ECE240 Introduction to Linear Dynamical Systems

Lecture 7: Least-squares: Part I

Dr. Yu Zhang

ECE Department, UC Santa Cruz

Fall 2025

Outline

Overdetermined Linear Equations

Consider

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is strictly skinny, i.e., $m > n$.

- This is an overdetermined set of linear equations (more equations than unknowns).
- For most $\mathbf{y}$, there is no exact solution $\mathbf{x}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$.

One approach to approximately solve $\mathbf{y} = \mathbf{A}\mathbf{x}$:

- Define the residual or error vector

$$\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{y}.$$

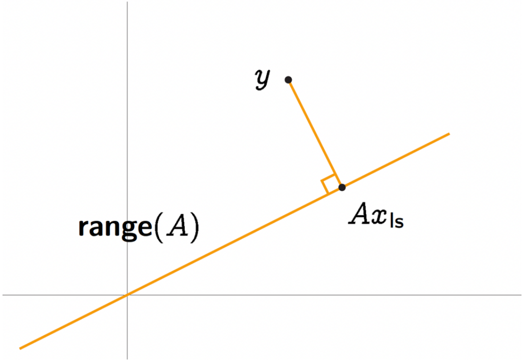
- Find $\mathbf{x}_{\mathrm{ls}}$ that minimizes the ℓ_2 norm: $\|\mathbf{r}\|$.

$\mathbf{x}_{\mathrm{ls}}$ is called the least squares (approximate) solution of $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Geometric Interpretation

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

- $\mathbf{A}\mathbf{x}_{\mathrm{ls}}$ is the point in $\mathrm{range}(\mathbf{A})$ closest to $\mathbf{y}$.
- $\mathbf{A}\mathbf{x}_{\mathrm{ls}}$ is the orthogonal projection of $\mathbf{y}$ on $\mathrm{range}(\mathbf{A})$.



Least Squares Solution: Normal Equations

Assume $\mathbf{A}$ is full column rank and skinny.

We minimize the squared norm of the residual

$$\|\mathbf{r}\|^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} - 2\mathbf{y}^\top \mathbf{A}\mathbf{x} + \mathbf{y}^\top \mathbf{y}.$$

Take gradient with respect to $\mathbf{x}$ and set to zero:

$$\nabla_{\mathbf{x}} \|\mathbf{r}\|^2 = 2\mathbf{A}^\top \mathbf{A}\mathbf{x} - 2\mathbf{A}^\top \mathbf{y} = \mathbf{0}.$$

This yields the **normal equations**

$$\boxed{\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y}}$$

Since $\mathbf{A}$ is full column rank, $\mathbf{A}^\top \mathbf{A}$ is invertible, we get

$$\boxed{\mathbf{x}_{\mathrm{ls}} = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{y}}$$

Least Squares Solution: Properties

- $\mathbf{x}_{\text{ls}}$ is a linear function of $\mathbf{y}$:
$$\mathbf{x}_{\text{ls}} = \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y}.$$
- If $\mathbf{A}$ is square and invertible, then
$$\mathbf{x}_{\text{ls}} = \mathbf{A}^{-1}\mathbf{y},$$
so the least squares solution coincides with the exact solution.
- If $\mathbf{y} \in \operatorname{range}(\mathbf{A})$, then $\mathbf{x}_{\text{ls}}$ solves $\mathbf{y} = \mathbf{A}\mathbf{x}$ exactly so that $\mathbf{r} = \mathbf{0}$.

Pseudoinverse for Full Column Rank Matrix

For skinny full rank $\mathbf{A}$, the pseudoinverse of $\mathbf{A}$ is

$$\mathbf{A}^{\dagger} = \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}.$$

Then the least squares solution can be written as

$$\mathbf{x}_{\mathrm{ls}} = \mathbf{A}^{\dagger}\mathbf{y}.$$

For skinny full rank $\mathbf{A}$, $\mathbf{A}^{\dagger}$ is a *left inverse*:

$$\mathbf{A}^{\dagger}\mathbf{A} = \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}_n.$$

Question: How about $\mathbf{A}\mathbf{A}^{\dagger}$?

