = \{A\hat{x} : \hat{x} \in \mathbb{R}^n\}$, so $A\hat{x}$ is just the closest vector to b in $V = \mathrm{Col}(A)$.

This is the orthogonal projection!

$$A\hat{x} = b_V \quad \text{for} \quad V = \mathrm{Col}(A)$$

How do we compute b_V and $\hat{x}$?

→ If $\hat{x}$ is any soln of $A^TA\hat{x} = A^Tb$, then $A\hat{x} = b_V$.

So $\hat{x}$ is a least-squares solution!

Procedure (Least Squares):

To find the least squares solution(s) of $Ax=b$:

(1) Solve the normal equation $A^TA\hat{x}=A^Tb$

(2) Any solution $\hat{x}$ is a least-squares soln, and $b_V = A\hat{x}$ for $V=\mathrm{Col}(A)$.

NB: The error is the distance from $A\hat{x}$ to b:

$$\text{error} = \|b-A\hat{x}\| = \|b-b_V\| = \|b_{V^\perp}\|$$

Recall that $\|b-A\hat{x}\| = \|b_{V^\perp}\|$ is minimized: if $b_{V^\perp} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$ then $\sqrt{a^2+b^2+c^2}$, or equivalently $b_{V^\perp}\cdot b_{V^\perp} = a^2+b^2+c^2$ is the minimized quantity.

$$\boxed{\text{The least-squares solution(s) minimize } b_{V^\perp}\cdot b_{V^\perp}}$$

This is why it's called a least squares solution: we're minimizing the sum of the squares of the entries of $b-A\hat{x}$.

Eg: Find the least-squares solution of $Ax=b$ for $A=\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}$ $b=\begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$.

(1) $A^TA = \begin{pmatrix} 5 & 3 \\ 3 & 3 \end{pmatrix}$ $A^Tb = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$

$$\left(\begin{array}{cc|c} 5 & 3 & 0 \\ 3 & 3 & 6 \end{array}\right) \xrightarrow{\text{rref}} \left(\begin{array}{cc|c} 1 & 0 & -3 \\ 0 & 1 & 5 \end{array}\right)$$

(2) $\hat{x} = \begin{pmatrix} -3 \\ 5 \end{pmatrix}$ ← the least-squares soln $\quad b_V = A\hat{x} = \begin{pmatrix} 5 \\ 2 \\ -1 \end{pmatrix}$

↳ for $V = \mathrm{Col}(A)$

The error is

$$\|b_{V^\perp}\| = \|b - b_V\| = \left\| \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 5 \\ 2 \\ -1 \end{pmatrix} \right\| = \left\| \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \right\|$$

$$= \sqrt{1^2 + (-2)^2 + 1^2} = \sqrt{6}.$$

[demo: where is $\hat{x}$?]

Eg: Find the least-squares solutions of $Ax = b$ for $A = \begin{pmatrix} 1 & -1 & -1 \\ 2 & 1 & 4 \\ 1 & -1 & -1 \end{pmatrix}$ and $b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

(1) $A^TA = \begin{pmatrix} 6 & 0 & 6 \\ 0 & 3 & 6 \\ 6 & 6 & 18 \end{pmatrix}$ $A^Tb = \begin{pmatrix} 4 \\ -1 \\ 2 \end{pmatrix}$

$$\left(\begin{array}{ccc|c} 6 & 0 & 6 & 4 \\ 0 & 3 & 6 & -1 \\ 6 & 6 & 18 & 2 \end{array}\right) \xrightarrow{\text{rref}} \left(\begin{array}{ccc|c} 1 & 0 & 1 & 2/3 \\ 0 & 1 & 2 & -1/3 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

(2) $\xrightarrow{\text{PVF}} \hat{x} = \begin{pmatrix} 2/3 \\ -1/3 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}$

In this case there are infinitely many least-squares solutions!

$b_V = A\hat{x}$ for any $\hat{x}$. Take $\hat{x} = \begin{pmatrix} 2/3 \\ -1/3 \\ 0 \end{pmatrix}$

$$\rightsquigarrow b_V = \begin{pmatrix} 1 & -1 & -1 \\ 2 & 1 & 4 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 2/3 \\ -1/3 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$b_{V^\perp} = b - b_V = 0$$

So the error is zero — the equation $Ax=b$ was consistent after all!

