

## Lecture 5: Point Estimation

Method of Moments · Maximum Likelihood · Why MLE Works

## Previously, on Lecture 4 . . .

**Likelihood:**  $L(\theta) = \prod f(X_i | \theta)$ . How well does  $\theta$  explain the data?

**Score:**  $s(\theta) = \frac{\partial}{\partial \theta} \log f(X | \theta)$ . How sensitive is the model to  $\theta$ ?

**Fisher information:**  $I(\theta) = \text{Var}[s(\theta)]$ . How much info does one observation carry?

**Cramér–Rao:**  $\text{Var}(\hat{\theta}) \geq 1/(nI(\theta))$ . The precision floor for unbiased estimators.

**Admissibility & Minimax:** Compare estimators by MSE across all  $\theta$ ; minimize worst-case risk.

**Today:** We know how to **judge** estimators. Now: how to **construct** them.

Two systematic recipes: **Method of Moments** and **Maximum Likelihood**.

# The Estimation Problem

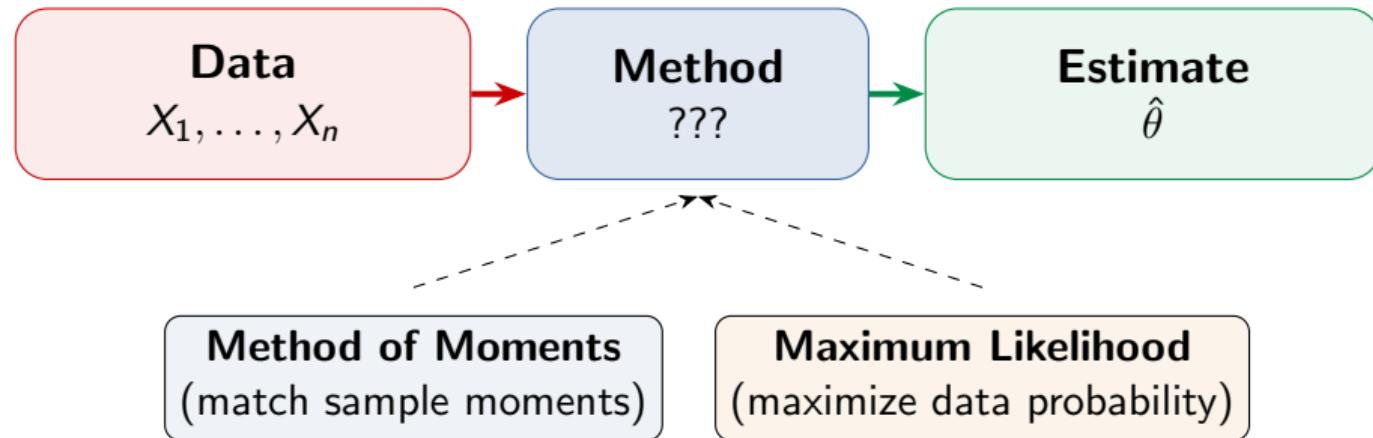
A factory produces lightbulbs. You test 50 and find a mean lifetime of 1,200 hours.

What can you say about the **true** mean lifetime?

In Lectures 3–4 we learned how to **judge** estimators (bias, variance, MSE, efficiency).

Today: how to **construct** them systematically.

# From Data to Parameters



# Method of Moments: The Key Insight

Every distribution's **population moments** are functions of its parameters:

## Population side

$$\text{1st moment: } \mu_1 = \mathbb{E}[X]$$

$$\text{2nd moment: } \mu_2 = \mathbb{E}[X^2]$$

$$k\text{-th moment: } \mu_k = \mathbb{E}[X^k]$$

← set equal →

## Sample side

$$\text{1st moment: } \hat{\mu}_1 = \frac{1}{n} \sum X_i = \bar{X}$$

$$\text{2nd moment: } \hat{\mu}_2 = \frac{1}{n} \sum X_i^2$$

$$k\text{-th moment: } \hat{\mu}_k = \frac{1}{n} \sum X_i^k$$

# Method of Moments: The Key Insight

Every distribution's **population moments** are functions of its parameters:

## Population side

$$\text{1st moment: } \mu_1 = \mathbb{E}[X]$$

$$\text{2nd moment: } \mu_2 = \mathbb{E}[X^2]$$

$$k\text{-th moment: } \mu_k = \mathbb{E}[X^k]$$

← set equal →

## Sample side

$$\text{1st moment: } \hat{\mu}_1 = \frac{1}{n} \sum X_i = \bar{X}$$

$$\text{2nd moment: } \hat{\mu}_2 = \frac{1}{n} \sum X_i^2$$

$$k\text{-th moment: } \hat{\mu}_k = \frac{1}{n} \sum X_i^k$$

## The MoM recipe:

1. Write population moments as functions of parameters
2. Replace with sample moments
3. Solve for  $\hat{\theta}$

$p$  unknown parameters  $\Rightarrow$  need  $p$  moment equations.

## MoM Example: Poisson (One Parameter)

**Setup:**  $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ . One unknown  $\Rightarrow$  one equation.

**Step 1.** Population moment:  $\mathbb{E}[X] = \lambda$  (the mean of a Poisson *is* the parameter)

**Concrete example:** A hospital records emergency arrivals per hour over  $n = 30$  hours:

$$3, 5, 2, 4, 6, 1, 3, 4, 2, 5, \dots \quad \text{with } \bar{X} = 3.8$$

$$\Rightarrow \hat{\lambda}_{\text{MoM}} = 3.8 \text{ arrivals per hour.}$$

**One parameter, one moment, one line of algebra.** That's MoM at its best.

## MoM Example: Poisson (One Parameter)

**Setup:**  $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ . One unknown  $\Rightarrow$  one equation.

**Step 1.** Population moment:  $\mathbb{E}[X] = \lambda$  (the mean of a Poisson *is* the parameter)

**Step 2.** Replace with sample moment:  $\bar{X} = \hat{\lambda}$

**Concrete example:** A hospital records emergency arrivals per hour over  $n = 30$  hours:

3, 5, 2, 4, 6, 1, 3, 4, 2, 5, ... with  $\bar{X} = 3.8$

$\Rightarrow \hat{\lambda}_{\text{MoM}} = 3.8$  arrivals per hour.

