Maximum Likelihood Estimation

Your first consulting job

- *Billionaire*: I have special coin, if I flip it, what's the probability it will be heads?
- *You*: Please flip it a few times:

flip $n=5$ times

HHTHT

- *You*: The probability is: $3/5$

$$\frac{k=3 \text{ heads}}{n=5 \text{ flips}}$$

- *Billionaire:* Why?

Coin – Binomial Distribution

- **Data**: sequence *D= (HHTHT ...)*, **k heads** out of **n flips**

Observed

- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ
  - Flips are i.i.d.: If $X_t \in \{0,1\}$ denoting tth flip
    - Independent events $\mathbb{P}(X_t = a, X_s = b) = \mathbb{P}(X_t = a)\,\mathbb{P}(X_s = b)$
    - Identically distributed according to Binomial distribution

$$\mathbb{P}(X_t = 0) = \mathbb{P}(X_s = 0) = 1 - \theta$$

$$\mathbb{P}(X_t = 1) = \theta$$

- $P(\mathcal{D}|\theta) = P(HHTHT|\theta)$

$$= P(H|\theta)\,P(H|\theta)\,P(T|\theta)\,P(H|\theta)\,P(T|\theta)$$

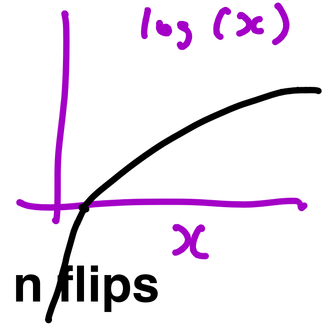
$$= \theta^3 (1-\theta)^2$$

In general,

$$P(D|\theta) = \theta^k (1-\theta)^{n-k}$$

Maximum Likelihood Estimation

$\mathbb{P}(\text{head}) = \theta^*$



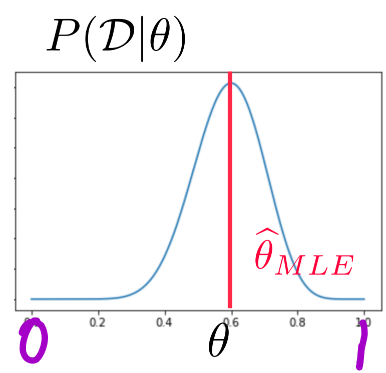
- **Data**: sequence *D= (HHTHT…)*, **k heads** out of **n flips**
- **Hypothesis:** P(Heads) = θ, P(Tails) = 1-θ

likelihood

$$\boxed{P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{n-k}}$$

- Maximum likelihood estimation (MLE): Choose θ that maximizes the probability of observed data:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(\mathcal{D}|\theta)$$
$$= \arg\max_{\theta} \; \log P(\mathcal{D}|\theta)$$

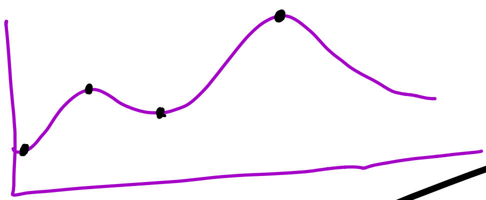


Your first learning algorithm

$\log(ab) = \log(a) + \log(b)$

for any $a, b > 0$

$$\widehat{\theta}_{MLE} := \arg\max_{\theta} \; \log\left(P(\mathcal{D}|\theta)\right)$$



$$= \arg\max_{\theta} \; \log\left(\theta^k (1-\theta)^{n-k}\right)$$

- Set derivative to zero: $\boxed{\dfrac{d}{d\theta} \log P(\mathcal{D}|\theta) = 0}$

$$= \arg\max_{\theta} \; k \log(\theta) + (n-k) \log(1-\theta)$$

$$\frac{\partial}{\partial\theta}[\,\cdot\,] = \frac{k}{\theta} + \frac{n-k}{1-\theta} \cdot (-1) = 0$$

(multiply $\theta(1-\theta)$ on both sides)

$$\rightarrow (1-\theta)k - \theta(n-k) = 0 \implies \widehat{\theta}_{MLE} = k/n$$

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{k}{n}$$

- *You*: flip the coin 5 times. *Billionaire*: I got 3 heads.

$$\widehat{\theta}_{MLE} = 3/5$$

- *You*: flip the coin 50 times. *Billionaire*: I got 20 heads.

$$\widehat{\theta}_{MLE} = 20/50 = 2/5$$

- *Billionaire:* Which one is right? Why?

