# ECE240 Introduction to Linear Dynamical Systems

## Lecture 8: Least-squares: Part II

Dr. Yu Zhang
ECE Department, UC Santa Cruz

Fall 2025

# Outline

# Least-norm Solution: Overdetermined Case

**Least squares problem:**

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2.$$

## Case 1: Overdetermined, full column rank ($m > n$, $\text{rank}(\mathbf{A}) = n$)

- Unique minimizer

$$\mathbf{x}_{\text{ls}} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y}.$$

- No ambiguity: there is only one least squares solution.

# Least-norm Solution: Underdetermined Case

## Case 2: Underdetermined, full row rank ($m < n$, $\text{rank}(\mathbf{A}) = m$)

- Infinitely many solutions of $\mathbf{Ax} = \mathbf{y}$ when $\mathbf{y} \in \text{range}(\mathbf{A})$.
- General solution: $\mathbf{x} = \mathbf{x}_p + \mathbf{z}$, $\mathbf{z} \in \text{null}(\mathbf{A})$
- We pick the *least-norm* one:

$$\mathbf{x}_{\text{ln}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_2 \qquad \text{s.t. } \mathbf{Ax} = \mathbf{y}.$$

- This choice corresponds to the solution with *minimum energy, minimum control effort, minimum parameter magnitude, or minimum complexity,* depending on the application.
- It is far from a theoretical curiosity: it appears often in engineering and data science.

# Least-norm Solution: Underdetermined Case

## Case 2: Least-norm solution

If $\mathbf{A}$ has full row rank,

$$\mathbf{x}_{\text{ln}} = \mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{y} = \mathbf{A}^{\dagger}\mathbf{y}.$$

This is the unique solution with smallest Euclidean norm among all $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x} = \mathbf{y}$.

Geometrically, it is the perpendicular projection of the origin onto the affine solution set.

# Simple Orthogonality Proof

Let $\mathbf{x}^\dagger := \mathbf{A}^\dagger \mathbf{y}$. Since $\mathbf{A}^\dagger$ is the *right inverse* of $\mathbf{A}$, we have

**Key facts:**

$$\mathbf{A}\mathbf{x}^\dagger = \mathbf{y}, \qquad \mathbf{x}^\dagger \in \text{range}(\mathbf{A}^\mathsf{T}) = \text{null}(\mathbf{A})^\perp.$$

Thus any solution is

$$\mathbf{x} = \mathbf{x}^\dagger + \mathbf{z}, \qquad \mathbf{z} \in \text{null}(\mathbf{A}).$$

Because $\mathbf{x}^\dagger \perp \mathbf{z}$,

$$\|\mathbf{x}\|_2^2 = \|\mathbf{x}^\dagger\|_2^2 + \|\mathbf{z}\|_2^2,$$

minimized when $\mathbf{z} = 0$. Thus,

$$\boxed{\mathbf{x}_{\text{ln}} = \mathbf{x}^\dagger = \mathbf{A}^\dagger \mathbf{y}}$$

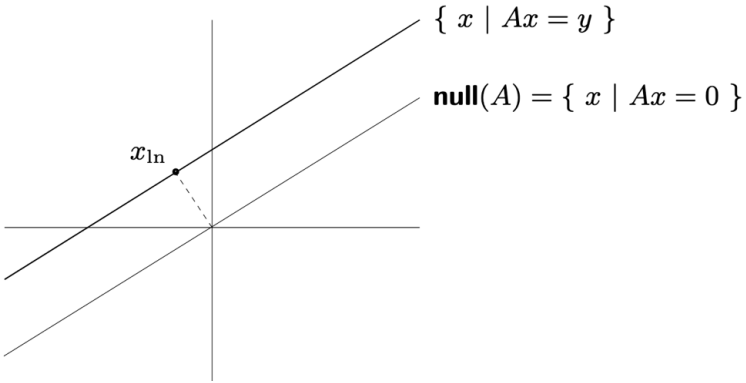
# Geometric Interpretation (1/2)

