7.1, 7.2 Maximum Likelihood Estimation

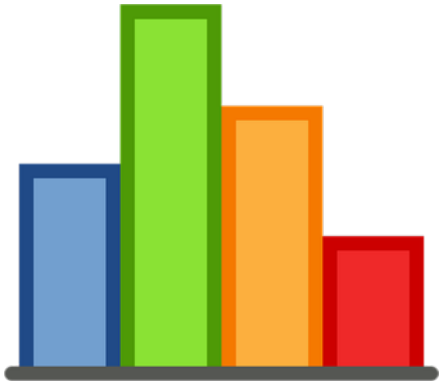
Anna Karlin
Most slides by Alex Tsun

Agenda

- Probability vs Statistics
- Likelihood
- Maximum Likelihood Estimation (MLE)
- MLE Example (Poisson)
- MLE Example (Normal)

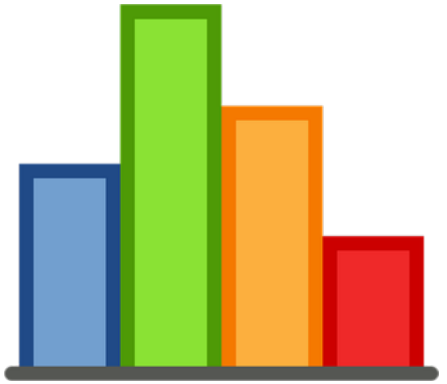
Probability vs Statistics

$Ber(p = 0.5)$ → **Probability**
given model, predict data → $P(THHTHH)$



Probability vs Statistics

$Ber(p = 0.5)$ → **Probability** given model, predict data → $P(THHTHH)$



$Ber(p = ???)$ ← **Statistics** given data, predict model ← $THHTHH$

Random Picture

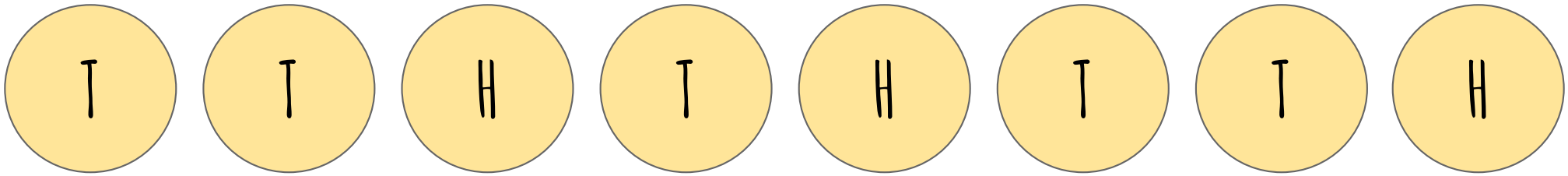
Likelihood (Intuition)

I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

Likelihood (Intuition)

I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

Flip it as many times as possible, and return # H / (# H + # T)!

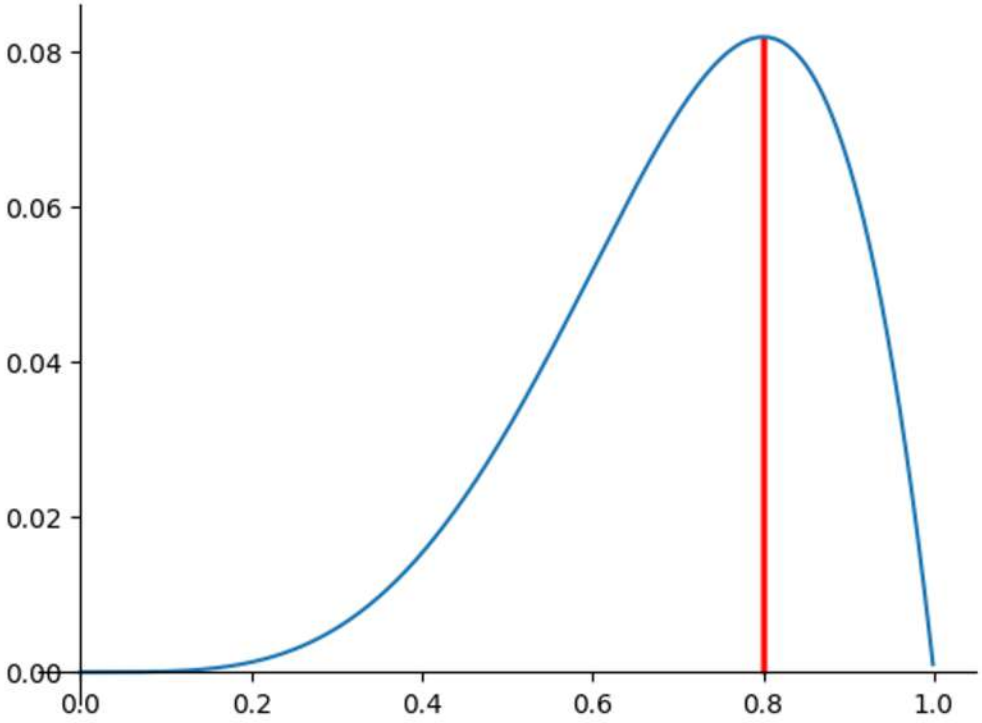


Likelihood (Intuition)

Let's say you saw 4 heads and 1 tail. You tell me $\hat{p} = \frac{4}{5}$.

