# Lecture 5: Point Estimation

Method of Moments · Maximum Likelihood · Why MLE Works

## Previously, on Lecture 4...

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

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

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

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

**Admissibility & Stein:** Biased estimators (shrinkage) can beat unbiased ones in $d \geq 3$.

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

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

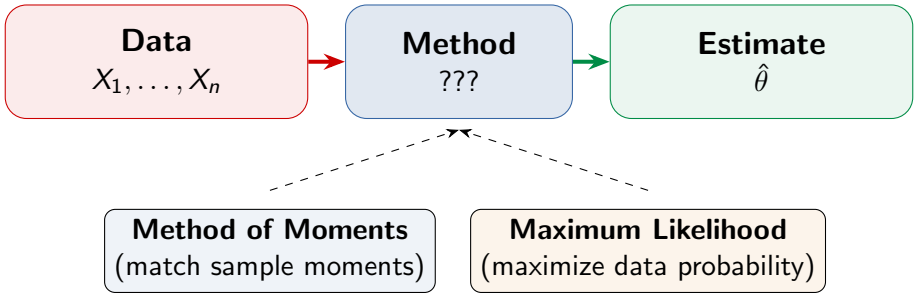
# The Estimation Problem

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 (MoM)

**Idea:** Set population moments equal to sample moments, then solve for the parameters.

$$\begin{aligned} \mathbb{E}[X] &= g_1(\theta) \\ \mathbb{E}[X^2] &= g_2(\theta) \\ &\vdots \end{aligned}$$

$\longrightarrow$ replace with

$$\begin{aligned} \bar{X} &= g_1(\hat{\theta}) \\ \tfrac{1}{n}\textstyle\sum X_i^2 &= g_2(\hat{\theta}) \\ &\vdots \end{aligned}$$

# Method of Moments (MoM)

**Idea:** Set population moments equal to sample moments, then solve for the parameters.

$$\begin{aligned} \mathbb{E}[X] &= g_1(\theta) \\ \mathbb{E}[X^2] &= g_2(\theta) \\ &\vdots \end{aligned}$$

$\longrightarrow$ replace with

$$\begin{aligned} \bar{X} &= g_1(\hat{\theta}) \\ \tfrac{1}{n}\textstyle\sum X_i^2 &= g_2(\hat{\theta}) \\ &\vdots \end{aligned}$$

**Pros:**

- Simple, quick to compute
- No distributional assumption needed for computation

**Cons:**

- Can give impossible values (e.g., $\hat{\sigma}^2 < 0$)
- Generally less efficient than MLE
- Awkward with many parameters

# MoM Example: Normal Distribution

**Model:** $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Two unknowns, need two equations.

**1st moment:** $\mathbb{E}[X] = \mu \quad \Rightarrow \quad \hat{\mu}_{\text{MoM}} = \bar{X}$

**2nd moment:** $\mathbb{E}[X^2] = \mu^2 + \sigma^2 \quad \Rightarrow \quad \hat{\sigma}^2_{\text{MoM}} = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$

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

# When MoM Goes Wrong

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

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

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

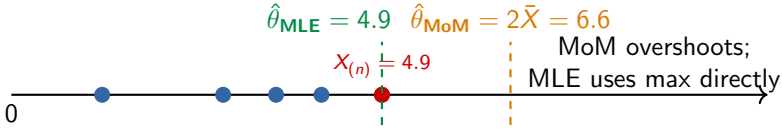
# When MoM Goes Wrong

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

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

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

**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?**

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

**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?**

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

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

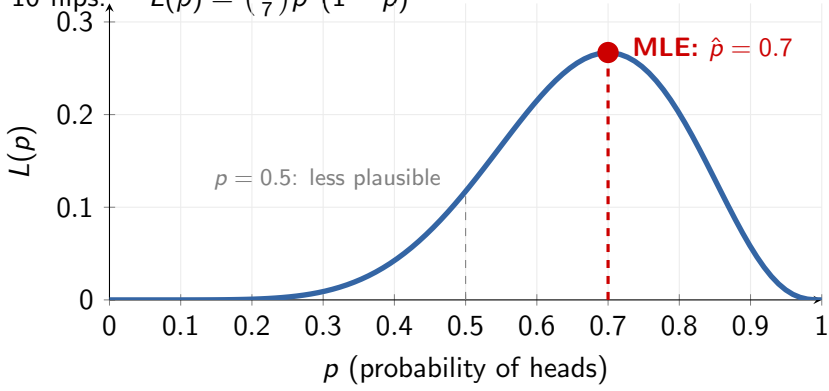
From Lecture 4, we already know:

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

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...

