

Lecture 4: Fisher Information & Cramér–Rao

Score Function · Fisher Information · CR Bound · Efficiency · Admissibility

Previously, on Lecture 3...

Bias: $\mathbb{E}[\hat{\theta}] - \theta$. Does it aim at the right place?

Variance: $\text{Var}(\hat{\theta})$. How much does it jump around?

MSE = $\text{Bias}^2 + \text{Var}$. Total error. Sometimes biased beats unbiased!

Consistency: $\hat{\theta}_n \xrightarrow{P} \theta$. Converges to truth with enough data.

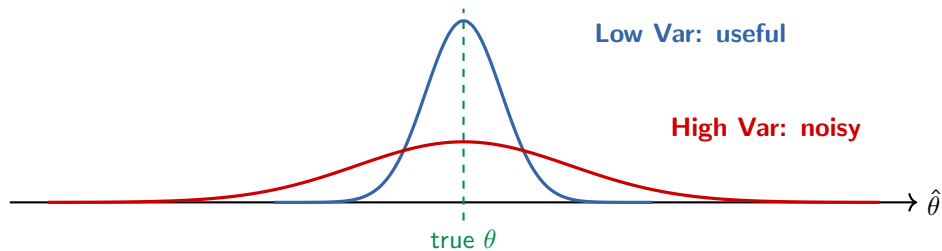
Sufficiency: $T(\mathbf{X})$ captures all info about θ . Rao–Blackwell improves estimators.

Today: Can we quantify the **best possible** precision?

Is there a fundamental **limit** on how good any estimator can be?

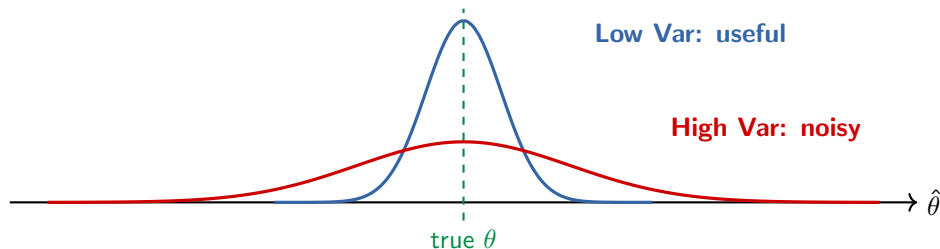
Why Does Lower Variance Matter?

From Lecture 3: an unbiased estimator **aims at the right place**. But if the variance is huge, individual estimates are all over the map.



Why Does Lower Variance Matter?

From Lecture 3: an unbiased estimator **aims at the right place**. But if the variance is huge, individual estimates are all over the map.



- ▶ Both estimators are **unbiased** — centered on the true θ
- ▶ But the **red one** often gives estimates **far from the truth**
- ▶ With **one** sample, you can't tell if you're close or not — lower variance = higher **confidence**

Among unbiased estimators, can we find the one with the smallest variance?

Can We Do Better? The Fundamental Question

For $X_i \sim N(\mu, \sigma^2)$, the sample mean \bar{X} estimates μ with $\text{Var}(\bar{X}) = \sigma^2/n$.

Can **any** unbiased estimator have **lower** variance?

Or is \bar{X} already the best we can do?

To answer this, we need to measure **how much information** one observation carries about θ .

Roadmap:

Why log? \rightarrow **Score function** (sensitivity of the model to θ) \rightarrow **Fisher information**
 \rightarrow **Cramér–Rao bound** (the variance floor)

From Data to Likelihood

Suppose we observe data X_1, X_2, \dots, X_n from some distribution $f(x \mid \theta)$.

Key assumption: observations are **i.i.d.** (independent and identically distributed).

From Data to Likelihood

Suppose we observe data X_1, X_2, \dots, X_n from some distribution $f(x | \theta)$.

Key assumption: observations are **i.i.d.** (independent and identically distributed).

Independence means the joint density **factors** into a product:

$$f(X_1, X_2, \dots, X_n | \theta) = f(X_1 | \theta) \cdot f(X_2 | \theta) \cdots f(X_n | \theta) = \prod_{i=1}^n f(X_i | \theta)$$

From Data to Likelihood

Suppose we observe data X_1, X_2, \dots, X_n from some distribution $f(x | \theta)$.