**One parameter, one moment, one line of algebra.** That's MoM at its best.

## MoM Example: Poisson (One Parameter)

**Setup:**  $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ . One unknown  $\Rightarrow$  one equation.

**Step 1.** Population moment:  $\mathbb{E}[X] = \lambda$  (the mean of a Poisson *is* the parameter)

**Step 2.** Replace with sample moment:  $\bar{X} = \hat{\lambda}$

**Step 3.** Solve:  $\hat{\lambda}_{\text{MoM}} = \bar{X}$  ✓ Done — nothing to solve!

## MoM Example: Poisson (One Parameter)

**Setup:**  $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ . One unknown  $\Rightarrow$  one equation.

**Step 1.** Population moment:  $\mathbb{E}[X] = \lambda$  (the mean of a Poisson *is* the parameter)

**Step 2.** Replace with sample moment:  $\bar{X} = \hat{\lambda}$

**Step 3.** Solve:  $\hat{\lambda}_{\text{MoM}} = \bar{X}$  ✓ Done — nothing to solve!

**Concrete example:** A hospital records emergency arrivals per hour over  $n = 30$  hours:

3, 5, 2, 4, 6, 1, 3, 4, 2, 5, ... with  $\bar{X} = 3.8$

$\Rightarrow \hat{\lambda}_{\text{MoM}} = 3.8$  arrivals per hour.

**One parameter, one moment, one line of algebra.** That's MoM at its best.

## MoM Example: Normal (Two Parameters — Full Derivation)

**Setup:**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ . Two unknowns  $\Rightarrow$  two equations.

**Step 1.** Write the first two population moments as functions of  $\mu$  and  $\sigma^2$ :

$$\mu_1 = \mathbb{E}[X] = \mu$$

$$\mu_2 = \mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2$$

## MoM Example: Normal (Two Parameters — Full Derivation)

**Setup:**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ . Two unknowns  $\Rightarrow$  two equations.

**Step 1.** Write the first two population moments as functions of  $\mu$  and  $\sigma^2$ :

$$\mu_1 = \mathbb{E}[X] = \mu$$

$$\mu_2 = \mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2$$

**Step 2.** Replace population moments with sample moments:

$$\bar{X} = \hat{\mu}$$

$$\frac{1}{n} \sum X_i^2 = \hat{\sigma}^2 + \hat{\mu}^2$$

## MoM Example: Normal (Two Parameters — Full Derivation)

**Setup:**  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ . Two unknowns  $\Rightarrow$  two equations.

**Step 1.** Write the first two population moments as functions of  $\mu$  and  $\sigma^2$ :

$$\mu_1 = \mathbb{E}[X] = \mu$$

$$\mu_2 = \mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2$$

**Step 2.** Replace population moments with sample moments:

$$\bar{X} = \hat{\mu}$$

$$\frac{1}{n} \sum X_i^2 = \hat{\sigma}^2 + \hat{\mu}^2$$

**Step 3.** Solve — first equation gives  $\hat{\mu}$  immediately; substitute into second:

$$\hat{\sigma}^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2 = \frac{1}{n} \sum X_i^2 - 2\bar{X}^2 + \bar{X}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

$$\boxed{\hat{\mu}_{\text{MoM}} = \bar{X}, \quad \hat{\sigma}_{\text{MoM}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Note: divides by  $n$ , not  $n-1$  — **biased!** Recall Bessel's correction from Lecture 3.

## MoM Example: Gamma Distribution

**Model:**  $X_1, \dots, X_n \sim \text{Gamma}(\alpha, \beta)$  (shape  $\alpha$ , rate  $\beta$ ). Two unknowns.

Population moments:

$$\mathbb{E}[X] = \alpha/\beta$$

$$\text{Var}(X) = \alpha/\beta^2$$

Set equal to sample moments and solve:

## MoM Example: Gamma Distribution

**Model:**  $X_1, \dots, X_n \sim \text{Gamma}(\alpha, \beta)$  (shape  $\alpha$ , rate  $\beta$ ). Two unknowns.

Population moments:

$$\mathbb{E}[X] = \alpha/\beta$$

$$\text{Var}(X) = \alpha/\beta^2$$

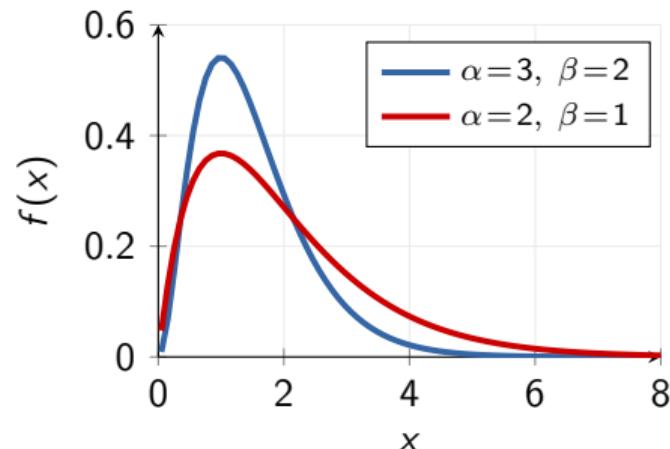
Set equal to sample moments and solve:

$$\hat{\beta}_{\text{MoM}} = \frac{\bar{X}}{S^2}, \quad \hat{\alpha}_{\text{MoM}} = \frac{\bar{X}^2}{S^2}$$

Simple algebra — done! ✓

MLE for Gamma requires the digamma function  $\psi(\alpha)$

— no closed form, numerical optimization only.



Gamma models waiting times, rainfall, income, insurance claims.

**Lesson:** MoM shines when MLE has no closed form. Quick, easy, often a good starting point.

## When MoM Goes Wrong

MoM can give **impossible** parameter values because it doesn't "know" the constraints.

**Example:** Fit a  $\text{Uniform}(0, \theta)$  distribution using MoM.

$$\text{Population mean: } \mathbb{E}[X] = \theta/2 \quad \Rightarrow \quad \hat{\theta}_{\text{MoM}} = 2\bar{X}$$

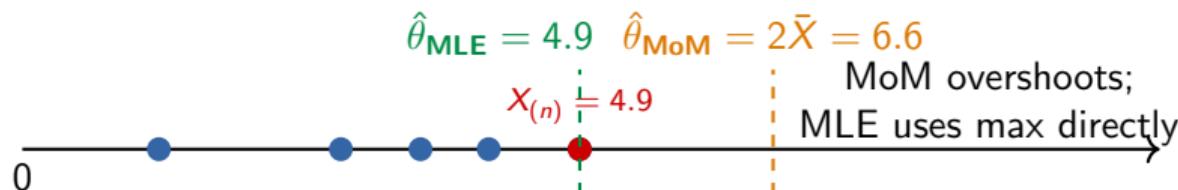
# When MoM Goes Wrong

MoM can give **impossible** parameter values because it doesn't "know" the constraints.

**Example:** Fit a  $\text{Uniform}(0, \theta)$  distribution using MoM.

$$\text{Population mean: } \mathbb{E}[X] = \theta/2 \Rightarrow \hat{\theta}_{\text{MoM}} = 2\bar{X}$$

**Problem:** We need  $\hat{\theta} \geq \max(X_i)$ , but MoM doesn't enforce this!



MoM doesn't use the data efficiently here — it ignores the maximum, which is the sufficient statistic.