How can you argue, ***objectively***, that this is the "best" estimate?

Is there some objective function it maximizes?



Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \ldots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$L(\boldsymbol{x}|\theta) = P(\text{seeing data} \mid \theta)$$

Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \dots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$\begin{aligned} L(\boldsymbol{x}|\theta) &= P(\text{seeing data} \mid \theta) \\ &= P(x_1, \dots, x_n \mid \theta) \end{aligned}$$

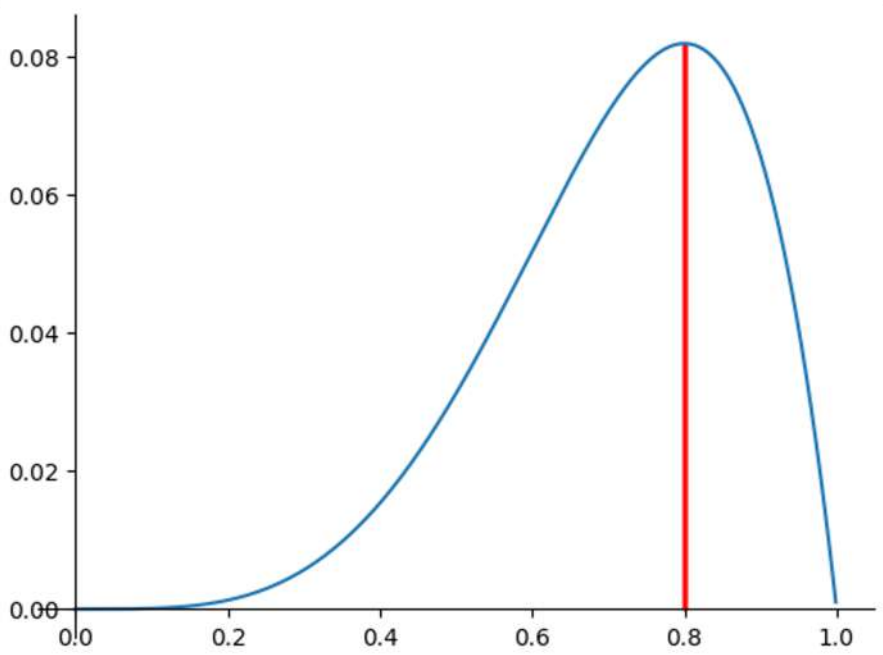
Likelihood (Intuition)

You assume a model (Bernoulli in our case) with unknown parameter θ, and receive iid samples $\boldsymbol{x} = (x_1, \dots, x_n) \sim Ber(\theta)$. The **likelihood** of the data given a parameter θ is

$$\begin{aligned} L(\boldsymbol{x}|\theta) &= P(\text{seeing data} \mid \theta) \\ &= P(x_1, \dots, x_n \mid \theta) \\ &= \prod_{i=1}^{n} p_X(x_i; \theta) \end{aligned}$$

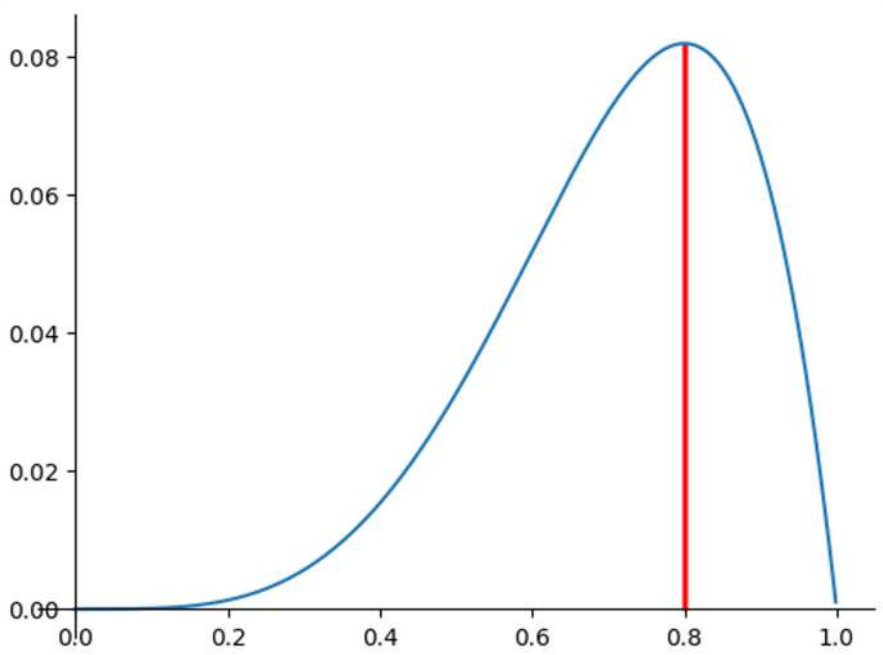
Maximum Likelihood Estimation (Bernoulli)