- The solution set is an affine subspace $\mathcal{S} = \mathbf{x}_0 + \text{null}(\mathbf{A})$, where $\text{null}(\mathbf{A})$ gives all directions inside this affine subspace.
- Least-norm solution means we want the point in $\mathcal{S}$ *closest to the origin*.
- *Key geometric fact:* The closest point to the origin in an affine space is the one whose vector is orthogonal to all directions in that space. Thus, the least–norm solution must satisfy

$$\mathbf{x}_{\text{ln}} \perp \text{null}(\mathbf{A}) \implies \mathbf{x}_{\text{ln}} \in \text{range}(\mathbf{A}^{\mathsf{T}}).$$

- Since $\mathbf{A}^{\dagger}\mathbf{y} = \mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{y} \in \text{range}(\mathbf{A}^{\mathsf{T}})$, it is indeed the least-norm solution $\mathbf{x}_{\text{ln}}$.

# Geometric Interpretation (2/2)




- *orthogonality condition:* $\mathbf{x}_{\mathrm{ln}} \perp \mathrm{null}(\mathbf{A})$
- *projection interpretation:* $\mathbf{x}_{\mathrm{ln}}$ is projection of $\mathbf{0}$ onto the solution set $\{ \mathbf{x} \mid \mathbf{Ax} = \mathbf{y} \}$

# Least-norm Solution via QR Factorization

Consider an *underdetermined system* $\mathbf{A}\mathbf{x} = \mathbf{y}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$ being full row rank ($m < n$).

- Compute QR factorization of $\mathbf{A}^{\mathsf{T}}$:

$$\mathbf{A}^{\mathsf{T}} = \mathbf{Q}\mathbf{R},$$

with

$$\mathbf{Q} \in \mathbb{R}^{n \times m}, \quad \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}_m, \quad \mathbf{R} \in \mathbb{R}^{m \times m} \text{ upper triangular, nonsingular.}$$

- Least-norm solution:

$$\mathbf{x}_{\mathsf{ln}} = \mathbf{A}^{\dagger}\mathbf{y} = \mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{A}^{\mathsf{T}})^{-1}\mathbf{y} = \mathbf{Q}\mathbf{R}^{-\mathsf{T}}\mathbf{y}.$$

- Its norm is

$$\|\mathbf{x}_{\mathsf{ln}}\|_2 = \|\mathbf{R}^{-\mathsf{T}}\mathbf{y}\|_2.$$

# Least-norm Solution: General Case

## Case 3: Rank deficient or general case

Even when $\mathbf{A}$ is not full column rank, there can be infinitely many least-squares minimizers:

$$\mathbf{x}_{\text{ls}} = \mathbf{x}_0 + \mathbf{z},\ \mathbf{z} \in \text{null}(\mathbf{A}).$$

The *least-norm LS solution* is still: $\boxed{\mathbf{x}_{\text{ln}} = \mathbf{A}^{\dagger}\mathbf{y}}$.

**Summary:**

- *Overdetermined, full column rank:* unique least squares solution, no need for least-norm selection.
- *Underdetermined or rank deficient:* infinitely many solutions, the pseudoinverse picks the minimum norm one.

# Least–norm LS solution: why $\mathbf{A}^\dagger\mathbf{y}$? (1/2)

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(\mathbf{A}) = r$. Use the full SVD:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}},$$

with

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r),\ \sigma_i > 0.$$

The pseudoinverse is

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^{\mathsf{T}}, \qquad \mathbf{\Sigma}^\dagger = \begin{bmatrix} \mathbf{\Sigma}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

# Least–norm LS solution: why $\mathbf{A}^\dagger\mathbf{y}$? (2/2)

Let $\mathbf{x} = \mathbf{V}\mathbf{z}$ and $\mathbf{y} = \mathbf{U}\mathbf{c}$, and write $\mathbf{z} = \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}$, $\mathbf{c} = \begin{bmatrix}\mathbf{c}_1\\ \mathbf{c}_2\end{bmatrix}$ with $\mathbf{z}_1, \mathbf{c}_1 \in \mathbb{R}^r$. Then