## The Likelihood Function (Recap from Lecture 4)

**Given the data I observed, how plausible is each parameter value?**

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

**Data is fixed,  $\theta$  varies.** Log turns the product into a sum (same maximizer).

# The Likelihood Function (Recap from Lecture 4)

**Given the data I observed, how plausible is each parameter value?**

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

**Data is fixed,  $\theta$  varies.** Log turns the product into a sum (same maximizer).

From Lecture 4, we already know:

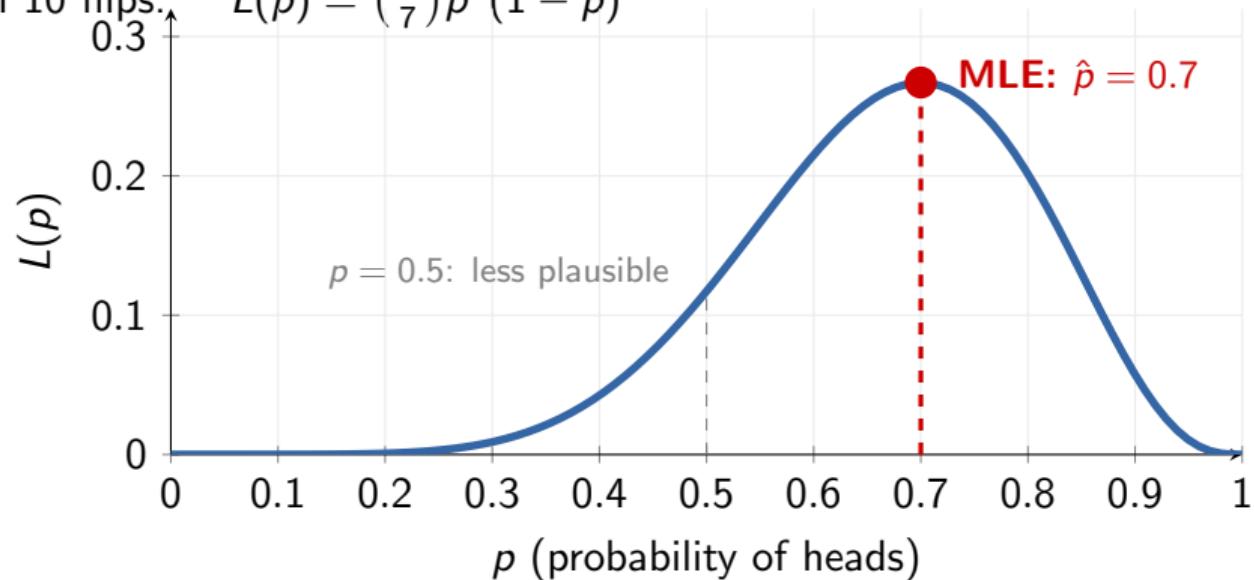
- ▶ The **score**  $s(\theta) = \ell'(\theta)$  measures sensitivity to  $\theta$ ;  $\mathbb{E}[s] = 0$
- ▶ **Fisher information**  $I(\theta) = \text{Var}[s] = -\mathbb{E}[\ell'']$  measures the curvature
- ▶ **Cramér–Rao:** no unbiased estimator can have  $\text{Var} < 1/(nI(\theta))$

Now: how to **use** the likelihood to actually **construct** estimators.

## Likelihood: Coin Flip Example

Data: 7 heads in 10 flips.

$$L(p) = \binom{10}{7} p^7 (1-p)^3$$



## The MLE Idea: What Would the Data Choose?

Imagine you could ask the data: “Which parameter value explains you best?”

The **Maximum Likelihood Estimator** picks the  $\theta$  that makes the observed data **most probable**:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$$

# The MLE Idea: What Would the Data Choose?

Imagine you could ask the data: “Which parameter value explains you best?”

The **Maximum Likelihood Estimator** picks the  $\theta$  that makes the observed data **most probable**:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$$

**Intuition:** If you flip a coin 10 times and get 7 heads...

- ▶ Is  $p = 0.5$  plausible? Somewhat.
- ▶ Is  $p = 0.7$  plausible? Very — it predicts exactly what you saw.
- ▶ Is  $p = 0.99$  plausible? Not really — you'd expect more heads.

MLE picks  $\hat{p} = 0.7$  because it maximizes the likelihood  $L(p) = \binom{10}{7} p^7 (1-p)^3$ .

At the MLE:  $s(\hat{\theta}) = 0$  (score equals zero — first-order condition from Lecture 4).

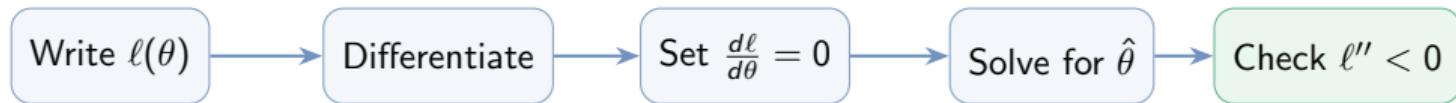
# Maximum Likelihood

Pick the parameter that makes the observed data most probable.

Four closed-form examples, then the ML connection.

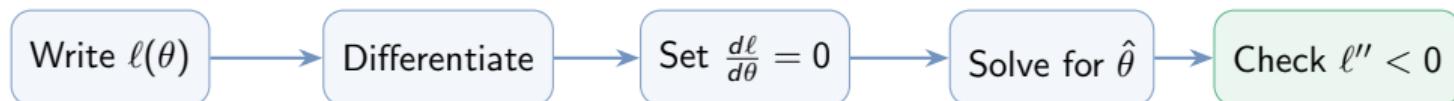
# MLE Recipe: Step by Step

In practice, finding the MLE is a calculus exercise:



# MLE Recipe: Step by Step

In practice, finding the MLE is a calculus exercise:



**When it's easy** (closed form):

- ▶ Exponential families
- ▶ Normal, Bernoulli, Poisson, Exp
- ▶ Solve  $s(\hat{\theta}) = 0$  by hand

**When it's hard** (numerical):

- ▶ Mixture models
- ▶ Logistic regression
- ▶ Use gradient ascent, Newton's method, or EM algorithm

Let's work through four closed-form examples.

## MLE: Bernoulli (Coin Fairness)

**Model:**  $X_i \sim \text{Bernoulli}(p)$ , observe  $k$  successes in  $n$  trials.

$$\ell(p) = k \log p + (n - k) \log(1 - p)$$

$$\frac{\partial \ell}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0$$

## MLE: Bernoulli (Coin Fairness)

**Model:**  $X_i \sim \text{Bernoulli}(p)$ , observe  $k$  successes in  $n$  trials.

$$\ell(p) = k \log p + (n - k) \log(1 - p)$$

$$\frac{\partial \ell}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0$$

$$\hat{p}_{\text{MLE}} = \frac{k}{n} = \bar{X}$$

The sample proportion — exactly what you'd guess intuitively.