Projection onto $\mathrm{range}(\mathbf{A})$

By definition, $\mathbf{A}\mathbf{x}_{\mathrm{ls}}$ is the point in $\mathrm{range}(\mathbf{A})$ closest to $\mathbf{y}$.

$$\mathbf{A}\mathbf{x}_{\mathrm{ls}} = \mathbf{A}\mathbf{A}^{\dagger}\mathbf{y} =: \mathbf{P}_{\mathrm{range}(\mathbf{A})}\mathbf{y}.$$

- The projection operator onto $\mathrm{range}(\mathbf{A})$ is linear.
- It is represented by the projection matrix

$$\mathbf{P}_{\mathrm{range}(\mathbf{A})} = \mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}\left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}.$$

- $\mathbf{P}_{\mathrm{range}(\mathbf{A})}$ is symmetric and idempotent:

$$\mathbf{P}^{\mathsf{T}} = \mathbf{P}, \quad \mathbf{P}^2 = \mathbf{P}.$$

Projection Matrix Must Be Idempotent

1. **What a projection does.**

- A projection maps any vector $\mathbf{x}$ to its component in a subspace $\mathcal{S}$:

$$\mathbf{x} \longmapsto \mathbf{P}\mathbf{x} \in \mathcal{S}.$$

- Once a vector is already in $\mathcal{S}$, projecting again should not move it:

$$\mathbf{P}\mathbf{x} \in \mathcal{S} \;\Rightarrow\; \mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{x}.$$

2. **This behavior implies idempotence.**

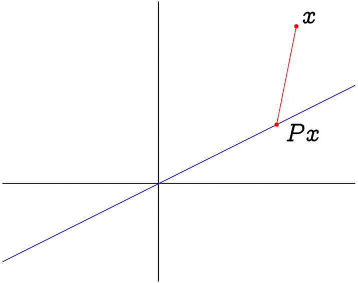
- Since projecting twice must equal projecting once for *all* $\mathbf{x}$:

$$\mathbf{P}^2\mathbf{x} = \mathbf{P}\mathbf{x} \quad \forall\,\mathbf{x} \iff \mathbf{P}^2 = \mathbf{P}.$$

- Idempotence encodes the idea: **"once projected, always projected."**

Projection vs. Orthogonal Projection

- However, idempotence alone does not guarantee the projection is orthogonal.
- Such a **P** may be:
  - an **oblique** projection, or
  - an **orthogonal** projection.
- We need an additional condition: **symmetry**.



Adding Symmetry Makes the Projection Orthogonal

Assume $\mathbf{P}$ is a projection $\mathbf{P}^2 = \mathbf{P}$. Add symmetry: $\mathbf{P}^\top = \mathbf{P}$.

Let $\mathbf{x} \in \mathbb{R}^n$, and decompose:

$$\mathbf{x} = \mathbf{P}\mathbf{x} + (\mathbf{x} - \mathbf{P}\mathbf{x}) = (\text{projected part}) + (\text{error part}).$$

To be an orthogonal projection, we need:

$$(\mathbf{x} - \mathbf{P}\mathbf{x}) \perp \operatorname{range}(\mathbf{P}).$$

Let $\mathbf{y} \in \operatorname{range}(\mathbf{P})$ such that $\mathbf{y} = \mathbf{P}\mathbf{z}$. Using $\mathbf{P}^\top = \mathbf{P}$, we have

$$(\mathbf{x} - \mathbf{P}\mathbf{x})^\top \mathbf{P}\mathbf{z} = \mathbf{x}^\top \mathbf{P}\mathbf{z} - (\mathbf{P}\mathbf{x})^\top \mathbf{P}\mathbf{z} = \mathbf{x}^\top (\mathbf{P} - \mathbf{P}^2)\mathbf{z} = 0 \Rightarrow (\mathbf{x} - \mathbf{P}\mathbf{x}) \perp \operatorname{range}(\mathbf{P}).$$

$$\boxed{\text{idempotent} + \text{symmetric} \Longrightarrow \text{orthogonal projection}}$$

Characterization of Orthogonal Projection Matrices

$$\boxed{\mathbf{P} \text{ is an orthogonal projection} \iff \mathbf{P}^2 = \mathbf{P} \text{ and } \mathbf{P}^\mathsf{T} = \mathbf{P}.}$$

Consequences:

- $\mathbf{P}$ is diagonalizable with eigenvalues 0 or 1.
- $\mathbf{I} - \mathbf{P}$ is orthogonal projection onto $\mathrm{Null}(\mathbf{P})$.
- The decomposition

$$\mathbb{R}^n = \mathrm{range}(\mathbf{P}) \oplus \mathrm{null}(\mathbf{P})$$

is an orthogonal direct sum.

