Matrix Computations
Chapter 3: Least-squares Problems and QR Decomposition

Section 3.2 Least-squares Solution

Jie Lu
ShanghaiTech University

LS Solution

Theorem (LS Optimality Condition)

$\mathbf{x}_{\text{LS}} \in \mathbb{R}^n$ is an optimal solution to the LS problem $\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ if and only if it satisfies the following normal equation:

$$\mathbf{A}^T\mathbf{A}\mathbf{x}_{\text{LS}} = \mathbf{A}^T\mathbf{y}. \qquad (*)$$

- The optimality condition $(*)$ is true for any $\mathbf{A}$, not limited to full-column rank $\mathbf{A}$
- When $\mathbf{A}$ has full-column rank,
  - $\mathbf{A}^T\mathbf{A}$ is nonsingular
  - $\mathbf{x}_{\text{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$ is the *unique* solution to $(*)$
- Same result holds for the complex case

$$\mathbf{A}^H\mathbf{A}\mathbf{x}_{\text{LS}} = \mathbf{A}^H\mathbf{y}$$

Proof using the Projection Theorem

The above Theorem can be proved using the Projection Theorem

Let $\mathbf{x}_{\mathrm{LS}}$ be an LS solution. Then,

$$\Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = \arg \min_{\mathbf{z} \in \mathcal{R}(\mathbf{A})} \|\mathbf{z} - \mathbf{y}\|_2^2 = \mathbf{A}\mathbf{x}_{\mathrm{LS}}$$

From the Projection Theorem (Section 1.2),

$$\begin{aligned} \Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = \mathbf{A}\mathbf{x}_{\mathrm{LS}} \quad &\Longleftrightarrow \quad \mathbf{z}^T(\mathbf{A}\mathbf{x}_{\mathrm{LS}} - \mathbf{y}) = 0 \text{ for all } \mathbf{z} \in \mathcal{R}(\mathbf{A}) \\ &\Longleftrightarrow \quad \mathbf{x}^T\mathbf{A}^T(\mathbf{A}\mathbf{x}_{\mathrm{LS}} - \mathbf{y}) = 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n \\ &\Longleftrightarrow \quad \mathbf{A}^T(\mathbf{A}\mathbf{x}_{\mathrm{LS}} - \mathbf{y}) = \mathbf{0} \end{aligned}$$

Orthogonal Projections

Suppose $\mathbf{A}$ has full column rank

- The projections of $\mathbf{y}$ onto $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A})^{\perp}$ are given by

$$\Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = \mathbf{A}\mathbf{x}_{\mathrm{LS}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$
$$\Pi_{\mathcal{R}(\mathbf{A})^{\perp}}(\mathbf{y}) = \mathbf{y} - \Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = (\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T)\mathbf{y}$$

- The orthogonal projector of $\mathbf{A}$ is defined as

$$\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

- The orthogonal complement projector of $\mathbf{A}$ is defined as

$$\mathbf{P}_{\mathbf{A}}^{\perp} = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

- $\Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = \mathbf{P}_{\mathbf{A}}\mathbf{y}$, $\Pi_{\mathcal{R}(\mathbf{A})^{\perp}}(\mathbf{y}) = \mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{y}$

Orthogonal Projections

Properties of $\mathbf{P_A}$ (same to $\mathbf{P_A^\perp}$):

- $\mathbf{P_A}$ is idempotent, i.e., $\mathbf{P_A P_A = P_A}$
- $\mathbf{P_A}$ is symmetric

Some other properties (will be revealed later):

- The eigenvalues of $\mathbf{P_A}$ are either zero or one
- $\mathbf{P_A}$ can be written as $\mathbf{P_A} = \mathbf{U}_1\mathbf{U}_1^T$ for some semi-orthogonal $\mathbf{U}_1$

Sketch of Proof: There always exists a semi-orthogonal $\mathbf{U}_1$ such that $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{U}_1)$, so that $\Pi_{\mathcal{R}(\mathbf{A})}(\mathbf{y}) = \Pi_{\mathcal{R}(\mathbf{U}_1)}(\mathbf{y})$ for all $\mathbf{y}$.
Also note that $\Pi_{\mathcal{R}(\mathbf{U}_1)}(\mathbf{y}) = \mathbf{U}_1\mathbf{U}_1^T\mathbf{y}$.
It follows that $(\mathbf{P_A} - \mathbf{U}_1\mathbf{U}_1^T)\mathbf{y} = \mathbf{0}$ for all $\mathbf{y}$. Therefore, $\mathbf{P_A} = \mathbf{U}_1\mathbf{U}_1^T$.

Pseudo-Inverse

The pseudo-inverse of a full-column-rank $\mathbf{A}$ is defined as

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

- $\mathbf{A}^{\dagger}$ satisfies $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}$, but not necessarily $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$
- $\mathbf{A}^{\dagger}\mathbf{y}$ is the unique LS solution
- We will study pseudo-inverse for general matrices later

LS by Convex Optimization

The LS optimality condition can also be proved via convex optimization

Definitions:

- The gradient of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any $\alpha \in [0, 1]$,

$$f(\alpha \mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y})$$

Fact: Consider an unconstrained optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \; f(\mathbf{x})$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable

- For convex f, $\mathbf{x}^\star$ is an optimal solution if and only if $\nabla f(\mathbf{x}^\star) = \mathbf{0}$
- For non-convex f, any point $\hat{\mathbf{x}}$ satisfying $\nabla f(\hat{\mathbf{x}}) = \mathbf{0}$ is a stationary point

LS by Convex Optimization (cont'd)

Fact: Consider a quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{R}\mathbf{x} + \mathbf{q}^T\mathbf{x} + c$$

where $\mathbf{R} = \mathbf{R}^T \in \mathbb{R}^{n\times n}$

- $\nabla f(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{q}$
- f is convex if $\mathbf{R}$ is positive semi-definite

The LS objective function is

$$f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2(\mathbf{A}^T\mathbf{y})^T\mathbf{x} + \|\mathbf{y}\|_2^2$$

Using the above fact, $\mathbf{x}_{\rm LS}$ is an LS optimal solution if and only if

$$\mathbf{A}^T\mathbf{A}\mathbf{x}_{\rm LS} - \mathbf{A}^T\mathbf{y} = \mathbf{0}$$

LS by Convex Optimization (cont'd)

Example: Consider a regularized LS problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2, \quad \text{for some constant (weight) } \lambda > 0$$

- Solution by optimization:

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

  The optimal solution is

$$\mathbf{x}_{\text{RLS}} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y}$$

- Solution by the Projection Theorem: Rewrite the problem as

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \mathbf{x} \right\|_2^2,$$

  and then use the Projection Theorem to get the same result