## MLE for Normal: Full Derivation

**Model:**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown.

**Step 1.** Write the likelihood (product of  $n$  Gaussian densities):

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

## MLE for Normal: Full Derivation

**Model:**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown.

**Step 1.** Write the likelihood (product of  $n$  Gaussian densities):

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

**Step 2.** Take the log (product  $\rightarrow$  sum):

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

## MLE for Normal: Full Derivation

**Model:**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown.

**Step 1.** Write the likelihood (product of  $n$  Gaussian densities):

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

**Step 2.** Take the log (product  $\rightarrow$  sum):

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

**Step 3.** Set  $\frac{\partial \ell}{\partial \mu} = 0$ :  $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{X}$

## MLE for Normal: Full Derivation

**Model:**  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown.

**Step 1.** Write the likelihood (product of  $n$  Gaussian densities):

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

**Step 2.** Take the log (product  $\rightarrow$  sum):

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

**Step 3.** Set  $\frac{\partial \ell}{\partial \mu} = 0$ :  $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{X}$

**Step 4.** Set  $\frac{\partial \ell}{\partial (\sigma^2)} = 0$ :  $-\frac{n}{2\sigma^2} + \frac{\sum (X_i - \bar{X})^2}{2\sigma^4} = 0$

$$\implies \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

# How Good Is the Normal MLE?

**For**  $\hat{\mu} = \bar{X}$ :

- ▶ Bias = 0 (unbiased)
- ▶ Var =  $\sigma^2/n$
- ▶ MSE =  $\sigma^2/n$
- ✓ = CR bound — **efficient!**

**For**  $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ :

- ▶ Bias =  $-\sigma^2/n$  (biased!)
- ▶ Var =  $2(n-1)\sigma^4/n^2$
- ▶ MSE =  $(2n-1)\sigma^4/n^2$

# How Good Is the Normal MLE?

For  $\hat{\mu} = \bar{X}$ :

- ▶ Bias = 0 (unbiased)
- ▶ Var =  $\sigma^2/n$
- ▶ MSE =  $\sigma^2/n$
- ✓ = CR bound — efficient!

For  $\hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ :

- ▶ Bias =  $-\sigma^2/n$  (biased!)
- ▶ Var =  $2(n-1)\sigma^4/n^2$
- ▶ MSE =  $(2n-1)\sigma^4/n^2$

Compare with Bessel's  $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$  (unbiased):

	$\hat{\sigma}_{\text{MLE}}^2$ (divide by $n$ )	$S^2$ (divide by $n-1$ )
Bias	$-\sigma^2/n$	0
MSE	$(2n-1)\sigma^4/n^2$	$2\sigma^4/(n-1)$

$\text{MSE}(\hat{\sigma}_{\text{MLE}}^2) < \text{MSE}(S^2)$  always! The biased MLE wins on MSE (Lecture 3 tradeoff).

## From MLE to Machine Learning

In ML, we model:  $y_i = f(\mathbf{x}_i; \mathbf{w}) + \varepsilon_i$ ,  $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

So  $y_i | \mathbf{x}_i \sim \mathcal{N}(f(\mathbf{x}_i; \mathbf{w}), \sigma^2)$ . The log-likelihood of  $\mathbf{w}$ :

$$\ell(\mathbf{w}) = \underbrace{-\frac{n}{2} \log(2\pi\sigma^2)}_{\text{const w.r.t. } \mathbf{w}} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

# From MLE to Machine Learning

In ML, we model:  $y_i = f(\mathbf{x}_i; \mathbf{w}) + \varepsilon_i$ ,  $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

So  $y_i | \mathbf{x}_i \sim \mathcal{N}(f(\mathbf{x}_i; \mathbf{w}), \sigma^2)$ . The log-likelihood of  $\mathbf{w}$ :

$$\ell(\mathbf{w}) = \underbrace{-\frac{n}{2} \log(2\pi\sigma^2)}_{\text{const w.r.t. } \mathbf{w}} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

$$\max_{\mathbf{w}} \ell(\mathbf{w}) \iff \min_{\mathbf{w}} \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2 = \text{MSE loss!}$$

## Gaussian noise + MLE = Least Squares

The MSE loss in machine learning is not arbitrary —  
it is exactly **maximum likelihood under Gaussian noise**.

Linear regression, neural nets with MSE loss, OLS — all are doing MLE.

Not just Gaussian — every noise model gives a different loss function...

## MLE and Cross-Entropy

Now:  $y_i \in \{0, 1\}$  (spam/not spam, click/no click, disease/healthy).

**Model:**  $P(y_i = 1 | \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$  where  $\sigma(z) = \frac{1}{1+e^{-z}}$  (logistic function)

## MLE and Cross-Entropy

Now:  $y_i \in \{0, 1\}$  (spam/not spam, click/no click, disease/healthy).

**Model:**  $P(y_i = 1 | \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$  where  $\sigma(z) = \frac{1}{1+e^{-z}}$  (logistic function)

The log-likelihood:

$$\ell(\mathbf{w}) = \sum_{i=1}^n [y_i \log \hat{p}_i + (1-y_i) \log(1-\hat{p}_i)] \quad \hat{p}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i)$$

## MLE and Cross-Entropy

Now:  $y_i \in \{0, 1\}$  (spam/not spam, click/no click, disease/healthy).

**Model:**  $P(y_i = 1 | \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$  where  $\sigma(z) = \frac{1}{1+e^{-z}}$  (logistic function)

The log-likelihood:

$$\ell(\mathbf{w}) = \sum_{i=1}^n [y_i \log \hat{p}_i + (1-y_i) \log(1-\hat{p}_i)] \quad \hat{p}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i)$$

$$\max_{\mathbf{w}} \ell(\mathbf{w}) \iff \min_{\mathbf{w}} \underbrace{- \sum [y_i \log \hat{p}_i + (1-y_i) \log(1-\hat{p}_i)]}_{\text{binary cross-entropy loss}}$$

**Bernoulli outcome + MLE = Cross-Entropy Loss**

Logistic regression, neural nets with sigmoid output — all doing MLE.

