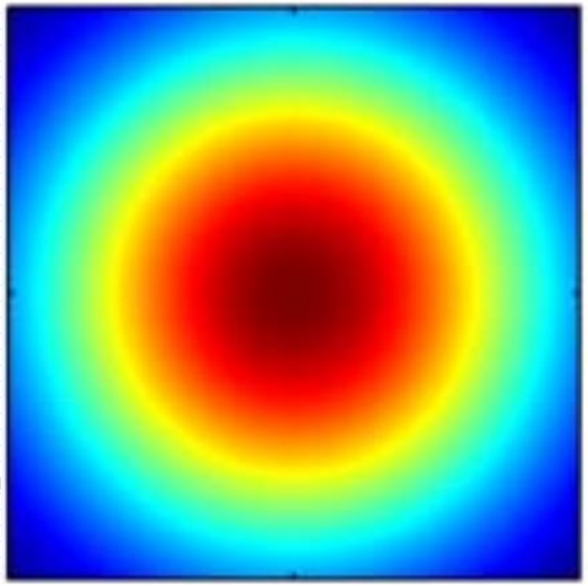
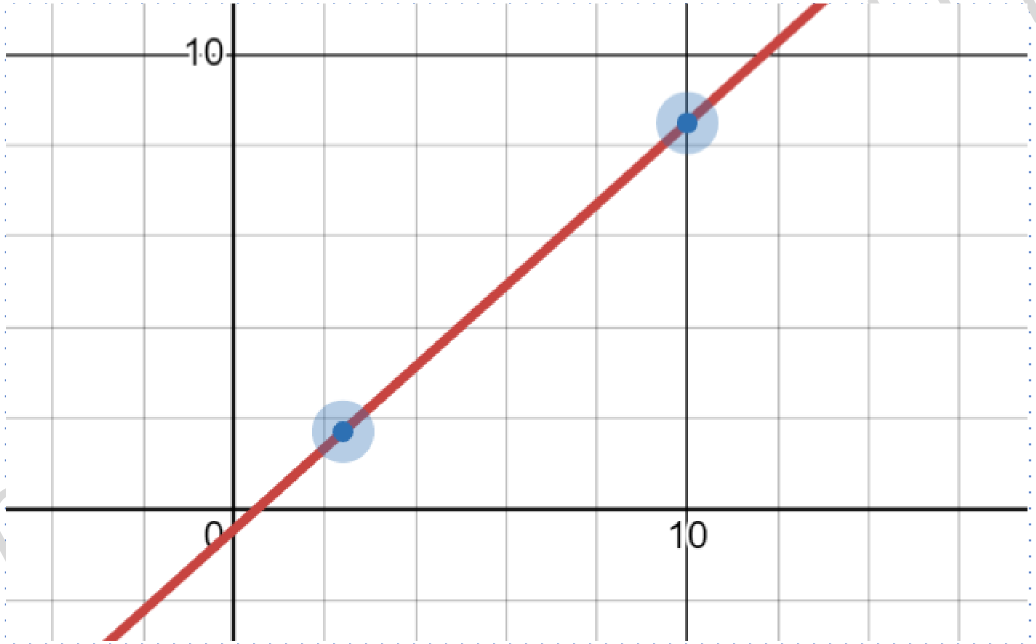
MLE vs MAP
Estimation



Prof. Arun Chauhan
Graphic Era University Dehradun

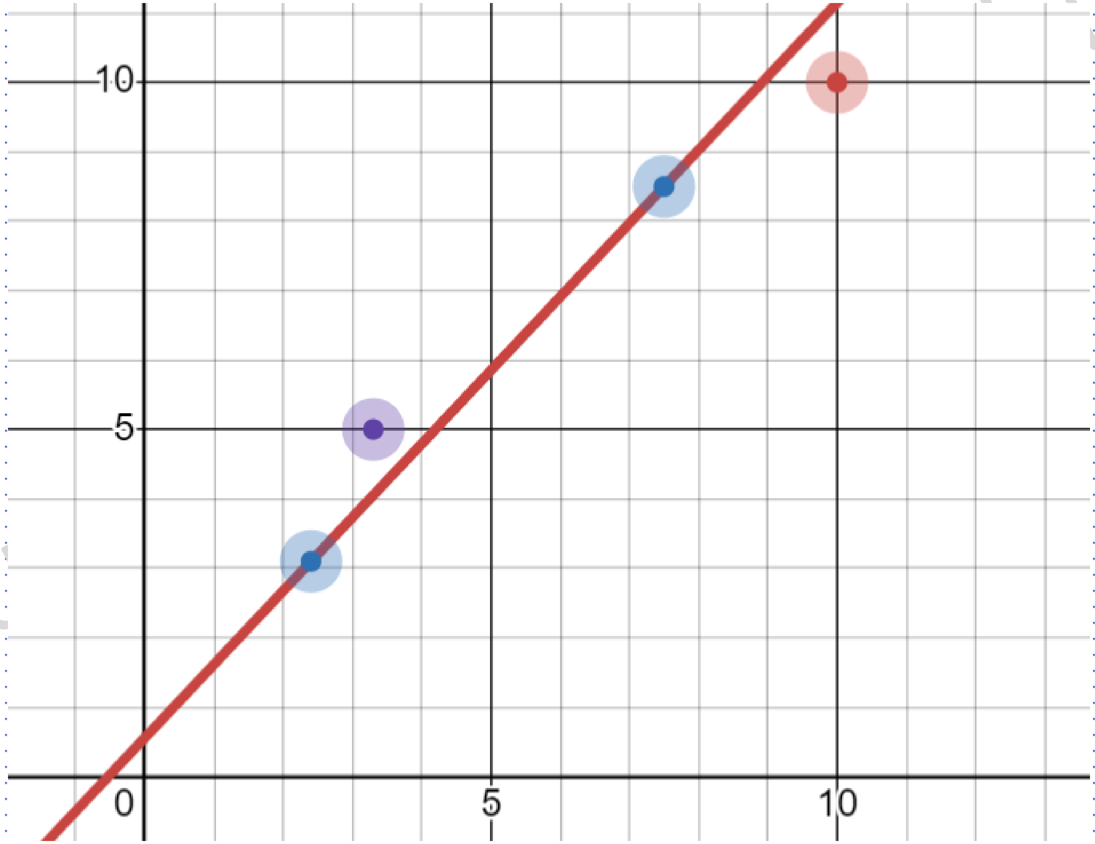
Learning target function from given data set D

$f : X \rightarrow Y$



Learning probabilistic function from noisy data

P(Y|X)



Two most common approaches to estimate P(Y|X)

Estimating Probabilities

Two Intuitive Algorithms

- **Algorithm 1**
- **Algorithm 2**

Problem at Hand

1

0

Estimate the Probability of Coin

Algorithm 1/ Algorithm 2

$\theta \longrightarrow \hat{\theta}$

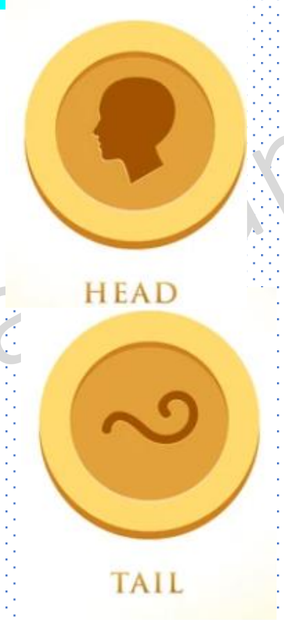
How to estimate the Probability of Coin?

Defining Mathematical Model of the problem at hand

Binary Random Variable

$X = 1$

$X = 0$



θ = True Probability

$\hat{\theta}$ = Estimated Probability

α_1 = # of Heads

α_2 = # of Tails

Algorithm 1

$$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

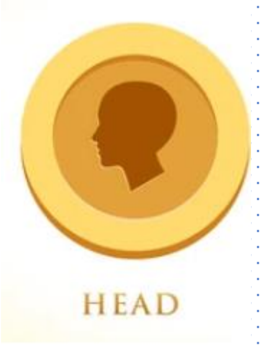
Examples:

$$\hat{\theta} = \frac{24}{24 + 26}$$

$$\hat{\theta} = \frac{1}{1 + 2}$$

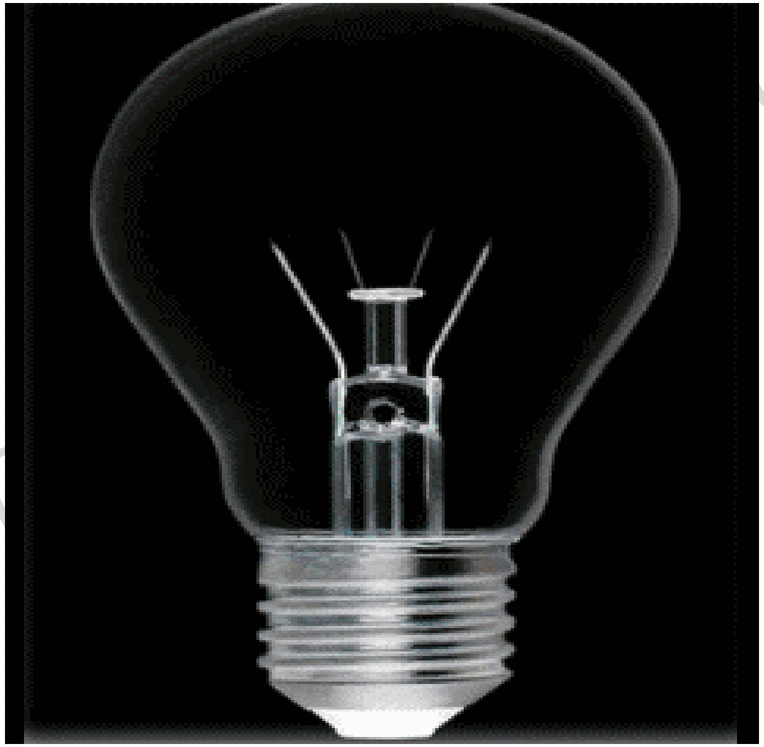
Limitation of Algorithm 1

Scarcity of the DATA



$$\hat{\theta} = \frac{1}{1+0}$$

What of we have prior knowledge ?



If the coin is government minted ?

Algorithm 2

Allow us to incorporate our prior knowledge

Re-Defining Mathematical Model of the problem at hand

Binary Random Variable

X = 1

X = 0

θ = True Probability

$\hat{\theta}$ = Estimated Probability

α_1 = # of Heads

α_2 = # of Tails

γ_1 = # of Imaginary Heads

γ_2 = # of Imaginary Tails

Algorithm 2

$$\hat{\theta} = \frac{\alpha_1 + \gamma_1}{\alpha_1 + \gamma_1 + \alpha_2 + \gamma_2}$$

Examples: $Let \;\; \gamma_1 = \gamma_2 = 100$

$$\hat{\theta} = \frac{24 + 100}{24 + 100 + 26 + 100}$$

$$\hat{\theta} = \frac{1 + 100}{1 + 100 + 2 + 100}$$

Estimating Probabilities

Algorithm 1

Estimate of $\hat{\theta}$ that maximizes the probability of the observed data.

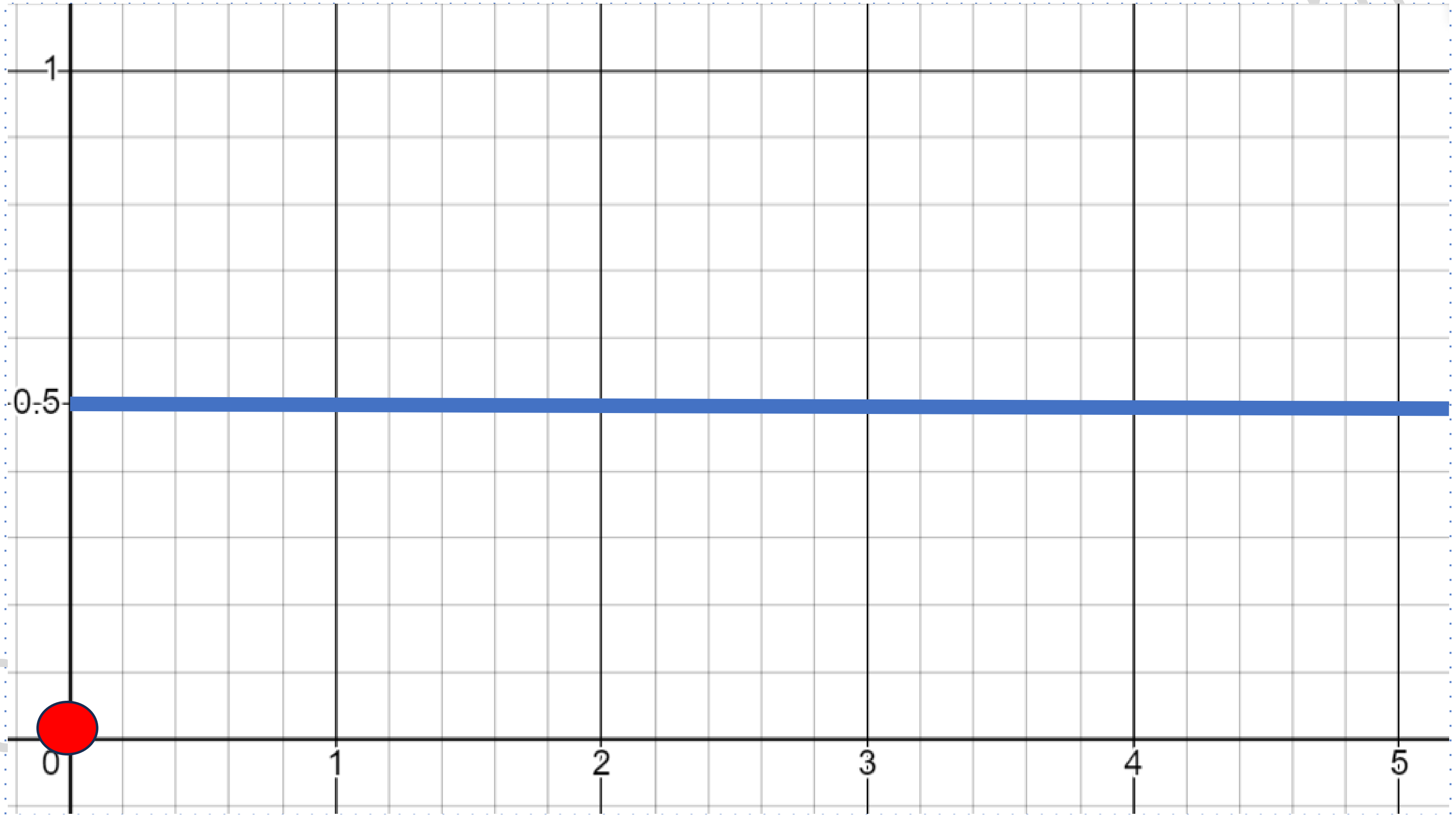
Vs

Algorithm 2

Estimate of $\hat{\theta}$ that is most probable, given observed data plus assumption

