Lecture 4: Fisher Information & Cramér–Rao

Score Function · Fisher Information · CR Bound · Admissibility · Stein's Paradox

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.

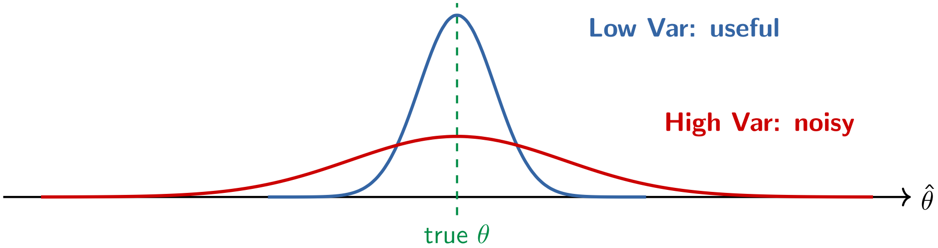
Sufficiency: $T(\mathbf{X})$ captures all info about θ. Rao–Blackwell improves estimators.

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

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

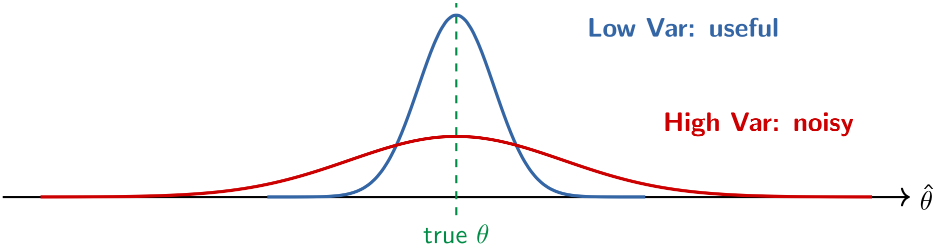
Why Does Lower Variance Matter?

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

We know $\mathrm{Var}(\bar{X}) = \sigma^2/n$ for estimating the mean.

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

From Data to Likelihood

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

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

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

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

From Data to Likelihood

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

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

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

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

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

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

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**:

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

Products are painful:

- Multiplying tiny numbers → 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 $f(x \mid \theta)$, the **score** measures how the log-probability changes with θ:

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

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

$\log f(x \mid p) = x \log p + (1{-}x) \log(1{-}p)$

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

The Score Function: How Sensitive Is the Model?

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

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

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

$\log f(x \mid p) = x \log p + (1{-}x) \log(1{-}p)$

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

- If we observe $x = 1$ and p is small, the score is **large positive** → "p should be higher"
- If we observe $x = 0$ and p is large, the score is **large negative** → "p should be lower"
- On average: $\mathbb{E}[s(p)] = 0$ — the score points in the right direction but **averages out**

Fisher Information: How Informative Is One Observation?

The score averages to zero, but it **varies**. More variation $=$ more information:

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

Fisher Information: How Informative Is One Observation?

The score averages to zero, but it **varies**. More variation = more information:

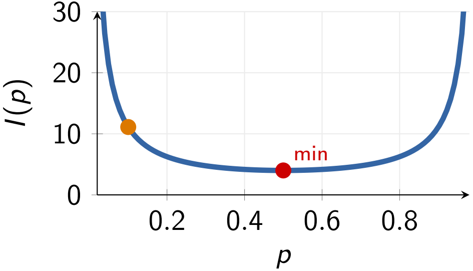
$$\boxed{I(\theta) = \mathrm{Var}[s(\theta)] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]}$$

Bernoulli derivation: We found $s(p) = \frac{X-p}{p(1-p)}$.

Since $\mathbb{E}[s] = 0$:

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

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

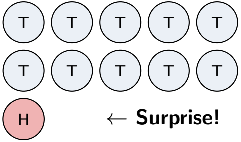


p near 0 or 1: very informative. $p = 0.5$: max noise, min info.