**Gaussian** → MSE — **Bernoulli** → Cross-Entropy — **Laplace** → MAE

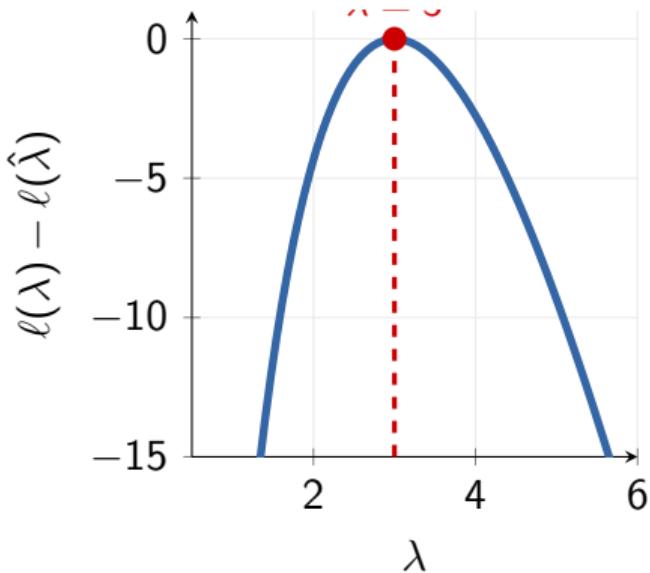
# MLE: Poisson (Rare Events)

**Model:**  $X_i \sim \text{Pois}(\lambda)$

(goals/match, earthquakes/yr, typos/pg)

$$\ell(\lambda) = (\sum X_i) \log \lambda - n\lambda + c$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum X_i}{\lambda} - n = 0$$



Example:  $n = 20$ ,  $\sum X_i = 60$

# MLE: Poisson (Rare Events)

**Model:**  $X_i \sim \text{Pois}(\lambda)$

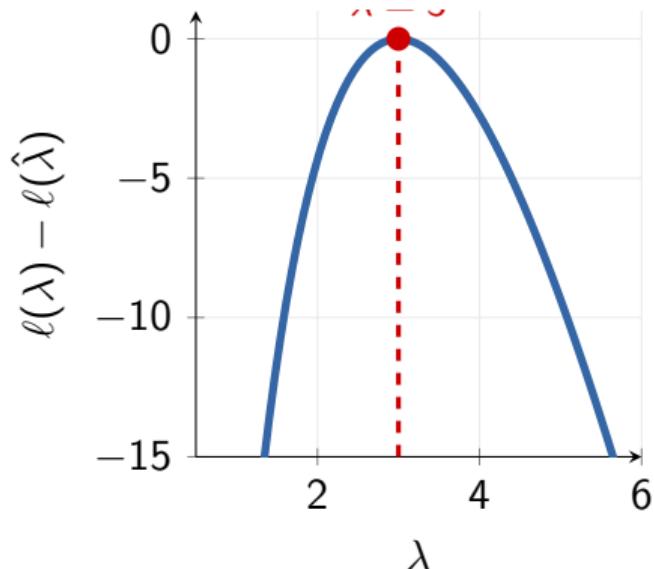
(goals/match, earthquakes/yr, typos/pg)

$$\ell(\lambda) = (\sum X_i) \log \lambda - n\lambda + c$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{\sum X_i}{\lambda} - n = 0$$

$$\hat{\lambda}_{\text{MLE}} = \bar{X}$$

Sample mean estimates the *rate*.



Example:  $n = 20$ ,  $\sum X_i = 60$

## MLE: Exponential (Waiting Times)

**Model:**  $X_i \sim \text{Exp}(\lambda)$  (time between arrivals, device lifetimes)

$$f(x | \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n X_i$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum X_i = 0$$

## MLE: Exponential (Waiting Times)

**Model:**  $X_i \sim \text{Exp}(\lambda)$  (time between arrivals, device lifetimes)

$$f(x | \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n X_i$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum X_i = 0$$

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{X}}$$

The reciprocal of the sample mean — intuitive since  $\mathbb{E}[X] = 1/\lambda$ .

## MLE: Summary of Examples

Distribution	Parameter	MLE	Real-world use
Bernoulli( $p$ )	$p$	$\bar{X}$	Coin fairness, conversion rates
Normal( $\mu, \sigma^2$ )	$\mu, \sigma^2$	$\bar{X}, \frac{1}{n} \sum (X_i - \bar{X})^2$	Measurement error
Poisson( $\lambda$ )	$\lambda$	$\bar{X}$	Count data, rare events
Exponential( $\lambda$ )	$\lambda$	$1/\bar{X}$	Waiting times, lifetimes

Notice: for exponential families, MLE often equals MoM! We'll see why in the "Why MLE Works" section.

# MoM vs MLE: When to Use Which?

	Method of Moments	Maximum Likelihood
Idea	Match sample moments	Maximize data probability
Computation	Usually algebraic	May need optimization
Efficiency	Generally less efficient	Asymptotically optimal
Impossible values?	Can happen ( $\hat{\sigma}^2 < 0$ )	Respects constraints
Invariance	No	Yes ( $g(\hat{\theta})$ is MLE of $g(\theta)$ )
Exp. family	Often same as MLE	Always uses suff. stat

**Rule of thumb:** Use MLE when you can (it's optimal).  
Use MoM as a quick starting point, or when MLE has no closed form.

# Properties of the MLE

Invariance, identifiability, and what can go wrong.

## Invariance Property

If  $\hat{\theta}_{\text{MLE}}$  is the MLE of  $\theta$ , then for any function  $g$ :

$$\widehat{g(\theta)}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}})$$

## Invariance Property

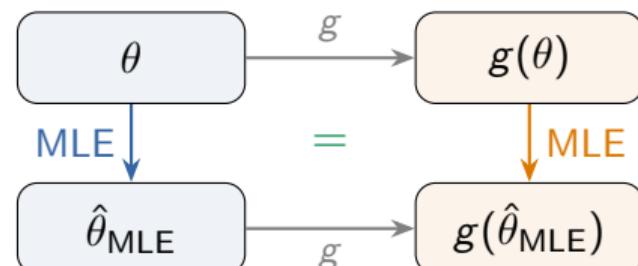
If  $\hat{\theta}_{MLE}$  is the MLE of  $\theta$ , then for any function  $g$ :

$$\widehat{g(\theta)}_{MLE} = g(\hat{\theta}_{MLE})$$

### Example:

- MLE of  $\lambda$  for Exp is  $\hat{\lambda} = 1/\bar{X}$
- Want MLE of mean  $\mu = 1/\lambda$ ?
- Apply  $g(\lambda) = 1/\lambda$ :  $\hat{\mu} = \bar{X}$  ✓

This doesn't hold for MoM or other estimators in general.



# Identifiability

Can we even hope to recover  $\theta$ ?