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{\Sigma}\mathbf{z} - \mathbf{c}\|_2^2 = \|\mathbf{\Sigma}_r\mathbf{z}_1 - \mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2.$$

Minimizing it gives $\mathbf{z}_1 = \mathbf{\Sigma}_r^{-1}\mathbf{c}_1$, while $\mathbf{z}_2$ is arbitrary. We have

$$\|\mathbf{x}\|_2 = \|\mathbf{z}\|_2 = \|\mathbf{\Sigma}_r^{-1}\mathbf{c}_1\|_2^2 + \|\mathbf{z}_2\|_2^2,$$

so the least-norm of $\mathbf{x}$ occurs at $\mathbf{z}_2 = \mathbf{0}$. Thus,

$$\mathbf{z}_{\text{ln}} = \mathbf{\Sigma}^\dagger\mathbf{c}, \qquad \mathbf{x}_{\text{ln}} = \mathbf{V}\mathbf{z}_{\text{ln}} = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{y} = \mathbf{A}^\dagger\mathbf{y}.$$

**Conclusion:** $\mathbf{A}^\dagger\mathbf{y}$ is the unique least–norm LS solution.

# Least Squares Estimation (LSE) Model

Many inversion, estimation, and reconstruction problems have the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v},$$

where

- $\mathbf{x}$ is the unknown signal or parameter vector that we wish to estimate.
- $\mathbf{y}$ is the measurement vector.
- $\mathbf{v}$ is an unknown noise or measurement error, typically small.
- The $i$th row of $\mathbf{A}$ characterizes the $i$th sensor.

# Least Squares Estimator (LSE)

Least squares estimator:

- Choose

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2,$$

  i.e., we fit the noiseless model $\mathbf{A}\mathbf{x}$ to the observed data $\mathbf{y}$.

- The solution is

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

# Linear Estimators and Unbiasedness

Suppose $\mathbf{A}$ is full rank and skinny, and we observe $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$.

Consider a linear estimator of the form $\hat{\mathbf{x}} = \mathbf{C}\mathbf{y}$.

- The estimator is unbiased if $\hat{\mathbf{x}} = \mathbf{x}$ whenever $\mathbf{v} = \mathbf{0}$.
- This holds if and only if $\mathbf{C}$ is a left inverse of $\mathbf{A}$; i.e.,

$$\mathbf{C}\mathbf{A} = \mathbf{I}_n$$

- For an unbiased linear estimator, the estimation error is

$$\hat{\mathbf{x}} - \mathbf{x} = \mathbf{C}(\mathbf{A}\mathbf{x} + \mathbf{v}) - \mathbf{x} = \mathbf{C}\mathbf{v}.$$

We would like $\mathbf{C}$ to be small while satisfying the constraint $\mathbf{C}\mathbf{A} = \mathbf{I}_n$.

# BLUE Property in Noiseless Setting

## Theorem (Minimum-Frobenius-Norm Left Inverse)

*Assume the noiseless linear model* $\mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathrm{rank}(\mathbf{A}) = n$.

*Consider linear estimators* $\hat{\mathbf{x}} = \mathbf{C}\mathbf{y}$.

- *(***Unbiasedness***) Requiring* $\hat{\mathbf{x}} = \mathbf{x}$ *for all* $\mathbf{x}$ *is equivalent to* $\mathbf{C}\mathbf{A} = \mathbf{I}_n$.
- *(***Best in Frobenius norm***) Among all matrices* $\mathbf{C} \in \mathbb{R}^{n \times m}$ *satisfying* $\mathbf{C}\mathbf{A} = \mathbf{I}_n$, *the matrix*

$$\mathbf{C}_{\mathrm{LS}} := \mathbf{A}^{\dagger} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}$$

*uniquely minimizes the Frobenius norm:*

$$\|\mathbf{C}\|_F \;\geq\; \|\mathbf{C}_{\mathrm{LS}}\|_F, \qquad \text{for all } \mathbf{C} \text{ with } \mathbf{C}\mathbf{A} = \mathbf{I}_n,$$

*with equality if and only if* $\mathbf{C} = \mathbf{C}_{\mathrm{LS}}$.