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$



Maximum Likelihood Estimation (Bernoulli)

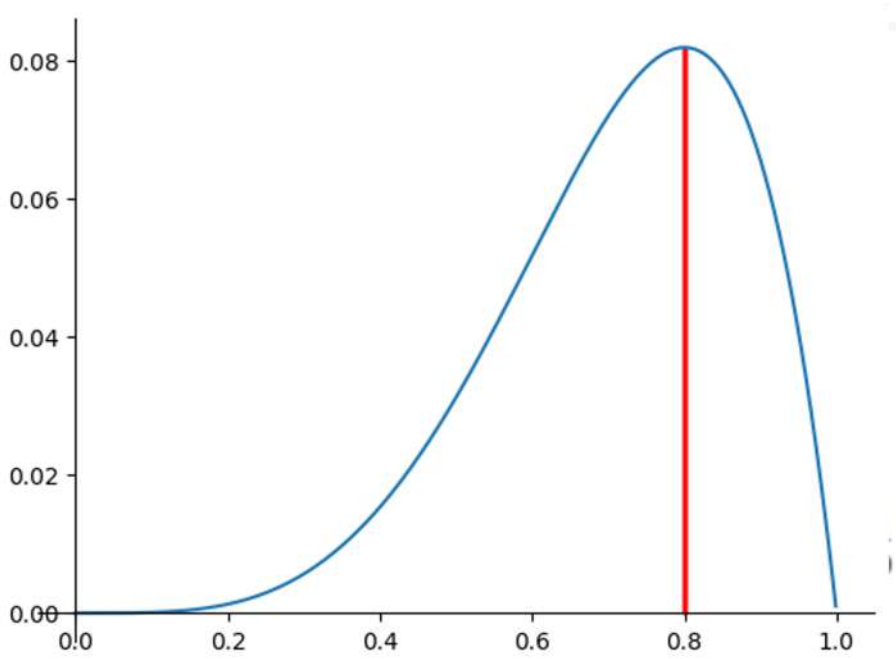
$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$



Maximum Likelihood Estimation (Bernoulli)

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

$$\frac{\partial}{\partial\theta} L(x|\theta) = 4\theta^3 - 5\theta^4 = \theta^3(4-5\theta)$$

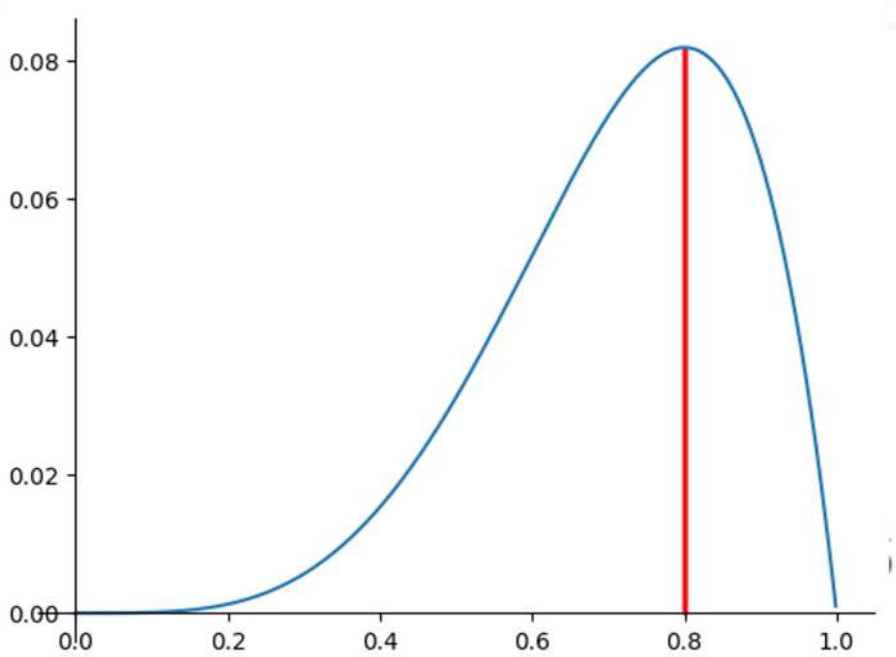


Maximum Likelihood Estimation (Bernoulli)