Fisher Information: The Coin Flip Intuition

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

Biased coin ($p = 0.01$)

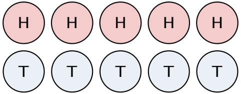


← **Surprise!**

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 ($p = 0.5$)



← Nothing surprising

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:

$$\boxed{I(\theta) = \mathbb{E}\left[s(\theta)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]}$$

Why are these the same? Start from $\mathbb{E}[s(\theta)] = 0$ and differentiate both sides w.r.t. θ:

$$0 = \frac{\partial}{\partial\theta}\mathbb{E}[s] = \mathbb{E}\left[\frac{\partial s}{\partial\theta}\right] + \mathbb{E}[s \cdot s] = \mathbb{E}[\ell''] + \mathbb{E}[s^2]$$

So: $\mathbb{E}[s^2] = -\mathbb{E}[\ell'']$. ✓

Verify for Bernoulli: $\ell(p) = x\log p + (1-x)\log(1-p)$

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

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:

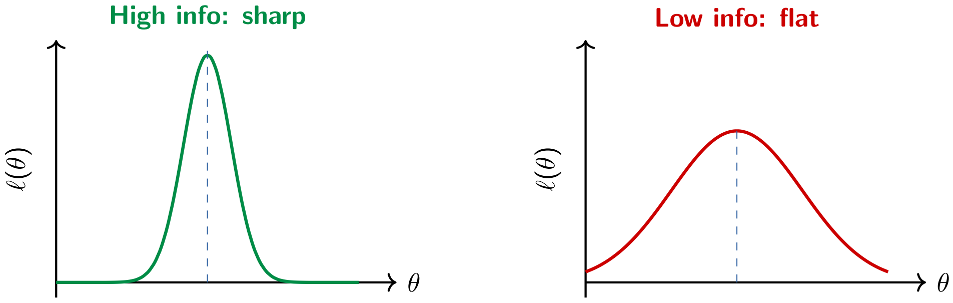
Distribution	$\ell''(\theta)$	$I(\theta)$	**Intuition**
Bern(p)	$-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$	$\frac{1}{p(1-p)}$	Fair coin = hardest to pin down
$N(\mu, \sigma_0^2)$	$-\frac{1}{\sigma_0^2}$	$\frac{1}{\sigma_0^2}$	Low noise → more info
Pois(λ)	$-x/\lambda^2$	$\frac{1}{\lambda}$	Rare events →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

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

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

Intuition: Sharp vs Flat Log-Likelihood



High info: sharp

Precise estimate

Low info: flat

Uncertain estimate

$I(\theta)$ 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.

Cramér–Rao Lower Bound

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

$$\boxed{\mathrm{Var}(\hat{\theta}) \;\geq\; \frac{1}{n \cdot I(\theta)}}$$

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:

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

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

- **More observations (n large)** ⇒ bound gets smaller ⇒ can estimate more precisely
- **More informative data ($I(\theta)$ large)** ⇒ bound gets smaller ⇒ 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!

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**	$\mathrm{Var}(\hat{\theta})$	**CR bound**	**Efficient?**
$\mathrm{Bern}(p)$	$\hat{p} = \bar{X}$	$\frac{p(1-p)}{n}$	$\frac{p(1-p)}{n}$	Yes
$N(\mu, \sigma_0^2)$	$\hat{\mu} = \bar{X}$	$\frac{\sigma_0^2}{n}$	$\frac{\sigma_0^2}{n}$	Yes
$\mathrm{Exp}(\lambda)$	$\hat{\lambda} = 1/\bar{X}$	$\frac{\lambda^2}{n}$	$\frac{\lambda^2}{n}$	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 $f(x \mid \theta) > 0$ 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.

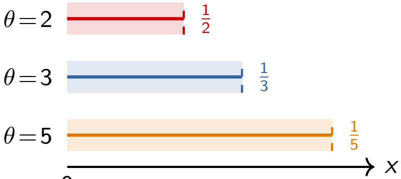
When CR Fails: The Uniform Distribution

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

- 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.