Key assumption: observations are **i.i.d.** (independent and identically distributed).

Independence means the joint density **factors** into a product:

$$f(X_1, X_2, \dots, X_n | \theta) = f(X_1 | \theta) \cdot f(X_2 | \theta) \cdots f(X_n | \theta) = \prod_{i=1}^n f(X_i | \theta)$$

We call this the **likelihood function** — the same product, viewed as a function of θ :

$$L(\theta) = \prod_{i=1}^n f(X_i | \theta)$$

Same formula, different perspective:

As a function of x : it's the joint density (probability of the data).

As a function of θ : it's the likelihood (how well θ explains the data).

But products of many small numbers are messy to work with...

Why the Logarithm? From Products to Sums

The likelihood is a product of n terms — and those terms can be tiny.

Taking the log turns this **product into a sum**:

$$L(\theta) = \prod_{i=1}^n f(X_i | \theta) \quad \xrightarrow{\log} \quad \ell(\theta) = \sum_{i=1}^n \log f(X_i | \theta)$$

Products are painful:

- ▶ Multiplying tiny numbers \rightarrow underflow
- ▶ Product rule for derivatives is messy
- ▶ Hard to work with analytically

Sums are friendly:

- ▶ Numerically stable
- ▶ Derivative of a sum = sum of derivatives
- ▶ LLN, CLT apply directly

Key fact: log is monotonically increasing, so $\arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$. Same maximizer!

The Score Function: How Sensitive Is the Model?

Given a model $f(x | \theta)$, the **score** measures how the log-probability changes with θ :

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X | \theta)$$

Concrete example: $X \sim \text{Bernoulli}(p)$.

$$\log f(x | p) = x \log p + (1-x) \log(1-p)$$

$$s(p) = \frac{\partial}{\partial p} [x \log p + (1-x) \log(1-p)] = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}$$

Key property: $\mathbb{E}[s(\theta)] = 0$ at the true θ .

The score points in the right direction on average, but **cancels out**.

What matters is how much it **varies** — that's Fisher information.

Why does the score average to zero? Not because of unbiasedness — it's a property of the model itself. Any density integrates to 1: $\int f(x | \theta) dx = 1$.

Differentiate both sides w.r.t. θ : $\int \frac{\partial f}{\partial \theta} dx = 0$.

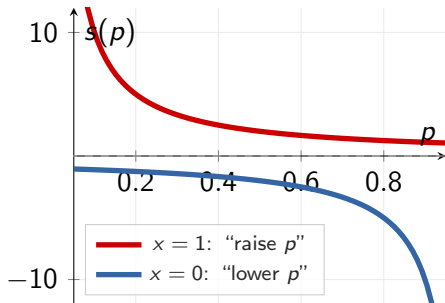
Since $\frac{\partial f}{\partial \theta} = f \frac{\partial \log f}{\partial \theta} = f s(\theta)$, this gives $\int f s(\theta) dx = \mathbb{E}[s(\theta)] = 0$. ✓

Reading the Score Function

For Bernoulli, $s(p) = \frac{x-p}{p(1-p)}$:

- ▶ When $x = 1$: $s(p) = \frac{1}{p}$
Score is positive: “ p should be **higher**”
- ▶ When $x = 0$: $s(p) = \frac{-1}{1-p}$
Score is negative: “ p should be **lower**”
- ▶ Near $p = 0$ or $p = 1$: score is **huge**
The data is very “surprising” → strong signal
- ▶ Near $p = 0.5$: score is **moderate**
Neither outcome is very surprising

The score is like a **compass needle**: it always points toward the true p , but swings more when the data is surprising.



Fisher Information: How Informative Is One Observation?

The score averages to zero, but it **varies**. More variation means different values of θ produce **noticeably different** data — making θ easier to pinpoint:

$$I(\theta) = \text{Var}[s(\theta)] = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f(X | \theta) \right)^2 \right]$$

Fisher Information: How Informative Is One Observation?