Quantifying Uncertainty

- For **n flips** and **k heads** the MLE is **unbiased** for true θ*:

$$\widehat{\theta}_{MLE} = \frac{k}{n} \qquad \mathbb{E}[\widehat{\theta}_{MLE}] = \theta^*$$

- **Expectation** describes how the estimator behaves *on average.*

$$\underbrace{\hat{\theta}_{MLE} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{X_t = H\}}$$

$X_t \in \{H, T\}$

$$\mathbb{1}\{\Omega\} = \begin{cases} 1 & \text{if } \Omega = \text{true} \\ 0 & \text{o.w.} \end{cases}$$

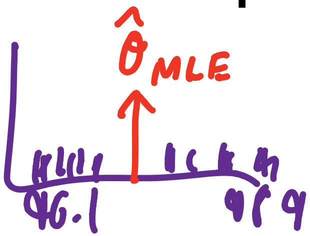
By assumption, $\exists\, \theta^* : \mathbb{P}(X_t = H) = \theta^*$

$$= \mathbb{E}[\mathbb{1}\{X_t = H\}]$$

$$\mathbb{E}[\hat{\theta}_{MLE}] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\mathbb{1}\{X_t = H\}] = \frac{1}{n}\sum_{t} \theta^* = \theta^*$$

Quantifying Uncertainty

- For **n flips** and **k heads** the MLE is **unbiased** for true θ*:



$$\widehat{\theta}_{MLE} = \frac{k}{n} \qquad \mathbb{E}[\widehat{\theta}_{MLE}] = \theta^*$$

Random Deterministic

- **Expectation** describes how the estimator behaves *on average.*
- The **Variance** is the expected squared deviation from the mean:

$$\text{Variance}(\widehat{\theta}_{MLE}) := \mathbb{E}\left[\left(\widehat{\theta}_{MLE} - \mathbb{E}[\widehat{\theta}_{MLE}]\right)^2\right]$$

- As a rule of thumb:

$$\widehat{\theta}_{MLE} \approx \mathbb{E}[\widehat{\theta}_{MLE}] \pm \sqrt{\text{Variance}(\widehat{\theta}_{MLE})}$$

- **Exercise**: compute the $\text{Variance}(\widehat{\theta}_{MLE})$

Expectation versus High Probability

- For **n flips** and **k heads** the MLE is **unbiased** for true θ^*:

$$\widehat{\theta}_{MLE} = \frac{k}{n} \qquad \mathbb{E}[\widehat{\theta}_{MLE}] = \theta^*$$

- Expectation describes how the estimator behaves *on average.*
- For any ε>0 can we bound $\mathbb{P}(|\widehat{\theta}_{MLE} - \mathbb{E}[\widehat{\theta}_{MLE}]| \geq \epsilon)$?

Markov's inequality
For any $t > 0$ and non-negative random variable X

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

- **Exercise**: Apply Markov's inequality to obtain bound.
(Hint: set $X = |\widehat{\theta}_{MLE} - \theta^*|^2$)

Maximum Likelihood Estimation

Each X_i is iid from $f(x;\theta)$

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

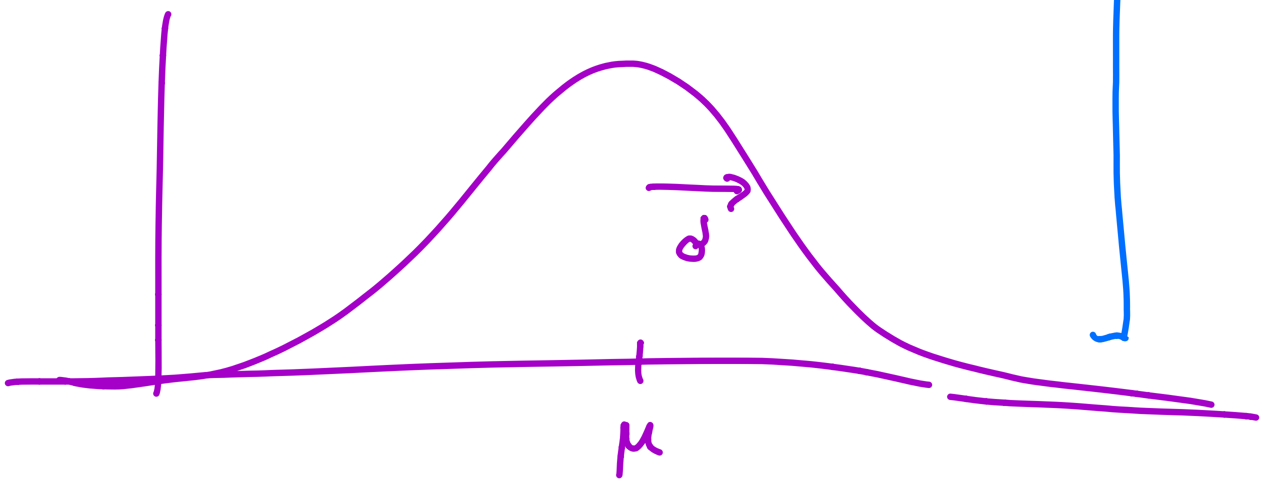
Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

What about continuous variables?

