Lecture 4: Fisher Information & Cramér–Rao

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

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

Why the Logarithm? From Products to Sums

The likelihood for i.i.d. data is a **product**:

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

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

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

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 $f(x \mid \theta)$, 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** $\rightarrow$ "p should be higher"
- If we observe $x = 0$ and p is large, the score is **large negative** $\rightarrow$ "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:

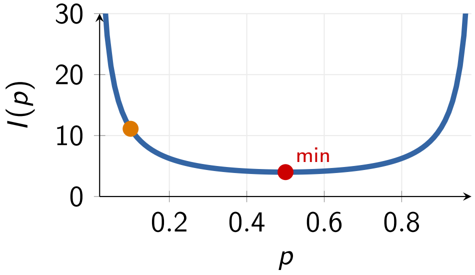
$$\boxed{I(\theta) = \mathrm{Var}[s(\theta)] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]}$$

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: 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:

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

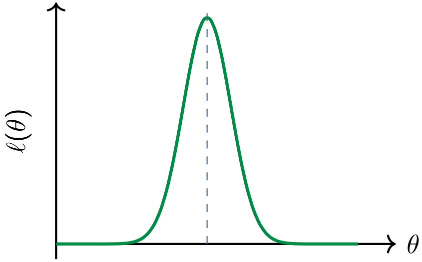
So: $\mathbb{E}[s^2] = -\mathbb{E}[\ell'']$. ✓

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

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

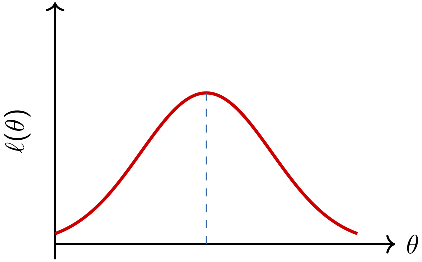
Intuition: Sharp vs Flat Log-Likelihood

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

Verify for Bernoulli:

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

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

Cramér–Rao: Efficiency and Practical Use

What it says:
There is a **floor** on how precise any unbi-
ased estimator can be

Efficient estimator:
Achieves the bound —
the **best possible**

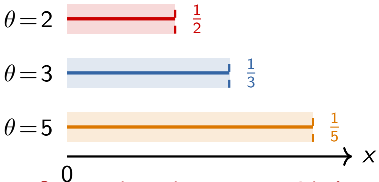
Practical use:
Tells you whether to keep searching for a better estimator

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 requires **regularity conditions**:

1. **Support** of $f(x \mid \theta)$ doesn't depend on θ
2. θ in the **interior** of the parameter space
3. Can **differentiate under the integral** sign (swap $\frac{\partial}{\partial\theta}$ and $\int$)
4. $0 < I(\theta) < \infty$ (finite, positive info)



Support boundary **moves** with θ

Counterexample: Uniform$(0, \theta)$

- Support $[0, \theta]$ depends on θ! (violates #1)
- Suff. stat: $X_{(n)} = \max_i X_i$
- $\mathrm{Var}(X_{(n)}) \sim 1/n^2$ — **faster** than CR!
  (CR would give $1/n$, but $1/n^2$ is possible here)

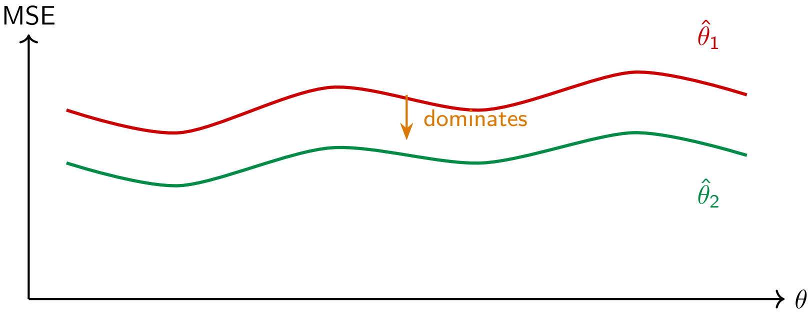
Good news: All exponential family distributions automatically satisfy
the regularity conditions. The CR bound always applies to them.