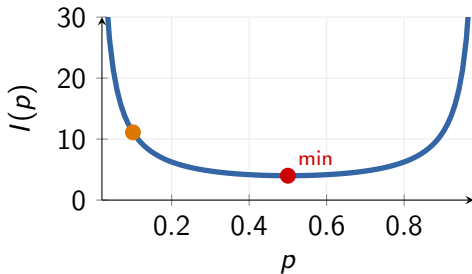
The score averages to zero, but it **varies**. More variation means different values of θ produce **noticeably different** data — making θ easier to pinpoint:

$$I(\theta) = \text{Var}[s(\theta)] = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f(X | \theta) \right)^2 \right]$$

Bernoulli derivation:

We found $s(p) = \frac{X-p}{p(1-p)}$. Since $\mathbb{E}[s] = 0$:

$$\begin{aligned} I(p) &= \mathbb{E}[s^2] = \mathbb{E} \left[\frac{(X-p)^2}{p^2(1-p)^2} \right] \\ &= \frac{\text{Var}(X)}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \boxed{\frac{1}{p(1-p)}} \end{aligned}$$

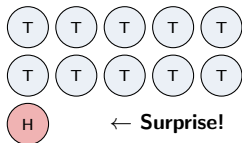


p near 0 or 1: very informative. $p = 0.5$: max noise, min info.

Fisher Information: The Coin Flip Intuition

Why is $I(p) = \frac{1}{p(1-p)}$ shaped like a U?

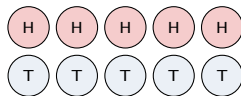
Biased coin ($p = 0.01$)



Almost every flip is Tails.
Seeing Heads is **very surprising** —
tells you a lot about p .

$I(0.01) \approx 100$ **high info**

Fair coin ($p = 0.5$)



H and T equally likely.
Neither outcome is surprising —
each flip tells you **very little**.

$I(0.5) = 4$ **low info**

Key insight: Fisher information measures how **surprised** you are by the data.

More surprise = more information = easier to pinpoint θ .

Fisher Information: Two Equivalent Forms

Under regularity conditions, there is an equivalent formula that's often easier to compute:

$$I(\theta) = \underbrace{\mathbb{E}[s(\theta)^2]}_{\text{variance of the score}} = -\underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X | \theta)\right]}_{\text{expected curvature of } \ell}$$

Fisher Information: Two Equivalent Forms

Under regularity conditions, there is an equivalent formula that's often easier to compute:

$$I(\theta) = \underbrace{\mathbb{E}[s(\theta)^2]}_{\text{variance of the score}} = -\underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X | \theta)\right]}_{\text{expected curvature of } \ell}$$

Intuition: Why should score variance = curvature?

- ▶ The score $s = \ell'$ is the **slope** of the log-likelihood
- ▶ At the true θ , the slope averages to 0: $\mathbb{E}[\ell'] = 0$
- ▶ A **sharply curved** ℓ (large $|\ell''|$) \rightarrow slope swings far from 0 \rightarrow high $\text{Var}(s)$
- ▶ A **flat** ℓ (small $|\ell''|$) \rightarrow slope barely moves \rightarrow low $\text{Var}(s)$

Fisher Information: Two Equivalent Forms

Under regularity conditions, there is an equivalent formula that's often easier to compute:

$$I(\theta) = \underbrace{\mathbb{E}[s(\theta)^2]}_{\text{variance of the score}} = -\underbrace{\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X | \theta)\right]}_{\text{expected curvature of } \ell}$$

Intuition: Why should score variance = curvature?