Maximum Likelihood Estimation

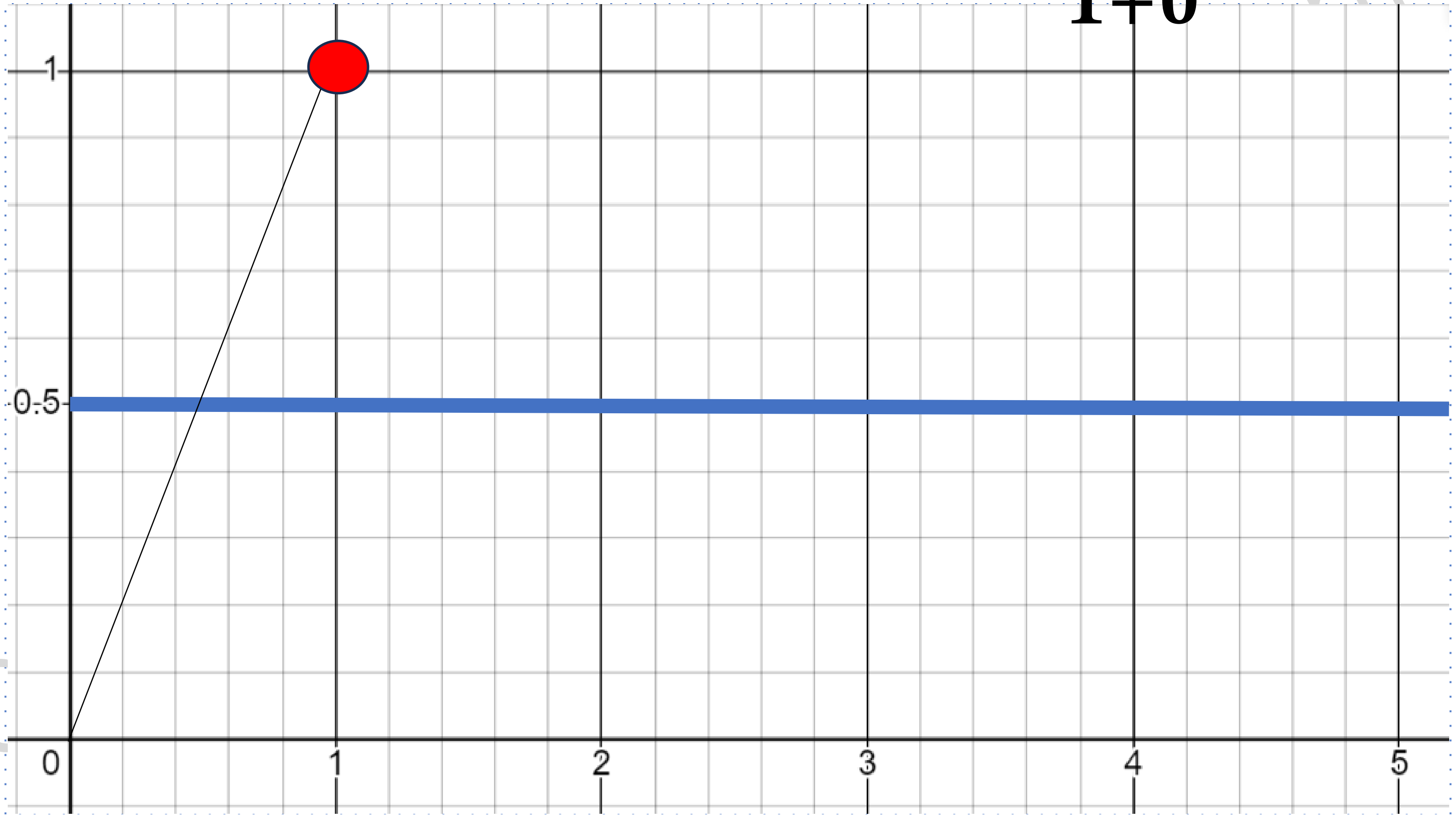
Trail = Not Available $\hat{\theta} = 0$