Support boundary **moves** with θ

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

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:

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

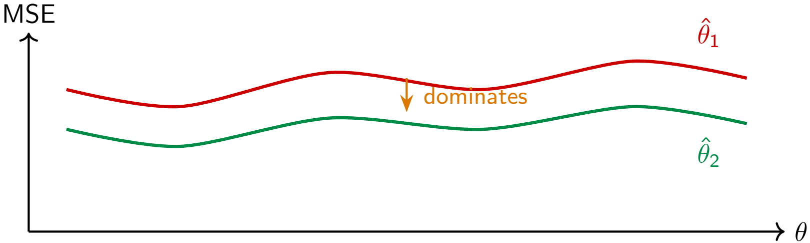
An estimator is **admissible** if no other estimator dominates it.

Admissibility

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

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

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

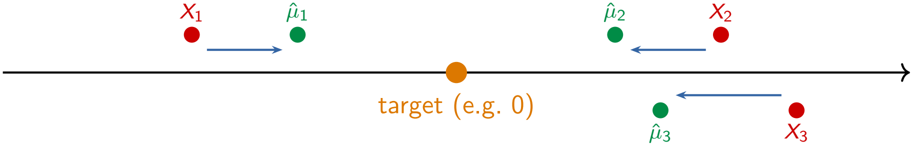


$\hat{\theta}_1$ is **inadmissible** — $\hat{\theta}_2$ is at least as good everywhere, and strictly better somewhere.

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.

What Is Shrinkage?

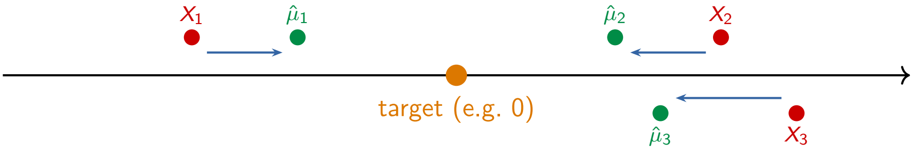
Idea: Instead of using the raw estimate, **pull it toward a fixed target** (often 0 or the grand mean).



The shrinkage estimator has the form: $\hat{\mu}_i^{\text{shrunk}} = (1-c)\cdot X_i + c \cdot \text{target}, \quad 0 < c < 1$

What Is Shrinkage?

Idea: Instead of using the raw estimate, **pull it toward a fixed target** (often 0 or the grand mean).



The shrinkage estimator has the form: $\hat{\mu}_i^{\text{shrunk}} = (1 - c) \cdot X_i + c \cdot \text{target}, \quad 0 < c < 1$

Why does this help?

- Raw estimates X_i are **noisy** — they overshoot in random directions
- Pulling toward a target **cancels some noise** (reduces variance)
- Yes, it introduces **bias** — but the variance reduction can more than compensate
- Net effect: **lower MSE** $= \text{Bias}^2 + \text{Var}$ (the bias-variance tradeoff!)

Stein's Paradox (1956)

Setup: Estimate d **unrelated** means simultaneously. One noisy measurement each: $X_i \sim N(\mu_i, 1)$.

$\mu_1 =$ avg temperature in Yerevan, $\mu_2 =$ price of tea in China, $\mu_3 =$ height of Eiffel Tower

Stein's Paradox (1956)

Setup: Estimate d **unrelated** means simultaneously. One noisy measurement each: $X_i \sim N(\mu_i, 1)$.

$\mu_1 =$ avg temperature in Yerevan, $\mu_2 =$ price of tea in China, $\mu_3 =$ height of Eiffel Tower

The paradox:

The "obvious" estimator $\hat{\mu}_i = X_i$ is **inadmissible** when $d \geq 3$!

There exists a *single* estimator that is better for **all three** simultaneously.

Stein's Paradox (1956)

Setup: Estimate d **unrelated** means simultaneously. One noisy measurement each: $X_i \sim N(\mu_i, 1)$.