- ▶ The score $s = \ell'$ is the **slope** of the log-likelihood
- ▶ At the true θ , the slope averages to 0: $\mathbb{E}[\ell'] = 0$
- ▶ A **sharply curved** ℓ (large $|\ell''|$) \rightarrow slope swings far from 0 \rightarrow high $\text{Var}(s)$
- ▶ A **flat** ℓ (small $|\ell''|$) \rightarrow slope barely moves \rightarrow low $\text{Var}(s)$

Formally: Differentiate $\mathbb{E}[s] = 0$ w.r.t. θ . Using the product rule under \int (note: $\frac{\partial f}{\partial \theta} = s \cdot f$):

$$0 = \int \underbrace{\frac{\partial s}{\partial \theta}}_{\ell''} f \, dx + \int s \cdot \underbrace{\frac{\partial f}{\partial \theta}}_{s \cdot f} \, dx = \mathbb{E}[\ell''] + \mathbb{E}[s^2] \quad \implies \quad \mathbb{E}[s^2] = -\mathbb{E}[\ell''] \quad \checkmark$$

Verifying the Two Forms: Bernoulli

Let's check both formulas give the same answer for $X \sim \text{Bernoulli}(p)$.

Form 1: Variance of score

$$s(p) = \frac{X-p}{p(1-p)}, \quad \mathbb{E}[s] = 0$$

$$I(p) = \mathbb{E}[s^2] = \frac{\text{Var}(X)}{p^2(1-p)^2}$$

$$= \frac{p(1-p)}{p^2(1-p)^2} = \boxed{\frac{1}{p(1-p)}}$$

Form 2: Expected curvature

$$\ell(p) = x \log p + (1-x) \log(1-p)$$

$$\ell''(p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

$$-\mathbb{E}[\ell''] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \boxed{\frac{1}{p(1-p)}}$$

Both forms give $I(p) = \frac{1}{p(1-p)}$. ✓

In practice, Form 2 ($-\mathbb{E}[\ell'']$) is usually easier to compute.

Fisher Information: Beyond Bernoulli

Using the second-derivative form $I(\theta) = -\mathbb{E}[\ell'']$, we can compute Fisher information for any distribution:

Distribution	$\ell''(\theta)$	$I(\theta)$	Intuition
Bern(p)	$-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$	$\frac{1}{p(1-p)}$	Fair coin = hardest to pin down
$N(\mu, \sigma_0^2)$	$-\frac{1}{\sigma_0^2}$	$\frac{1}{\sigma_0^2}$	Low noise \rightarrow more info
Pois(λ)	$-x/\lambda^2$	$\frac{1}{\lambda}$	Rare events \rightarrow more info
Exp(λ)	$-1/\lambda^2$	$\frac{1}{\lambda^2}$	Fast decay \rightarrow more info

Fisher Information: Beyond Bernoulli

Using the second-derivative form $I(\theta) = -\mathbb{E}[\ell'']$, we can compute Fisher information for any distribution:

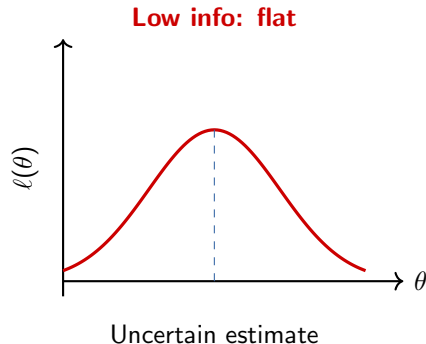
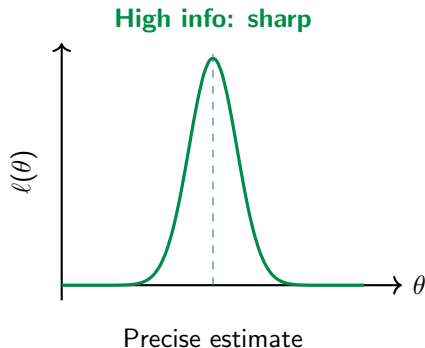
Distribution	$\ell''(\theta)$	$I(\theta)$	Intuition
Bern(p)	$-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$	$\frac{1}{p(1-p)}$	Fair coin = hardest to pin down
$N(\mu, \sigma_0^2)$	$-\frac{1}{\sigma_0^2}$	$\frac{1}{\sigma_0^2}$	Low noise \rightarrow more info
Pois(λ)	$-x/\lambda^2$	$\frac{1}{\lambda}$	Rare events \rightarrow more info
Exp(λ)	$-1/\lambda^2$	$\frac{1}{\lambda^2}$	Fast decay \rightarrow more info

For n i.i.d. observations: the score is a sum $s_n = \sum_{i=1}^n s_i$ of i.i.d. terms, so:

$$I_n(\theta) = \text{Var}(s_n) = n \text{Var}(s_1) = n \cdot I(\theta)$$

Fisher information is additive: $I_n(\theta) = n \cdot I(\theta)$.
More observations = proportionally more information.