Maximum Likelihood Estimation

Trail = H

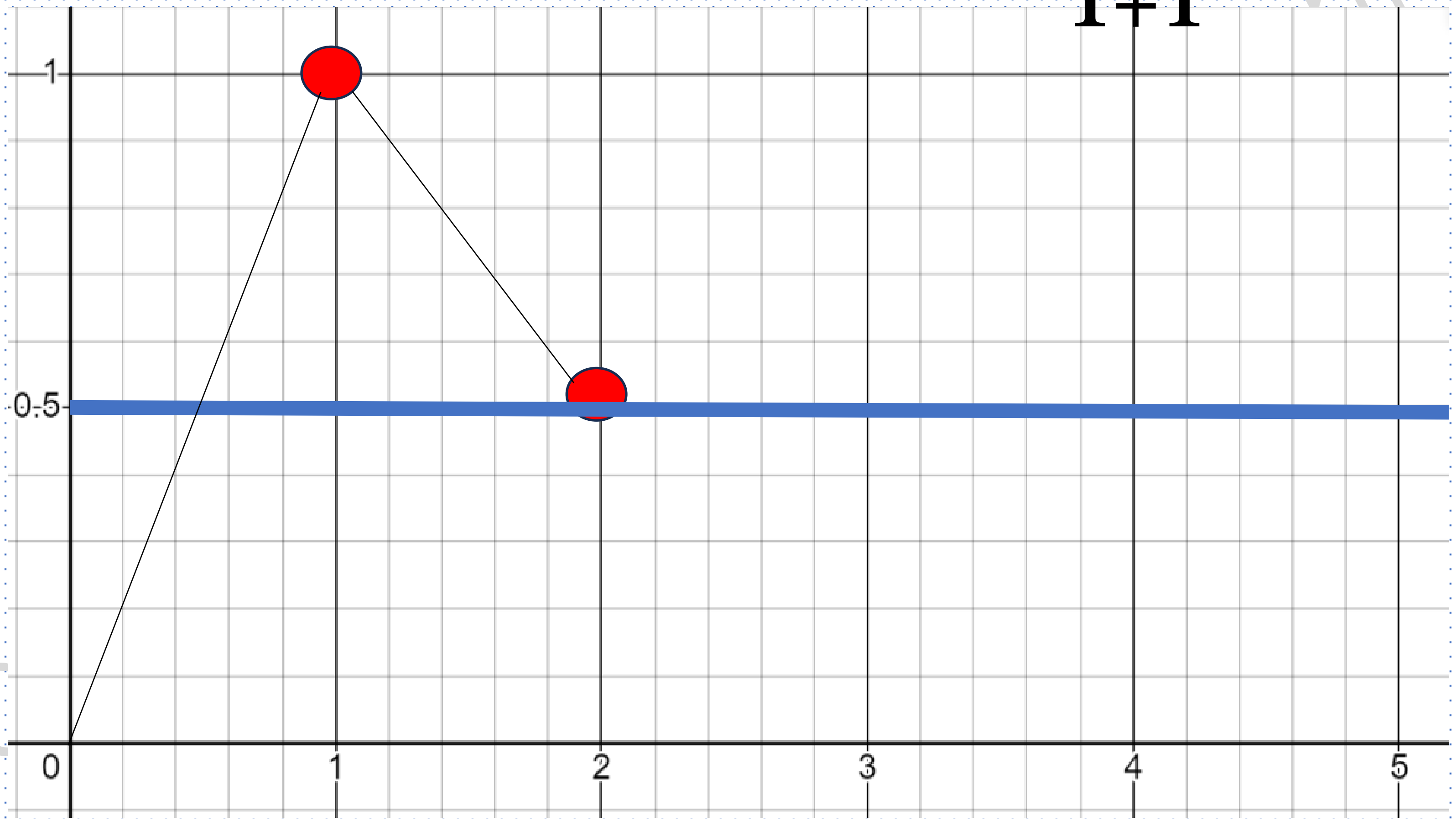
$$\hat{\theta} = \frac{1}{1+0} = 1$$



Maximum Likelihood Estimation

Trail = H T

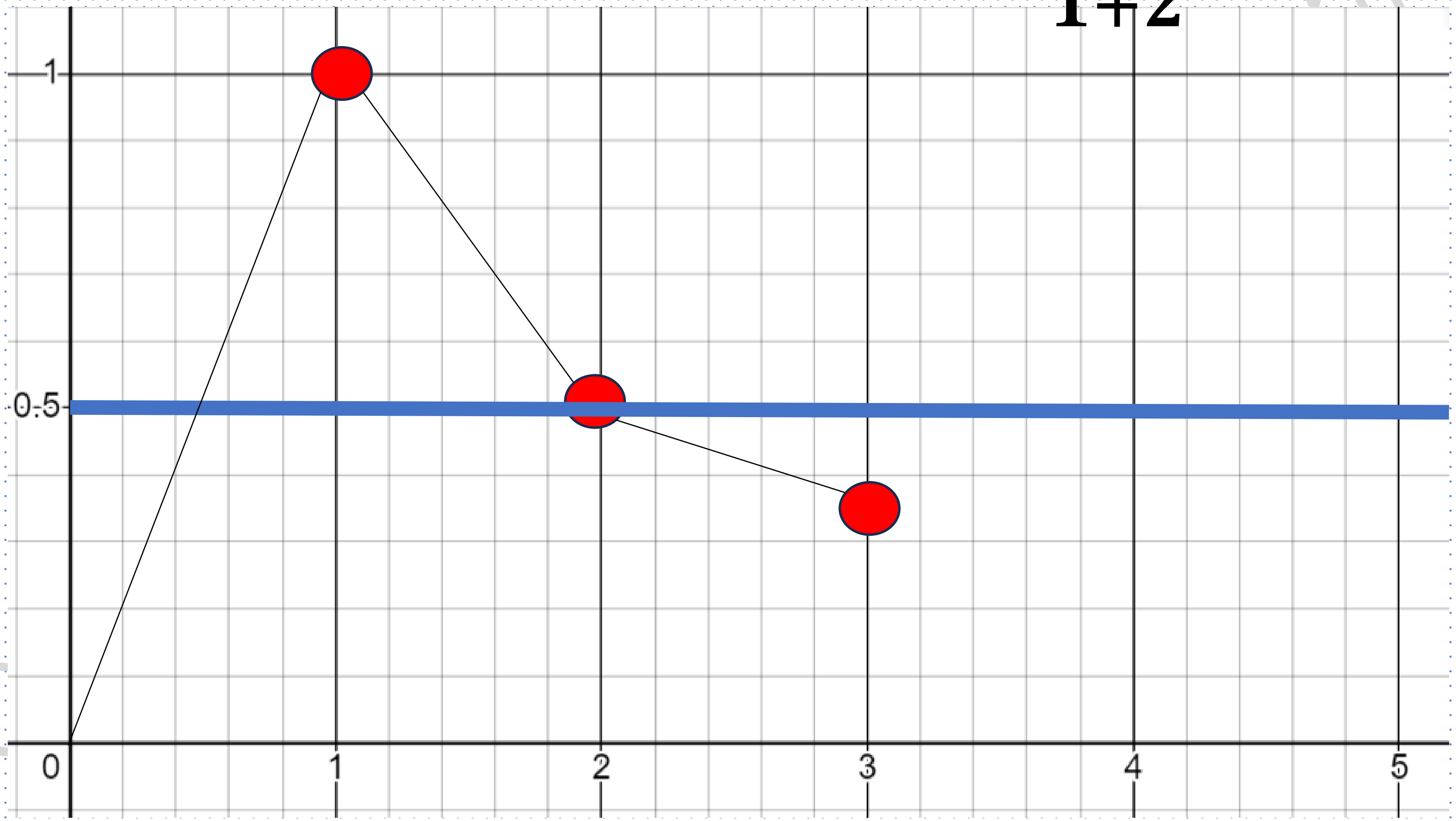
$$\hat{\theta} = \frac{1}{1+1} = 0.5$$



Maximum Likelihood Estimation

Trail = H T T

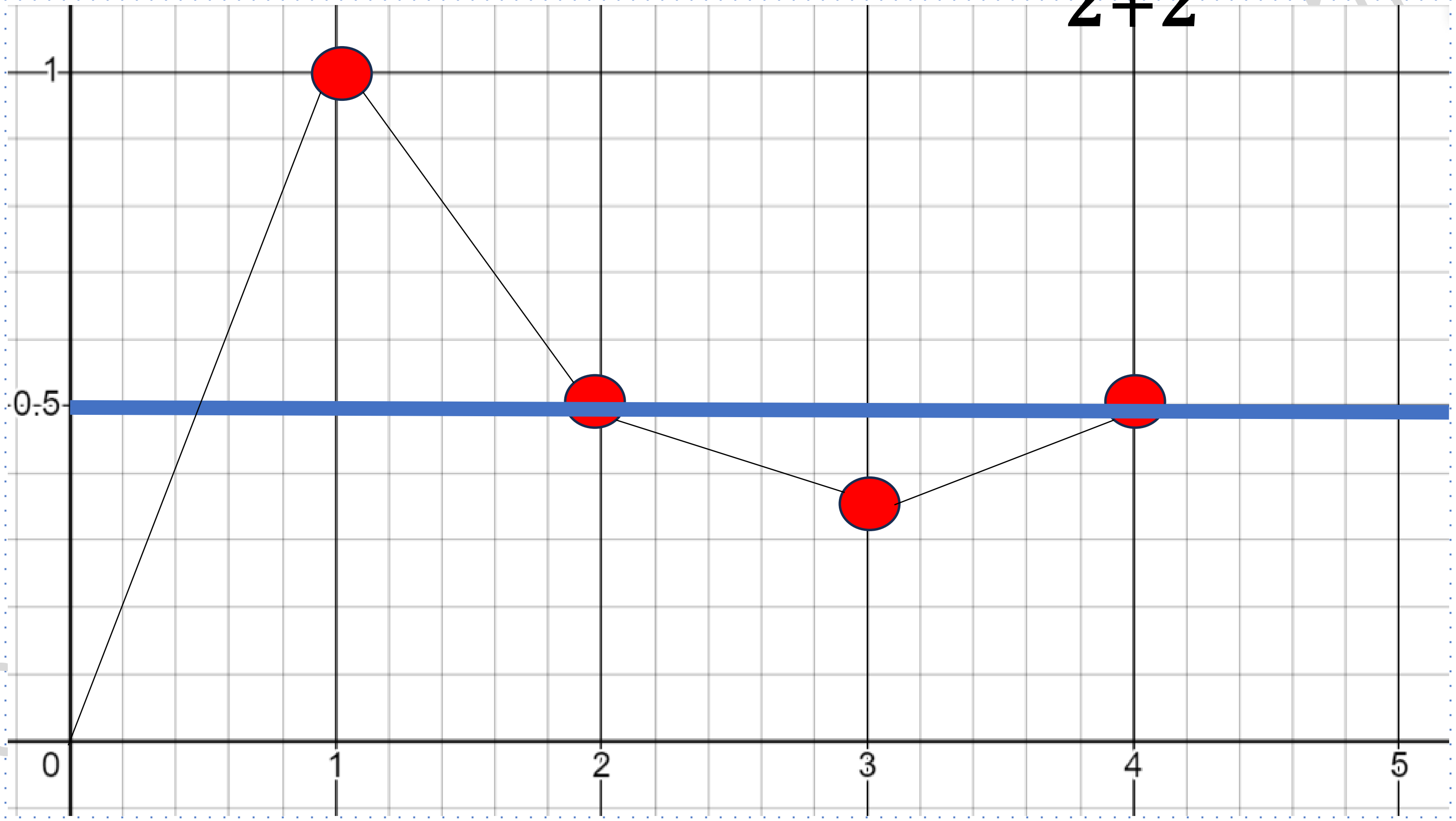
$$\hat{\theta} = \frac{1}{1+2} = 0.33$$



Maximum Likelihood Estimation

Trail = H T T H

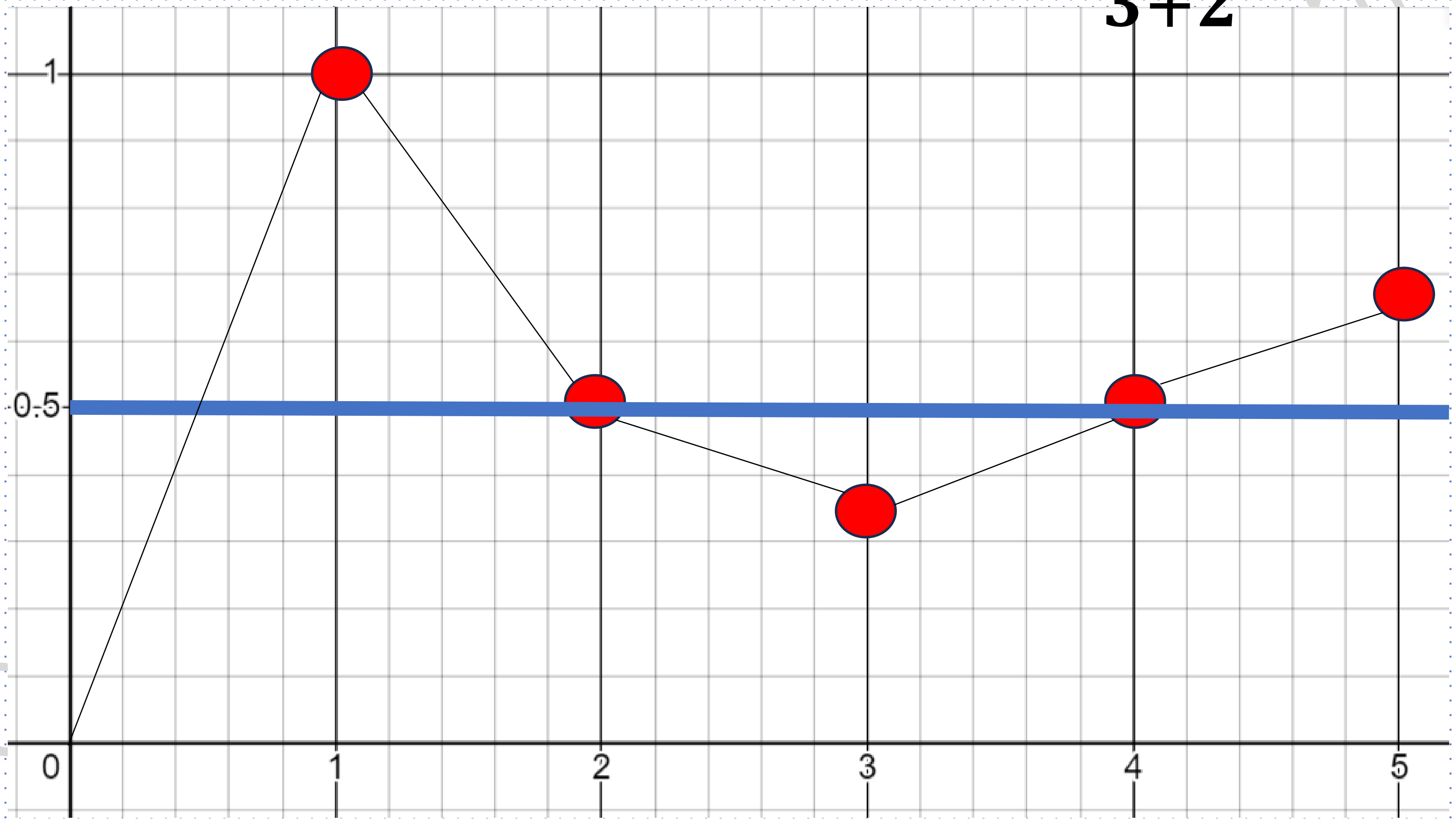
$$\hat{\theta} = \frac{2}{2+2} = 0.5$$



Maximum Likelihood Estimation

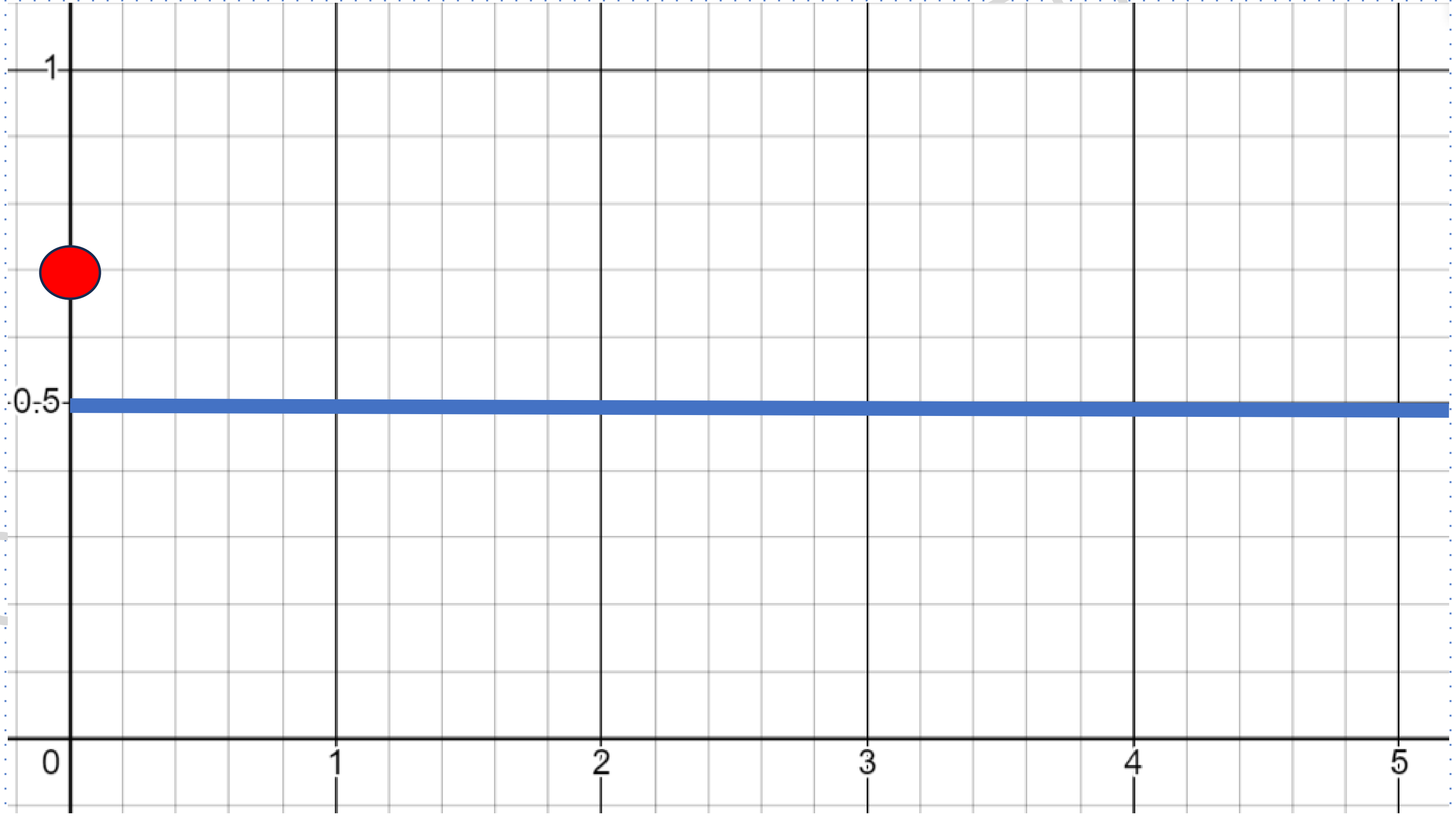
Trail = H T T H H

$$\hat{\theta} = \frac{3}{3+2} = 0.66$$



Maximum a Posteriori (MAP)

Trail = Not Available $\hat{\theta} = \frac{7}{7+3} = 0.7$ $\gamma_1 = 7$ $\gamma_2 = 3$

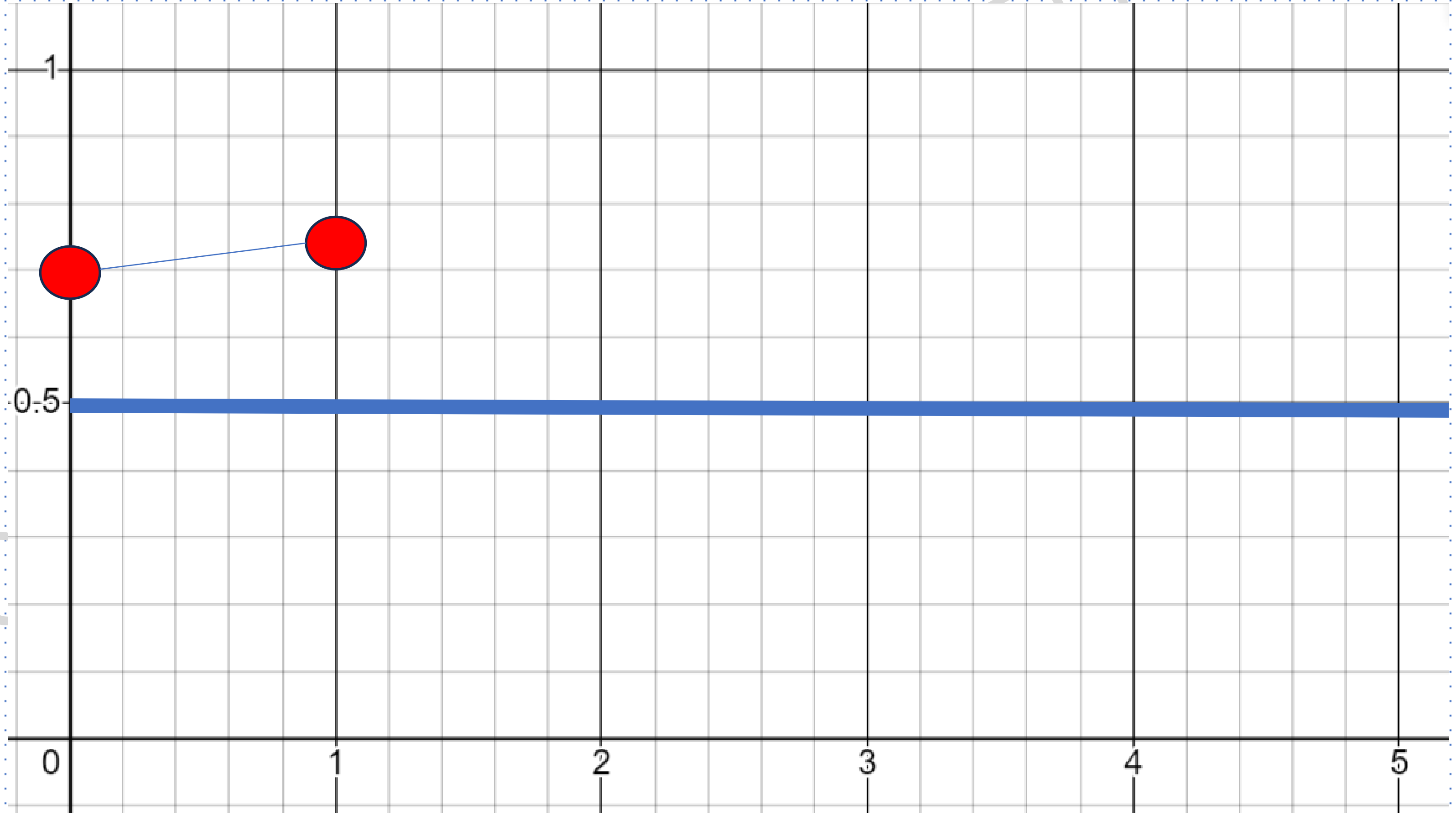


Maximum a Posteriori (MAP)