A model is **identifiable** if different parameter values give different distributions:

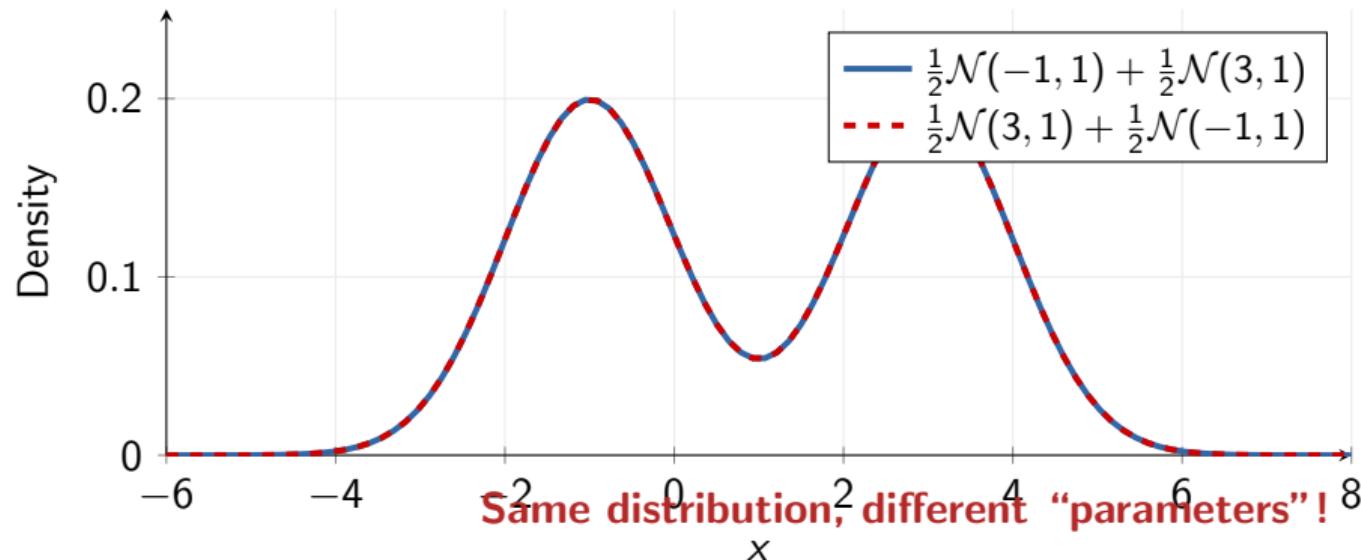
$$\theta_1 \neq \theta_2 \Rightarrow f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$$

## When it fails:

- ▶ **Mixtures:**  $\frac{1}{2}N(-1, 1) + \frac{1}{2}N(3, 1)$  — swap components, same distribution
- ▶ **Overparameterized:**  $X \sim N(\alpha + \beta, 1)$  — data reveals  $\alpha + \beta$ , not each separately
- ▶ **Neural nets:** Swap two hidden neurons and their weights — same function, different  $\theta$

If the model isn't identifiable, no amount of data will pin down  $\theta$ .

# Visualizing Non-Identifiability



# Why MLE Works

Connecting MLE to sufficiency, exponential families,  
and the Cramér–Rao bound from Lectures 3–4.

# MLE and Sufficient Statistics

In Lecture 3 we learned: a **sufficient statistic**  $T(\mathbf{X})$  captures everything about  $\theta$ .

**Key fact:** The MLE depends on the data **only through** the sufficient statistic.

If  $T(\mathbf{X})$  is sufficient for  $\theta$ , then the MLE  $\hat{\theta}$  is a function of  $T$ .

Check our examples:

Model	Suff. stat $T$	MLE	Function of $T$ ?
$Bern(p)$	$\sum X_i$	$\bar{X} = T/n$	✓
$N(\mu, \sigma_0^2)$	$\sum X_i$	$\bar{X} = T/n$	✓
$Pois(\lambda)$	$\sum X_i$	$\bar{X} = T/n$	✓
$Exp(\lambda)$	$\sum X_i$	$1/\bar{X} = n/T$	✓

No coincidence — MLE **always** uses sufficient statistics. No information is wasted.

## MLE in Exponential Families: The General Recipe

Recall from Lecture 3:  $f(x | \theta) = h(x) \exp(\eta(\theta) T(x) - A(\theta))$

For  $n$  i.i.d. observations, the log-likelihood depends on the data **only through**  $\sum T(X_i)$ :

$$\ell(\theta) = \eta(\theta) \sum_{i=1}^n T(X_i) - nA(\theta) + \text{const}$$

Setting  $\ell'(\theta) = 0$  gives a universal MLE formula for natural families ( $\eta = \theta$ ):

$$A'(\hat{\theta}_{\text{MLE}}) = \frac{1}{n} \sum_{i=1}^n T(X_i)$$

## MLE in Exponential Families: The General Recipe

Recall from Lecture 3:  $f(x | \theta) = h(x) \exp(\eta(\theta) T(x) - A(\theta))$

For  $n$  i.i.d. observations, the log-likelihood depends on the data **only through**  $\sum T(X_i)$ :

$$\ell(\theta) = \eta(\theta) \sum_{i=1}^n T(X_i) - nA(\theta) + \text{const}$$

Setting  $\ell'(\theta) = 0$  gives a universal MLE formula for natural families ( $\eta = \theta$ ):

$$A'(\hat{\theta}_{\text{MLE}}) = \frac{1}{n} \sum_{i=1}^n T(X_i)$$

Let's verify with Poisson:

$$\text{Poisson: } T(x) = x, \quad \eta = \log \lambda, \quad A(\eta) = e^\eta, \quad A'(\eta) = e^\eta = \lambda$$

Formula says:  $\hat{\lambda} = A'(\hat{\eta}) = \frac{1}{n} \sum X_i = \bar{X}$  ✓ — same answer we derived by hand!

**One formula, every exponential family.** Plug in  $T$  and  $A$ , get the MLE.

And since  $\mathbb{E}[T(X)] = A'(\eta)$ , the MLE is **exactly the MoM estimator**.

