8. Least squares

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

given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, find vector $x \in \mathbb{R}^n$ 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:

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

- also called linear least squares problem

- nonlinear least squares problem is to minimize $\sum_i r_i(x)^2$ with general $r_i$
Example

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

- 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:
  \[ 10x_1 - 2x_2 - 4 = 0, \quad -2x_1 + 10x_2 + 4 = 0 \]

solution is \( \hat{x}_1 = 1/3, \hat{x}_2 = -1/3 \)
Example: polynomial approximation

fit a polynomial of degree less than \( n \)

\[
p(t) = x_1 + x_2 t + x_3 t^2 + \cdots + x_n t^{n-1}
\]

to \( m \) data points \((t_1, y_1), \ldots, (t_m, y_m)\)

- least squares formulation: choose coefficients \( x_1, \ldots, x_n \) that minimize

\[
(p(t_1) - y_1)^2 + (p(t_2) - y_2)^2 + \cdots + (p(t_m) - y_m)^2
\]

- in matrix form: minimize \( \|Ax - b\|^2 \) with

\[
A = \begin{bmatrix}
1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\
1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\
\vdots & \vdots & \vdots & & \vdots \\
1 & t_m & t_m^2 & \cdots & t_m^{n-1}
\end{bmatrix}, \quad b = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_m
\end{bmatrix}
\]
Example

$m = 50$ points, polynomials of degree $10$ ($n = 11$) and $15$ ($n = 16$)
Least squares and linear equations

any \( \hat{x} \) that satisfies

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

is a solution of the least squares problem

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

least squares problem in terms of rows $\tilde{a}_1^T$, $\tilde{a}_2^T$, \ldots, $\tilde{a}_m^T$ of $A$

$$\text{minimize} \quad \|Ax - b\|^2 = (\tilde{a}_1^T x - b_1)^2 + \cdots + (\tilde{a}_m^T x - b_m)^2$$

• if $\tilde{a}_i \neq 0$, distance of $x$ to hyperplane $H_i = \{y \mid \tilde{a}_i^T y = b_i\}$ is

$$d_i(x) = \frac{|\tilde{a}_i^T x - b_i|}{\|\tilde{a}_i\|}$$

• least squares solution minimizes weighted sum of squared distances

$$\|Ax - b\|^2 = \sum_{i=1}^{m} w_i d_i(x)^2 \quad \text{with weights} \ w_i = \|\tilde{a}_i\|^2$$

• if row norms are equal, LS solution minimizes sum of squared distances
Example

\[ A = \begin{bmatrix} 2 & 0 \\ 1 & -1 \\ 0 & -1 \end{bmatrix} \]

\[ b = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \]

\[ \|Ax - b\|^2 = 4d_1(x)^2 + 2d_2(x)^2 + d_3(x)^2 \]

\(d_1(x)\) is distance to \(H_1\), \(d_2(x)\) is distance to \(H_2\), \(d_3(x)\) is distance to \(H_3\)
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

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

$$\hat{x} = A^\dagger b$$

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 4-23 that

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

is the \textit{pseudo-inverse} of a left invertible matrix
Proof

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

\[
\|Ax - b\|^2 = \|A(x - \hat{x}) + (A\hat{x} - b)\|^2 \\
= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 \\
> \|A\hat{x} - b\|^2
\]

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

- residual vector 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 $A(A^TA)^{-1}A^T$

range($A$) = span($a_1, \ldots, a_n$)
Outline

• least squares problem

• solution of a least squares problem

• solving least squares problems
Normal equations

\[ A^T Ax = A^T b \]

- these equations are called the *normal equations* of the LS problem
- coefficient matrix is the Gram matrix \( A^T A \) 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 $A = QR$

$$\hat{x} = (A^T A)^{-1} A^T b = ((QR)^T (QR))^{-1} (QR)^T b$$

$$= (R^T Q^T QR)^{-1} R^T Q^T b$$

$$= (R^T R)^{-1} R^T Q^T b$$

$$= R^{-1} R^{-T} R^T Q^T b$$

$$= R^{-1} Q^T b$$

Algorithm

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

complexity: $2mn^2$ flops
Example

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

1. QR factorization: \( A = QR \) with

\[
Q = \begin{bmatrix}
3/5 & 0 \\
4/5 & 0 \\
0 & 1
\end{bmatrix}, \quad R = \begin{bmatrix}
5 & -10 \\
0 & 1
\end{bmatrix}
\]

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 $3 \times 2$ matrix with ‘almost linearly dependent’ columns

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

we round intermediate results to 8 significant decimal digits
Solving the normal equations

**Method 1:** form Gram matrix $A^T A$ and solve normal equations

$$A^T A = \begin{bmatrix} 1 & -1 \\ -1 & 1 + 10^{-10} \end{bmatrix} \sim \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad A^T b = \begin{bmatrix} 0 \\ 10^{-10} \end{bmatrix}$$

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

**Method 2:** QR factorization of $A$ is

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

rounding does not change any values (in this example)

- QR factorization method is more stable
- problem with method 1 occurs when forming Gram matrix $A^T A$