Trail = H

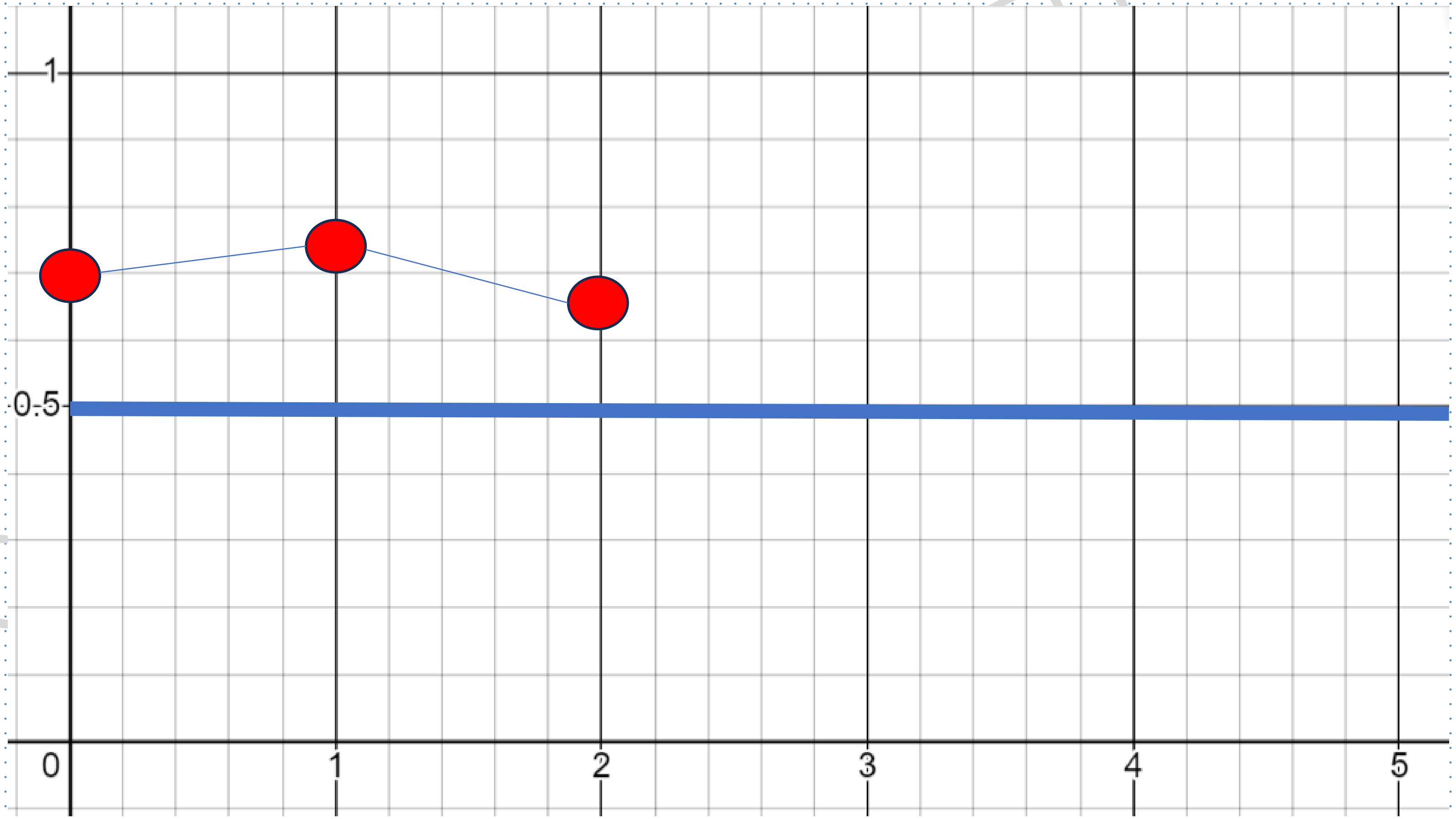
$$\hat{\theta} = \frac{7+1}{7+1+3} = 0.72$$

$\gamma_1 = 7$ $\gamma_2 = 3$



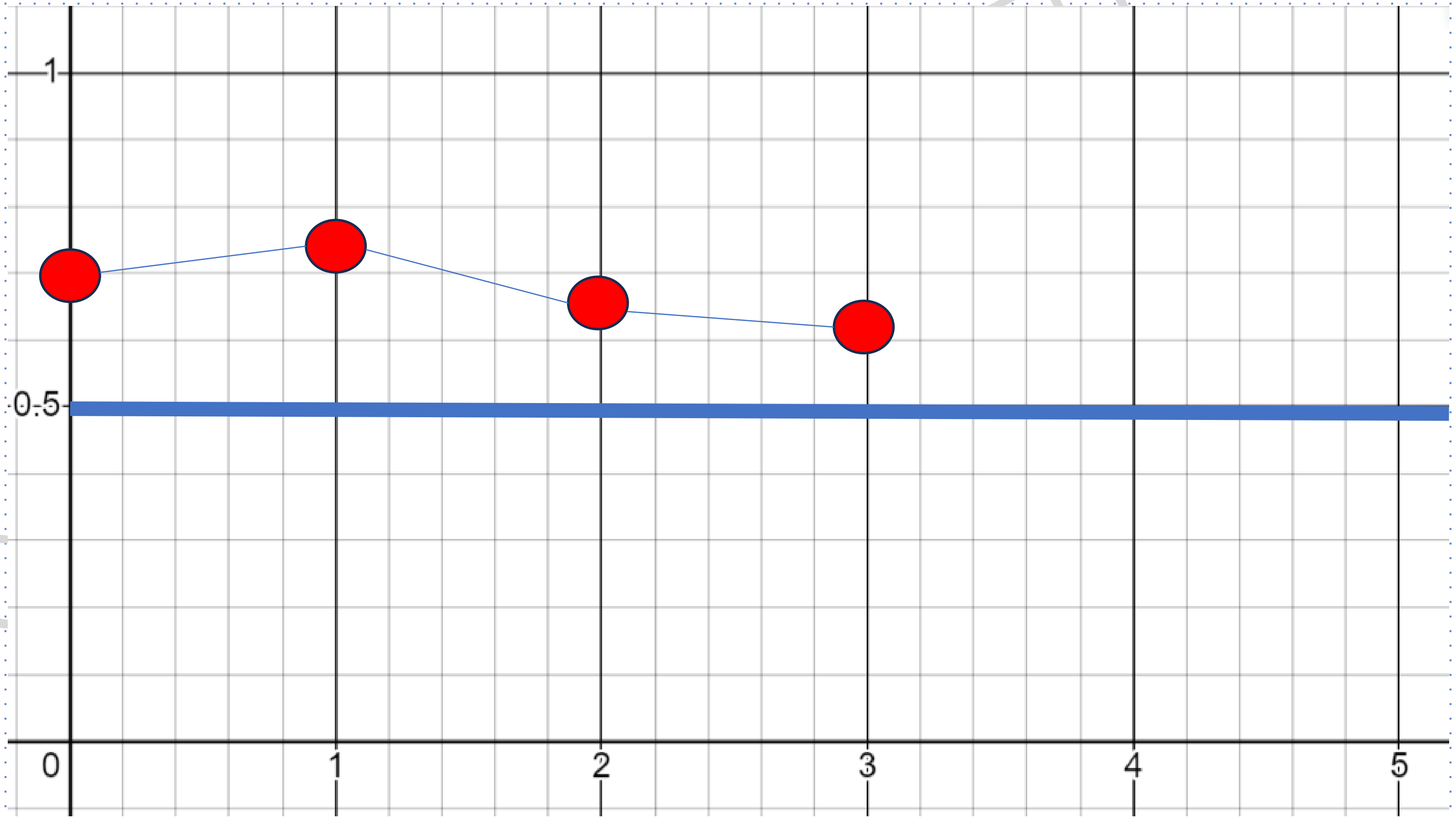
Maximum a Posteriori (MAP)

Trail = HT $\hat{\theta} = \frac{7+1}{7+1+3+1} = 0.66$ $\gamma_1 = 7$ $\gamma_2 = 3$



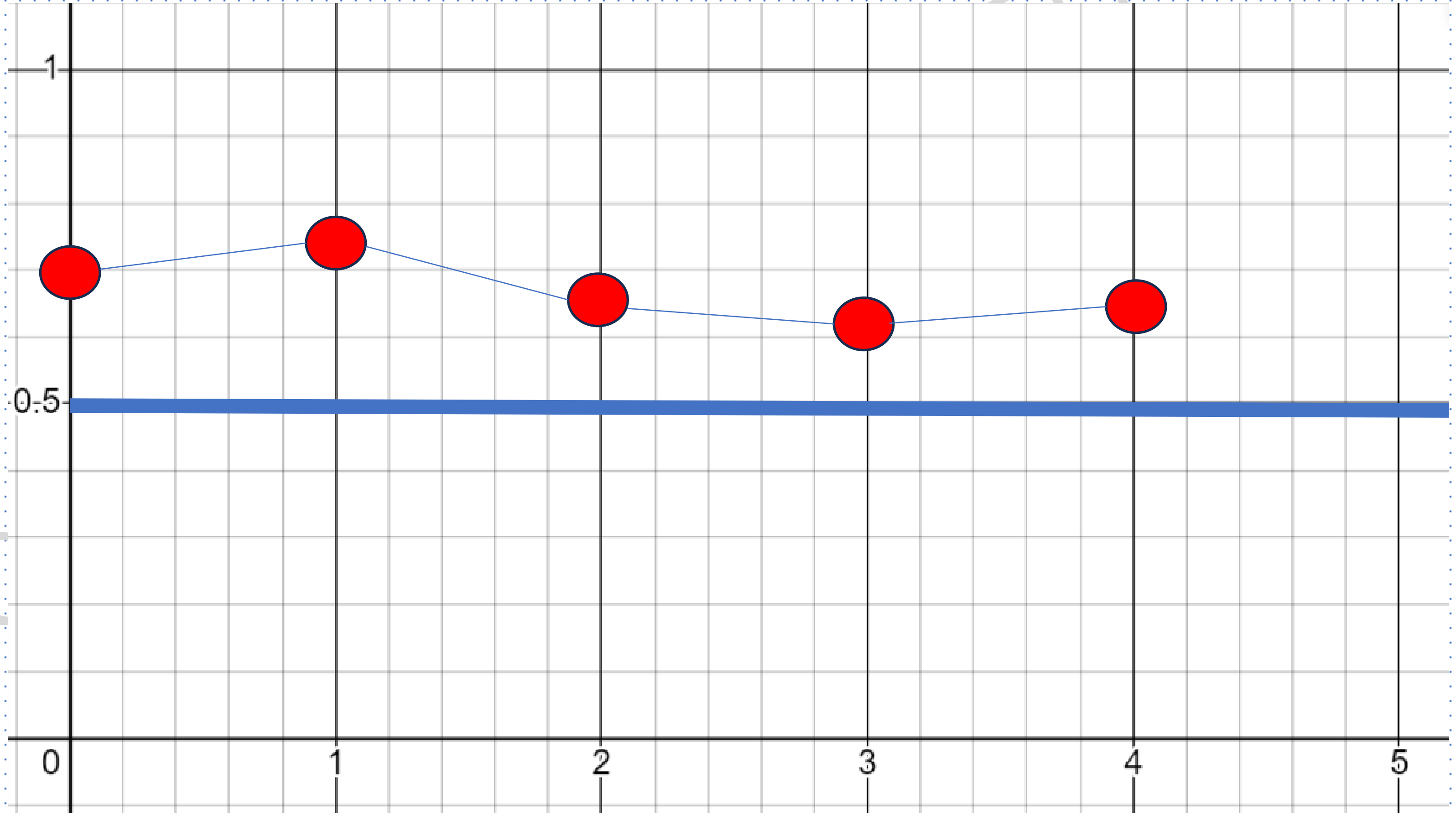
Maximum a Posteriori (MAP)

Trail = HT T $\hat{\theta} = \frac{7+1}{7+1+3+2} = 0.61$ $\gamma_1 = 7$ $\gamma_2 = 3$



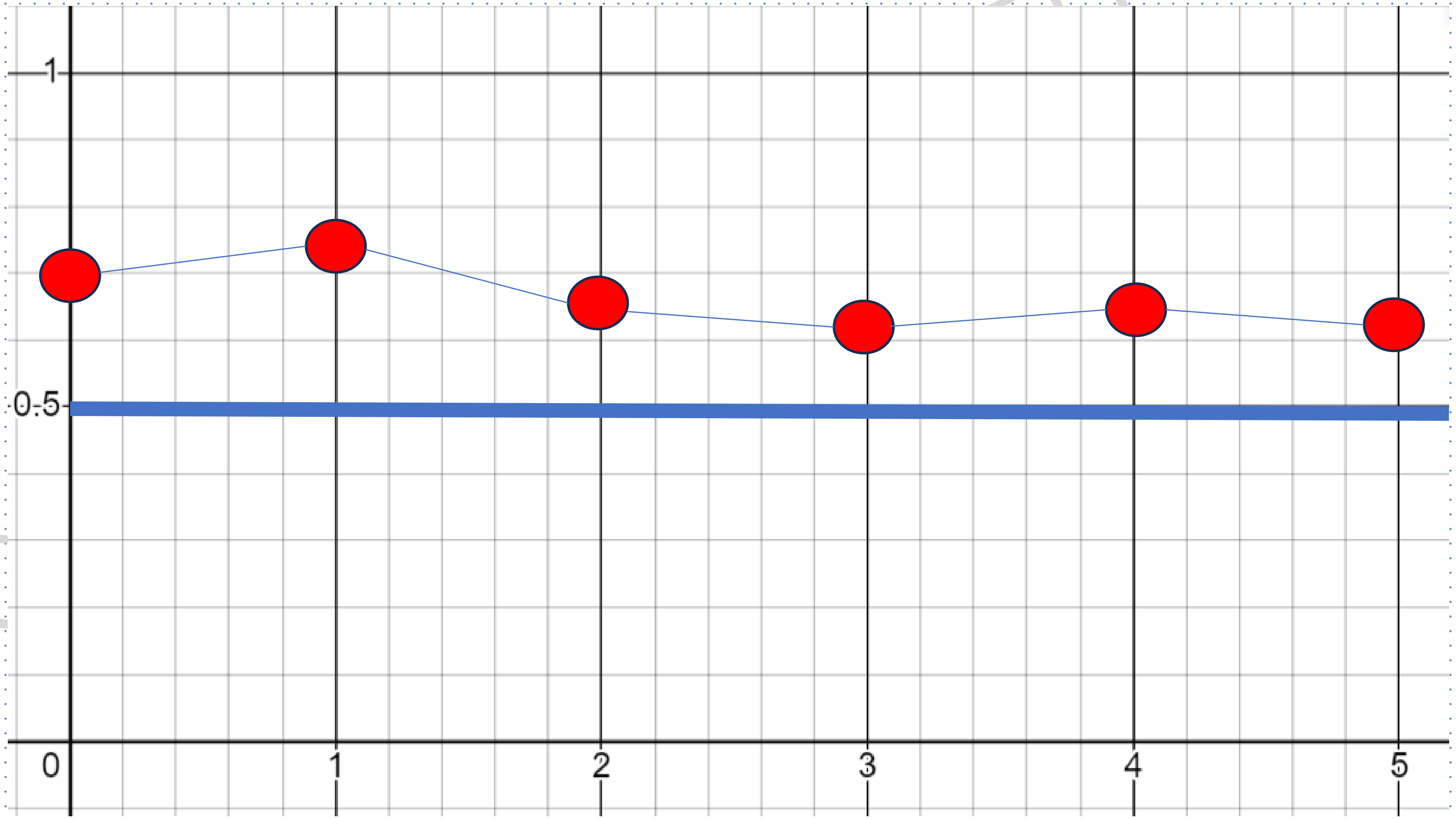
Maximum a Posteriori (MAP)