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

$$\frac{\partial}{\partial\theta} L(x|\theta) = 4\theta^3 - 5\theta^4 = \theta^3(4-5\theta)$$

$$\hat{\theta}^3\left(4 - 5\hat{\theta}\right) = 0 \rightarrow$$

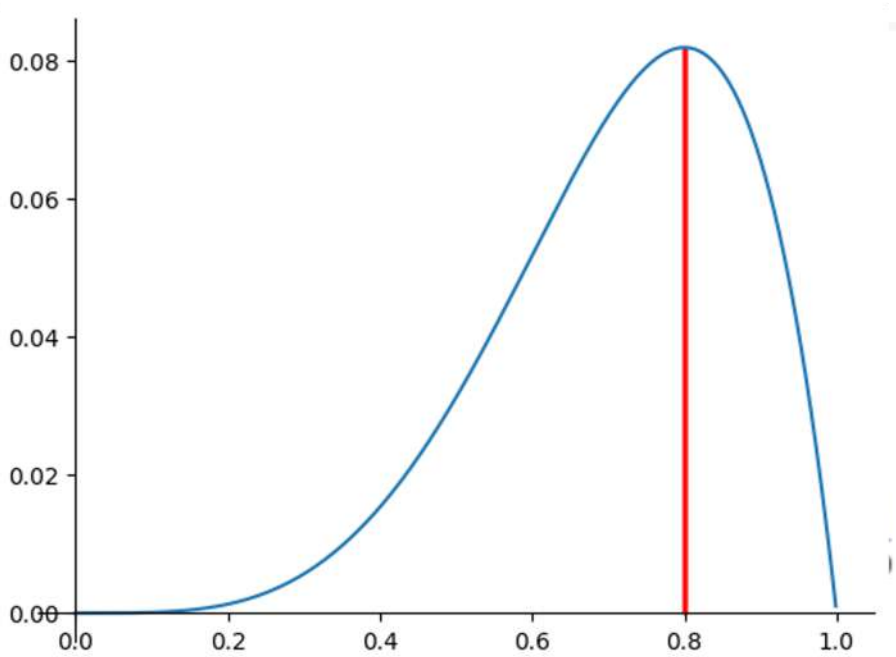


Maximum Likelihood Estimation (Bernoulli)

$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

$$\frac{\partial}{\partial\theta} L(x|\theta) = 4\theta^3 - 5\theta^4 = \theta^3(4-5\theta)$$

$$\hat{\theta}^3\left(4-5\hat{\theta}\right) = 0 \rightarrow \hat{\theta} = \frac{4}{5} \text{ or } 0$$

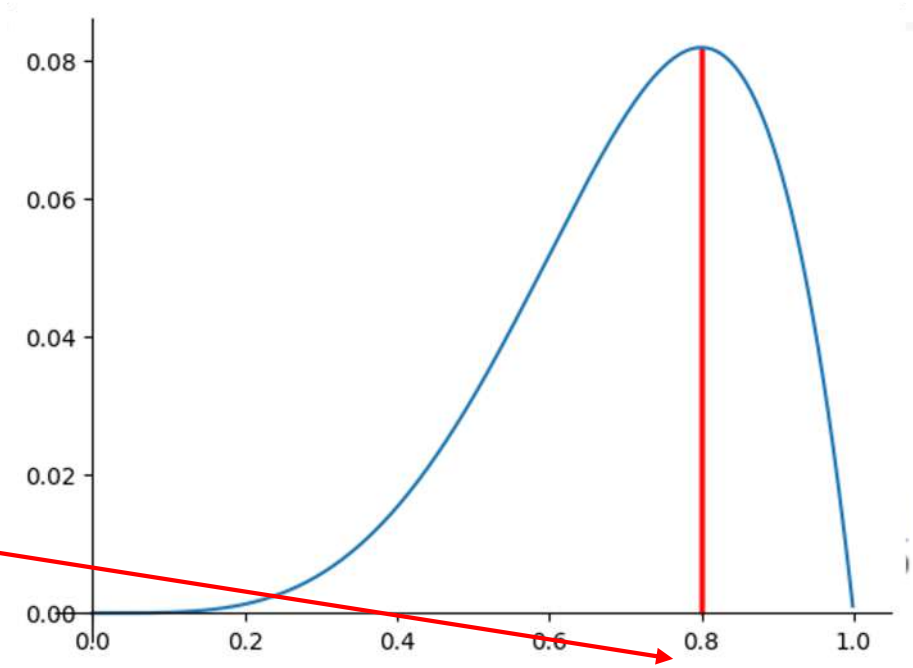


Maximum Likelihood Estimation (Bernoulli)