# Why MLE Works: The Big Theoretical Guarantees

Under regularity conditions (Lecture 4), MLE has remarkable properties:

**1. Consistent:**  $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0$  as  $n \rightarrow \infty$  (gets the right answer eventually)

**2. Asymptotically Normal:**  $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$

**3. Asymptotically Efficient:** achieves the **Cramér–Rao bound** as  $n \rightarrow \infty$

**4. Invariant:** MLE of  $g(\theta)$  is  $g(\hat{\theta}_{\text{MLE}})$  for any function  $g$

# Why MLE Works: The Big Theoretical Guarantees

Under regularity conditions (Lecture 4), MLE has remarkable properties:

**1. Consistent:**  $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0$  as  $n \rightarrow \infty$  (gets the right answer eventually)

**2. Asymptotically Normal:**  $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$

**3. Asymptotically Efficient:** achieves the **Cramér–Rao bound** as  $n \rightarrow \infty$

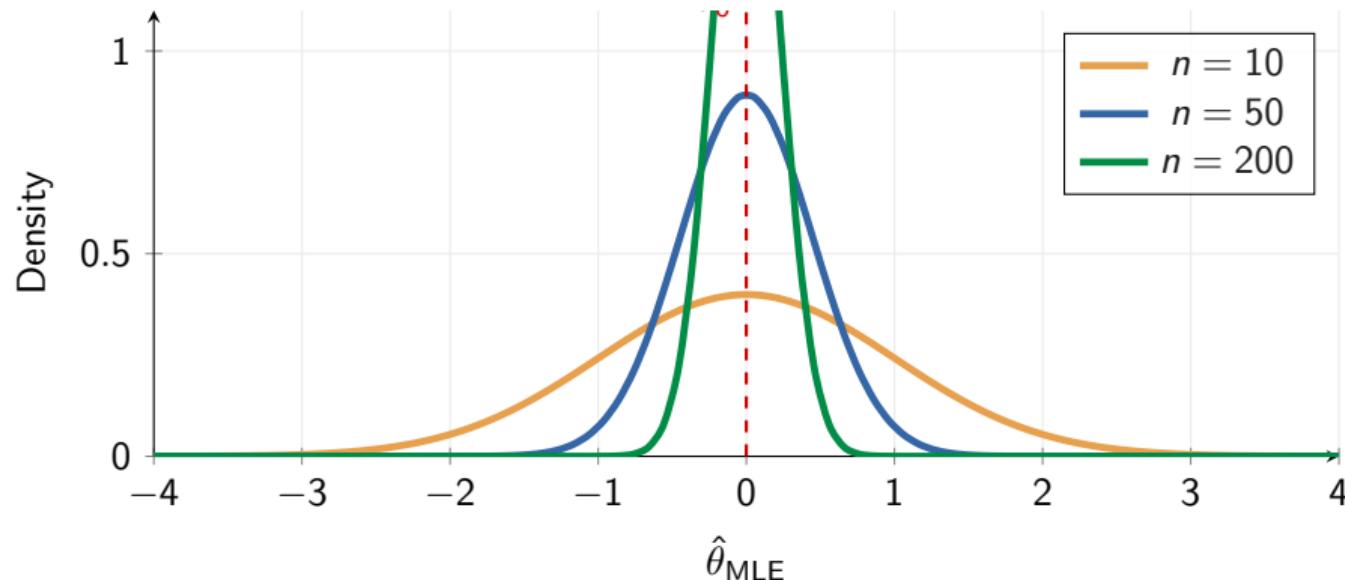
**4. Invariant:** MLE of  $g(\theta)$  is  $g(\hat{\theta}_{\text{MLE}})$  for any function  $g$

**Translation:** With enough data, MLE is approximately unbiased, approximately normal, and **no other estimator can do better.**

This is why MLE is the default method in statistics and machine learning.

## Asymptotic Normality: Seeing It

As  $n$  grows, the sampling distribution of the MLE converges to a Normal centered at the truth:



Variance shrinks as  $\frac{1}{n I(\theta_0)}$ : more data  $\Rightarrow$  tighter bell  $\Rightarrow$  more precise estimate.

With  $n = 200$  observations, MLE is practically pinpointed at the truth.

## From MLE to Standard Errors

Asymptotic normality says:  $\hat{\theta}_{\text{MLE}} \sim N\left(\theta_0, \frac{1}{n I(\theta_0)}\right)$

**Problem:** We don't know  $\theta_0$  — that's what we're estimating!

## From MLE to Standard Errors

Asymptotic normality says:  $\hat{\theta}_{\text{MLE}} \sim N\left(\theta_0, \frac{1}{n I(\theta_0)}\right)$

**Problem:** We don't know  $\theta_0$  — that's what we're estimating!

**Solution:** Plug in  $\hat{\theta}$  to get the **standard error**:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{n I(\hat{\theta})}}$$

or equivalently:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{J_n(\hat{\theta})}}$$

where  $J_n(\hat{\theta}) = -\ell_n''(\hat{\theta})$  is the **observed** Fisher information (the actual curvature at the MLE).

## From MLE to Standard Errors

Asymptotic normality says:  $\hat{\theta}_{\text{MLE}} \sim N\left(\theta_0, \frac{1}{n I(\theta_0)}\right)$

**Problem:** We don't know  $\theta_0$  — that's what we're estimating!

**Solution:** Plug in  $\hat{\theta}$  to get the **standard error**:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{n I(\hat{\theta})}}$$

or equivalently:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{J_n(\hat{\theta})}}$$

where  $J_n(\hat{\theta}) = -\ell_n''(\hat{\theta})$  is the **observed** Fisher information (the actual curvature at the MLE).

**Example:** Bernoulli,  $\hat{p} = 0.3$ ,  $n = 100$ .

$$I(p) = \frac{1}{p(1-p)} \Rightarrow \text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.3 \cdot 0.7}{100}} = 0.046$$

**This is how statistical software reports standard errors.**

Every time you see  $\hat{\theta} \pm \text{SE}$  in R, Python, or a paper,  
it's using Fisher information under the hood. More in Lecture 9.

# MLE Achieves the Cramér–Rao Bound

From Lecture 4, the **CR bound**:  $\text{Var}(\hat{\theta}) \geq \frac{1}{n I(\theta)}$  for unbiased estimators.

Does MLE hit this bound?

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

For **exponential families**, the MLE of the natural parameter is efficient (hits the CR bound exactly). For other models, MLE is **asymptotically** efficient — it approaches the bound as  $n \rightarrow \infty$ .

## When MLE Goes Wrong

MLE has great asymptotic theory, but several things can go wrong:

- **Small samples:** MLE is asymptotic — can be poor for small  $n$ .

Example: 0 heads in 3 flips  $\Rightarrow \hat{p}_{\text{MLE}} = 0$ . Surely too extreme!

## When MLE Goes Wrong

MLE has great asymptotic theory, but several things can go wrong:

- ▶ **Small samples:** MLE is asymptotic — can be poor for small  $n$ .

Example: 0 heads in 3 flips  $\Rightarrow \hat{p}_{\text{MLE}} = 0$ . Surely too extreme!

- ▶ **Boundary of parameter space:**  $\text{Uniform}(0, \theta) \Rightarrow \hat{\theta} = X_{(n)}$ .

Always underestimates:  $\text{bias} = -\theta/(n+1)$ . Regularity conditions fail (Lecture 4), CR doesn't apply.