Trail = HT TH $\hat{\theta} = \frac{7+2}{7+2+3+2} = 0.64$ $\gamma_1 = 7$ $\gamma_2 = 3$



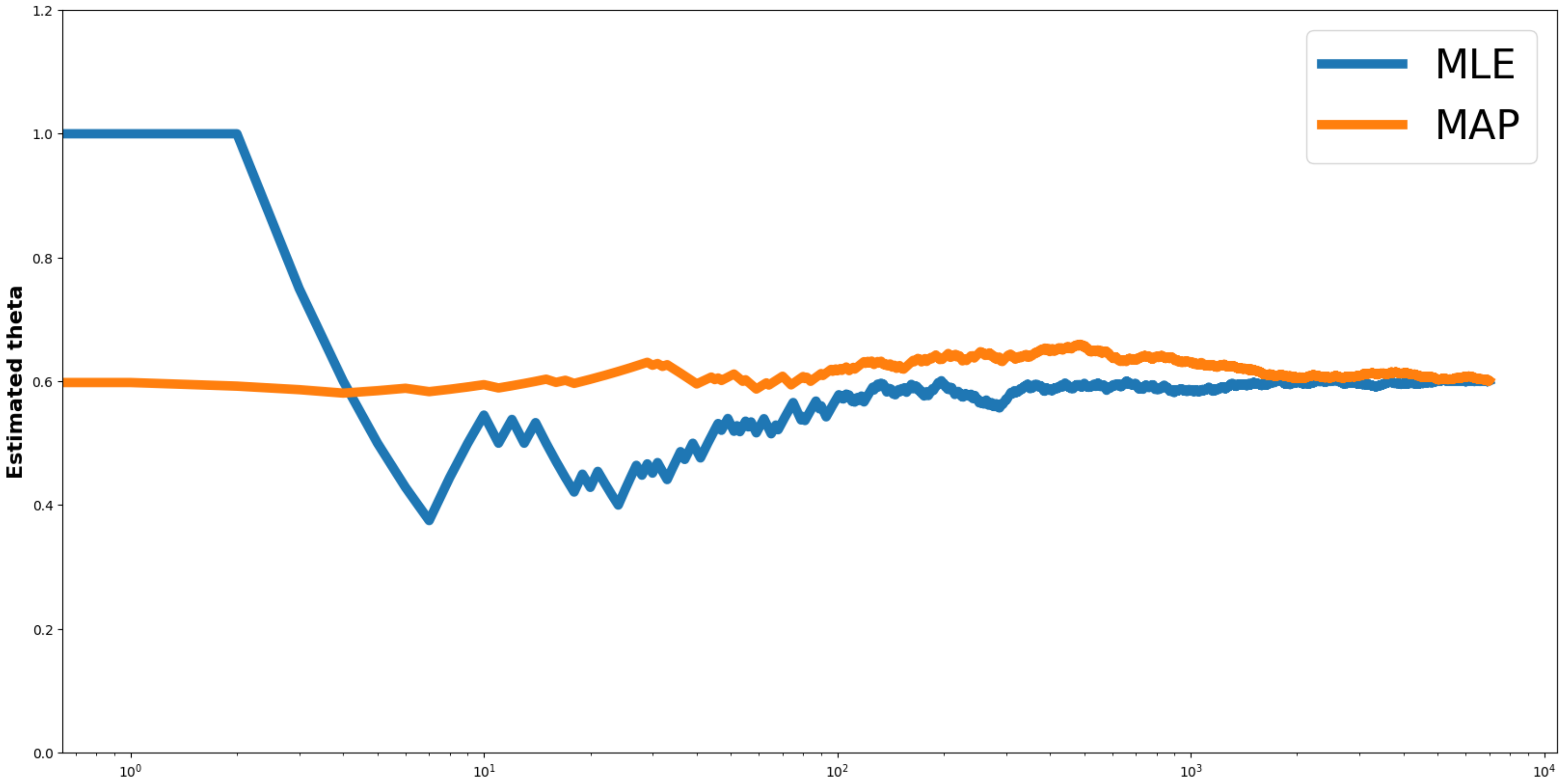
Maximum a Posteriori (MAP)

Trail = HT THH $\hat{\theta} = \frac{7+3}{7+3+3+2} = 0.66$ $\gamma_1 = 7$ $\gamma_2 = 3$



Correct MAP priors

$\gamma_1 = 60$ and $\gamma_2 = 40$

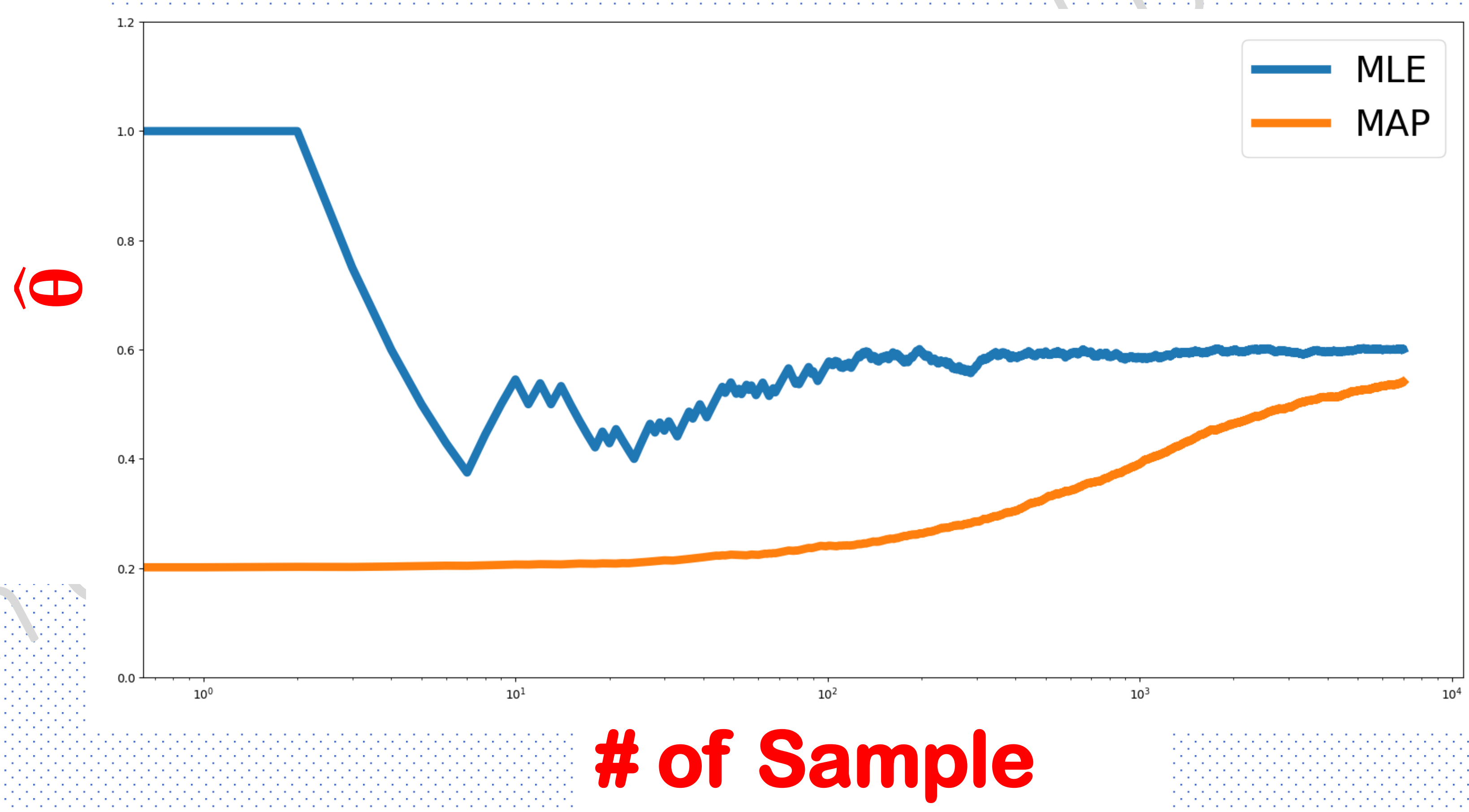


$\hat{\theta}$

# of Sample

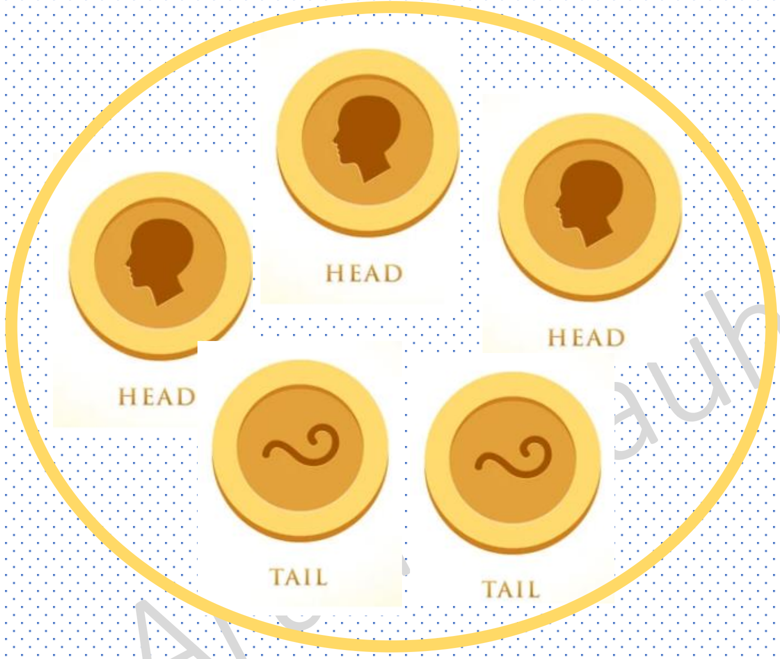
Incorrect Strong MAP priors

$\gamma_1 = 200$ and $\gamma_2 = 800$



Maximum Likelihood Estimation (MLE)

Data Set (D)



$P($  $|\theta) = \theta$

$P($  $|\theta) = 1 - \theta$

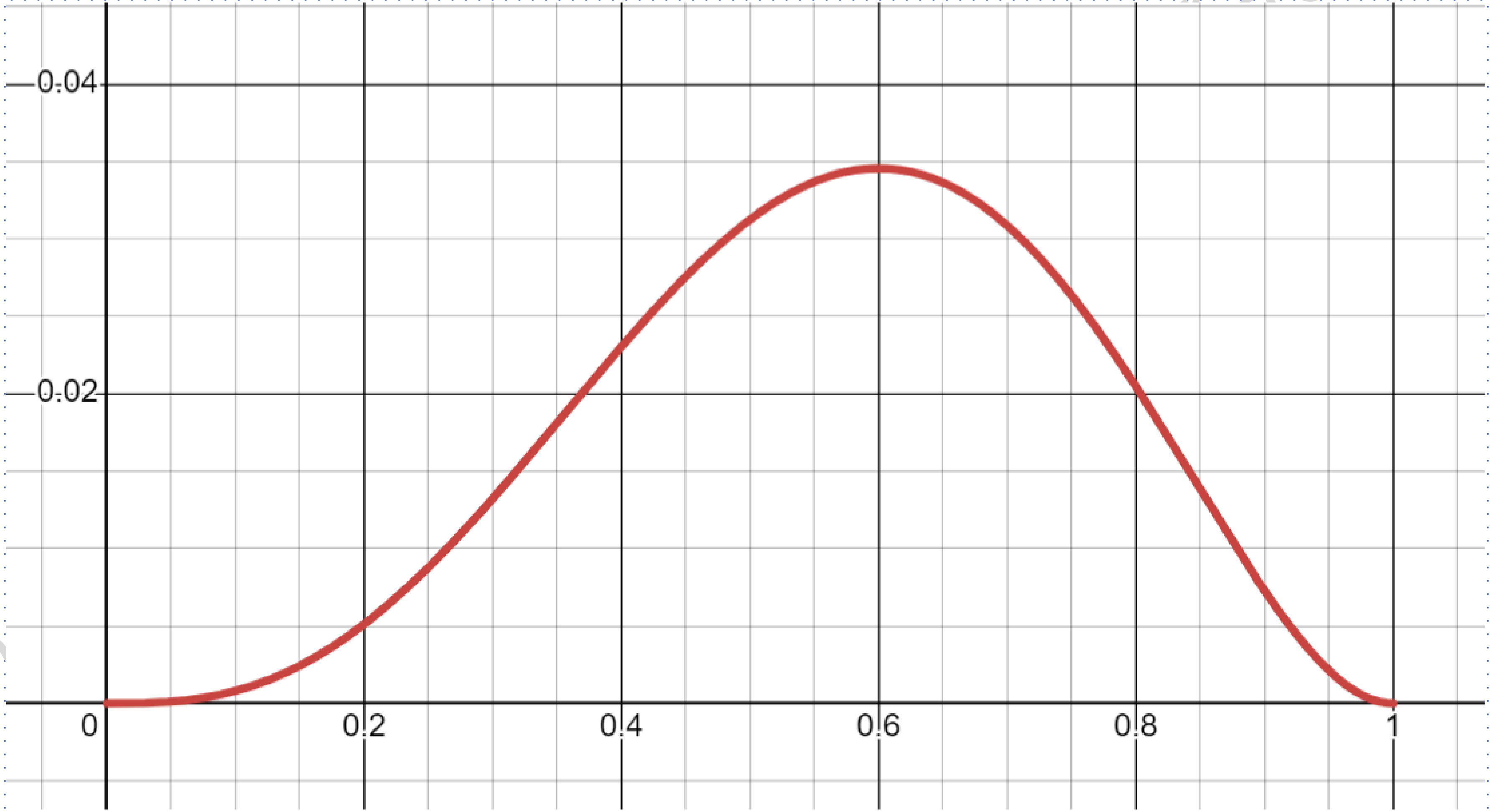
$$P(D\,|\theta) = \theta \cdot \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta)$$

$$P(D\,|\theta) = \theta^{3}(1-\theta)^{2}$$

Likelihood Function ?

