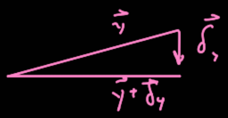
Today: Ridge Regression

- Noise
- Sensitivity / Perturbation Analysis / Condition Number
- Ridge Regression

Consider $\underbrace{A}_{\text{Data}} \underbrace{\vec{x}}_{\text{unknown}} = \underbrace{\vec{y}}_{\text{measurement}}$ $\quad A \in \mathbb{R}^{n \times n}$ invertible

- say we have a perturbed $\vec{y}$ s.t.



$\vec{y} \to \vec{y} + \vec{\delta}_y$ and this caused $\vec{x}$ to be perturbed ($\vec{x} \to \vec{x} + \vec{\delta}_x$)

Q/ How big is $\delta \vec{x}$?

↳ let's think about $\| \vec{\delta}_x \|_2$?

its relation to $\vec{x}$, so we look at this ratio:

$$\frac{\| \vec{\delta}_x \|}{\| \vec{x} \|_2}$$

↳ want to also connect this to $\vec{y}$

$$A(\vec{x} + \vec{\delta}_x) = \vec{y} + \vec{\delta}_y$$

↳ know $A\vec{x} = \vec{y}$ (by def.)

$$\Downarrow$$

$$A\vec{\delta}_x = \vec{\delta}_y$$

$$\vec{\delta}_x = A^{-1}\vec{\delta}_y \quad \text{(A invertible, square)}$$

$$\| \vec{\delta}_x \|_2 = \| A^{-1} \vec{\delta}_y \|_2$$

Recall:

$$\| A \|_2 = \max_{\|\vec{y}\|_2 = 1} \| A\vec{y} \|_2$$

$$= \max_{y} \frac{\| A\vec{y} \|_2}{\| \vec{y} \|_2}$$

$$\|\vec{\delta}_x\|_2 = \|A^{-1}\vec{\delta}_y\|_2 \underset{1}{\leq} \|A^{-1}\|_2 \|\vec{\delta}_y\|_2$$ (by def of spectral norm)

↳ want to now bound $\|\vec{x}\|_2$

$A\vec{x} = \vec{y}$

→ want a lower bound bc we have $\dfrac{\|\vec{\delta}_x\|}{\|\vec{x}\|}$ and we want this ratio as small as possible. (so want $\|\vec{x}\|$ as big as possible)

$$\|\vec{y}\|_2 = \|A\vec{x}\|_2$$
$$\leq \|A\|_2 \|\vec{x}\|_2$$
$$\|\vec{x}\|_2 \underset{2}{\geq} \frac{\|\vec{y}\|_2}{\|A\|_2}$$

Combining 1 & 2

$$\|\vec{\delta}_x\|_2 \leq \|A^{-1}\|_2 \|\vec{\delta}_y\|_2$$

$$\frac{\|\vec{\delta}_x\|_2}{\|\vec{x}\|_2} \leq \frac{\|A^{-1}\|_2 \|\vec{\delta}_y\|_2}{\|\vec{y}\| / \|A\|_2} = \|A\|_2 \|A^{-1}\| \left(\frac{\|\vec{\delta}_y\|_2}{\|\vec{y}\|_2} \right)$$

[Recall: $\|A\|_2 = \sigma_{max}$

$\|A^{-1}\|_2 = \dfrac{1}{\sigma_{min}}$

↳ A^{-1}'s max sigma is the reciprocal of A]

$$= \|A\|_2 \|A^{-1}\| \left(\frac{\|\vec{\delta}_y\|_2}{\|\vec{y}\|_2} \right)$$

$$= \boxed{\frac{\sigma_{max}}{\sigma_{min}}} \left(\frac{\|\vec{\delta}_y\|_2}{\|\vec{y}\|_2} \right)$$

★ Condition Number★ of a matrix A

⇒ fully defined by properties of A not by $\vec{y}$

Condition Number in the Context of Least Squares

$$\vec{x} = (A^TA)^{-1}A^T\vec{b}$$

Normal equations : $\underbrace{(A^TA)}_{\text{square!}}\vec{x} = A^T\vec{b}$

↳ instead of thinking abt the condition number of A (if it's non-square), we can think about the condition number of (A^TA)

↳ singular values of A & A^TA are related!