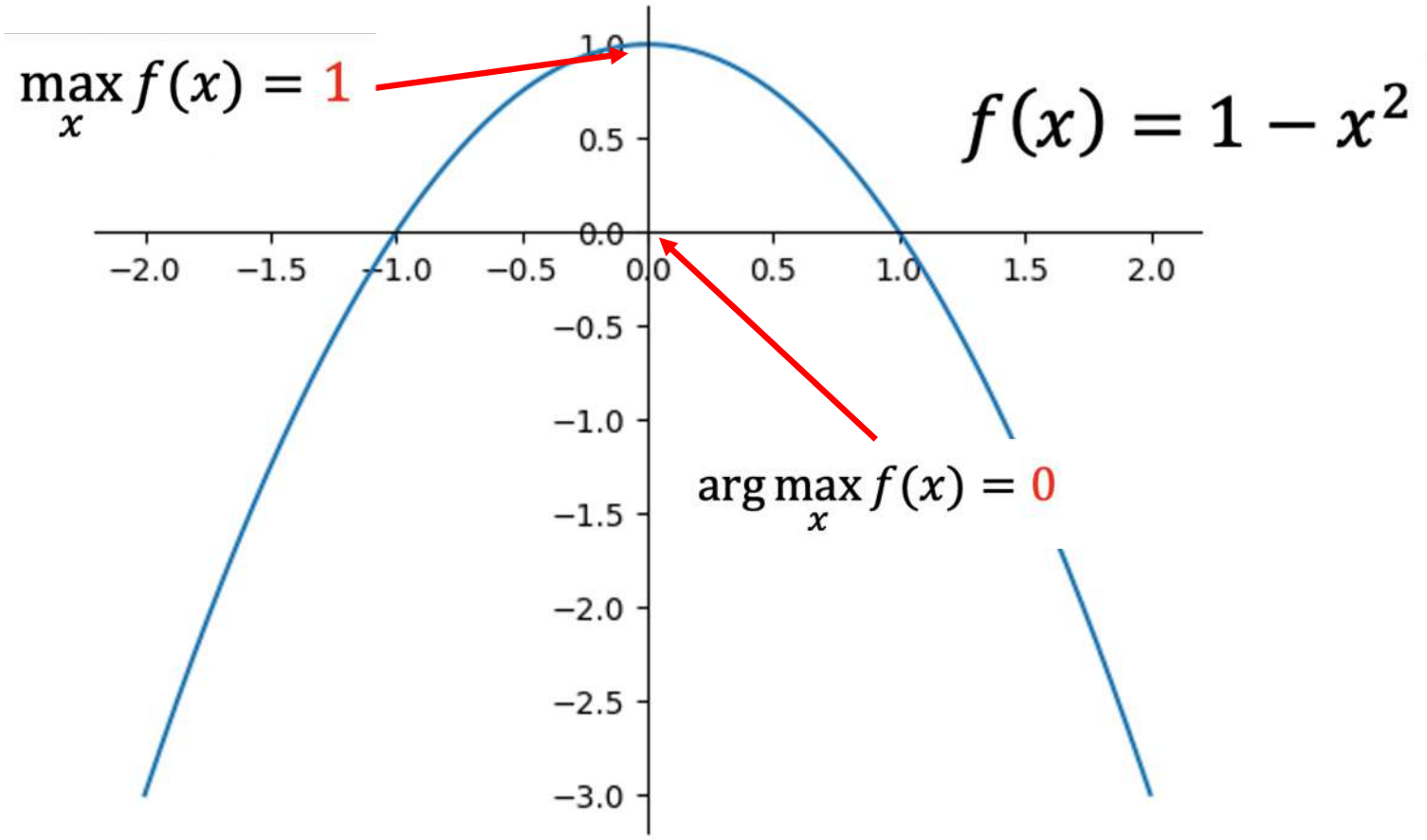
$$L(HHHHT \mid \theta) = \theta^4(1-\theta) = \theta^4 - \theta^5$$

$$\frac{\partial}{\partial\theta} L(x|\theta) = 4\theta^3 - 5\theta^4 = \theta^3(4-5\theta)$$

$$\hat{\theta}^3(4-5\hat{\theta}) = 0 \rightarrow \hat{\theta} = \boxed{\frac{4}{5}} \text{ or } 0$$



Max vs Argmax



Maximum Likelihood Estimation (Poisson)