$$P(D \mid \theta) = \theta^{3}(1-\theta)^{2}$$

$P(D \mid \theta)$

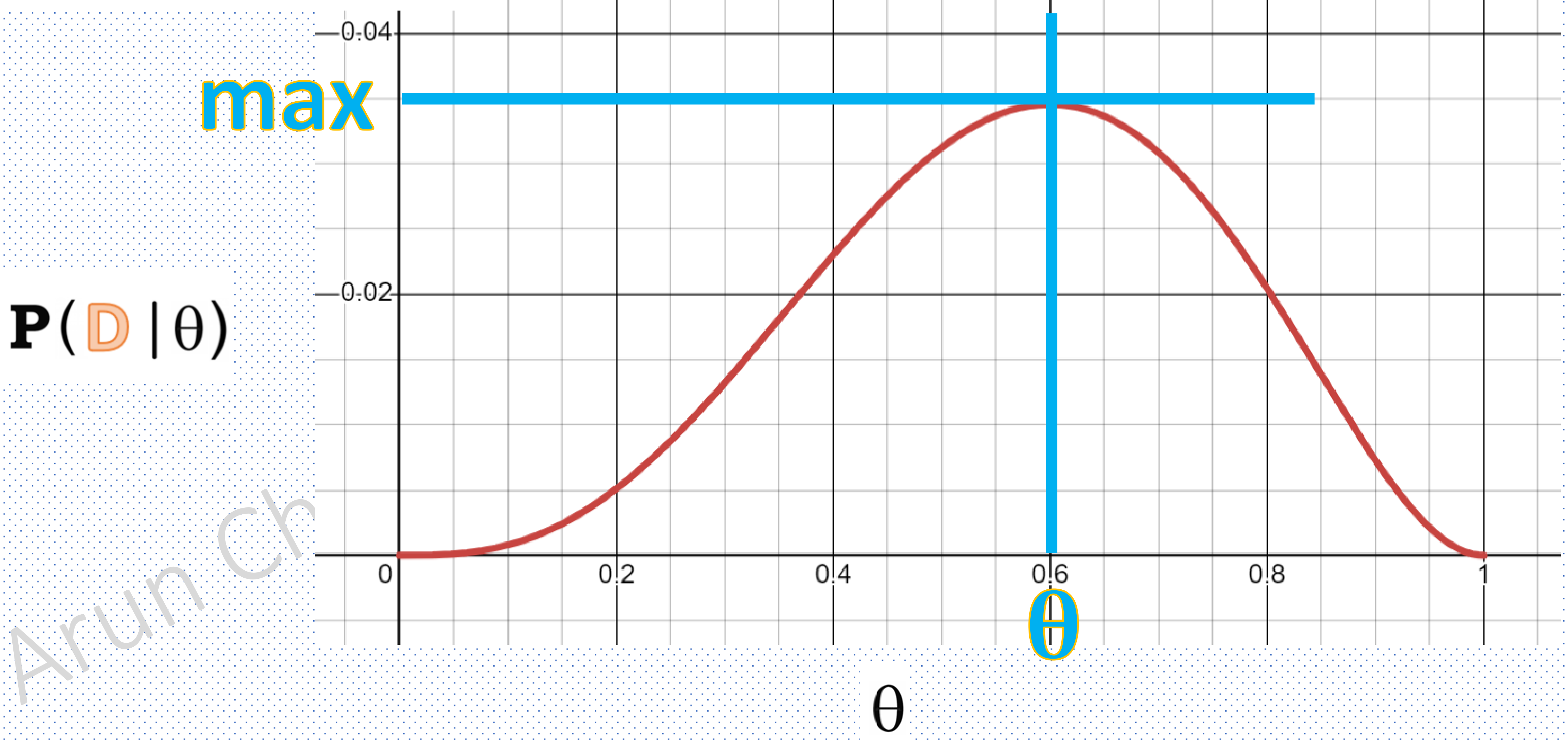


θ

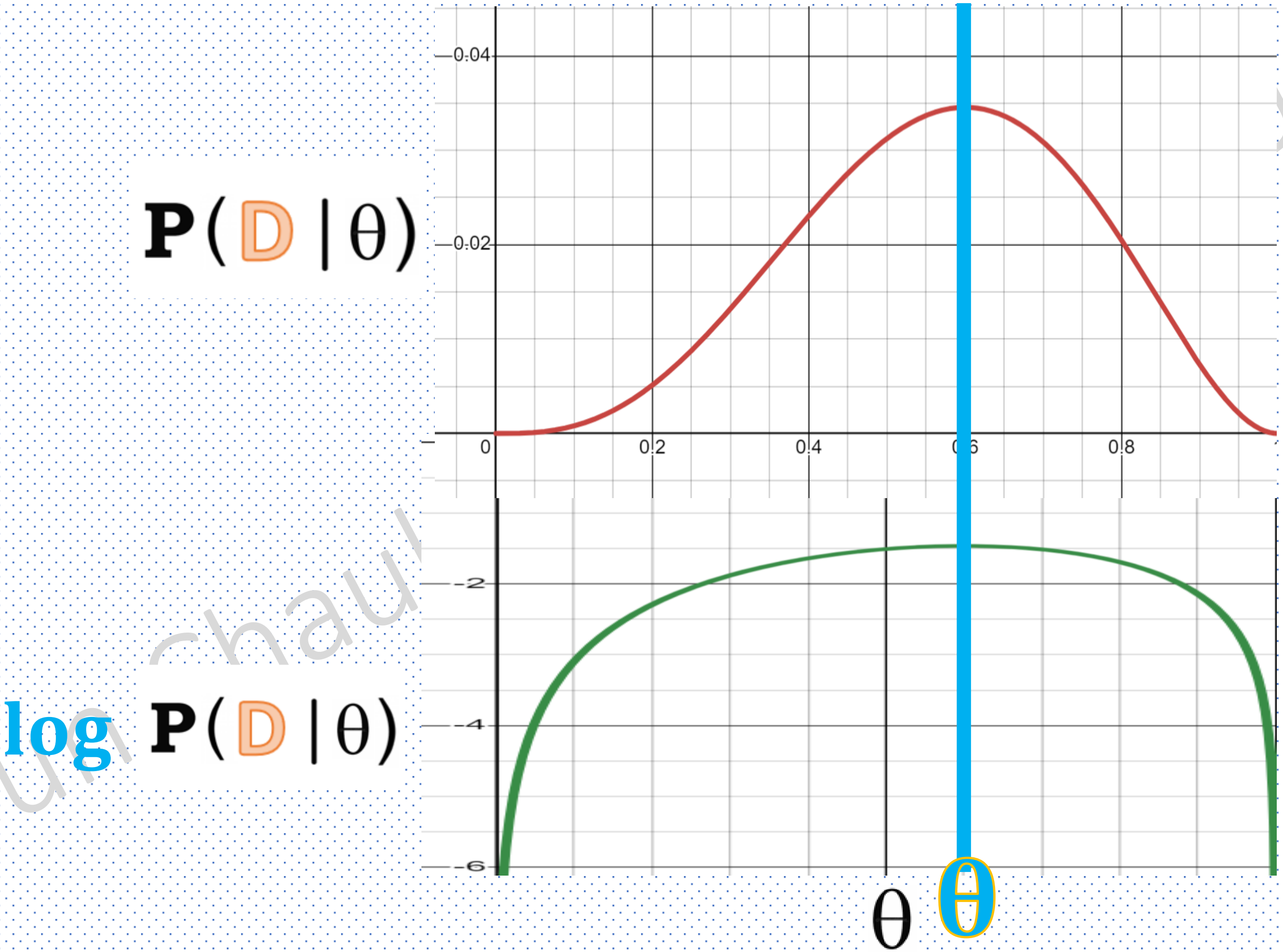
Maximum Likelihood Function ?

$\max_{\theta} P(D|\theta)$

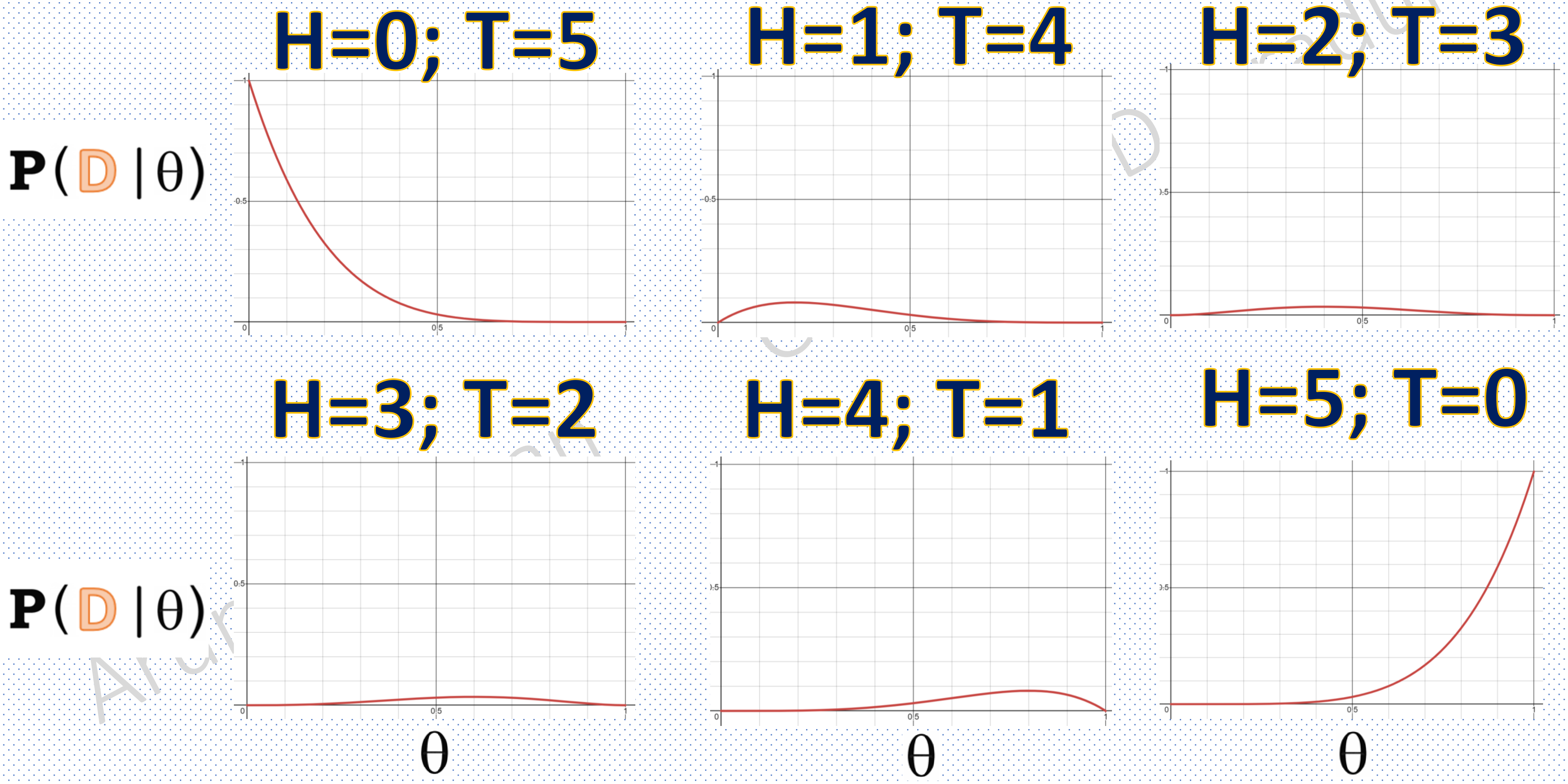
$P(D|\theta) = \theta^3 (1-\theta)^2$



Find the maximum value of θ ?

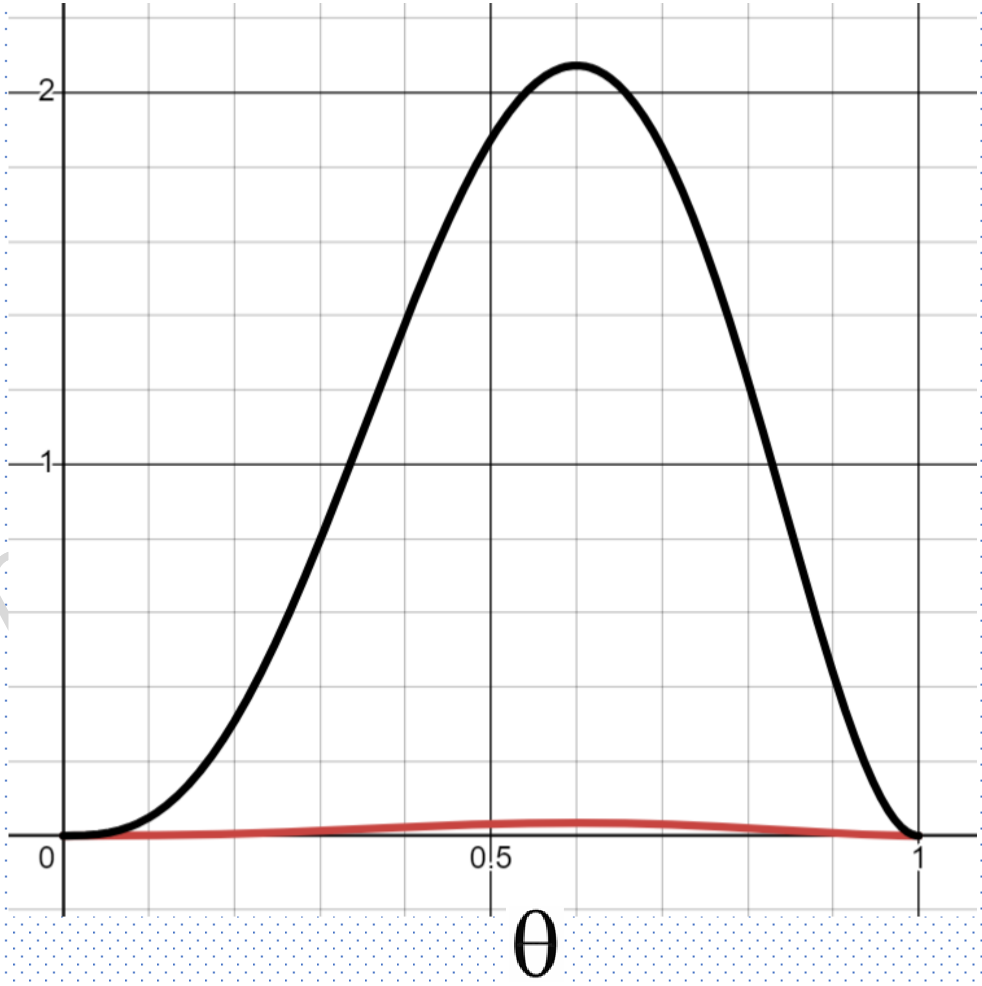


Likelihood Function for different Data Sets ?