Intuition: Sharp vs Flat Log-Likelihood



$I(\theta)$ measures the **curvature** of the log-likelihood at the true θ .

Sharp curve \Rightarrow high $I(\theta)$ \Rightarrow data is very informative \Rightarrow estimator is precise.

This connects the two forms: $I(\theta) = -\mathbb{E}[\ell'']$ is literally the expected curvature.

The Cramér–Rao Bound

We've quantified how much **information** each observation carries.

Now: is there a **fundamental limit** on how precise
any estimator can be?

Cramér–Rao Lower Bound

Now we can answer the fundamental question. For any **unbiased** estimator $\hat{\theta}$ based on n i.i.d. observations:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)}$$

Intuition: Why $\frac{1}{n \cdot I(\theta)}$?

- ▶ **More observations (n large)** \Rightarrow bound gets smaller \Rightarrow can estimate more precisely
- ▶ **More informative data ($I(\theta)$ large)** \Rightarrow bound gets smaller \Rightarrow each observation tells us more
- ▶ The bound is **tight** for many models — it's the actual achievable precision

Cramér–Rao Lower Bound

Now we can answer the fundamental question. For any **unbiased** estimator $\hat{\theta}$ based on n i.i.d. observations:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)}$$

Intuition: Why $\frac{1}{n \cdot I(\theta)}$?

- ▶ **More observations (n large)** \Rightarrow bound gets smaller \Rightarrow can estimate more precisely
- ▶ **More informative data ($I(\theta)$ large)** \Rightarrow bound gets smaller \Rightarrow each observation tells us more
- ▶ The bound is **tight** for many models — it's the actual achievable precision

Verify for Bernoulli:

$$I(p) = \frac{1}{p(1-p)} \quad \Rightarrow \quad \text{CR bound: } \text{Var}(\hat{p}) \geq \frac{1}{n \cdot \frac{1}{p(1-p)}} = \frac{p(1-p)}{n}$$

Actual variance of $\hat{p} = \bar{X}$: $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$ ✓ Hits the bound exactly!

Efficient Estimators

Definition: An unbiased estimator $\hat{\theta}$ is **efficient** if

$$\text{Var}(\hat{\theta}) = \frac{1}{n \cdot I(\theta)}$$

i.e., it **achieves the Cramér–Rao lower bound exactly**.

What does efficiency mean in practice?

- ▶ You **cannot do better** among unbiased estimators — it extracts all available information
- ▶ No data is “wasted” — every observation contributes maximally
- ▶ If your estimator is efficient, **stop looking** for a better unbiased one
- ▶ If it's **not** efficient, there might be room for improvement

Not all models have efficient estimators!

But when one exists, it's usually the MLE (for large n , the MLE is **asymptotically efficient**).

Cramér–Rao: Efficiency and Practical Use

What it says:

A **floor** on how precise any unbiased estimator can be

Efficient estimator:

Achieves the bound — the **best possible**

Practical use:

Tells you whether to keep searching for a better one

Model	Estimator	$\text{Var}(\hat{\theta})$	CR bound	Efficient?
$\text{Bern}(p)$	$\hat{p} = \bar{X}$	$\frac{p(1-p)}{n}$	$\frac{p(1-p)}{n}$	Yes
$N(\mu, \sigma_0^2)$	$\hat{\mu} = \bar{X}$	$\frac{\sigma_0^2}{n}$	$\frac{\sigma_0^2}{n}$	Yes
$\text{Exp}(\lambda)$	$\hat{\lambda} = 1/\bar{X}$	$\frac{\lambda^2}{n}$	$\frac{\lambda^2}{n}$	Yes

Regularity Conditions: When Does CR Apply?

