Lecture 5

Least-Square

⇒ Least-Square is about approximate solution of overdetermined eqandons.

★ **Overdetermined linear equations**

⇒ Consider $y = Ax$ where $A \in R^{m \times n}$ is (strictly) skinny i.e. $m > n$.

- ↳ Called overdetermined set of linear equations (more equations than unknowns)
- ↳ For most y, cannot solve for x

⇒ One approach to approximately solve $y = Ax$.

→ define residual or error

$$r = Ax - y$$

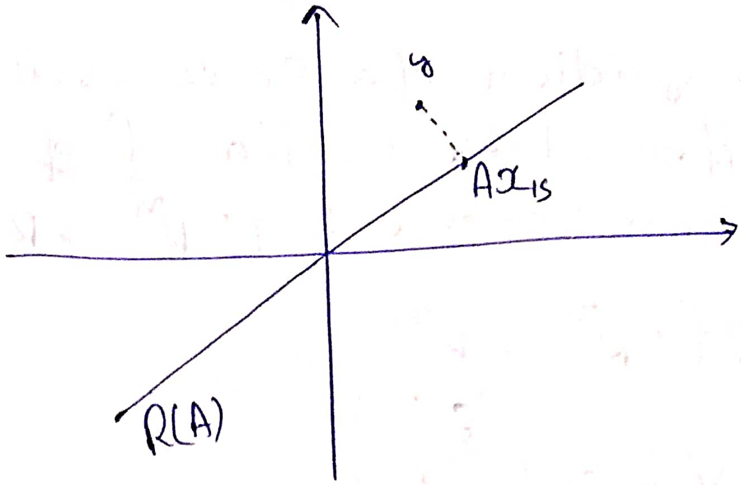
→ find $x = x_{ls}$ that minimizes $\|r\|$

⇒ x_{ls} called least-square (approximate) solution of $y = Ax$.

★ Geometric interpretation

⇒ Ax_{ls} is point in $\mathcal{R}(A)$ closest to y.

{ Ax_{ls} is projection of y onto $R(A)$ }



$$\boxed{Ax_{ls} = P_{R(A)}(y)}$$ { Notation for projection }

★ Least-square (approximate) solution

⇒ Assume A is full rank, skinny.

⇒ to find x_{ls} we'll minimize norm of residual squared.

$$\|r\|^2 = (Ax-y)^T(Ax-y)$$
$$= (x^TA^T - y^T)(Ax-y)$$
$$= x^TA^TAx - y^TAx - \underbrace{x^TA^Ty}_{y^TAx} + y^Ty$$

{ Since they are 1x1 transpose of each other. }

$$= \underbrace{x^TA^TAx}_{\text{Quadratic}} - \underbrace{2y^TAx}_{\text{linear}} + \underbrace{y^Ty}_{\text{Constant}}$$

⇒ Set gradient wrt x to zero:

$$\nabla_x \|r\|^2 = 2A^TAx - 2A^Ty = 0$$

gradient definition

The gradient of a scalar-valued differentiable function f of several variables: $f: R^n \rightarrow R$

$$\nabla f: R^n \rightarrow R^n$$

$$\nabla f = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix}$$

⇒ yields the normal equations:

$$A^TA x = A^Ty$$

{This is an invertible $n \times n$ matrix as A is full rank and skinny}

⇒ $\boxed{x_{ls} = (A^TA)^{-1}A^Ty}$ {Very famous formula}

$\Rightarrow x_{ls} = A^{-1}y$ if A is Square

$\Rightarrow x_{ls}$ Solves $y = Ax_{ls}$ if $y \in R(A)$

$\Rightarrow A^{+} = (A^TA)^{-1}A^T$ is called <u>pseudo-inverse</u> of A

or

<u>Moore-Penrose inverse</u>

$\Rightarrow A^{+}$ is a left inverse of (full rank, skinny) A.

$$A^{+}A = \left[(A^TA)^{-1}A^T\right]A$$
$$= (A^TA)^{-1}(A^TA)$$
$$= I_n$$

★ <u>Projection on $R(A)$</u>

$\Rightarrow Ax_{ls}$ is (by defination) the point in $R(A)$ that is closest to y, i.e. it is the projection of y onto $R(A)$.

$$Ax_{ls} = P_{R(A)}(y)$$

$\Rightarrow$ The projection function $P_{R(A)}$ is linear and given by

$$P_{R(A)}(y) = Ax_{ls} = \underbrace{\left(A(A^TA)^{-1}A^T\right)}_{\text{Called projection matrix associated with } R(A)}y$$

★ Orthogonality principle

⇒ Optimal residual

$$r = Ax_{ls} - y = (A(A^TA)^{-1}A^T - I)y$$

is orthogonal to $R(A)$.

