8. Least squares

- least squares problem
- solution of a least squares problem
- solving least squares problems

Least squares problem

given $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$, find vector $x \in \mathbf{R}^n$ that minimizes

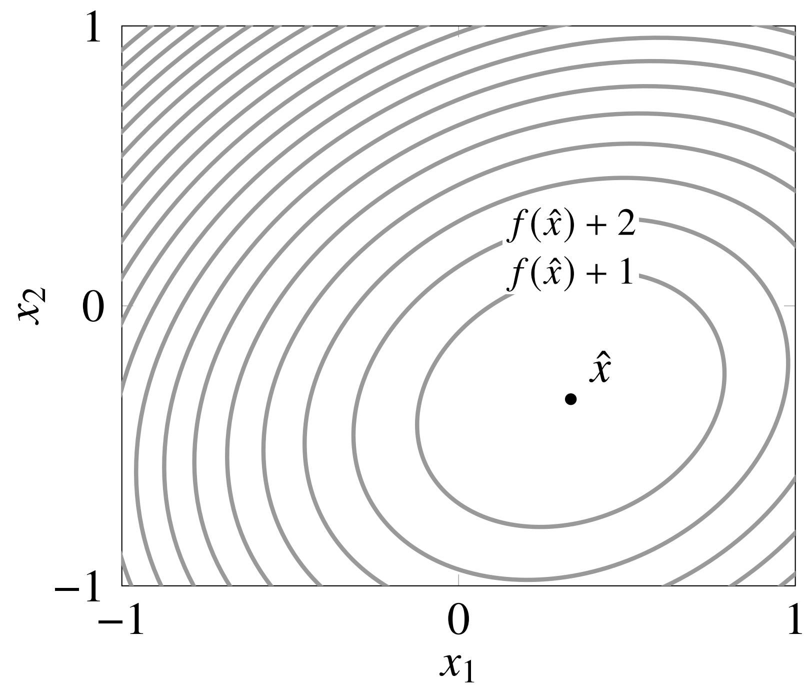
$$\|Ax - b\|^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij}x_j - b_i \right)^2$$

- "least squares" because we minimize a sum of squares of affine functions:

$$\|Ax - b\|^2 = \sum_{i=1}^{m} r_i(x)^2, \qquad r_i(x) = \sum_{j=1}^{n} A_{ij}x_j - b_i$$

- the problem is also called the *linear* least squares problem

Example



$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

- the least squares solution $\hat{x}$ minimizes

$$f(x) = \|Ax - b\|^2 = (2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2$$

- to find $\hat{x}$, set derivatives with respect to x_1 and x_2 equal to zero:

$$10x_1 - 2x_2 - 4 = 0, \qquad -2x_1 + 10x_2 + 4 = 0$$

solution is $(\hat{x}_1, \hat{x}_2) = (1/3, -1/3)$

Least squares and linear equations

$$\text{minimize} \quad \|Ax - b\|^2$$

- solution of the least squares problem: any $\hat{x}$ that satisfies

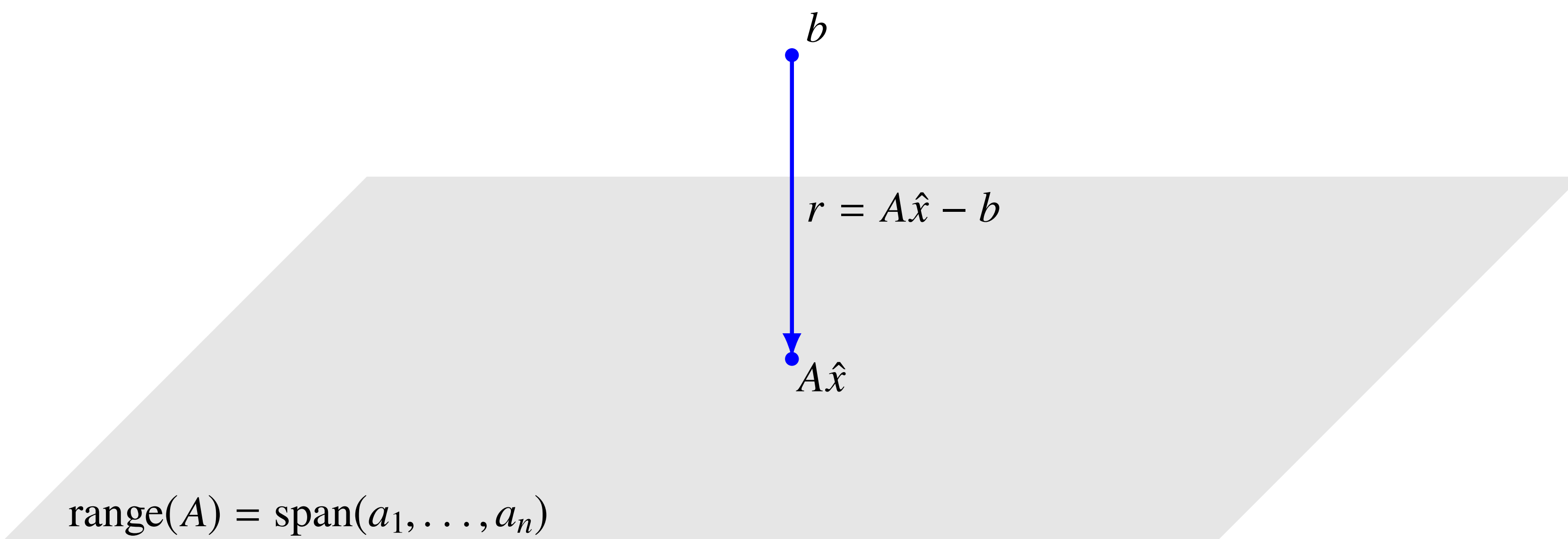
$$\|A\hat{x} - b\| \leq \|Ax - b\| \quad \text{for all } x$$

- $\hat{r} = A\hat{x} - b$ is the *residual vector*
- if $\hat{r} = 0$, then $\hat{x}$ solves the linear equation $Ax = b$
- if $\hat{r} \neq 0$, then $\hat{x}$ is a *least squares approximate solution* of the equation
- in most least squares applications, $m > n$ and $Ax = b$ has no solution

Column interpretation

least squares problem in terms of columns $a_1, a_2, \ldots, a_n$ of A:

