MLE, MAP & BAYESIAN

PART I: COIN FLIP

Bayes Rule

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

A = Parameters (θ)

B = Data (D)

Likelihood

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

Prior

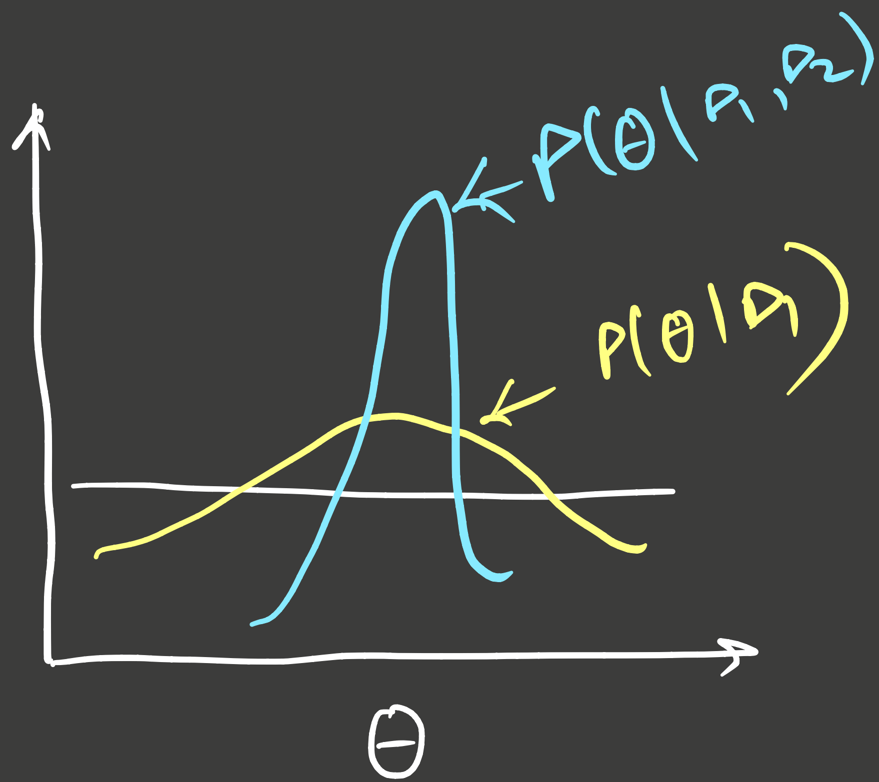
Evidence

Online Learning

No data $\qquad P(\theta)$

Observed D_1 $\qquad P(\theta|D_1)$

Observed D_2 $\qquad P(\theta|D_2) = \dfrac{P(D_2|\theta)\ \text{Prior}}{P(D_2)}$



Assume Coin TOSS MULTIPLE TIMES

OBSERVATION := {H, H}

Q) What is p(Head) ?

Maximum likelihood Estimation (MLE)

6H, 4T

$P(H) = .6$; $P(T) = .4$

n_H, n_T

$$P(H) = \frac{n_H}{n_H + n_T}$$ ← Comes from an MLE estimate

Let $P(H) = \theta$

likelihood $= P(D|\theta)$

$$\boxed{\hat{\theta}_{MLE} = \underset{\theta}{\arg\max}\ P(D|\theta)}$$

To prove

$$\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}$$

$D = n_H, n_T$ where $D = \{D_1, \ldots D_m\}$ {H or T}

$P(H) = \theta \Rightarrow P(T) = 1 - \theta$

$$P(D_1, D_2 \ldots D_m | \theta) = P(D_1|\theta) P(D_2|\theta) \ldots P(D_m|\theta)$$

Q) What is $P(D_1|\theta)$? $= \begin{cases} P(D_1 = H) \text{ given } \theta \text{ if } D_1 = H = \theta \\ P(D_1 = T) \text{ given } \theta \text{ if } D_1 = T = 1 - \theta \end{cases}$

$$P(D \mid \theta) = (P(\text{Head}))^{n_H} \, (P(\text{Tail}))^{n_T}$$ ← Bernoulli (Not binomial

$$P(D \mid \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

$$\underset{(LL)}{\text{Log-likelihood}} = \log P(D \mid \theta) = n_H \log \theta + n_T \log(1-\theta)$$

$$\frac{\partial LL}{\partial \theta} = 0 \Rightarrow \frac{n_H}{\theta} + \frac{n_T}{1-\theta} = 0$$

$$\Rightarrow \theta_{MLE} = \frac{n_H}{n_H + n_T}$$

θ) $2H, 0T$

$P(H) = 1; \; P(T) = 0$

Maximum A Posteriori (MAP)

* MLE doesn't have a notion of prior knowledge

* Overfitting

MAP overcomes these

$P(\theta)$: Prior probability model.