(Compare W7L2 pp. 4-5)

Observation 1:

$Ax=b$ has a unique least-squares soln

$\Longleftrightarrow$ A has full column rank!

$\rightarrow$ This is exactly when $A^TA\hat{x} = A^Tb$ has a unique solution (A^TA is invertible).

Otherwise, there are infinitely many least-squares solns. This means $\|b - A\hat{x}\|$ is minimized for any such $\hat{x}$:

$$b_V = A\hat{x} \text{ for any solution } \hat{x}.$$

(There can't be zero least-squares solutions! $A^TA\hat{x} = A^Tb$ is always consistent.)

Observation 2: If $Ax=b$ is consistent, then

$$\begin{pmatrix}\text{least squares}\\ \text{solutions}\end{pmatrix} = \begin{pmatrix}\text{ordinary}\\ \text{solutions}\end{pmatrix}.$$

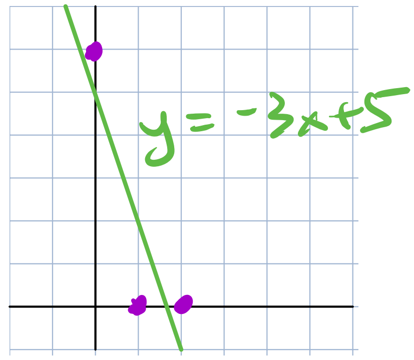
Indeed, a least-squares soln is just a soln of $A\hat{x} = b_V$ $(V = \mathrm{Col}(A))$, and $b = b_V \iff b \in \mathrm{Col}(A) \iff Ax=b$ is consistent.

Least-squares is often useful for fitting data to a model.

Eg (linear regression):

Find the best-fit line $y = Cx + D$ thru the data points $(0,6), (1,0), (2,0)$.

If $(0,6)$ lies on $y=Cx+D$ then substituting $x=0, y=6$ would give $6=C\cdot 0+D$. So we want to solve:



$$\begin{aligned} (0,6)&: & 6 &= C\cdot 0+D \\ (1,0)&: & 0 &= C\cdot 1+D \\ (2,0)&: & 0 &= C\cdot 2+D \end{aligned}$$

in the unknowns C & D

ie $Ax=b$ for $A=\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}$ $x=\begin{pmatrix} C \\ D \end{pmatrix}$ $b=\begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$

NB: the data points are not collinear ⇝ no exact solution! (maybe measurement error).

We found a least-squares solution before:

$\hat{x}=\begin{pmatrix} -3 \\ 5 \end{pmatrix}$ ⇝ best-fit line $y=-3x+5$

Important Question: [demo]

What quantity did we minimize?

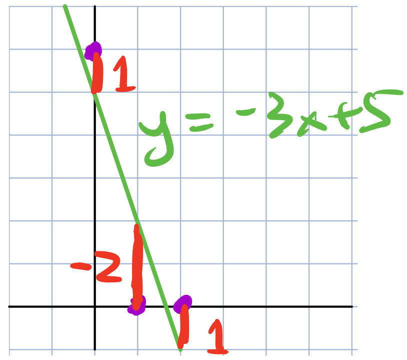
We minimized $\|b-A\hat{x}\|^2=\|b_{V^\perp}\|^2$.

$b=\begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$ = y-values of data pts.

$A\hat{x}=\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} -3 \\ 5 \end{pmatrix}=\begin{pmatrix} -3\cdot 0+5 \\ -3\cdot 1+5 \\ -3\cdot 2+5 \end{pmatrix}=$ y-values of $y=-3x+5$ at x-values $0,1,2$.

So $b_{V^\perp} = b - A\hat{x} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$

$= \Big($ vertical distances from $y=-3x+5$ to the data points $\Big)$

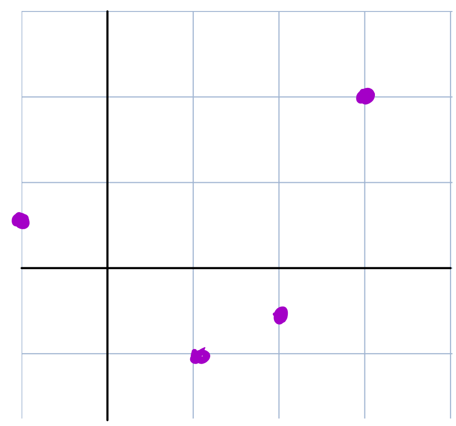


We minimized the sum of the squares of the vertical distances (the error).