$\mu_1 =$ avg temperature in Yerevan, $\mu_2 =$ price of tea in China, $\mu_3 =$ height of Eiffel Tower

The paradox:

The "obvious" estimator $\hat{\mu}_i = X_i$ is **inadmissible** when $d \geq 3$!

There exists a *single* estimator that is better for **all three** simultaneously.

The **James–Stein estimator** shrinks every X_i toward zero:

$$\hat{\mu}_i^{JS} = \underbrace{\left(1 - \frac{d-2}{\|\mathbf{X}\|^2}\right)}_{\text{shrinkage factor } c} \cdot X_i$$

Why is this shocking? These quantities are **completely unrelated**!

Yet estimating them **jointly** beats estimating each one separately.

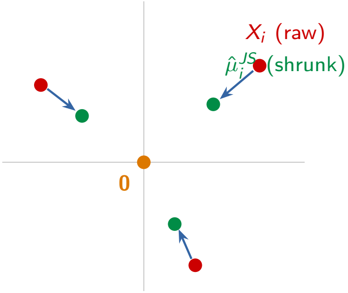
Why Does Stein's Paradox Work?

The MSE comparison:

- Raw: total MSE $= d$ (1 per coordinate)
- James–Stein: total MSE $< d$
  (provably, for *any* $\boldsymbol{\mu}$, when $d \geq 3$)

Why $d \geq 3$?

- The shrinkage factor $c = 1 - \frac{d-2}{\|\mathbf{X}\|^2}$ must be estimated from data
- In $d = 1, 2$: estimating c is too noisy — the error wipes out the gain
- In $d \geq 3$: $\|\mathbf{X}\|^2$ concentrates enough → net win



Each arrow shrinks toward $\mathbf{0}$.
On average, the shrunk points are **closer to the true $\boldsymbol{\mu}$**.

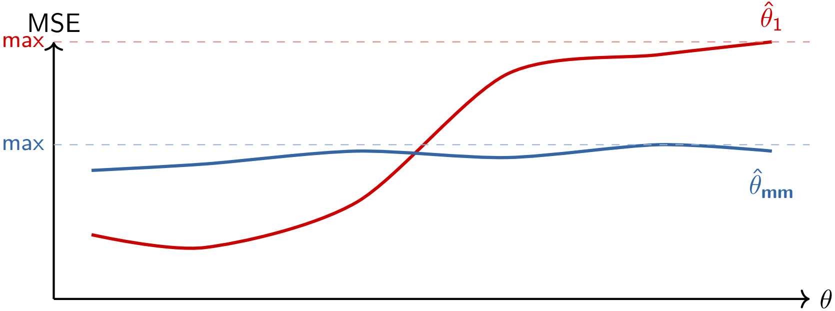
Connection to ML: James–Stein is an early form of **regularization**. Ridge regression (L^2 penalty) does exactly this: shrink coefficients toward zero.

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**.

Three Philosophies of Estimation

Plug-in (unbiased)

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

Admissible in $d = 1$
Inadmissible in $d \geq 3$

Shrinkage

Pull estimates toward a central value (e.g. 0)

Biased but lower MSE (James–Stein)

Minimax

Minimize worst-case risk

Conservative guarantee
No single θ can hurt you badly

Takeaway: In high dimensions ($d \geq 3$), shrinkage estimators are provably better
than using each sample statistic on its own.

MAP estimation (Lecture 6) formalizes shrinkage: a **prior** = automatic regularization.

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

MSE $= \mathrm{Bias}^2 + \mathrm{Var}$. 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: $\mathrm{Var} \geq 1/(n \cdot I(\theta))$. The efficiency floor.

Admissibility: No other estimator dominates it everywhere.

Minimax: Best worst-case guarantee. Shrinkage often wins.

Homework

1. Compute the Fisher information $I(\lambda)$ for Poisson(λ).
   Find the Cramér–Rao lower bound for estimating λ. Is $\hat{\lambda} = \bar{X}$ efficient?

2. For $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$: 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?