The Cramér–Rao bound doesn't hold for every model. It requires these **regularity conditions**:

1. **Fixed support:** the set of x values where $f(x | \theta) > 0$ doesn't depend on θ
2. **Interior parameter:** θ is in the **interior** of the parameter space (not at a boundary)
3. **Differentiation under the integral:** we can swap $\frac{\partial}{\partial \theta}$ and \int
(this is how we proved $\mathbb{E}[s(\theta)] = 0$ and derived the two forms of $I(\theta)$)
4. **Finite information:** $0 < I(\theta) < \infty$

Good news: All **exponential family** distributions (Normal, Bernoulli, Poisson, Exponential, Gamma, ...) automatically satisfy these conditions.
The CR bound always applies to them.

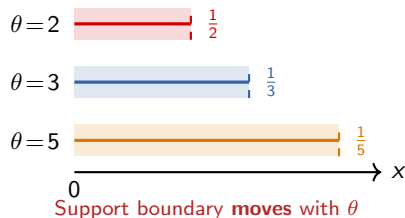
When CR Fails: The Uniform Distribution

Counterexample: $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$

- ▶ Support is $[0, \theta]$ — depends on θ !
(violates condition #1)
- ▶ The sufficient statistic is $X_{(n)} = \max_i X_i$
- ▶ Its variance: $\text{Var}(X_{(n)}) \sim \frac{1}{n^2}$

CR would predict a floor of $1/n$.
But $1/n^2$ is **much faster** — we beat
the “bound”!

The bound simply **doesn't apply**
here.



Lesson: Always check regularity conditions before applying CR.
When they fail, estimators can be *better* than the “bound” suggests.

Beyond Unbiasedness

The CR bound only applies to **unbiased** estimators.

What if we **allow bias** to reduce MSE?

We need new criteria to compare estimators. . .

Beyond Unbiasedness: What If We Allow Bias?

The Cramér–Rao bound tells us: among **unbiased** estimators, variance $\geq \frac{1}{nI(\theta)}$.

But from Lecture 3, we know biased estimators can have **lower MSE!**

If we drop the “unbiased” requirement,
how do we compare estimators?

We need a new criterion that works for **all** estimators — biased or not.

Two approaches:

Admissibility: Is there *any* estimator that beats yours everywhere?

Minimax: Which estimator has the best *worst-case* performance?

Admissibility

Definition: $\hat{\theta}_1$ is **inadmissible** if $\exists \hat{\theta}_2$ that **dominates** it:

$$\text{MSE}(\hat{\theta}_2, \theta) \leq \text{MSE}(\hat{\theta}_1, \theta) \quad \forall \theta, \quad \text{with strict inequality for some } \theta$$

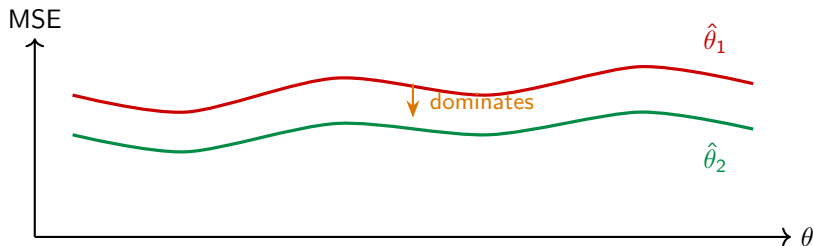
An estimator is **admissible** if no other estimator dominates it.

Admissibility

Definition: $\hat{\theta}_1$ is **inadmissible** if $\exists \hat{\theta}_2$ that **dominates** it:

$$\text{MSE}(\hat{\theta}_2, \theta) \leq \text{MSE}(\hat{\theta}_1, \theta) \quad \forall \theta, \quad \text{with strict inequality for some } \theta$$

An estimator is **admissible** if no other estimator dominates it.



$\hat{\theta}_1$ is **inadmissible** — $\hat{\theta}_2$ is at least as good everywhere, and strictly better somewhere.