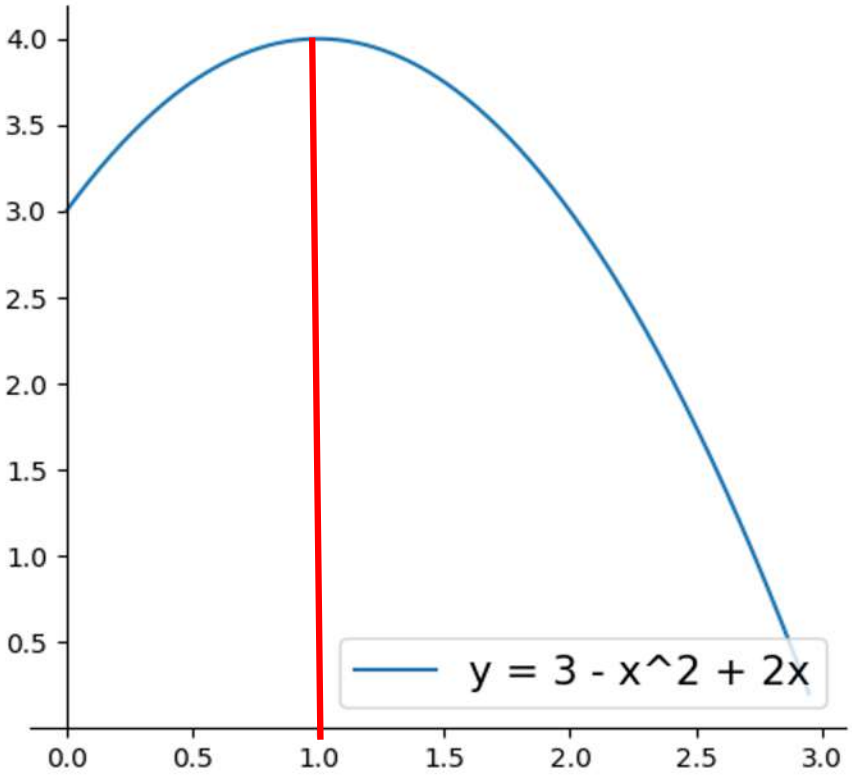
Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

Maximum Likelihood Estimation (Poisson)

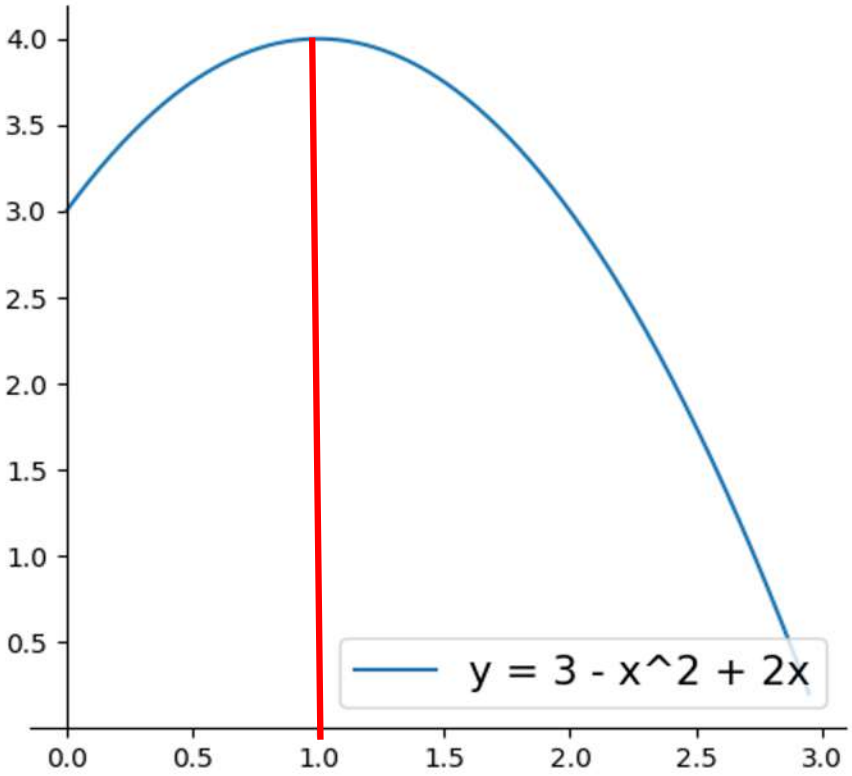
Let's say $x_1, x_2, \dots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

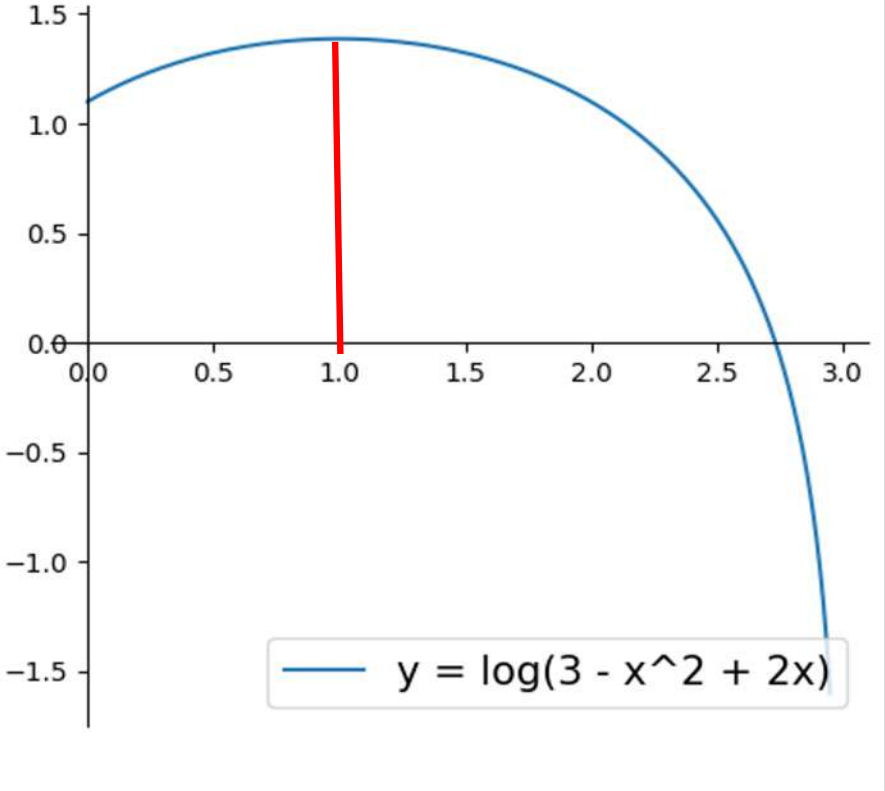
$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