# Proof of BLUE Property

**Proof.**

Let $\mathbf{C}$ be any matrix with $\mathbf{CA} = \mathbf{I}_n$. Write $\mathbf{C} = \mathbf{C}_{\mathrm{LS}} + \mathbf{D}$, for some $\mathbf{D} \in \mathbb{R}^{n \times m}$. Then

$$\mathbf{CA} = (\mathbf{C}_{\mathrm{LS}} + \mathbf{D})\mathbf{A} = \mathbf{C}_{\mathrm{LS}}\mathbf{A} + \mathbf{DA} = \mathbf{I}_n + \mathbf{DA} \implies$$

$$\boxed{\mathbf{DA} = \mathbf{0}, \quad \mathbf{D}\mathbf{C}_{\mathrm{LS}}^{\mathsf{T}} = \mathbf{C}_{\mathrm{LS}}\mathbf{D}^{\mathsf{T}} = \mathbf{0}}$$

Next, $\|\mathbf{C}\|_F^2 = \|\mathbf{C}_{\mathrm{LS}} + \mathbf{D}\|_F^2 = \|\mathbf{C}_{\mathrm{LS}}\|_F^2 + \|\mathbf{D}\|_F^2 + 2\langle \mathbf{C}_{\mathrm{LS}}, \mathbf{D} \rangle_F$, where $\langle \mathbf{C}_{\mathrm{LS}}, \mathbf{D} \rangle_F = \mathrm{trace}(\mathbf{D}^{\mathsf{T}}\mathbf{C}_{\mathrm{LS}}) = \mathrm{trace}(\mathbf{C}_{\mathrm{LS}}\mathbf{D}^{\mathsf{T}}) = 0$.

Therefore

$$\|\mathbf{C}\|_F^2 = \|\mathbf{C}_{\mathrm{LS}}\|_F^2 + \|\mathbf{D}\|_F^2 \ \geq \ \|\mathbf{C}_{\mathrm{LS}}\|_F^2,$$

with equality if and only if $\mathbf{D} = \mathbf{0}$, that is, $\mathbf{C} = \mathbf{C}_{\mathrm{LS}}$. This proves the claim. □

# BLUE Property in Noisy Setting

We consider the linear model with additive white Gaussian noise

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v},$$

where $\mathbb{E}[\mathbf{v}] = \mathbf{0}$, $\text{cov}(\mathbf{v}) = \sigma^2\mathbf{I}$, and $\text{rank}(\mathbf{A}) = n$.

A *linear estimator* has the form $\hat{\mathbf{x}} = \mathbf{C}\mathbf{y}$.

- Unbiasedness requires that $\boxed{\mathbf{x} = \mathbb{E}[\hat{\mathbf{x}}]}$ for all $\mathbf{x}$
- $\mathbb{E}[\hat{\mathbf{x}}] = \mathbf{C}\,\mathbb{E}[\mathbf{y}] = \mathbf{C}\mathbf{A}\mathbf{x} \implies \mathbf{C}\mathbf{A} = \mathbf{I}_n$.

**Set of linear unbiased estimators**

$$\mathcal{U} := \{\mathbf{C} \in \mathbb{R}^{n\times m} : \mathbf{C}\mathbf{A} = \mathbf{I}_n\}.$$

# Least-Squares Estimator and Its Covariance

Properties:

- $\mathbf{C}_{\text{LS}}\mathbf{A} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}_n$, so $\hat{\mathbf{x}}_{\text{LS}}$ is unbiased.
- Since $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, we have $\text{cov}(\hat{\mathbf{x}}_{\text{LS}}) = \sigma^2\,\mathbf{C}_{\text{LS}}\mathbf{C}_{\text{LS}}^{\mathsf{T}}$.