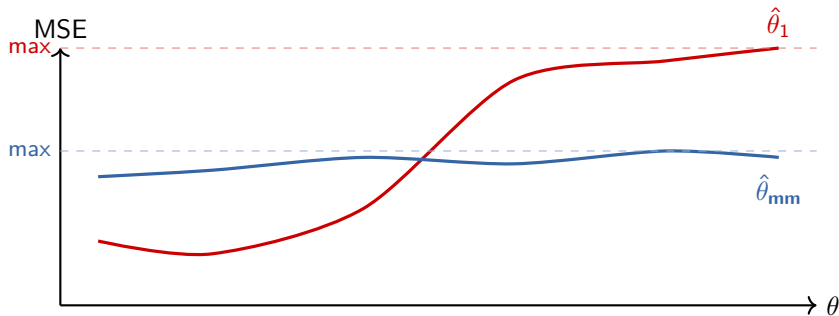
Familiar? This is exactly **Pareto dominance** from multi-criteria optimization!
 $\hat{\theta}_2$ Pareto-dominates $\hat{\theta}_1$: better on some criteria (values of θ), no worse on any.
Admissible estimators = the **Pareto front** of the MSE landscape.

Minimax Estimators

Analogy: You don't know tomorrow's weather (θ). A minimax thinker picks the option whose **worst outcome is least bad**.

A **minimax** estimator minimizes the **worst-case** risk:

$$\hat{\theta}_{\text{minimax}} = \arg \min_{\hat{\theta}} \max_{\theta} \text{MSE}(\hat{\theta}, \theta)$$



$\hat{\theta}_1$ can be great for some θ , but terrible for others. $\hat{\theta}_{\text{mm}}$ is never great, but **never terrible either**.

Two Approaches to Estimation

Plug-in (Unbiased)

Use sample statistic directly
 (\bar{X}, S^2, \hat{p})

Simple and intuitive
CR bound measures precision
Efficient = best possible

Minimax

Minimize worst-case risk
across all values of θ

May introduce bias
Conservative guarantee
No single θ can hurt you badly

Coming soon: Bayesian estimation adds a **prior** on θ , which acts as automatic **regularization** — pulling estimates toward a central value.

This is a principled way to trade bias for lower variance.

What We Haven't Covered (Yet)

Lectures 3–4 focused on **point estimation** — producing a single “best guess” for θ . But there's much more to statistical inference:

Point estimation: How to *construct* estimators — MoM, MLE (Lecture 5)

Bayesian estimation: Priors, posteriors, MAP, regularization (Lecture 6)

Computational methods: EM algorithm, MCMC for complex models (Lectures 7–8)

Confidence intervals: How uncertain is our estimate? (Lecture 9)

Bootstrap: Resampling to estimate uncertainty without formulas (Lecture 10)

Hypothesis testing: Is the effect real or just noise? (Lectures 11–12)

Our tools (bias, MSE, CR bound, sufficiency) will be the **foundation** for all of these.

Summary: How to Judge an Estimator

Bias: $\mathbb{E}[\hat{\theta}] - \theta$. Does it aim at the right place?

Variance: $\text{Var}(\hat{\theta})$. How much does it jump around?

MSE = $\text{Bias}^2 + \text{Var}$. Total error. Biased can beat unbiased!

Consistency: $\hat{\theta}_n \xrightarrow{P} \theta$. Converges to truth with enough data.

Sufficiency: $T(\mathbf{X})$ captures everything about θ . Compress without loss.

Cramér–Rao: $\text{Var} \geq 1/(n \cdot I(\theta))$. The efficiency floor.

Efficiency: Achieves the CR bound. Best possible among unbiased.

Admissibility / Minimax: Compare estimators even when biased.

Homework

1. Compute the Fisher information $I(\lambda)$ for $\text{Poisson}(\lambda)$.
Find the Cramér–Rao lower bound for estimating λ . Is $\hat{\lambda} = \bar{X}$ efficient?
2. For $X_1, \dots, X_n \sim \text{Exp}(\lambda)$: compute $I(\lambda)$ using both
the variance-of-score and second-derivative formulas. Verify they agree.
3. Three estimators T_1, T_2, T_3 have MSE curves as functions of $\theta \in [0, 1]$.
Sketch an example where T_1 and T_2 are admissible but T_3 is not.
Then sketch an example where T_1 is the minimax estimator.

Questions?