Eg (best-fit parabola):

Find the best-fit parabola $y = Bx^2+Cx+D$ thru the data points $(-1,\tfrac{1}{2}), (1,-1), (2,-\tfrac{1}{2}), (3,2)$

Substitute the data points for x & y we want to solve



$(-1,\tfrac{1}{2})$: $\quad \tfrac{1}{2} = B(-1)^2 + C(-1) + D$

$(1,-1)$: $\quad -1 = B(1)^2 + C(1) + D$

$(2,-\tfrac{1}{2})$: $\quad -\tfrac{1}{2} = B(2)^2 + C(2) + D$

$(3,2)$: $\quad 2 = B(3)^2 + C(3) + D$

$\rightsquigarrow Ax=b$ for $A = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}$ $\quad x = \begin{pmatrix} B \\ C \\ D \end{pmatrix}$ $\quad b = \begin{pmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{pmatrix}$

Let's find the least-squares solution.

$$A^TA = \begin{pmatrix} 99 & 35 & 15 \\ 35 & 15 & 5 \\ 15 & 5 & 4 \end{pmatrix} \qquad A^Tb = \begin{pmatrix} 31/2 \\ 7/2 \\ 1 \end{pmatrix}$$

$$\left(\begin{array}{ccc|c} 99 & 35 & 15 & 31/2 \\ 35 & 15 & 5 & 7/2 \\ 15 & 5 & 4 & 1 \end{array}\right) \xrightarrow{\text{RREF}} \left(\begin{array}{ccc|c} 1 & 0 & 0 & 53/88 \\ 0 & 1 & 0 & -379/440 \\ 0 & 0 & 1 & -41/44 \end{array}\right)$$

$$\hat{x} = \begin{pmatrix} 53/88 \\ -379/440 \\ -41/44 \end{pmatrix} \rightsquigarrow y = \frac{53}{88}x^2 - \frac{379}{440}x - \frac{41}{44}$$

[demo]

Question: What did we minimize? always $\|b - A\hat{x}\|$

$$b = \begin{pmatrix} 1/2 \\ -1 \\ -1/2 \\ 2 \end{pmatrix} = \text{y-values of data pts.}$$

$$A\hat{x} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \begin{pmatrix} 53/88 \\ -379/440 \\ -41/44 \end{pmatrix} = \begin{pmatrix} \frac{53}{88}(-1)^2 + \frac{379}{440}(-1) - \frac{41}{44} \\ \frac{53}{88}(1)^2 + \frac{379}{440}(1) - \frac{41}{44} \\ \frac{53}{88}(2)^2 + \frac{379}{440}(2) - \frac{41}{44} \\ \frac{53}{88}(3)^2 + \frac{379}{440}(3) - \frac{41}{44} \end{pmatrix}$$

= y-values of $y = \frac{53}{88}x^2 - \frac{379}{440}x - \frac{41}{44}$ at x-values $-1, 1, 2, 3$

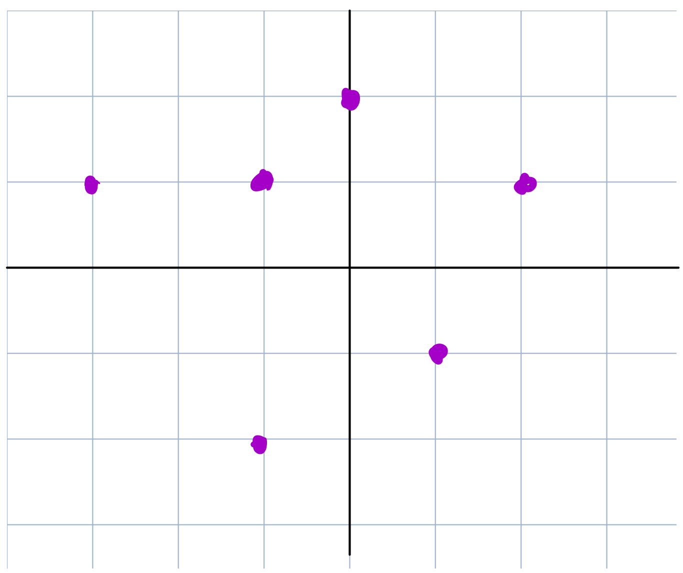
So $b_{V^\perp} = b - A\hat{x}$ = vertical distances from the graph to the data points, like before.