$$\text{minimize} \quad \|Ax - b\|^2 = \|\sum_{j=1}^{n} a_j x_j - b\|^2$$

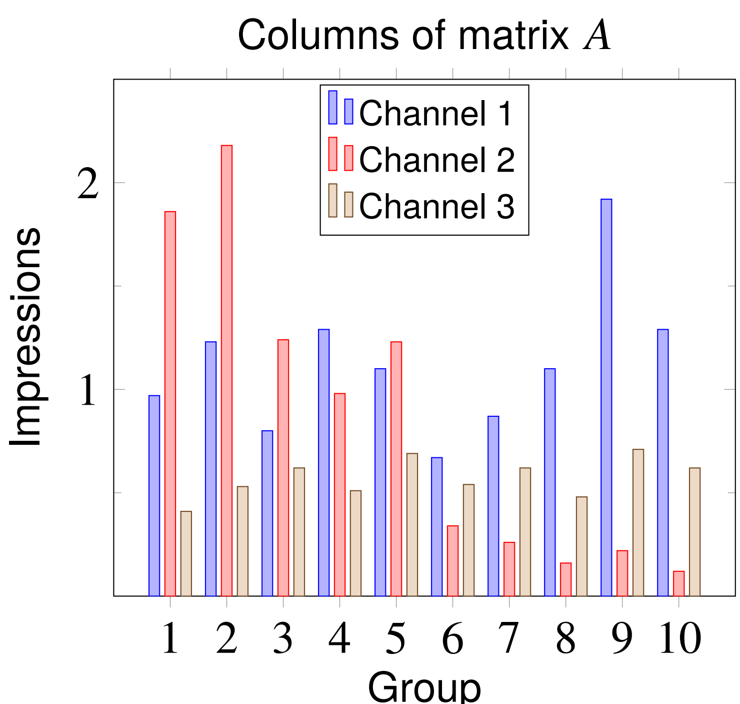


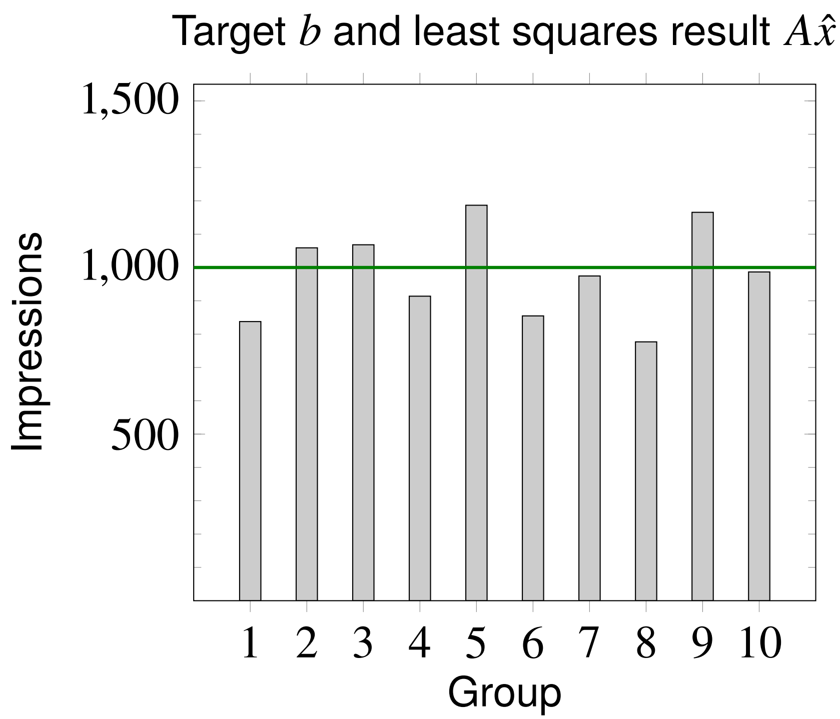
- $A\hat{x}$ is the vector in $\text{range}(A) = \text{span}(a_1, a_2, \ldots, a_n)$ closest to b
- geometric intuition suggests that $\hat{r} = A\hat{x} - b$ is orthogonal to $\text{range}(A)$

Example: advertising purchases

- m demographic groups; n advertising channels
- A_{ij} is # impressions (views) in group i per dollar spent on ads in channel j
- x_j is amount of advertising purchased in channel j
- $(Ax)_i$ is number of impressions in group i
- b_i is target number of impressions in group i