Orthogonality Principle

- The optimal residual is $\mathbf{r}^* = \mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y} = -(\underbrace{\mathbf{I} - \mathbf{P}_{\mathrm{range}(\mathbf{A})}}_{\mathbf{Q}})\mathbf{y}$.
- $\mathbf{Q}$ is the orthogonal projection onto the subspace $\mathrm{range}(\mathbf{A})^\perp = \mathrm{null}(\mathbf{A}^\mathsf{T})$
- **Orthogonality principle:** $\mathbf{r}^*$ is orthogonal to $\mathrm{range}(\mathbf{A})$; equivalently,

$$\langle \mathbf{r}^*, \mathbf{A}\mathbf{z} \rangle = -\mathbf{y}^\mathsf{T}\mathbf{Q}\mathbf{A}\mathbf{z} = 0$$

for all $\mathbf{z} \in \mathbb{R}^n$ because $\mathbf{Q}\mathbf{A} = \mathbf{0}$.

Completion of Squares

For any $\mathbf{x}$, write

$$\mathbf{A}\mathbf{x} - \mathbf{y} = (\mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y}) + \mathbf{A}(\mathbf{x} - \mathbf{x}_{\mathrm{ls}}).$$

Since $\mathbf{A}(\mathbf{x} - \mathbf{x}_{\mathrm{ls}}) \in \mathrm{range}(\mathbf{A})$ is orthogonal to $\mathbf{r}^* = \mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y}$, we obtain

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 = \left\|\mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y}\right\|^2 + \left\|\mathbf{A}(\mathbf{x} - \mathbf{x}_{\mathrm{ls}})\right\|^2.$$

This shows that for any $\mathbf{x} \neq \mathbf{x}_{\mathrm{ls}}$,

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\| > \|\mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y}\|.$$

Least-squares via QR Factorization

- Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ is skinny and full rank. Factor as $\mathbf{A} = \mathbf{Q}_1 \mathbf{R}_1$ with

$$\mathbf{Q}_1 \in \mathbb{R}^{m \times n} \text{ has orthonormal columns; i.e., } \mathbf{Q}_1^\top \mathbf{Q}_1 = \mathbf{I}_n$$

$$\mathbf{R}_1 \in \mathbb{R}^{n \times n} \text{ invertible, upper triangular.}$$

- Pseudo inverse: $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top = (\mathbf{R}^\top \mathbf{Q}_1^\top \mathbf{Q}_1 \mathbf{R})^{-1} \mathbf{R}_1^\top \mathbf{Q}_1^\top = \mathbf{R}_1^{-1} \mathbf{Q}_1^\top$.
- Least-squares solution:

$$\mathbf{x}_{\mathrm{ls}} = \mathbf{R}_1^{-1} \mathbf{Q}_1^\top \mathbf{y},$$

  where back substitution can be used to quickly and accurately find $\mathbf{x}_{\mathrm{ls}}$ without explicitly inverting $\mathbf{R}_1$.
- Projection onto range($\mathbf{A}$):

$$\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top = \mathbf{A}\mathbf{R}_1^{-1} \mathbf{Q}_1^\top = \mathbf{Q}_1 \mathbf{Q}_1^\top.$$

Least-squares via Full QR Factorization (1/2)

- Full QR factorization:

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix},$$

  with

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \in \mathbb{R}^{m \times m} \text{ orthonormal} : \mathbf{Q}\mathbf{Q}^\mathsf{T} = \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}_m \implies \mathbf{Q}_1\mathbf{Q}_1^\mathsf{T} + \mathbf{Q}_2\mathbf{Q}_2^\mathsf{T} = \mathbf{I}_m$$

- where $\mathbf{Q}_2$ is obtained by completing a basis for $\mathbb{R}^m$; i.e.,

$$\mathbb{R}^m = \text{range}(\mathbf{Q}_1) \oplus \text{range}(\mathbf{Q}_2).$$

- See the example for $\mathbb{R}^3$ and the general procedure in Section 5 at the end of this slide deck.

Least-squares via Full QR Factorization (2/2)

- Multiplication by an orthonormal matrix does not change the norm, so

$$\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 = \left\|\mathbf{Q}^{\mathsf{T}}(\mathbf{A}\mathbf{x}-\mathbf{y})\right\|_2^2 = \left\|\mathbf{Q}^{\mathsf{T}}\mathbf{Q}\begin{bmatrix}\mathbf{R}_1\\ \mathbf{0}\end{bmatrix}\mathbf{x} - \begin{bmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{bmatrix}^{\mathsf{T}}\mathbf{y}\right\|_2^2$$

$$= \left\|\begin{bmatrix}\mathbf{R}_1\mathbf{x}-\mathbf{Q}_1^{\mathsf{T}}\mathbf{y}\\ -\mathbf{Q}_2^{\mathsf{T}}\mathbf{y}\end{bmatrix}\right\|_2^2 = \|\mathbf{R}_1\mathbf{x}-\mathbf{Q}_1^{\mathsf{T}}\mathbf{y}\|_2^2 + \|\mathbf{Q}_2^{\mathsf{T}}\mathbf{y}\|_2^2 \tag{1}$$