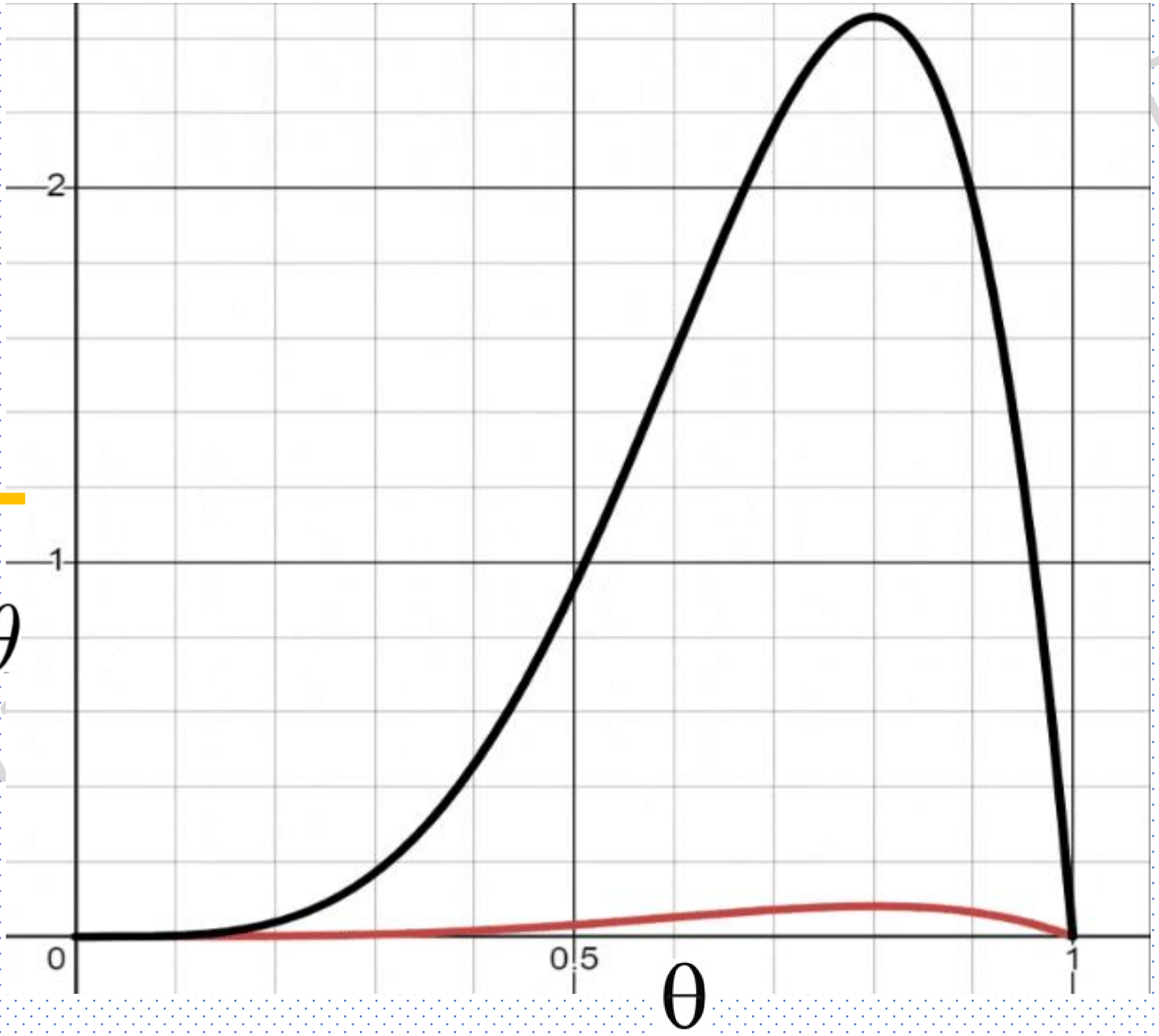
Normalized Likelihood Function !

$$\frac{\mathbf{P}(D \mid \theta)}{\int \mathbf{P}(D \mid \theta)\, d\theta}$$



Normalizing Likelihood function for different datasets.

$$\frac{\mathbf{P}(\mathbb{D} \mid \theta)}{\int \mathbf{P}(\mathbb{D} \mid \theta)\, d\theta}$$



H=0; T=5

H=1; T=4

H=2; T=3

H=3; T=2

H=4; T=1

H=5; T=0

Find the maximum value of $\log P(D|\theta)$?

Log likelihood function : $l(\theta) = \log P(D \mid \theta)$

Take the derivative of $l(\theta)$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \log P(D|\theta)}{\partial \theta} = \frac{\partial \log \left[\theta^{\alpha_1}(1-\theta)^{\alpha_2} \right]}{\partial \theta}$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \left[\alpha_1 \log \theta + \alpha_2 \log(1-\theta) \right]}{\partial \theta} = \alpha_1 \frac{\partial \log \theta}{\partial \theta} + \alpha_2 \frac{\partial \log(1-\theta)}{\partial \theta}$$

$$\frac{\partial l(\theta)}{\partial \theta} = \alpha_1 \frac{\partial \log \theta}{\partial \theta} + \alpha_2 \frac{\partial \log(1-\theta)}{\partial(1-\theta)} . \frac{\partial(1-\theta)}{\partial \theta}$$

$$\frac{\partial l(\theta)}{\partial \theta} = \alpha_1 \frac{1}{\theta} + \alpha_2 \frac{1}{(1-\theta)} . (-1)$$

Find the maximum value of $\log P(D|\theta)$?

Set derivative equals to zero

$$0 = \alpha_1 \frac{1}{\theta} - \alpha_2 \frac{1}{(1-\theta)}$$

$$\alpha_2 \frac{1}{(1-\theta)} = \alpha_1 \frac{1}{\theta}$$

$$\alpha_2\theta = \alpha_1(1-\theta)$$

$$\theta(\alpha_1 + \alpha_2) = \alpha_1$$

$$\theta = \frac{\alpha_1}{(\alpha_1 + \alpha_2)}$$

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta)$$

$$= \arg\max_{\theta} \log P(D|\theta)$$

$$= \frac{\alpha_1}{(\alpha_1 + \alpha_2)}$$

Maximum a Posteriori Probability Estimation (MAP)

$$P(\theta|D) = \frac{P(D|\theta) * P(\theta)}{P(D)}$$

(By Bayes Rule)

- $P(\theta)$ is the prior distribution over θ.
- $P(D|\theta)$ is the likelihood function.
- $P(\theta|D)$ is the posterior distribution over θ.
- $P(D)$ is the probability of the Data Set.

Prior Distribution: $P(\theta)$

- $P(\theta)$ is prior distribution over θ.

In Bayesian Inference we use **Conjugate Prior** .

$$P(\theta|D) = \frac{P(D|\theta) * P(\theta)}{P(D)}$$

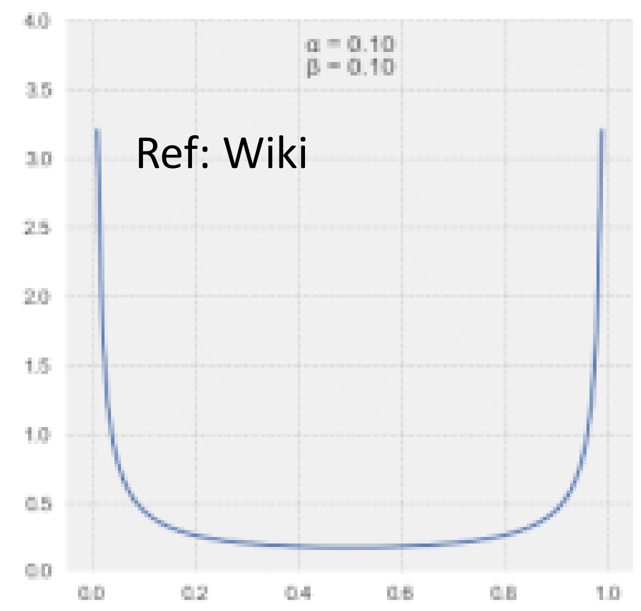
$$\theta^{A}(1-\theta)^{B} = \frac{\theta^{\alpha_1}(1-\theta)^{\alpha_2} * \theta^{M}(1-\theta)^{N}}{P(D)}$$

Prior Distribution: $P(\theta)$

θ is a binary random variable (∴ Binomial Distribution).

The natural choice is Beta Distribution

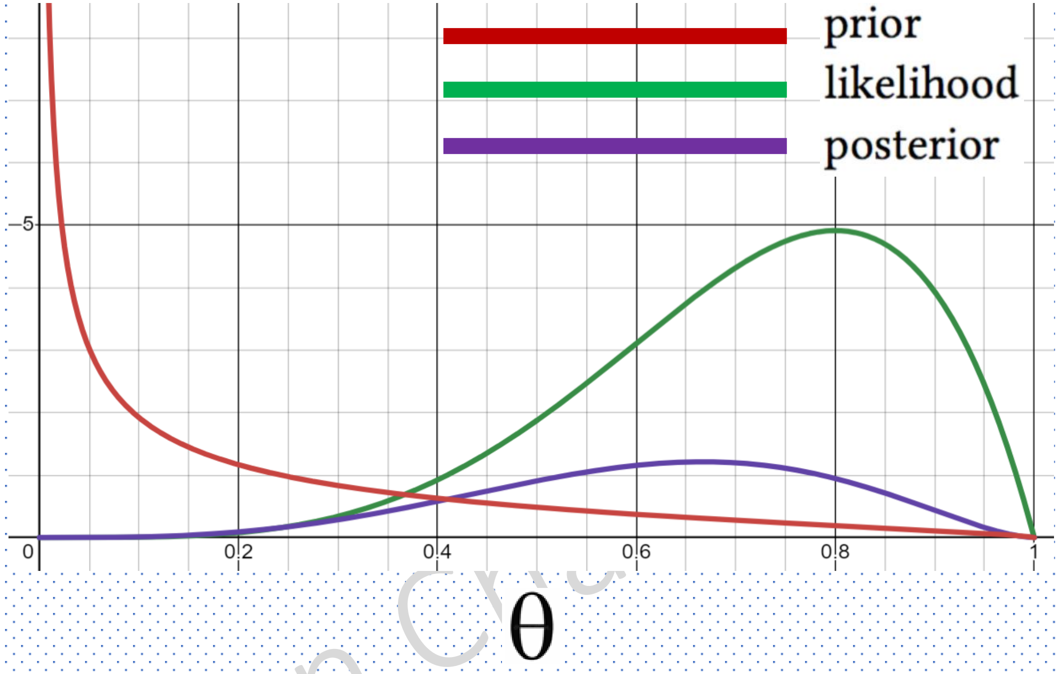
(Conjugate Prior of Binomial Distribution)