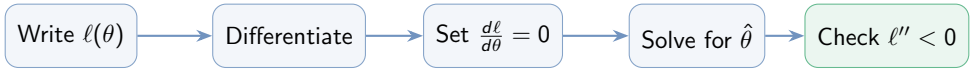
- Is $p = 0.5$ plausible? Somewhat.
- Is $p = 0.7$ plausible? Very — it predicts exactly what you saw.
- Is $p = 0.99$ plausible? Not really — you'd expect more heads.

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).

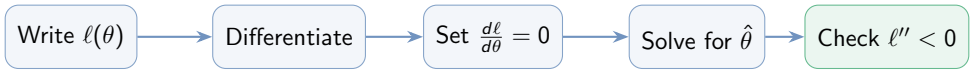
# MLE Recipe: Step by Step

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.

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

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

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

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

# MLE for Normal: Full Derivation

**Model:** $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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:** $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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 $\to$ sum):

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

# MLE for Normal: Full Derivation

**Model:** $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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 $\to$ sum):

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

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

# MLE for Normal: Full Derivation

**Model:** $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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 $\to$ sum):

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

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

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

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

# 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):**

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

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

# From MLE to Machine Learning

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

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

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

# From MLE to Machine Learning

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

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

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

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

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

The log-likelihood:

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

# MLE and Cross-Entropy

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

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

The log-likelihood:

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

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

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

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

**Gaussian** $\rightarrow$ MSE — **Bernoulli** $\rightarrow$ Cross-Entropy — **Laplace** $\rightarrow$ MAE

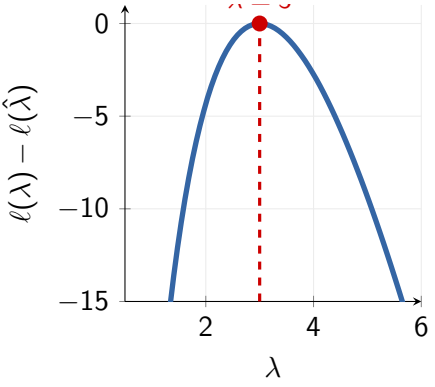
# MLE: Poisson (Rare Events)

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

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

$$\ell(\lambda) = \left(\sum X_i\right) \log \lambda - n\lambda + \text{c}$$

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




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

# MLE: Poisson (Rare Events)

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

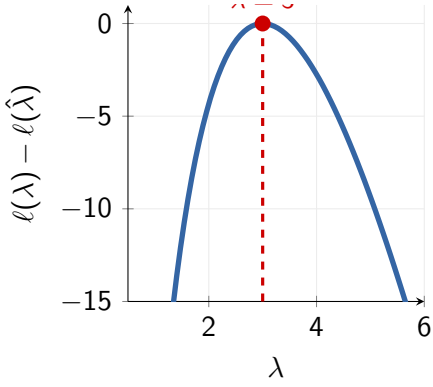
(goals/match, earthquakes/yr, typos/pg)

$$\ell(\lambda) = \left(\sum X_i\right) \log \lambda - n\lambda + \text{c}$$

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

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

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)

$f(x \mid \lambda) = \lambda e^{-\lambda x}$ 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)

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

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

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

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

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 shortly.

# 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.

# Invariance Property

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

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

# Invariance Property

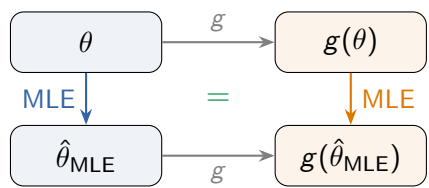
If $\hat{\theta}_{\text{MLE}}$ is the MLE of $\theta$, then for any function $g$:

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

**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:

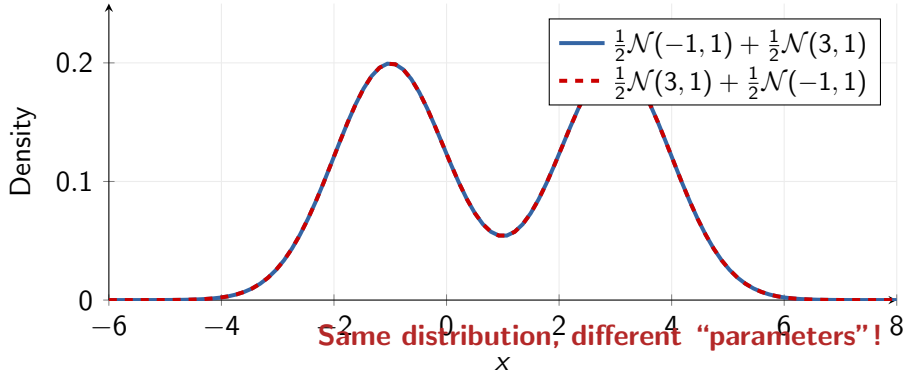
$$\theta_1 \neq \theta_2 \quad \Rightarrow \quad f(\cdot \mid \theta_1) \neq f(\cdot \mid \theta_2)$$

**When it fails:**

- **Mixture models:** swapping component labels gives the same distribution
- **Overparameterized models:** more parameters than the data can distinguish
- **Symmetric likelihoods:** multiple maxima, MLE is not unique

If the model isn't identifiable, no amount of data will help.