- Objective (1) is minimized by choosing $\mathbf{x}_{\mathrm{ls}} = \mathbf{R}_1^{-1}\mathbf{Q}_1^{\mathsf{T}}\mathbf{y}$, which makes the first term zero.
- Optimal residual $\mathbf{r}^* = \mathbf{A}\mathbf{x}_{\mathrm{ls}} - \mathbf{y} = (\mathbf{Q}_1\mathbf{Q}_1^{\mathsf{T}} - \mathbf{I})\mathbf{y} = -\mathbf{Q}_2\mathbf{Q}_2^{\mathsf{T}}\mathbf{y}$.
- $\mathbf{Q}_1\mathbf{Q}_1^{\mathsf{T}}$ is the projection onto range($\mathbf{A}$).
- $\mathbf{Q}_2\mathbf{Q}_2^{\mathsf{T}}$ is the projection onto range$(\mathbf{A})^{\perp}$ (orthogonal complement).

Growing Sets of Regressors

Consider the family of least-squares problems

$$\min_{\mathbf{x}} \left\| \sum_{i=1}^{p} x_i \mathbf{a}_i - \mathbf{y} \right\|_2 \quad \text{for } p = 1, \ldots, n,$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_p$ are called regressors (a.k.a. features/predictors/covariates).

- Let $\mathbf{A}_p = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{bmatrix} \in \mathbb{R}^{m \times p}$.
- Approximate $\mathbf{y}$ by a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_p$.
- Regress $\mathbf{y}$ on $\mathbf{a}_1, \ldots, \mathbf{a}_p$: Project $\mathbf{y}$ onto $\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_p\}$.
- As p increases, the fit generally improves and the optimal residual decreases.
- Let $\mathbf{A}_p = \mathbf{Q}_p \mathbf{R}_p$ denote its QR factorization. For each $p \le n$, the LS solution is

$$\mathbf{x}_{\text{ls}}^{(p)} = (\mathbf{A}_p^\top \mathbf{A}_p)^{-1} \mathbf{A}_p^\top \mathbf{y} = \mathbf{R}_p^{-1} \mathbf{Q}_p^\top \mathbf{y}.$$

Example of Completing a Basis in $\mathbb{R}^3$

Given $\mathbf{Q}_1$ with two orthonormal vectors

$$\mathbf{q}_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{bmatrix}, \qquad \mathbf{q}_2 = \begin{bmatrix} \frac{\sqrt{6}}{6} \\ -\frac{\sqrt{6}}{6} \\ \frac{\sqrt{6}}{3} \end{bmatrix},$$

we complete a basis for $\mathbb{R}^3$ by taking their *cross product*:

$$\mathbf{q}_3 = \mathbf{q}_1 \times \mathbf{q}_2 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} \end{bmatrix}.$$

Then, $\mathbf{Q} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix}$ is orthonormal, and the single column $\mathbf{Q}_2 = \mathbf{q}_3$ completes the basis of $\mathbb{R}^3$.

Cross Product in $\mathbb{R}^3$

Given two vectors

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

their cross product is defined as

$$\mathbf{a} \times \mathbf{b} = \begin{bmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{bmatrix}.$$

Mnemonic: Use the 3×3 determinant

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \mathbf{e}_1 \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - \mathbf{e}_2 \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \mathbf{e}_3 \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}.$$

Treat the first row as formal symbols and expand along the first row.

Completing a Basis in $\mathbb{R}^n$

Given an orthonormal set $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k \in \mathbb{R}^n, \quad k < n$, we complete it to an orthonormal basis of $\mathbb{R}^n$ by finding additional vectors in the orthogonal complement

$$\text{range}(\mathbf{Q}_1)^{\perp} = \{\, \mathbf{v} \in \mathbb{R}^n : \mathbf{q}_i^\mathsf{T} \mathbf{v} = 0 \text{ for all } i \,\}.$$

Procedure (Gram–Schmidt extension):

1. Choose any $\mathbf{v} \notin \text{range}(\mathbf{Q}_1)$.
2. Remove components along existing vectors:

$$\tilde{\mathbf{v}} = \mathbf{v} - \sum_{i=1}^{k} (\mathbf{q}_i^\mathsf{T} \mathbf{v})\, \mathbf{q}_i.$$

3. Normalize:

$$\mathbf{q}_{k+1} = \frac{\tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|_2}.$$

Repeating this until n vectors are obtained yields an orthonormal basis.