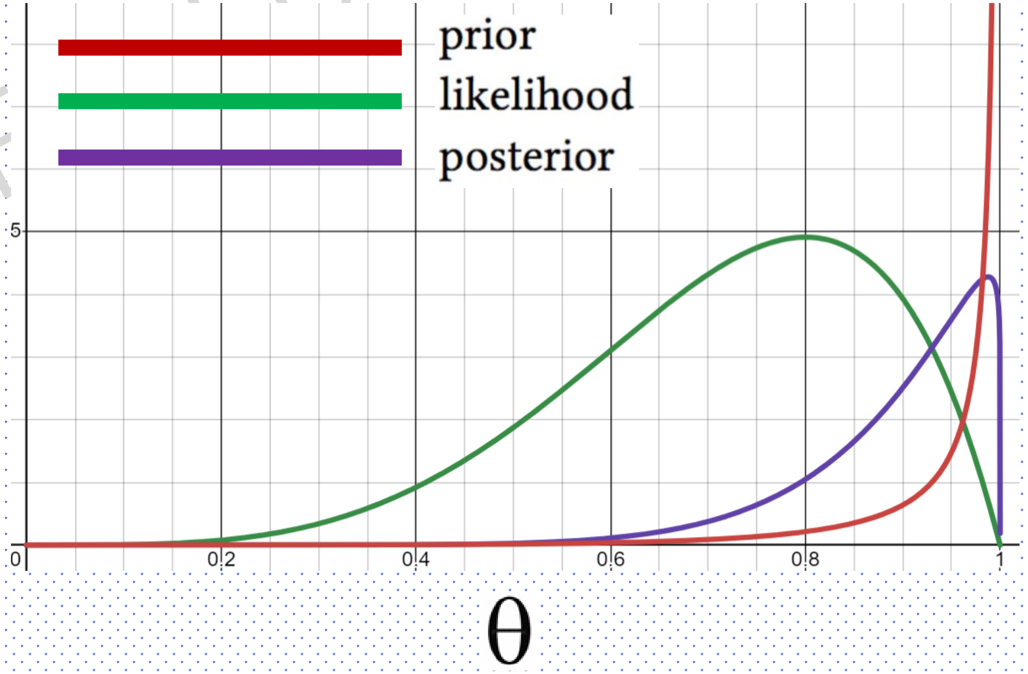


$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

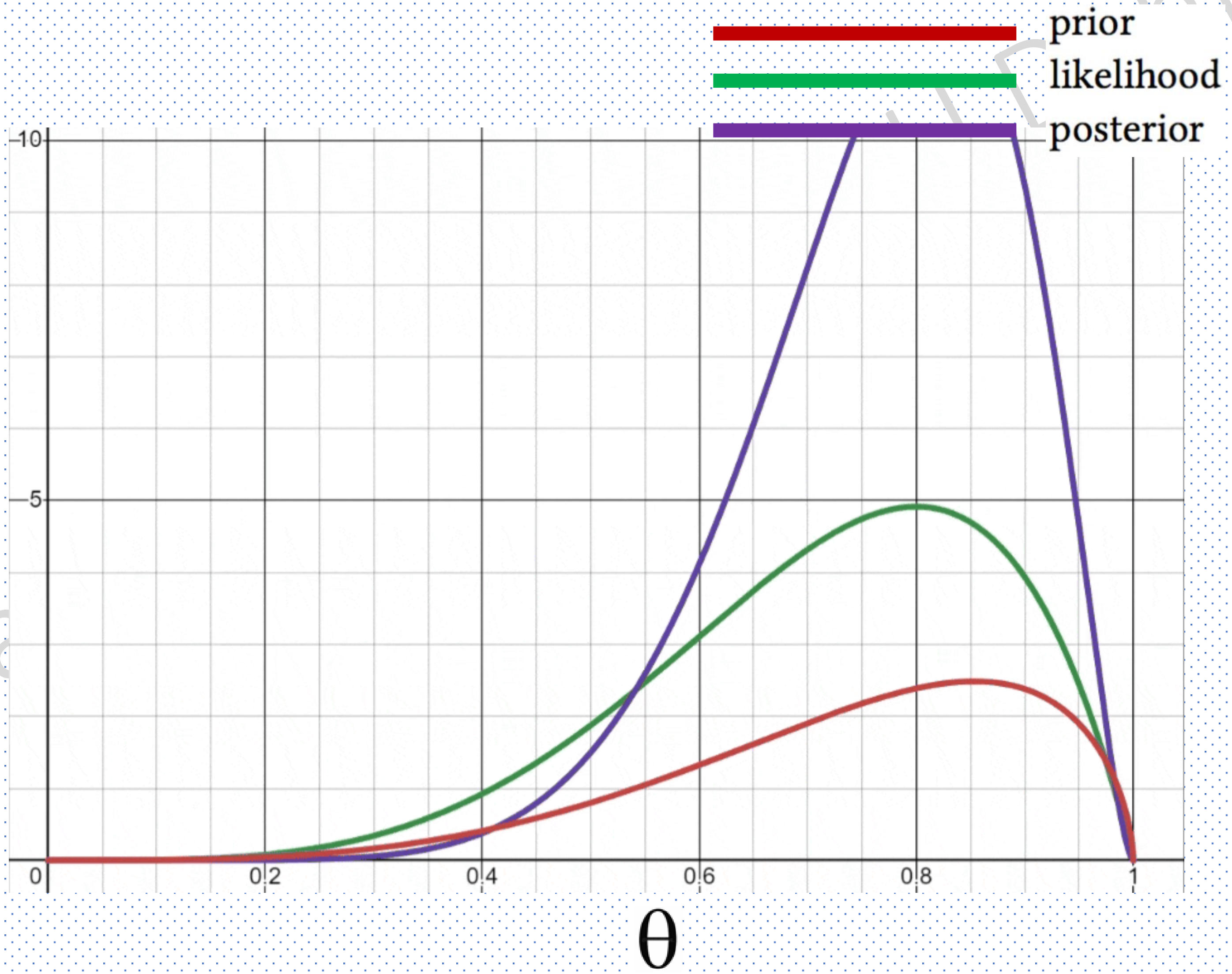
Normalizing Constant

P(θ|Data) is the posterior distribution over θ

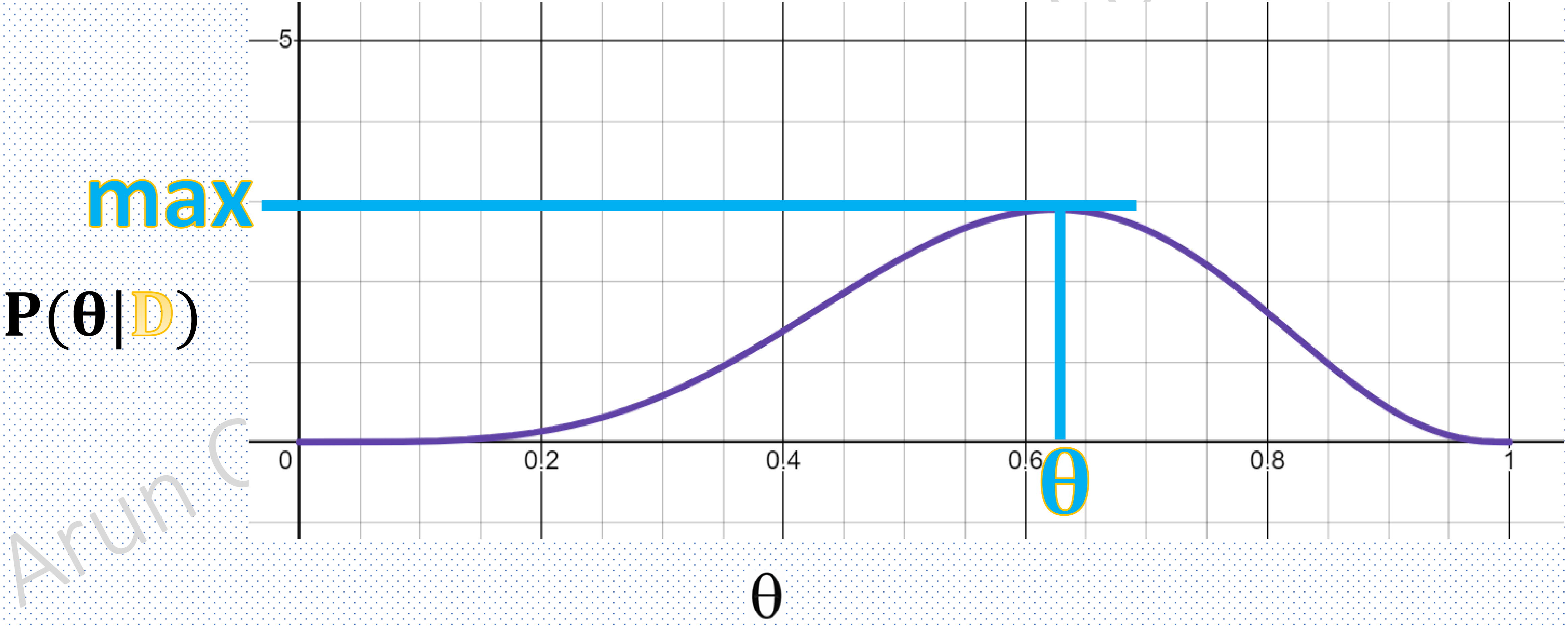




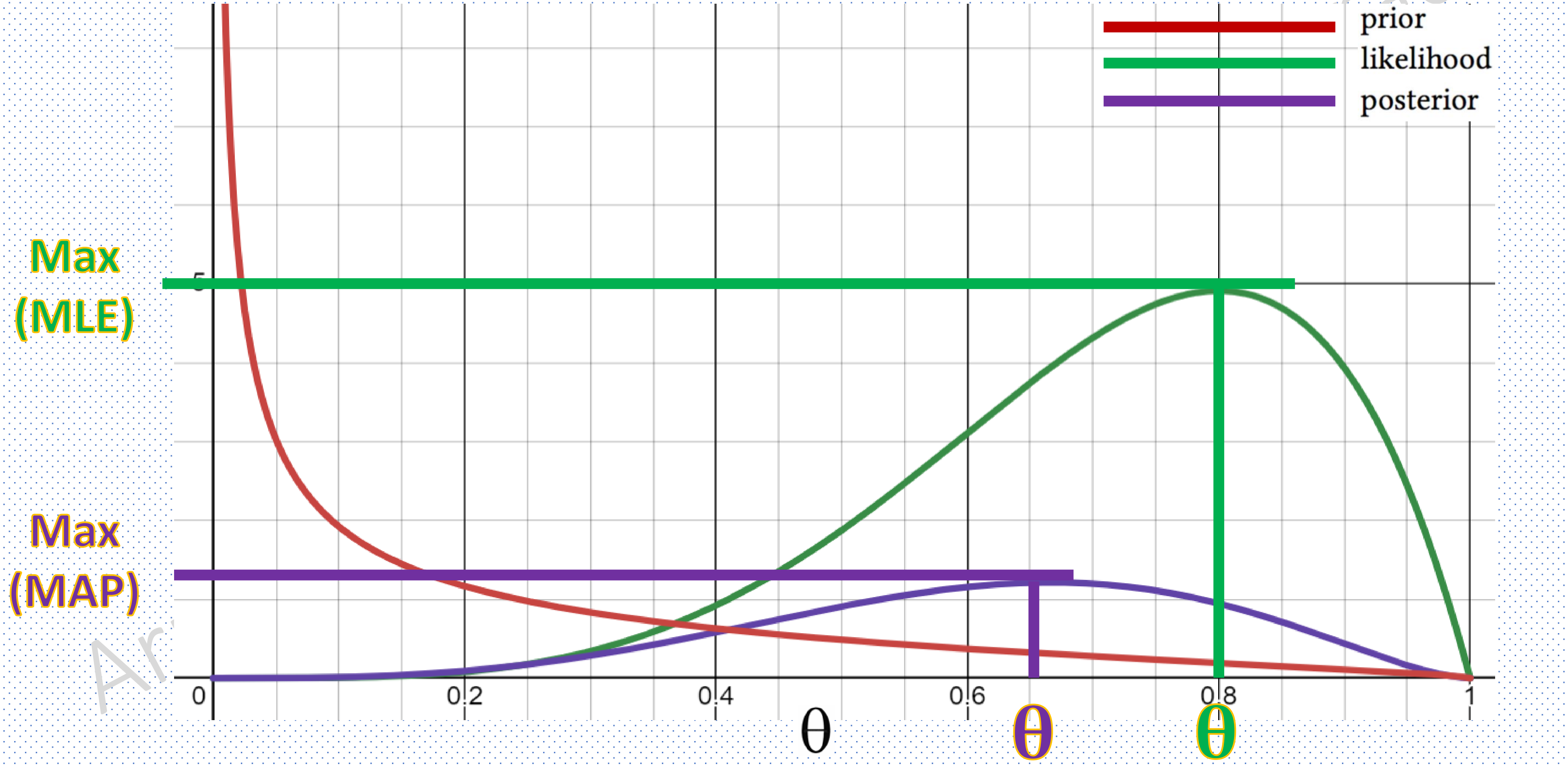
$P(\theta|Data)$ is the **posterior distribution over** θ



Maximum a Posteriori Probability Estimation (MAP)



MAP Vs MLE



prior
likelihood
posterior
Max (MLE)
Max (MAP)
5
0
0.2
0.4
0.6
0.8
1
θ
θ
θ

Maximum a Posteriori Probability Estimation (MAP)

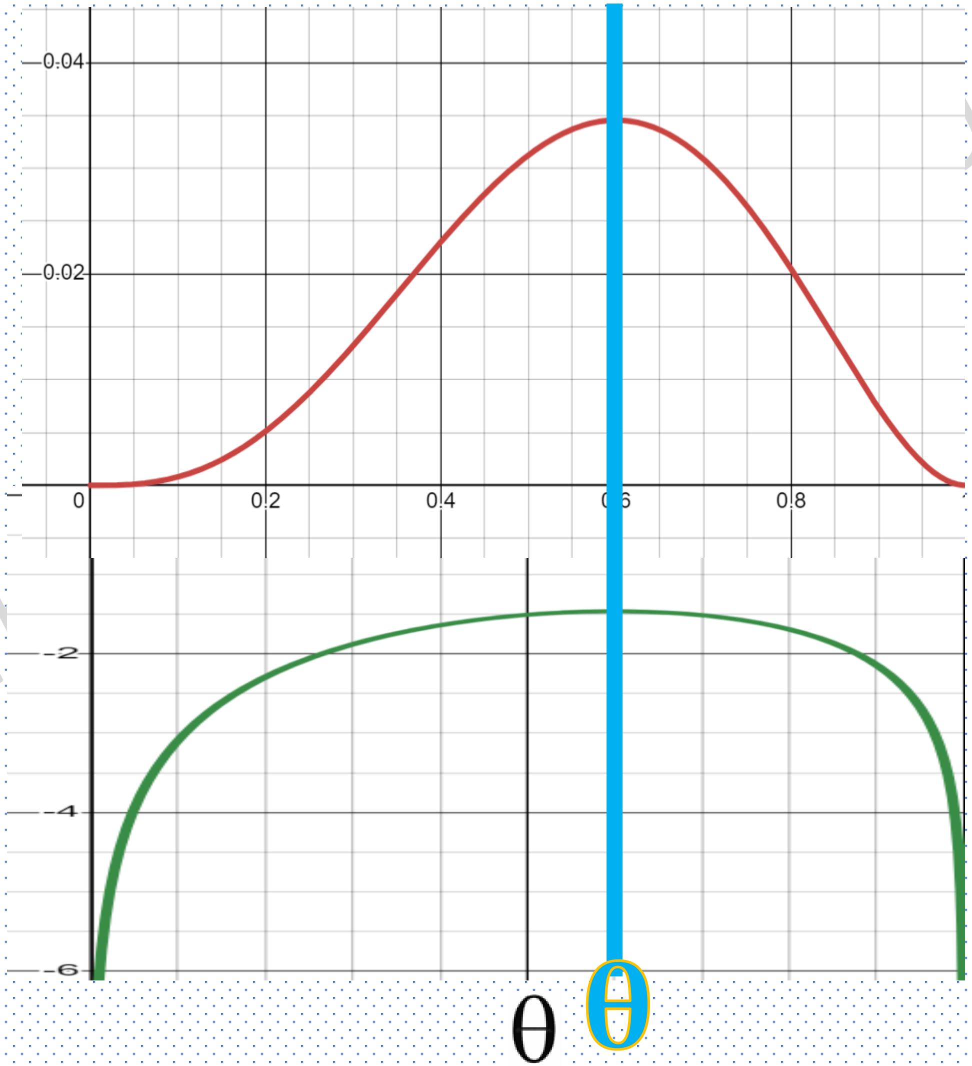
$$\hat{\theta} = \arg\max_{\theta} P(\theta|D)$$

$$\hat{\theta} = \arg\max_{\theta} \frac{P(D|\theta) * P(\theta)}{P(D)} = \arg\max_{\theta} P(D|\theta) * P(\theta)$$

Because θ does not depend on $P(D)$

Maximum of f(x) Vs log f(x) ?

$P(\theta|D)$



log $P(\theta|D)$

Find the maximum value of $\log P(\theta|D)$?

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log \mathbf{P(Data|\theta) * P(\theta)}$$

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log \quad \theta^{\alpha_1}(1-\theta)^{\alpha_2} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log \ \theta^{\alpha_1+\alpha-1}(1-\theta)^{\alpha_2+\beta-1}$$

$$\hat{\theta}^{MAP} = \frac{\alpha_1+\alpha-1}{(\alpha_1+\alpha-1)+(\alpha_2+\beta-1)}$$

By Setting derivative equals to zero

References

- http://www.cs.cmu.edu/~tom/mlbook/Joint_MLE_MAP.pdf