This same method works to find a best-fit function of the form $y = Af + Bg + Ch + \cdots$ where $f, g, h, \ldots$ are really any functions! Just plug the x-values of your data points into $f, g, h \rightsquigarrow$ linear equations for $A, B, C, \ldots$

Eg (best-fit trigonometric function):
see §6.5 in ILA for an example.

This "real-life" example of Gauss was in the first lecture:

Eg: An asteroid has been observed at coordinates:
$(0,2)$, $(2,1)$, $(1,-1)$, $(-1,-2)$, $(-3,1)$, $(-1,1)$



Question: What is the most likely orbit? Will the asteroid crash into the Earth?

Fact: The orbit is an ellipse.

Equation for an ellipse:

$$x^2 + By^2 + Cxy + Dx + Ey + F = 0$$

For our points to lie on the ellipse, substitute the coordinates into (x,y) $\leadsto$ these should hold:

$$\begin{array}{ll}
 & x \;\; y \\
(0,2): & 0 + 4B + 0 + 0 + 2E + F = 0 \\
(2,1): & 4 + B + 2C + 2D + E + F = 0 \\
(1,-1): & 1 + B - C + D - E + F = 0 \\
(-1,-2): & 1 + 4B + 2C - D - 2E + F = 0 \\
(-3,1): & 9 + B - 3C + D - 3E + F = 0 \\
(-1,1): & 1 + B - C - D + E + F = 0
\end{array}$$

constants

Move the constants to the RHS $\leadsto$ this is $Ax=b$

for $A = \begin{pmatrix} 4 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 4 & 2 & -1 & -2 & 1 \\ 1 & -3 & 1 & -3 & 1 \\ 1 & -1 & -1 & 1 & 1 \end{pmatrix}$ $\quad x = \begin{pmatrix} B \\ C \\ D \\ E \\ F \end{pmatrix}$ $\quad b = \begin{pmatrix} 0 \\ -4 \\ -1 \\ -1 \\ -9 \\ -1 \end{pmatrix}$

Least-squares solution: [demo]

$$\hat{x} = \left(\frac{405}{266}, -\frac{89}{133}, \frac{201}{133}, -\frac{123}{266}, -\frac{687}{133}\right)$$

$$\leadsto x^2 + \frac{405}{266}y^2 - \frac{89}{133}xy + \frac{201}{133}x - \frac{123}{266}y - \frac{687}{133} = 0$$

What quantity did we minimize? $\|b - A\hat{x}\|$ or $\|A\hat{x} - b\|$

$$A\hat{x} - b = \begin{pmatrix} 4 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 4 & 2 & -1 & -2 & 1 \\ 1 & -3 & 1 & -3 & 1 \\ 1 & -1 & -1 & 1 & 1 \end{pmatrix} \hat{x} - \begin{pmatrix} 0 \\ -4 \\ -1 \\ -1 \\ -9 \\ -1 \end{pmatrix}$$

$$= \begin{pmatrix} 0^2 + \frac{405}{266}(2)^2 - \frac{89}{133}(0)(2) + \frac{201}{133}(0) - \frac{123}{266}(2) - \frac{687}{133} \\ 2^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(2)(1) + \frac{201}{133}(2) - \frac{123}{266}(1) - \frac{687}{133} \\ 1^2 + \frac{405}{266}(-1)^2 - \frac{89}{133}(1)(-1) + \frac{201}{133}(1) - \frac{123}{266}(-1) - \frac{687}{133} \\ (-1)^2 + \frac{405}{266}(-2)^2 - \frac{89}{133}(-1)(-2) + \frac{201}{133}(-1) - \frac{123}{266}(-2) - \frac{687}{133} \\ (-3)^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(-3)(1) + \frac{201}{133}(-3) - \frac{123}{266}(1) - \frac{687}{133} \\ (-1)^2 + \frac{405}{266}(1)^2 - \frac{89}{133}(-1)(1) + \frac{201}{133}(-1) - \frac{123}{266}(1) - \frac{687}{133} \end{pmatrix}$$

This was what you get by substituting the x- and y- values of the data points into the LHS of

$$x^2 + \frac{405}{266}y^2 - \frac{89}{133}xy + \frac{201}{133}x - \frac{123}{266}y - \frac{687}{133} = 0$$

It's the distance from zero. [demo]

Upshot: You're minimizing $\|b - A\hat{x}\|$; it's up to you to interpret that quantity in your original problem.