## Gauss–Markov Theorem (BLUE)

Among all linear unbiased estimators $\hat{\mathbf{x}} = \mathbf{C}\mathbf{y}$ with $\mathbf{C}\mathbf{A} = \mathbf{I}_n$, the least-squares estimator

$$\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{A}^{\mathsf{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y}$$

has the minimum covariance matrix. It is the *Best Linear Unbiased Estimator (BLUE)*.

# Proof of BLUE: Covariance Comparison

For $\hat{\mathbf{x}} = \mathbf{C}\mathbf{y} = (\mathbf{C}_{\text{LS}} + \mathbf{D})\mathbf{y}$, we have

$$\begin{aligned}
\text{cov}(\hat{\mathbf{x}}) &= \sigma^2(\mathbf{C}_{\text{LS}} + \mathbf{D})(\mathbf{C}_{\text{LS}} + \mathbf{D})^{\mathsf{T}} \\
&= \sigma^2 \left( \mathbf{C}_{\text{LS}}\mathbf{C}_{\text{LS}}^{\mathsf{T}} + \mathbf{D}\mathbf{D}^{\mathsf{T}} + \mathbf{C}_{\text{LS}}\mathbf{D}^{\mathsf{T}} + \mathbf{D}\mathbf{C}_{\text{LS}}^{\mathsf{T}} \right) \\
&= \sigma^2 \left( \mathbf{C}_{\text{LS}}\mathbf{C}_{\text{LS}}^{\mathsf{T}} + \mathbf{D}\mathbf{D}^{\mathsf{T}} \right) \\
&= \text{cov}(\hat{\mathbf{x}}_{\text{LS}}) + \sigma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}}.
\end{aligned}$$

$$\implies \text{cov}(\hat{\mathbf{x}}) - \text{cov}(\hat{\mathbf{x}}_{\text{LS}}) = \sigma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}} \succeq \mathbf{0}.$$

# Additive White Gaussian Noise (AWGN) Model

- Assume the linear model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.
- For fixed $\mathbf{x}$, measurement vector $\mathbf{y}$ is simply a shifted Gaussian vector; i.e.,

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x},\, \sigma^2\mathbf{I}).$$

- Then, the conditional density of $\mathbf{y}$ given $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) &= \frac{1}{(2\pi)^{m/2}\det(\sigma^2\mathbf{I})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x})^{\mathsf{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{A}\mathbf{x})\right) \\ &= \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2\right) \end{aligned}$$

# Intuition of Likelihood

- For each candidate $\mathbf{x}$, the likelihood assigns a score $L(\mathbf{x}) := p(\mathbf{y} \mid \mathbf{x})$. The maximum likelihood estimator (MLE) picks the $\mathbf{x}$ with the highest score.
- The likelihood $p(\mathbf{y} \mid \mathbf{x})$ measures *how well a guess of* $\mathbf{x}$ *explains the observed data* $\mathbf{y}$.

- The formula is the same as a conditional probability, but the *interpretation* is different:
  - As a probability, $p(\mathbf{y} \mid \mathbf{x})$ is a function of $\mathbf{y}$ with $\mathbf{x}$ fixed.
  - As a likelihood, $L(\mathbf{x}) = p(\mathbf{y} \mid \mathbf{x})$ is viewed as a function of $\mathbf{x}$ with $\mathbf{y}$ fixed.
- Probability asks: "Given the parameter, how probable is this data?"
  Likelihood asks: "Given the data, how plausible is each parameter value?"

# Maximum Likelihood Estimator (MLE)

- For the AWGN model, we have

$$\log p(\mathbf{y} \mid \mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \text{const.}$$

- Maximizing the likelihood is equivalent to minimizing $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2$.
  Thus, the maximum likelihood estimator is

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \arg\max_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x}) = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2.$$

- Therefore

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \hat{\mathbf{x}}_{\mathrm{ls}} = \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y}.$$

  Least squares estimator equals the maximum likelihood estimator when the noise is Gaussian with covariance $\sigma^2\mathbf{I}$.