# Visualizing Non-Identifiability



# 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

Recall the **exponential family** form from Lecture 3: $f(x \mid \theta) = h(x) \exp\big(\eta(\theta)\, T(x) - A(\theta)\big)$

For $n$ i.i.d. observations, the log-likelihood is:

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

Setting the derivative to zero:

$$\eta'(\theta) \sum_{i=1}^{n} T(X_i) = nA'(\theta)$$

**For natural exponential families** ($\eta = \theta$):

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

The MLE is **always** a function of the sufficient statistic $\sum T(X_i)$,
and it **equals the MoM estimator** in natural form!

# 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 $n \to \infty$ (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 $n \to \infty$

**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 $n \to \infty$ (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 $n \to \infty$

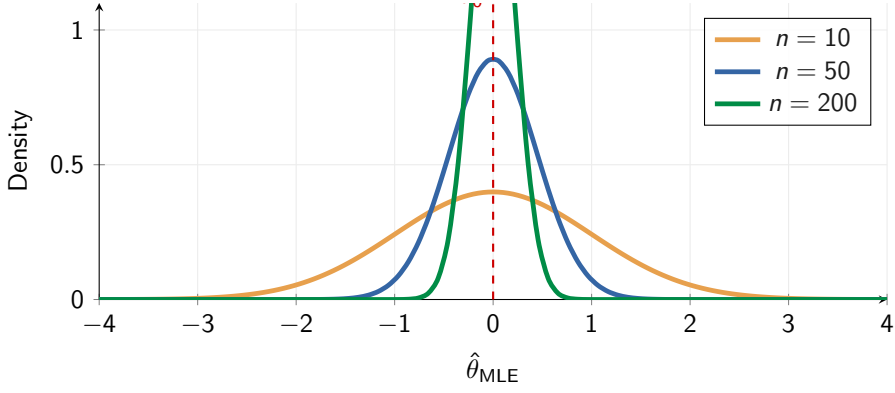
**4. Invariant:** MLE of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$ for any function $g$

**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.

# 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 $n \to \infty$.

# 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: bias $= -\theta/(n+1)$, not the usual $O(1/n)$.

# 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:** Uniform$(0, \theta) \Rightarrow \hat{\theta} = X_{(n)}$.
  Always underestimates: bias $= -\theta/(n+1)$, not the usual $O(1/n)$.
- **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

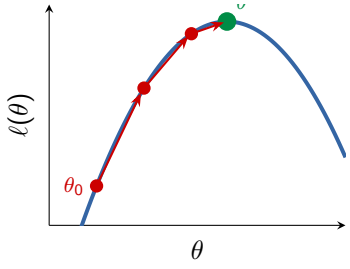
MLE has great asymptotic theory, but several things can go wrong:

- **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:** Uniform$(0, \theta) \Rightarrow \hat{\theta} = X_{(n)}$.
  Always underestimates: bias $= -\theta/(n+1)$, not the usual $O(1/n)$.
- **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`

# 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 $\hat{\sigma}^2_{\text{MLE}}$ (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 Geometric($p$): $f(x \mid p) = (1-p)^{x-1}p, \;\; x = 1, 2, \ldots$
   Is this MLE unbiased? Is it efficient (check against the CR bound)?

2. For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, show that $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2$
   equals the MoM estimator. Why is this not a coincidence? (Hint: exponential family.)

3. Show that the MLE for Uniform$(0, \theta)$ is $\hat{\theta} = X_{(n)} = \max(X_1, \ldots, X_n)$.
   Is this unbiased? Is it consistent? (Note: this is **not** an exponential family!)

4. Simulate $n = 50$ samples from Poisson($\lambda = 3$) and compute the MLE.
   Repeat 10,000 times. Verify: (a) $\hat{\lambda}$ is approximately unbiased, (b) $\text{Var}(\hat{\lambda}) \approx \lambda/n$.

# Questions?

Next: MAP estimation, priors, and the Bayesian perspective