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

Posteriori

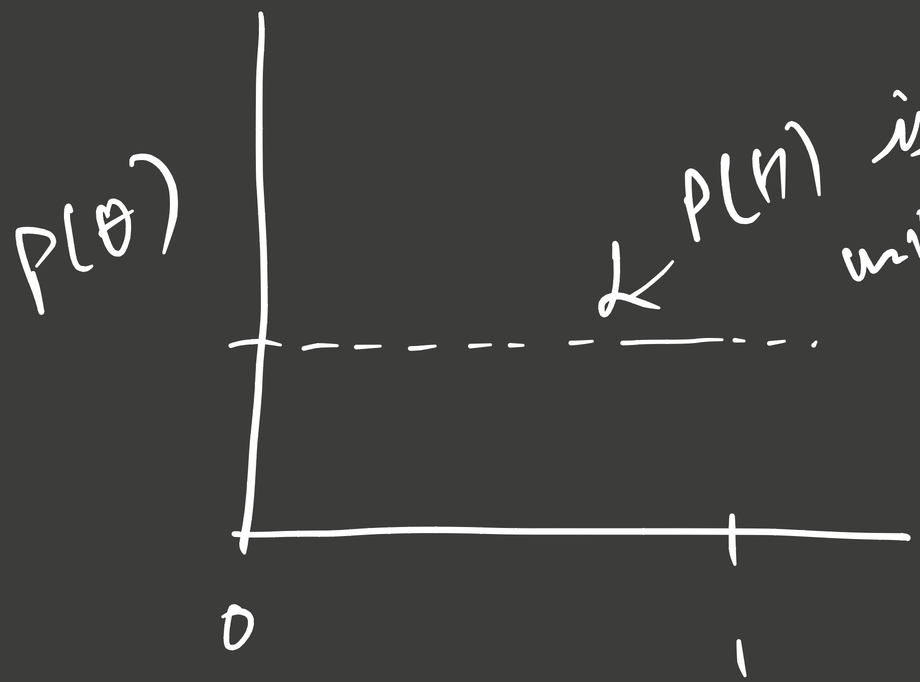
goal for MAP

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$

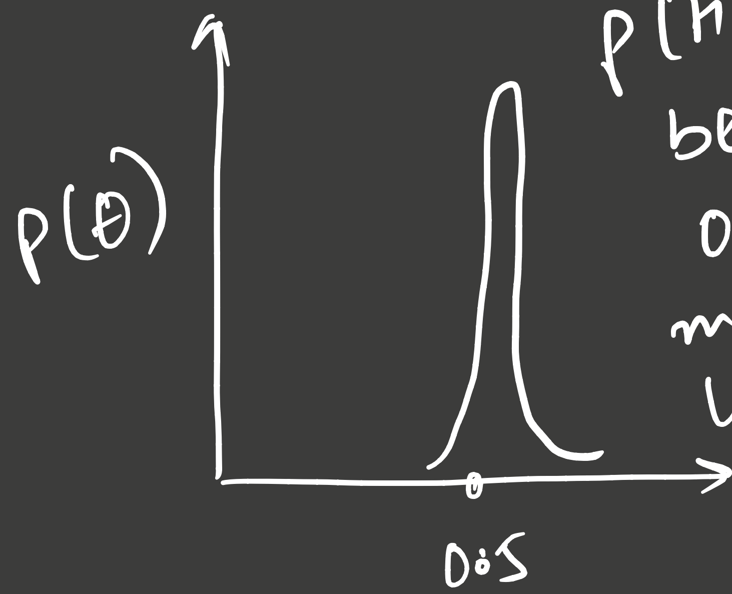
$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D \mid \theta)\, P(\theta)$$