★ Least-square via QR factorization

⇒ $A \in R^{m\times n}$ skinny, full rank.

⇒ factor $A = QR$ with $Q^TQ = I_n$, $R \in R^{n\times n}$ upper triangular invertible.

⇒ pseudo-inverse is

$$\begin{aligned}(A^TA)^{-1}A^T &= ((QR)^TQR)^{-1}(QR)^T \\ &= (R^T\underline{Q^TQ}R)^{-1}(R^TQ^T) \\ &= (R^TR)^{-1}R^TQ^T \\ &= R^{-1}\underline{R^{-T}R^T}Q^T \\ &= R^{-1}Q^T\end{aligned}$$

so $\boxed{x_{ls} = R^{-1}Q^Ty}$

⇒ Projection on $R(A)$ given by matrix

$$A(A^TA)^{-1}A^T = AR^{-1}Q^T = QRR^{-1}Q^T = QQ^T$$

★ Least-squares via full QR factorization

⇒ full QR factorization:

$$A = [Q_1\ Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

with $[Q_1\ Q_2] \in \mathbb{R}^{m \times m}$ orthogonal, $R_1 \in \mathbb{R}^{n \times n}$ upper triangular invertible.

⇒ multiplication by orthogonal matrix doesn't change norm, so

$$\|Ax - y\|^2 = \left\| [Q_1, Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - y \right\|^2$$

$$\Rightarrow \left\| [Q_1\ Q_2]^T [Q_1\ Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} - [Q_1\ Q_2]^T y \right\|^2$$

$$= \left\| \begin{bmatrix} R_1 x - Q_1^T y \\ -Q_2^T y \end{bmatrix} \right\|^2$$

$$= \|R_1 x - Q_1^T y\|^2 + \|Q_2^T y\|^2$$

⇒ to minimize this,

$$R_1 x - Q_1^T y = 0$$

$$\Rightarrow \boxed{x_{ls} = R_1^{-1} Q_1^T y}$$

⇒ residual with optimal x is

$$Ax_{ls} - y = Q_2 Q_2^T y$$

- $Q_1 Q_1^T$ gives projection onto $R(A)$
- $Q_2 Q_2^T$ gives projection onto $R(A)^\perp$

★ Least-squares estimation

$$y = Ax + V$$

→ x is what we want to estimate or reconstruct.

→ y is our sensor measurements

→ V is an unknown noise or measurement error (assumed small)

⇒ least-square estimation: Choose as estimate $\hat{x}$ that minimizes

$$\| A\hat{x} - y \|$$

least-square estimate is just $\hat{x} = (A^T A)^{-1} A^T y$

★ <u>BLUE Property</u> (Best Linear Unbiased Estimator)

⇒ Linear measurement with noise.

$$y = Ax + v \quad \{A \Rightarrow \text{Skinny \& full rank}\}$$

⇒ Consider a linear estimator of form

$$\hat{x} = By$$

{Called unbiased if $\hat{x} = x$ whenever $v = 0$}

B is left inverse of A

⇒ Estimation error of unbiased linear estimator is

$$\begin{aligned} x - \hat{x} &= x - B(Ax + v) \\ &= x - x - Bv \\ &= -Bv \end{aligned}$$

{Independent of x}

⇒ We want B "small".

⇒ <u>Fact:</u> $A^{+} = (A^TA)^{-1}A^T$ is the smallest left inverse of A.

Smallest left inverse means for any B with $BA = I$, we have

$$\sum_{ij} B_{ij}^2 \geq \sum_{i,j} A_{ij}^{+2}$$

{So least-square provides the best linear unbiased estimator (BLUE)}