Optimizing Function vs Log(Function)

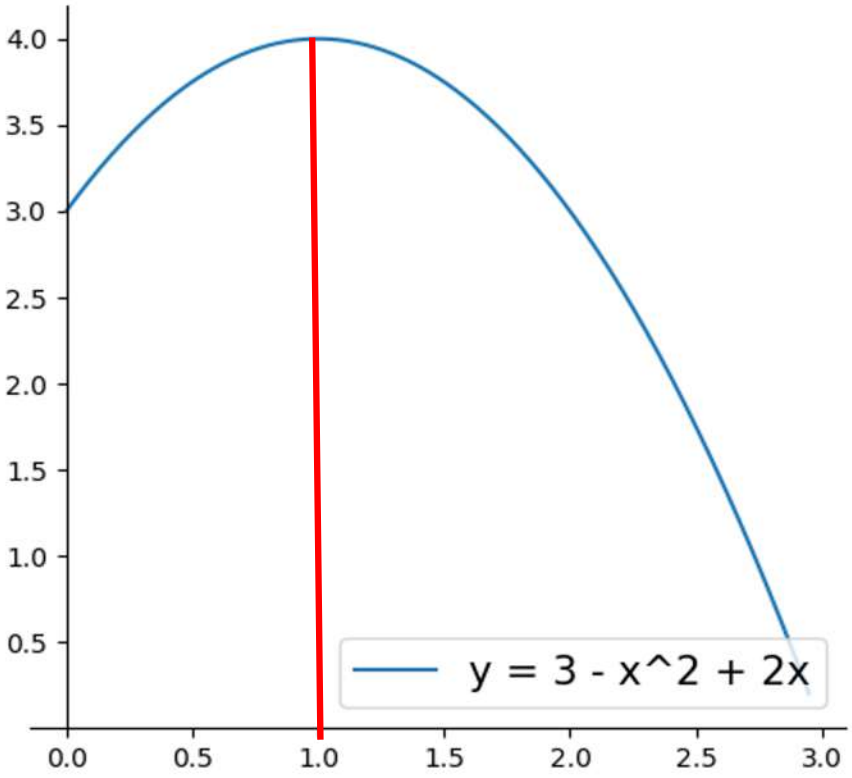


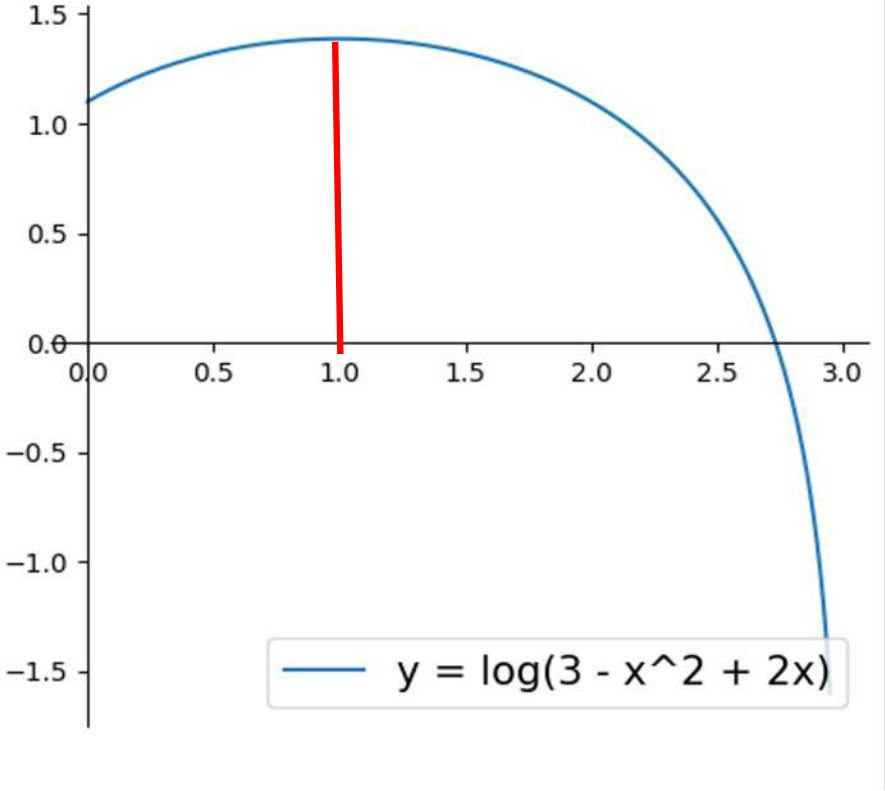
Optimizing Function vs Log(Function)





Optimizing Function vs Log(Function)





Since $g(x) = \log x$ **is strictly increasing**, it preserves order, and in particular, the argmax.