# Error Covariance and Mean Squared Error

Assume again $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and $\hat{\mathbf{x}}_{\text{ls}} = \mathbf{A}^\dagger\mathbf{y}$ with $\mathbf{A}^\dagger = \left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}$.

**Estimation error:**

$$\mathbf{r}_{\text{ls}} := \hat{\mathbf{x}}_{\text{ls}} - \mathbf{x} = \mathbf{A}^\dagger\mathbf{y} - \mathbf{x} = \mathbf{A}^\dagger(\mathbf{A}\mathbf{x} + \mathbf{v}) - \mathbf{x} = \mathbf{A}^\dagger\mathbf{v}.$$

**Error covariance (Inverse Fisher information matrix):**

$$\begin{aligned}\boldsymbol{\Sigma}_{\text{err}} = \mathbb{E}\left[\mathbf{r}_{\text{ls}}\mathbf{r}_{\text{ls}}^\mathsf{T}\right] &= \mathbf{A}^\dagger\mathbb{E}[\mathbf{v}\mathbf{v}^\mathsf{T}](\mathbf{A}^\dagger)^\mathsf{T} \\ &= \sigma^2\,\mathbf{A}^\dagger(\mathbf{A}^\dagger)^\mathsf{T} \\ &= \sigma^2\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}.\end{aligned}$$

**Mean squared error (MSE):**

$$\text{MSE}(\hat{\mathbf{x}}_{\text{ls}}) = \mathbb{E}\left[\|\mathbf{r}_{\text{ls}}\|_2^2\right] = \text{trace}\left(\boldsymbol{\Sigma}_{\text{err}}\right) = \sigma^2\,\text{trace}\left(\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\right).$$

# Connection to MMSE Under a Gaussian Prior (1/2)

Assume the AWGN model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$ with $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and impose a Gaussian prior $\mathbf{x} \sim \mathcal{N}(\mathbf{m}_x, \boldsymbol{\Sigma}_x)$.

**Posterior is Gaussian.** Because both the prior and likelihood are Gaussian, the posterior $p(\mathbf{x} \mid \mathbf{y})$ is also Gaussian, with

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right),$$

where

$$\boldsymbol{\Sigma}_{x|y} = \left(\mathbf{A}^\mathsf{T}\mathbf{A}/\sigma^2 + \boldsymbol{\Sigma}_x^{-1}\right)^{-1}, \quad \boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^\mathsf{T}\mathbf{y}/\sigma^2 + \boldsymbol{\Sigma}_x^{-1}\mathbf{m}_x\right).$$

**MMSE estimator.** The minimum mean squared error estimator is the posterior mean:

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathbb{E}[\mathbf{x} \mid \mathbf{y}] = \boldsymbol{\mu}_{x|y},$$

which is an affine (and often linear) function of $\mathbf{y}$.

# Connection to MMSE Under a Gaussian Prior (2/2)

**Connection to Least Squares.**

- If the prior becomes *noninformative* (very large $\boldsymbol{\Sigma}_x$), then $\boldsymbol{\Sigma}_x^{-1} \to \mathbf{0}$ and

$$\hat{\mathbf{x}}_{\text{MMSE}} \;\to\; \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y} = \hat{\mathbf{x}}_{\text{LS}} = \hat{\mathbf{x}}_{\text{ML}}.$$

  Thus, with no prior information, MMSE reduces to LSE and MLE.

- If $\mathbf{m}_x = \mathbf{0}$ and $\boldsymbol{\Sigma}_x$ is chosen appropriately, we recover the *Wiener filter*:

$$\hat{\mathbf{x}}_{\text{Wiener}} = \boldsymbol{\Sigma}_x\mathbf{A}^{\mathsf{T}}\left(\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^{\mathsf{T}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}.$$

**Takeaway.** Under Gaussian assumptions, least squares is closely linked to both maximum likelihood and MMSE estimation: MLE/LSE arises with no prior, while MMSE incorporates prior information through $\boldsymbol{\Sigma}_x$.