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.

Stein's Paradox (1956)

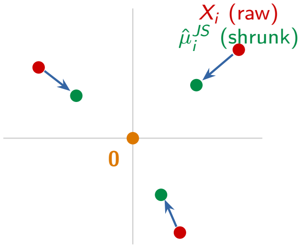
Surprising fact:

When estimating $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ from $X_i \sim N(\mu_i, 1)$, the sample mean $\hat{\mu}_i = X_i$ is **inadmissible** when $d \geq 3$!

The **James–Stein estimator** dominates it:

$$\hat{\mu}_i^{JS} = \left(1 - \frac{d-2}{\|\mathbf{X}\|^2}\right) X_i$$

- **Shrinks** each X_i toward 0
- Works even if μ_i's are unrelated!
- A little bias buys a lot of variance reduction



Paradox: estimating the average temperature in Yerevan *improves* if you jointly estimate it with the price of tea in China and the height of the Eiffel Tower.

Why Does Stein's Paradox Work?

The MSE comparison tells the whole story:

$\mathrm{MSE}(\mathbf{X}) = d$
(1 per coordinate)

$\xrightarrow{\text{shrinkage helps}}$

$\mathrm{MSE}(\hat{\boldsymbol{\mu}}^{JS}) < d$
(always, when $d \geq 3$)

Why $d \geq 3$? The shrinkage factor $\frac{d-2}{\|\mathbf{X}\|^2}$ needs to be estimated from data.

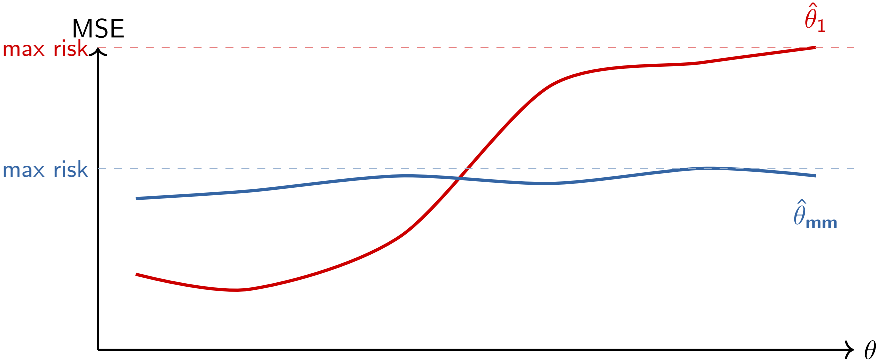
- In $d = 1$ or 2: not enough "room" — estimation error of the shrinkage factor wipes out the gain
- In $d \geq 3$: $\|\mathbf{X}\|^2$ concentrates well enough $\to$ shrinkage factor is accurate $\to$ net win

Connection to ML: James–Stein shrinkage is an early form of **regularization**.
Ridge regression (L^2 penalty) does the same thing: shrink coefficients toward zero. The bias-variance tradeoff in action: a little bias buys a lot of variance reduction.

Minimax Estimators

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

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



Minimax = **conservative**: protects against the worst θ. Minimax hedges.

Three Philosophies of Estimation

Plug-in (unbiased)	Shrinkage	Minimax
Use sample statistic directly ($\bar{X}$, S^2, $\hat{p}$) Admissible in $d = 1$ Inadmissible in $d \geq 3$	Pull estimates toward a central value (e.g. 0) Biased but lower MSE (James–Stein)	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. We'll see more of this in later lectures.

What We Haven't Covered (Yet)

This lecture focused on **point estimation** — producing a single "best guess" for θ. But there's much more to statistical inference:

Confidence intervals: How uncertain is our estimate? (Lectures 6–7)

Hypothesis testing: Is the effect real or just noise? (Lectures 8–9)

Bayesian estimation: Incorporating prior beliefs (Lecture 6)

Bootstrap: Resampling to estimate uncertainty without formulas (Lecture 7)

Asymptotic theory: What happens as $n \to \infty$ in general? (Lecture 6)

Nonparametric estimation: What if we don't assume a distribution at all?

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.

Admissibility: No other estimator dominates it everywhere.

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

Homework

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

Questions?