- *Billionaire*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Some properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \quad \rightarrow \quad Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow \quad Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

MLE for Gaussian

- Prob. of i.i.d. samples $\boxed{D=\{x_1,\ldots,x_n\}}$ (e.g., temperature):

$$P(\mathcal{D}|\mu,\sigma) = P(x_1,\ldots,x_n|\mu,\sigma) = \prod_{i=1}^{n} P(x_i|\mu,\sigma)$$

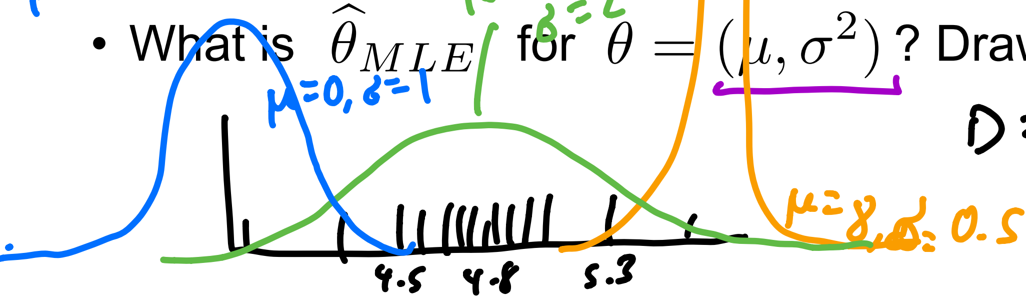
$$= \underbrace{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}$$

- Log-likelihood of data:

$$\log P(\mathcal{D}|\mu,\sigma) = -n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

- What is $\widehat{\theta}_{MLE}$ for $\theta = \underline{(\mu,\sigma^2)}$? Draw a picture!

$D = \{4.5, 5.3, 4.8, \ldots\ldots\}$



Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu} \log P(\mathcal{D}|\mu,\sigma) = \frac{d}{d\mu}\left[\underline{-n \log(\sigma\sqrt{2\pi})} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right]$$

$$= \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0$$

$$\Rightarrow \quad \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\log P(\mathcal{D}|\mu,\sigma) = \frac{d}{d\sigma}\left[\underline{-n\log(\sigma\sqrt{2\pi})} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -n \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \sqrt{2\pi} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2} \cdot (-2)\sigma^{-3} = 0$$

(multiply σ^3 on both sides after setting deriv. $=0$)

$$\rightarrow \quad -n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{MLE})^2$$

Learning Gaussian parameters

- MLE:

$$\left\{\begin{array}{l} \widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu}_{MLE})^2 \end{array}\right.$$

- MLE for the variance of a Gaussian is **biased** (Exercise)

$$\underbrace{\mathbb{E}[\widehat{\sigma^2}_{MLE}]} \neq \underline{\sigma}^2$$

- Unbiased variance estimator:

$$\widehat{\sigma^2}_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu}_{MLE})^2$$

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_{\theta} L_n(\theta)$

$X_i \sim \exp(\lambda)$

$P(X_i \geq t) = e^{-\lambda t}$

$\hat{\theta}_{MLE}$ is unbiased if

$\mathbb{E}[\hat{\theta}_{MLE}] = \theta^*$

Maximum Likelihood Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Likelihood function $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$

Log-Likelihood function $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i;\theta))$

Maximum Likelihood Estimator (MLE) $\widehat{\theta}_{MLE} = \arg\max_\theta L_n(\theta)$

$\sqrt{\text{Sample Variance}}$

Properties (under benign regularity conditions—smoothness, identifiability, etc.):

- Asymptotically consistent and normal: $\frac{\widehat{\theta}_{MLE} - \theta_*}{\widehat{se}} \sim \mathcal{N}(0,1)$
- Asymptotic Optimality, minimum variance (see Cramer-Rao lower bound)

Recap

- Learning is…
  - Collect some data
    - E.g., coin flips
  - Choose a hypothesis class or model
    - E.g., binomial
  - Choose a loss function
    - E.g., data likelihood
  - Choose an optimization procedure
    - E.g., set derivative to zero to obtain MLE
  - Justifying the accuracy of the estimate
    - E.g., Markov's inequality