Some examples of $P(\theta)$

$P(H) = \theta$



$P(H)$ is uniformly distributed in $[0,1]$



$P(H)$ being 0.5 more likely than other values

Beta distribution

$$\text{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

$\Downarrow$

Similar to

$$\theta^{n_H} (1-\theta)^{n_T}$$

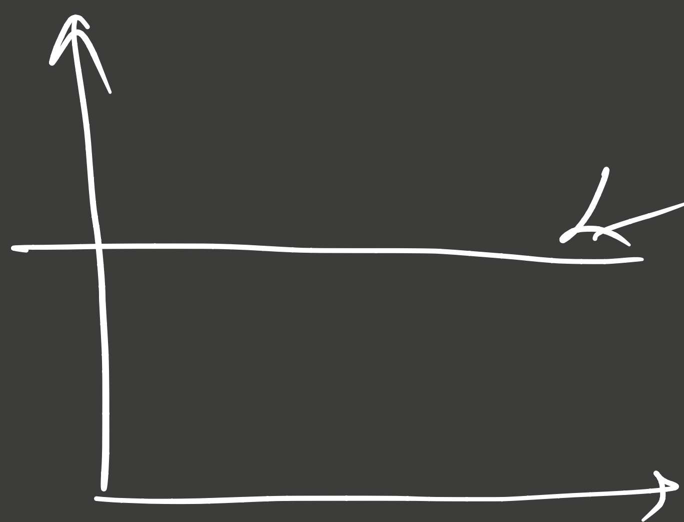
$$\Gamma(n) = (n-1)!$$

(Natural nos)

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$$

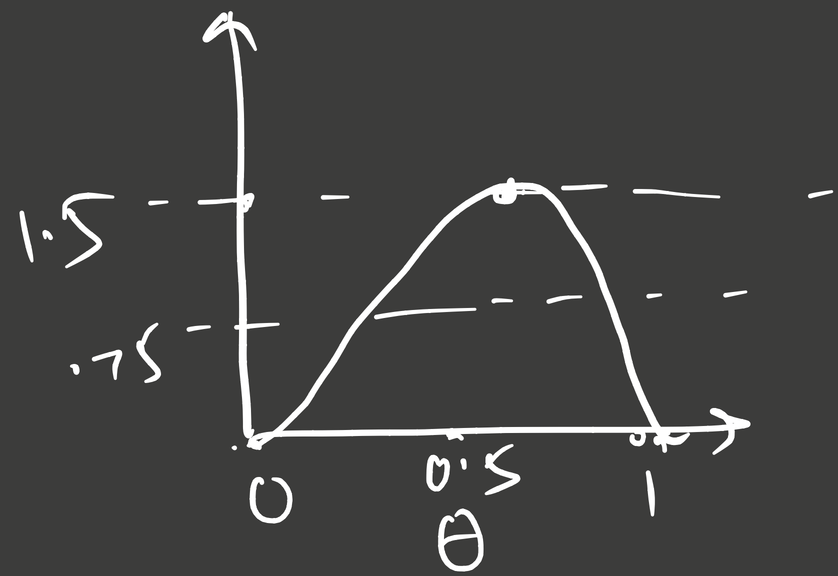
$$\mathrm{Beta}(\theta \mid 1, 1) = \frac{\Gamma(1+1)}{\Gamma(1)\,\Gamma(1)}\,\theta^{1-1}(1-\theta)^{1-1}$$

$$= 1$$



uniform distribution

$$\text{Beta}(\theta \mid 2,2) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} \theta(1-\theta) = 6\theta(1-\theta)$$



$\text{Beta}(\theta \mid a=9, b=1)$

$\Rightarrow$ Indicate higher probability of heads compared to tails.

$D = n_H, n_T$

$$P(\theta) = \text{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

θ).

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ P(D \mid \theta)\, P(\theta)$$

$$= \underset{\theta}{\text{argmax}}\ \theta^{n_H} (1-\theta)^{n_T}\, \theta^{a-1} (1-\theta)^{b-1} * K$$

constant

$$= \underset{\theta}{\text{argmax}}\ \theta^{n_H+a-1} (1-\theta)^{n_T+b-1}$$

$$\boxed{\hat{\theta}_{MAP} = \frac{n_H + a - 1}{n_H + n_T + a + b - 2}}$$

Conjugate Prior

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

$P(\theta)$ is CONJUGATE TO $P(D|\theta)$

if.

$P(\theta|D)$ and $P(\theta)$ are from same distribution family.