Example: $m = 10$, $n = 3$, $b = 10^3 \mathbf{1}$

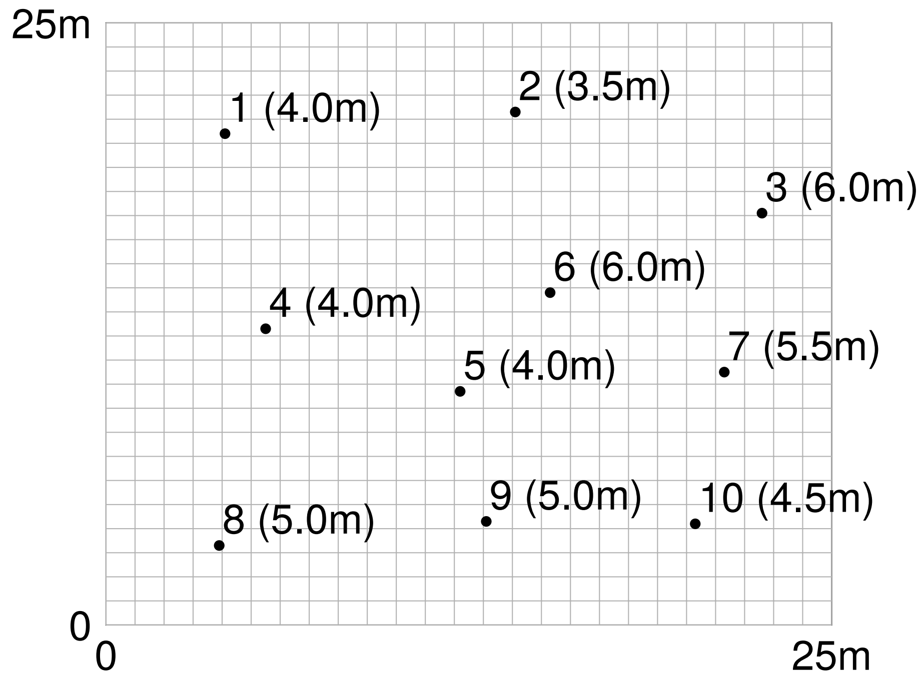




Example: illumination

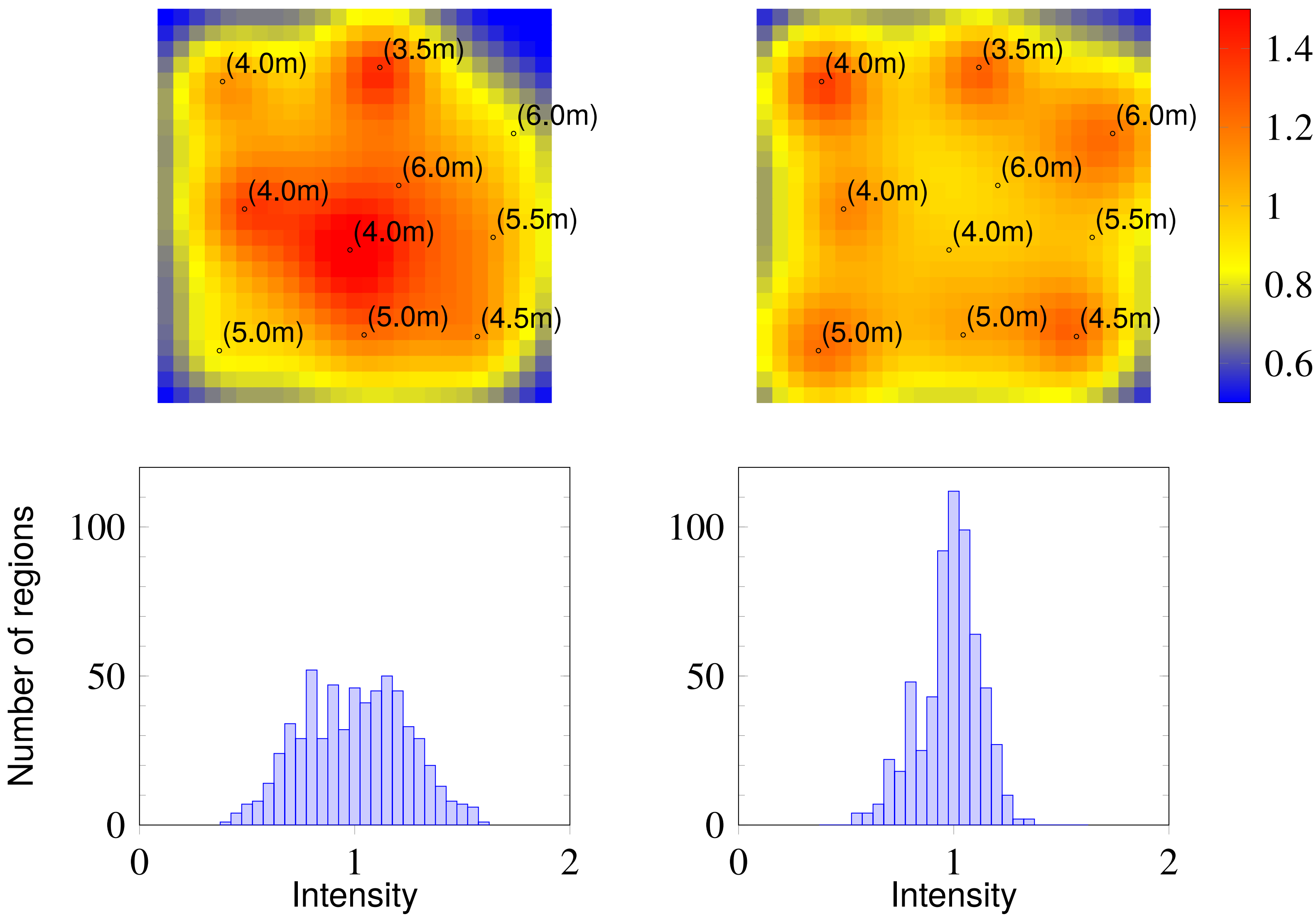
- n lamps at given positions above an area divided in m regions
- A_{ij} is illumination in region i if lamp j is on with power 1 and other lamps are off
- x_j is power of lamp j
- $(Ax)_i$ is illumination level at region i
- b_i is target illumination level at region i