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

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

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a^b) = b \log a$$

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \dots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

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

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a^b) = b \log a$$

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} [-\theta + x_i \ln \theta - \ln(x_i!)]$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} [-\theta + x_i \ln\theta - \ln(x_i!)]$$

$$\frac{\partial}{\partial\theta}\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}\left[-1 + \frac{x_i}{\theta}\right]$$

$$\sum_{i=1}^{n}\left[-1 + \frac{x_i}{\hat{\theta}}\right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximum Likelihood Estimation (Poisson)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $Poi(\theta)$. (might look like $x_1 = 3, x_2 = 5, x_3 = 4$, etc.) What is the MLE of θ?

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!}$$

$$\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n} [-\theta + x_i \ln\theta - \ln(x_i!)]$$

$$\frac{\partial}{\partial\theta}\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}\left[-1 + \frac{x_i}{\theta}\right]$$

$$\sum_{i=1}^{n}\left[-1 + \frac{x_i}{\hat{\theta}}\right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}}\sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial^2}{\partial\theta^2}\ln L(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{n}\left[-\frac{x_i}{\theta^2}\right] < 0 \rightarrow \text{concave down everywhere}$$

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood: Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x as the probability of seeing the data.

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood: Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x as the probability of seeing the data.

If X is discrete,

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta)$$

Likelihood

Realization/Sample: A realization/sample x of a random variable X is the value that is actually observed.

Likelihood: Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the likelihood of x as the probability of seeing the data.

If X is discrete,

$$L(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta)$$

Maximum Likelihood Estimation (MLE)

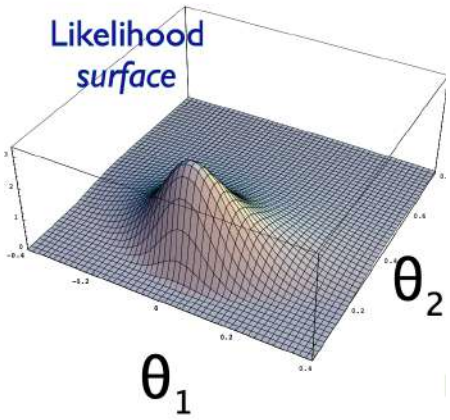
Maximum Likelihood Estimation (MLE): Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,;\theta)$ (if X discrete), or from density $f_X(t\,;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters). We define the maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ to be the parameter which maximizes the likelihood (or equivalently, the log-likelihood).

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{x} \mid \theta) = \arg\max_{\theta} \ln L(\boldsymbol{x} \mid \theta)$$

Random Picture

Maximum Likelihood Estimation (Normal)

$x_i \sim N(\mu, \sigma^2),\ \mu, \sigma^2$ both unknown



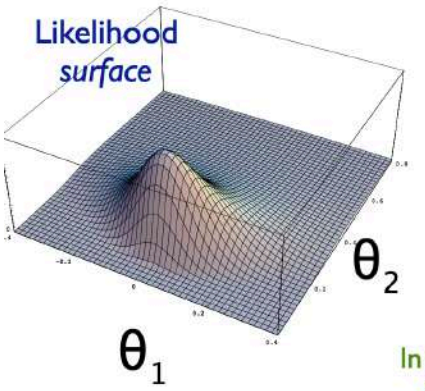
Maximum Likelihood Estimation (Normal)

$x_i \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown

$$\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_1}\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{i=1}^{n} \frac{(x_i - \theta_1)}{\theta_2} = 0$$

$$\widehat{\theta_1} = \left(\sum_{i=1}^{n} x_i\right)/n = \overline{x}$$



Sample mean is MLE of population mean, again

In general, a problem like this results in 2 equations in 2 unknowns. Easy in this case, since θ_2 drops out of the $\partial/\partial\theta_1 = 0$ equation

$x_i \sim N(\mu, \sigma^2),\ \mu, \sigma^2$ both unknown

$$\ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n} -\frac{1}{2}\frac{2\pi}{2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2} = 0$$

$$\widehat{\theta}_2 = \left(\sum_{i=1}^{n}(x_i - \widehat{\theta}_1)^2\right)/n = \overline{s}^2$$

Sample variance is MLE of *population* variance