Bernoulli likelihood, Gamma is conjugate

Relationship b/w MAP & MLE

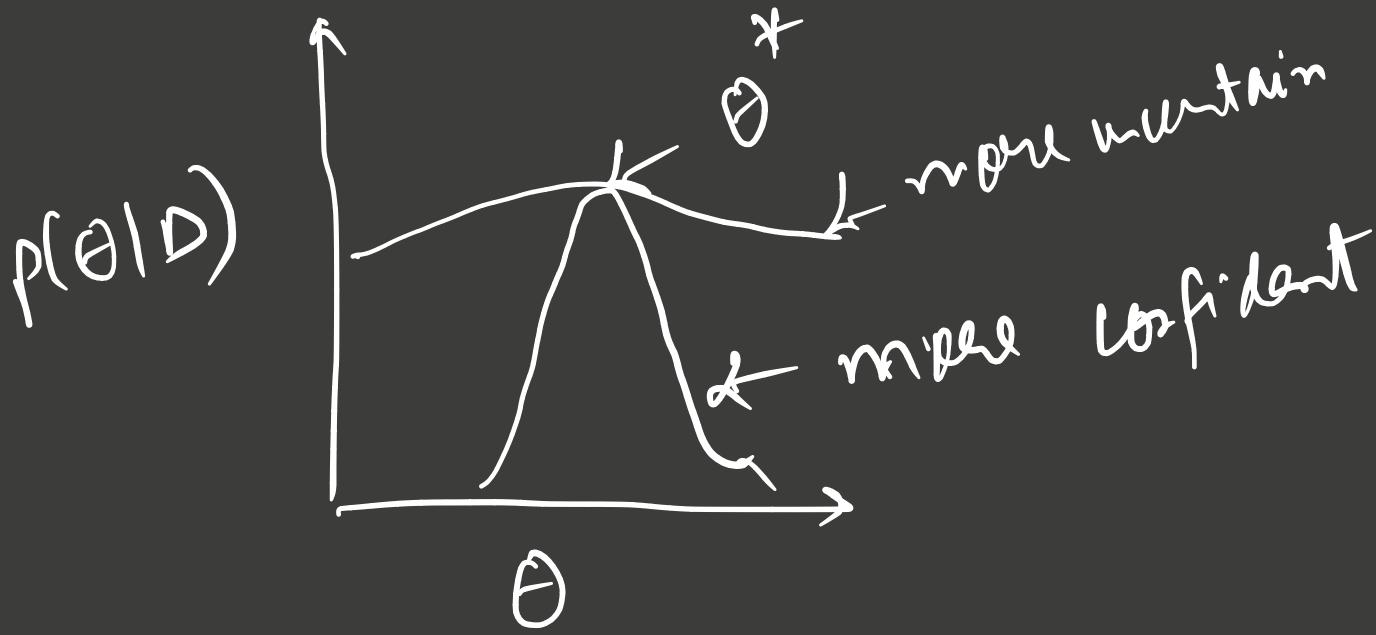
$\hat{\theta}_{MAP}$ $\hat{\theta}_{MLE}$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)\, \cancel{P(\theta)}$$ if this is uniform

$$= \hat{\theta}_{MLE} \text{ (for uniform prior)}$$

Fully Bayesian

* MLE & MAP don't give you uncertainty (or full distribut^n)



Predictive distributⁿ

$$p(\text{Next coin} = H \mid \text{Data})\ ?$$

① What θ to use?

use all possible values of θ

$$p(\text{Next} = H \mid D) = \int p(\text{Next} = H, \theta \mid D)\, d\theta$$

$H \to 1$

$T \to 0$

$$p(\text{Next} = c \mid \theta) = \theta^{c}(1-\theta)^{1-c}$$

or

$p(c \mid \theta)$

why this

$c = 0$

$p(\text{tails} \mid \theta) = \theta^{0}(1-\theta)^{1}$

$$p(\text{Next} = c \mid D, a, b) = \int p(\text{Next} = c, \theta \mid D, a, b)\, d\theta$$

Beta distribution for prior

why? SUM RULE

$\therefore p(x) = \int p(x, y)\, dy$

$$= \int p(N_{ext} = c \mid \theta)\, p(\theta \mid D, a, b)\, d\theta$$

why? ∵ once θ is known, D, a, b don't influence $p(N_{ext} = c)$

$$= \int \theta^{c} (1-\theta)^{1-c} \frac{\Gamma(n_H + n_T + a + b)\, \theta^{n_H + a - 1} (1-\theta)^{n_T + b - 1}}{\Gamma(n_H + a)\, \Gamma(n_T + b)}\, d\theta$$

$$= \frac{\Gamma(n_H + n_T + a + b)}{\Gamma(n_H + a)\, \Gamma(n_T + b)} \int \theta^{c + n_H + a - 1} (1-\theta)^{n_T + b - c}\, d\theta$$

$$= \frac{\Gamma(n_H + n_T + a + b)\, \Gamma(c + n_H + a)\, \Gamma(n_T + b - c + 1)}{\Gamma(n_H + a)\, \Gamma(n_T + b)\, \Gamma(1 + n_H + a + n_T + b)}$$