Example: $m = 25^2$, $n = 10$; figure shows position and height of each lamp



Example: illumination

- left: illumination pattern for equal lamp powers ($x = \mathbf{1}$)
- right: illumination pattern for least squares solution $\hat{x}$, with $b = \mathbf{1}$



Outline

- least squares problem
- **solution of a least squares problem**
- solving least squares problems

Solution of a least squares problem

if A has linearly independent columns (is left-invertible), then the vector

$$\begin{aligned} \hat{x} &= (A^T A)^{-1} A^T b \\ &= A^\dagger b \end{aligned}$$

is the unique solution of the least squares problem

$$\text{minimize} \quad \|Ax - b\|^2$$

- in other words, if $x \neq \hat{x}$, then $\|Ax - b\|^2 > \|A\hat{x} - b\|^2$
- recall from page 4.22 that

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

  is called the *pseudo-inverse* of a left-invertible matrix

Proof

we show that $\|Ax - b\|^2 > \|A\hat{x} - b\|^2$ for $x \neq \hat{x}$:

$$\begin{aligned} \|Ax - b\|^2 &= \|A(x - \hat{x}) + (A\hat{x} - b)\|^2 \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 \\ &> \|A\hat{x} - b\|^2 \end{aligned}$$

- 2nd step follows from $A(x - \hat{x}) \perp (A\hat{x} - b)$:

$$(A(x - \hat{x}))^T (A\hat{x} - b) = (x - \hat{x})^T (A^T A\hat{x} - A^T b) = 0$$

- 3rd step follows from linear independence of columns of A:

$$A(x - \hat{x}) \neq 0 \quad \text{if } x \neq \hat{x}$$

Derivation from calculus

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right)^2$$

- partial derivative of f with respect to x_k

$$\frac{\partial f}{\partial x_k}(x) = 2 \sum_{i=1}^{m} A_{ik} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right) = 2(A^T(Ax - b))_k$$