# When MLE Goes Wrong

MLE has great asymptotic theory, but several things can go wrong:

- ▶ **Small samples:** MLE is asymptotic — can be poor for small  $n$ .  
Example: 0 heads in 3 flips  $\Rightarrow \hat{p}_{\text{MLE}} = 0$ . Surely too extreme!
- ▶ **Boundary of parameter space:**  $\text{Uniform}(0, \theta) \Rightarrow \hat{\theta} = X_{(n)}$ .  
Always underestimates: bias  $= -\theta/(n+1)$ . Regularity conditions fail (Lecture 4), CR doesn't apply.
- ▶ **Neyman–Scott problem:** Too many nuisance parameters  $\Rightarrow$  **inconsistent MLE**.  
 $n$  groups with 2 obs each, own mean  $\mu_i$ : MLE of  $\sigma^2$  converges to  $\sigma^2/2$ !

# When MLE Goes Wrong

MLE has great asymptotic theory, but several things can go wrong:

- ▶ **Small samples:** MLE is asymptotic — can be poor for small  $n$ .

Example: 0 heads in 3 flips  $\Rightarrow \hat{p}_{\text{MLE}} = 0$ . Surely too extreme!

- ▶ **Boundary of parameter space:**  $\text{Uniform}(0, \theta) \Rightarrow \hat{\theta} = X_{(n)}$ .

Always underestimates: bias  $= -\theta/(n+1)$ . Regularity conditions fail (Lecture 4), CR doesn't apply.

- ▶ **Neyman–Scott problem:** Too many nuisance parameters  $\Rightarrow$  **inconsistent MLE**.

$n$  groups with 2 obs each, own mean  $\mu_i$ : MLE of  $\sigma^2$  converges to  $\sigma^2/2$ !

- ▶ **Overfitting:** Flexible models memorize noise.

Degree-20 polynomial through 25 points  $\Rightarrow$  wild oscillations.

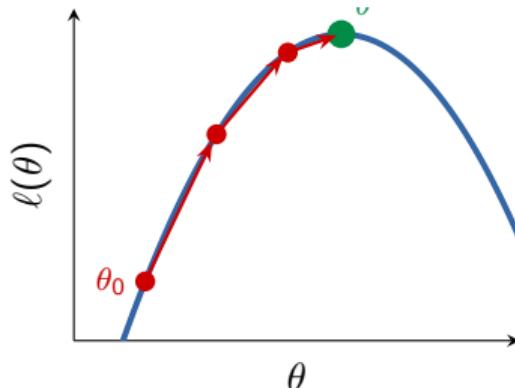
**Common cure:** Add a prior  $\rightarrow$  MAP estimation (Lecture 6).

Prior = regularization = controlled bias toward simpler models.

# When There's No Closed Form

Many models (logistic regression, mixtures, neural nets) require **numerical** optimization.

**Gradient ascent:**



$$\theta_{t+1} = \theta_t + \alpha \cdot \ell'(\theta_t)$$

Follow the slope uphill. The default in deep learning.

**Newton–Raphson:**

$$\theta_{t+1} = \theta_t - \frac{\ell'(\theta_t)}{\ell''(\theta_t)}$$

Uses curvature ( $\ell'' \leftrightarrow$  Fisher info) for smarter steps.

In Python: `scipy.optimize.minimize`

For latent variables: **EM algorithm** (Lecture 7)

# Summary: Constructing Estimators

**MoM:** Match sample moments to population moments. Simple but can give impossible values.

**MLE:** Maximize  $L(\theta) = \prod f(X_i | \theta)$ . The go-to method for estimation.

**Invariance:** MLE of  $g(\theta)$  is  $g(\hat{\theta}_{\text{MLE}})$ . Transform freely.

**Sufficiency:** MLE always uses the sufficient statistic — no information wasted.

**Consistency:**  $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta_0$ . Correct in the long run.

**Asymptotic normality:**  $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, 1/I(\theta_0))$ . Enables standard errors.

**Efficiency:** MLE achieves the CR bound (exactly for exp. families, asymptotically otherwise).

**ML connection:** Gaussian  $\rightarrow$  MSE, Bernoulli  $\rightarrow$  cross-entropy. Loss functions are MLE!

## Practical: Implement MLE

1. Implement MLE for a Gaussian **from scratch**:
  - ▶ Write the log-likelihood function
  - ▶ Optimize numerically (`scipy.optimize`) and compare with closed form
2. Compare  $\hat{\sigma}_{\text{MLE}}^2$  (divides by  $n$ ) with  $S^2$  (divides by  $n-1$ ). Verify the bias from Lecture 3 empirically with simulation.
3. Fit a Poisson to real count data. Check: is the MLE efficient? Compute the CR bound and compare with the observed variance.
4. Plot the log-likelihood surface — observe the peak at the MLE and relate its **curvature** to Fisher information.

## Homework

1. Derive the MLE for  $\text{Geometric}(p)$ :  $f(x | p) = (1 - p)^{x-1}p$ ,  $x = 1, 2, \dots$   
Is this MLE unbiased? Is it efficient (check against the CR bound)?
2. For  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ , show that  $\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$  equals the MoM estimator. Why is this not a coincidence? (Hint: exponential family.)
3. Show that the MLE for  $\text{Uniform}(0, \theta)$  is  $\hat{\theta} = X_{(n)} = \max(X_1, \dots, X_n)$ .  
Is this unbiased? Is it consistent? (Hint: not an exponential family; recall the MoM comparison.)
4. Simulate  $n = 50$  samples from  $\text{Poisson}(\lambda = 3)$  and compute the MLE.  
Repeat 10,000 times. Verify: (a)  $\hat{\lambda}$  is approximately unbiased, (b)  $\text{Var}(\hat{\lambda}) \approx \lambda/n$ .

# Recommended Visualizations & Resources

## Interactive: MLE & Likelihood

[rpsychologist.com/likelihood](http://rpsychologist.com/likelihood) — drag sliders to see how likelihood, score, and Fisher info change in real time. The best interactive MLE demo available.

## Interactive: Seeing Theory (Brown University)

[seeing-theory.brown.edu/frequentist-inference](http://seeing-theory.brown.edu/frequentist-inference) — beautiful animations for point estimation, confidence intervals, and bootstrap. Chapter 5 covers MLE.

## Video: StatQuest — MLE, clearly explained

[statquest.org/statquest-maximum-likelihood-clearly-explained](http://statquest.org/statquest-maximum-likelihood-clearly-explained) — 6-minute visual walkthrough of MLE with the Normal distribution. Great for review.

## Interactive: MLU-Explain (Amazon)

[mlu-explain.github.io](http://mlu-explain.github.io) — visual articles on logistic regression, cross-validation, and neural networks. See how MLE powers these ML methods.

## Reading: Penn State STAT 415

[online.stat.psu.edu/stat415/lesson/1/1.4](http://online.stat.psu.edu/stat415/lesson/1/1.4) — worked MoM examples with step-by-step solutions for Exponential, Gamma, and more.

# Questions?

Next: Lecture 6 — MAP estimation, priors, and the Bayesian perspective