Shift property of e-vals

$$(A+\lambda I)\vec{v}_1 = A\vec{v}_1 + \lambda\vec{v}_1$$
$$= \lambda_1\vec{v}_1 + \lambda\vec{v}_1$$
$$= (\lambda_1+\lambda)\vec{v}_1$$

eigenvectors of A

Ridge Regression

Consider

$$\min_{\vec{x}} \underbrace{\|A\vec{x}-\vec{b}\|}_{\text{from least squares}} + \lambda^2\underbrace{\|\vec{x}\|_2^2}_{\text{says that large } \vec{x}\text{'s are bad}} \qquad \lambda > 0$$

↳ this is convex, so we can find the (unique) minimum by finding the gradient (wrt x) and setting it equal to 0.

$$\nabla_x f(\vec{x}) = \nabla_x \left(\|A\vec{x} - \vec{b}\| + \lambda^2 \|\vec{x}\|_2^2 \right)$$

$$= \nabla_x \left(\|A\vec{x} - \vec{b}\| \right) + \nabla\left(\lambda^2 \|\vec{x}\|_2^2 \right)$$

Ayah attempt:

$$= \nabla_x (A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b}) + 2\lambda^2 \vec{x}$$

$$= \nabla_x \left(\vec{x}^T A^T A \vec{x} - \vec{x}^T A^T \vec{b} - \vec{b}^T A \vec{x} + \cancel{\vec{b}^T \vec{b}}^{\,0} \right) + 2\lambda^2 \vec{x}$$

$$= 2A^T A \vec{x} - A^T \vec{b} - \vec{b}^T A + 2\lambda^2 \vec{x}$$

$$= 2A^T A \vec{x} - 2(\vec{b}^T A)^T + 2\lambda^2 \vec{x}$$

Ranade:

$$F(\vec{x}) = (A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b}) + \lambda^2 \vec{x}^T \vec{x}$$

$$= \vec{x}^T A^T A \vec{x} - \vec{x}^T A^T \vec{b} - \vec{b}^T A \vec{x} + \lambda^2 \vec{x}^T \vec{x} + \vec{b}^T \vec{b}$$

$$\nabla f(\vec{x}) = 2A^T A \vec{x} - 2\vec{b}^T A^T + 2\lambda^2 \vec{x}$$

→ set equal to 0

$$0 = 2A^T A \vec{x} - 2\vec{b}^T A^T + 2\lambda^2 \vec{x}$$

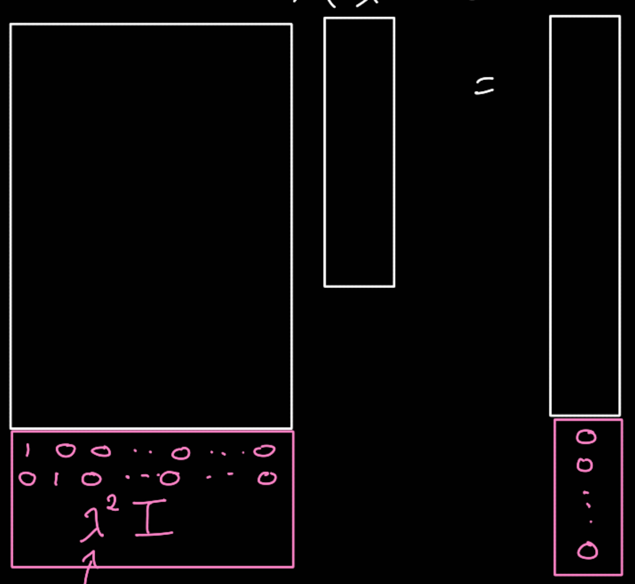
$$= A^T A \vec{x} - \vec{b}^T A^T + \lambda^2 \vec{x}$$

$$A^T \vec{b} = (A^T A + \lambda^2 I)\vec{x}$$

$$\vec{x}^* = \underbrace{(A^T A + \lambda^2 I)^{-1}}_{\text{guaranteed to be invertible because } \lambda^2 > 0} A^T \vec{b}$$

- say we know some side information about the vector $\vec{x}$, eg $\vec{x}$ is close to 0.

$$A\vec{x} = \vec{b}$$

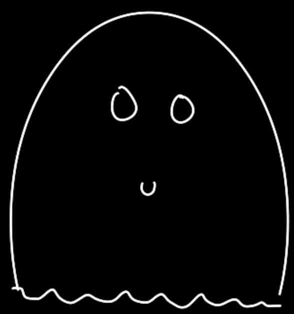


very large $\lambda \equiv$ "I'm VERY confident $\vec{x}$ is close to 0". λ small $\equiv$ "I'm not entirely sure

If $\vec{x}$ is close to $\vec{0}$ but I think it might be.

Reformulating:

$$\begin{bmatrix} A \\ \lambda I \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{b} \\ \vec{0} \end{bmatrix}$$



"ghost data"

$$\vec{x} = \left(\begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} A \\ \lambda I \end{bmatrix} \right)^{-1} \begin{bmatrix} A^T & \lambda I \end{bmatrix} \begin{bmatrix} b^T \\ \vec{0} \end{bmatrix}$$

$$= (A^T A + \lambda^2 I)^{-1} (A^T \vec{b} + \vec{0})$$

↳ this is again the ridge solution ♡