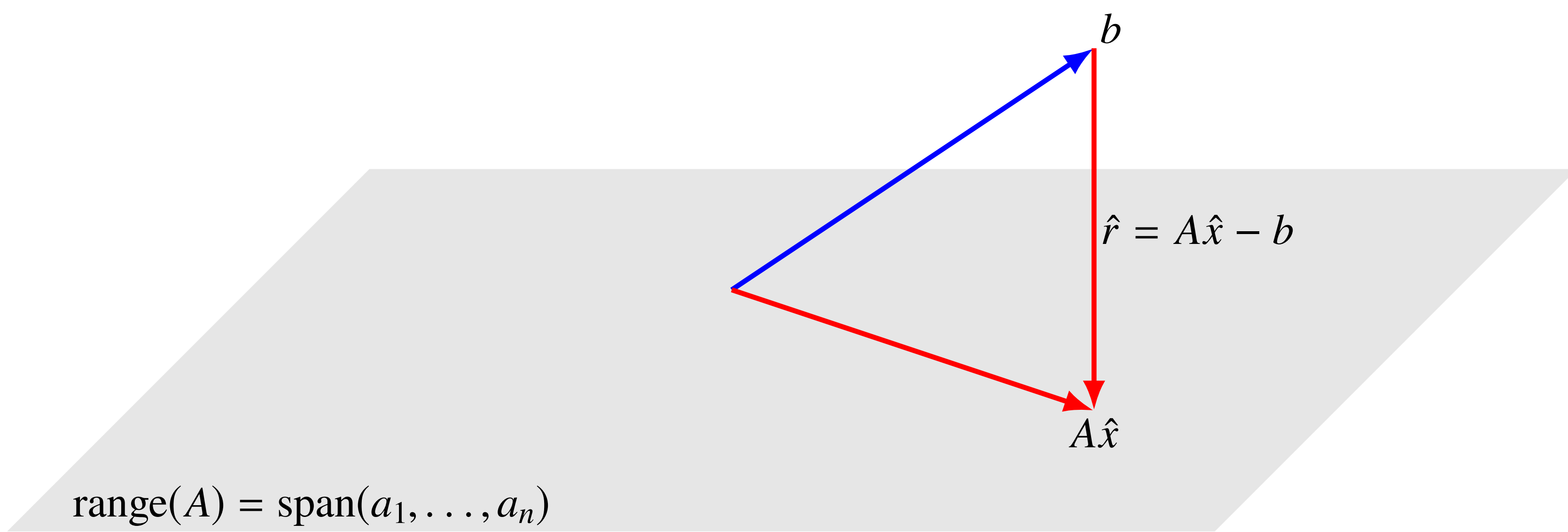
- gradient of f is

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) = 2A^T(Ax - b)$$

- minimizer $\hat{x}$ of $f(x)$ satisfies $\nabla f(\hat{x}) = 2A^T(A\hat{x} - b) = 0$

Geometric interpretation

residual vector $\hat{r} = A\hat{x} - b$ satisfies $A^T\hat{r} = A^T(A\hat{x} - b) = 0$



- residual vector $\hat{r}$ is orthogonal to every column of A; hence, to $\text{range}(A)$
- projection on $\text{range}(A)$ is a matrix-vector multiplication with the matrix

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

Outline

- least squares problem
- solution of a least squares problem
- **solving least squares problems**

Normal equations

$$A^T A x = A^T b$$

- these equations are called the *normal equations* of the least squares problem
- coefficient matrix $A^T A$ is the Gram matrix of A
- equivalent to $\nabla f(x) = 0$ where $f(x) = \|Ax - b\|^2$
- all solutions of the least squares problem satisfy the normal equations

if A has linearly independent columns, then:

- $A^T A$ is nonsingular
- normal equations have a unique solution $\hat{x} = (A^T A)^{-1} A^T b$

QR factorization method

rewrite least squares solution using QR factorization $A = QR$

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

Algorithm

1. compute QR factorization $A = QR$ ($2mn^2$ flops if A is $m \times n$)
2. matrix-vector product $d = Q^Tb$ ($2mn$ flops)
3. solve $Rx = d$ by back substitution (n^2 flops)

complexity: $2mn^2$ flops

Example

$$A = \begin{bmatrix} 3 & -6 \\ 4 & -8 \\ 0 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} -1 \\ 7 \\ 2 \end{bmatrix}$$

1. QR factorization: $A = QR$ with

$$Q = \begin{bmatrix} 3/5 & 0 \\ 4/5 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} 5 & -10 \\ 0 & 1 \end{bmatrix}$$

2. calculate $d = Q^T b = (5, 2)$
3. solve $Rx = d$

$$\begin{bmatrix} 5 & -10 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

solution is $x_1 = 5$, $x_2 = 2$

Solving the normal equations

why not solve the normal equations

$$A^T A x = A^T b$$

as a set of linear equations?

Example: a 3×2 matrix with "almost linearly dependent" columns

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 10^{-5} \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 10^{-5} \\ 1 \end{bmatrix},$$

we round intermediate results to 8 significant decimal digits

Solving the normal equations

Method 1: form Gram matrix A^TA and solve normal equations

$$A^TA = \begin{bmatrix} 1 & -1 \\ -1 & 1+10^{-10} \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad A^Tb = \begin{bmatrix} 0 \\ 10^{-10} \end{bmatrix}$$

after rounding, the Gram matrix is singular; hence method fails

Method 2: QR factorization of A is

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 1 & -1 \\ 0 & 10^{-5} \end{bmatrix}$$

rounding does not change any values (in this example)

- problem with method 1 occurs when forming Gram matrix A^TA
- QR factorization method is more stable because it avoids forming A^TA