Least-squares

- least-squares (approximate) solution of overdetermined equations
- projection and orthogonality principle
- least-squares estimation
- BLUE property

Overdetermined linear equations

consider $y = Ax$ where $A \in \mathbb{R}^{m \times n}$ 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 $y = Ax$:

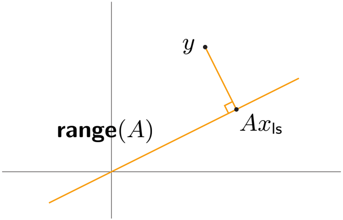
- define *residual* or error $r = Ax - y$
- find $x = x_{\rm ls}$ that minimizes $\|r\|$

$x_{\rm ls}$ called *least-squares* (approximate) solution of $y = Ax$

Geometric interpretation

Given $y \in \mathbb{R}^m$, find $x \in \mathbb{R}^n$ to minimize $\|Ax - y\|$

$Ax_{\rm ls}$ is point in **range**(A) closest to y ($Ax_{\rm ls}$ is *projection* of y onto **range**(A))



Least-squares (approximate) solution

- assume A is full rank, skinny
- to find $x_{\rm ls}$, we'll minimize norm of residual squared,

$$\|r\|^2 = x^T A^T A x - 2y^T A x + y^T y$$

- set gradient w.r.t. x to zero:

$$\nabla_x \|r\|^2 = 2A^T A x - 2A^T y = 0$$

- yields the *normal equations*: $A^T A x = A^T y$
- assumptions imply $A^T A$ invertible, so we have

$$x_{\rm ls} = (A^T A)^{-1} A^T y$$

...a very famous formula

Least-squares (approximate) solution

- $x_{\rm ls}$ is linear function of y
- $x_{\rm ls} = A^{-1}y$ if A is square
- $x_{\rm ls}$ solves $y = Ax_{\rm ls}$ if $y \in \textbf{range}(A)$

Least-squares (approximate) solution

for A skinny and full rank, the *pseudo-inverse* of A is

$$A^\dagger = (A^T A)^{-1} A^T$$

- for A skinny and full rank, $A^\dagger$ is a *left inverse* of A

$$A^\dagger A = (A^T A)^{-1} A^T A = I$$

- if A is not skinny and full rank then $A^\dagger$ has a different definition

Projection on **range**(A)

Ax_{ls} is (by definition) the point in **range**(A) that is closest to y, *i.e.*, it is the *projection* of y onto **range**(A)

$$Ax_{\text{ls}} = \mathcal{P}_{\mathbf{range}(A)}(y)$$

- the projection function $\mathcal{P}_{\mathbf{range}(A)}$ is linear, and given by

$$\mathcal{P}_{\mathbf{range}(A)}(y) = Ax_{\text{ls}} = A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}y$$

- $A(A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}$ is called the *projection matrix* (associated with **range**(A))

Orthogonality principle

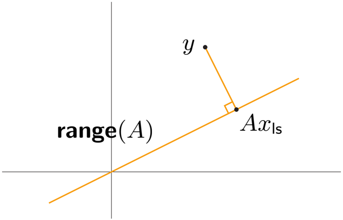
optimal residual

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

is orthogonal to **range**(A):

$$\langle r, Az\rangle = y^\mathsf{T}(A(A^\mathsf{T}A)^{-1}A^\mathsf{T} - I)^\mathsf{T}Az = 0$$

for all $z \in \mathbb{R}^n$



Completion of squares

since $r = Ax_{\mathsf{ls}} - y \perp A(x - x_{\mathsf{ls}})$ for any x, we have

$$\begin{aligned} \|Ax - y\|^2 &= \|(Ax_{\mathsf{ls}} - y) + A(x - x_{\mathsf{ls}})\|^2 \\ &= \|Ax_{\mathsf{ls}} - y\|^2 + \|A(x - x_{\mathsf{ls}})\|^2 \end{aligned}$$

this shows that for $x \neq x_{\mathsf{ls}}$, $\|Ax - y\| > \|Ax_{\mathsf{ls}} - y\|$

Least-squares via QR factorization

- $A \in \mathbb{R}^{m \times n}$ skinny, full rank
- factor as $A = QR$ with $Q^\mathsf{T} Q = I_n$, $R \in \mathbb{R}^{n \times n}$ upper triangular, invertible
- pseudo-inverse is

$$A^\dagger = (A^\mathsf{T} A)^{-1} A^\mathsf{T} = (R^\mathsf{T} Q^\mathsf{T} Q R)^{-1} R^\mathsf{T} Q^\mathsf{T} = R^{-1} Q^\mathsf{T}$$

so $x_{\rm ls} = R^{-1} Q^\mathsf{T} y$

- projection on **range**(A) given by matrix

$$A(A^\mathsf{T} A)^{-1} A^\mathsf{T} = A R^{-1} Q^\mathsf{T} = Q Q^\mathsf{T}$$

Least-squares via full QR factorization

- full QR factorization:

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

$\begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \in \mathbb{R}^{m \times m}$ orthogonal, $R_1 \in \mathbb{R}^{n \times n}$ upper triangular, invertible

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

$$\begin{aligned}
\|Ax - y\|^2 &= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - y \right\|^2 \\
&= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^{\mathsf{T}} y \right\|^2 \\
&= \left\| \begin{bmatrix} R_1 x - Q_1^{\mathsf{T}} y \\ -Q_2^{\mathsf{T}} y \end{bmatrix} \right\|^2 \\
&= \|R_1 x - Q_1^{\mathsf{T}} y\|^2 + \|Q_2^{\mathsf{T}} y\|^2
\end{aligned}$$

Least-squares via full QR factorization

so for any y,

$$\|Ax - y\|^2 = \|R_1 x - Q_1^\mathsf{T} y\|^2 + \|Q_2^\mathsf{T} y\|^2$$

- this is evidently minimized by choice $x_{\mathsf{ls}} = R_1^{-1} Q_1^\mathsf{T} y$ (which makes first term zero)
- residual with optimal x is

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

- $Q_1 Q_1^\mathsf{T}$ gives projection onto **range**(A)
- $Q_2 Q_2^\mathsf{T}$ gives projection onto **range**$(A)^\perp$

Least-squares estimation

many applications in inversion, estimation, and reconstruction problems have form

$$y = Ax + v$$

- x is what we want to estimate or reconstruct
- y is our sensor measurement(s)
- v is an unknown *noise* or *measurement error* (assumed small)
- ith row of A characterizes ith sensor

Least-squares estimation

least-squares estimation: choose as estimate $\hat{x}$ that minimizes

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

$i.e.$, deviation between

- what we actually observed (y), and
- what we would observe if $x = \hat{x}$, and there were no noise ($v = 0$)

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

BLUE property

suppose A full rank, skinny, and we have linear measurement with noise

$$y = Ax + v$$

consider a *linear estimator* of form $\hat{x} = By$

- B is called *unbiased* if $\hat{x} = x$ whenever $v = 0$
  - no estimation error when there is no noise
  - equivalent to left inverse property $BA = I$
- estimation error of unbiased linear estimator is

$$x - \hat{x} = x - B(Ax + v) = -Bv$$

- so we'd like B 'small' and $BA = I$

BLUE property

fact: $A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$ is the *smallest* left inverse of A, in the following sense:

for any B with $BA = I$, we have

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

$i.e.$, least-squares provides the *best linear unbiased estimator* (BLUE)