Limitation of MLE

- Over-reliance on data sample D. If data is limited, estimates can be very wrong.
    - Example, Bernoulli p could be zero if no 1s in 10 trials.

- No indication on the uncertainty of the estimated parameters.
    - Example, for a Bernoulli parameters whether estimation is made from two with 50% heads or 1000 examples with 50% heads, the estimated parameter is the same.

D_1 $|D_1| = 2$ $N_1(D_1) = 1,$ $N - N = 1$ $\hat{p}_1 = 0.5$

D_2 $|D_1| = 1000$ $N_1(D_2) = 500$ $\hat{p}_2 = 0.5$

- No mechanism to specify human's prior knowledge of the parameters.

Example of limitations of MLE

- Suppose a toss a coin 10 times and get

  H, H, H, H, H, H, H, H, H, H

Estimate p? MLE: $\hat{p} = \# \text{ones} / N = 1$

What is your guess on the probability p of head?

- Suppose you want to form a music band, and you are looking for bass guitarist. You ask 7 random batchmates: “Can you play the bass guitar” and you get answers

  N, N, N, N, N, N, N D

What fraction of batchmates play bass guitar?

MLE: 0

Do you have a different guess? 0.01

prior

Bayesian estimation

- Treat the parameters as a random variable which has a distribution.
- Step 1: Humans specify their prior knowledge of the values of the parameters as a distribution $\mathrm{f}_{\mathrm{pr}}(\theta)$
  - Example: $\mathrm{f}_{\mathrm{pr}}(\theta) \sim U(0,1)$ where θ denotes the parameter p of a Bernoulli
  - Example for Gaussian:

Temperature of CPU on your laptop $T \sim G(\theta, \sigma^2)$

$$f_{pr}(\theta) \sim \mathcal{N}(30, 10)$$

Also called prior probability

Bayesian estimation

- Calculate the posterior distribution of parameters after observing data D following Bayes rule

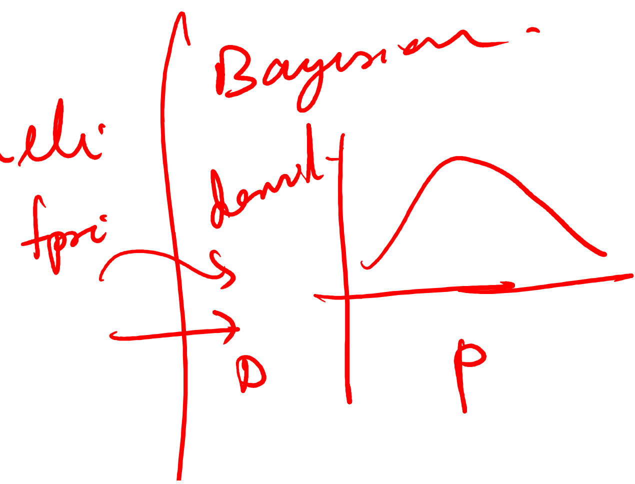
$$\cancel{f(D|\theta) = f(\theta)f(D|\theta)/\int_\theta f(\theta)f(D|\theta)}$$

$$f_{post}(\theta|D) = \frac{f(D|\theta)\,f_{pr}(\theta)}{\int_{\theta'} f(D|\theta')\,f_{pr}(\theta')} \quad ✓$$

Posterior probability

MLE p of Bernoulli

$$D \longrightarrow \hat{p} = 0.6$$



Using Bayesian estimates

$f_p(\theta|D) \equiv f_{po}(\theta|D)$

- Exact Bayesian probability computation:
  - Given a new x, calculate $f(X|D)$

$$f(x|D) = \int_\theta f(x|\theta)\, f_{po}(\theta|D)\, d\theta$$

Binomial eg:

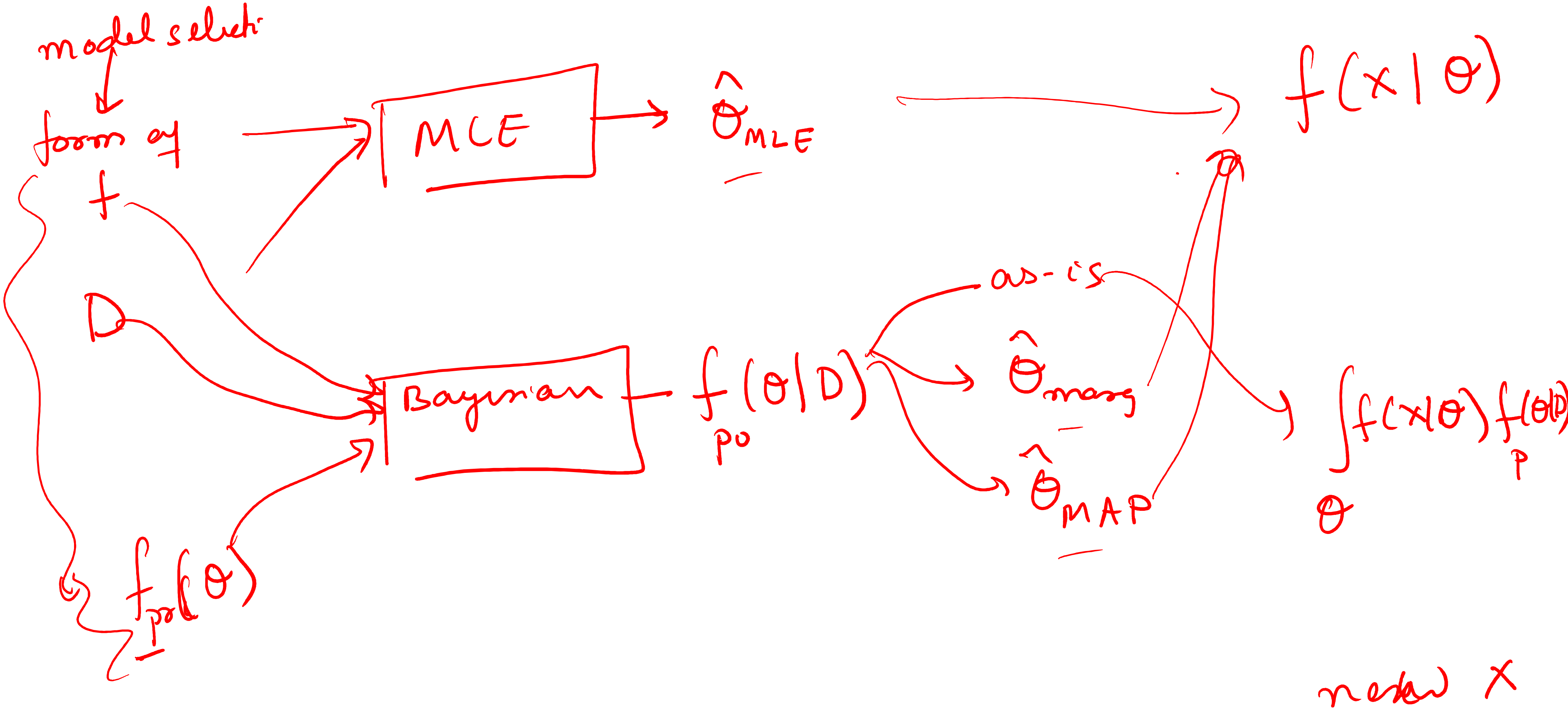
- Expected value of parameters: calculate expected value of $f_{pos}(\theta|D)$

$$\hat{\theta}_{mary} = E_{f_{pos}(\theta|D)}[\theta] = \int_\theta \theta \cdot f_{po}(\theta|D)\, d\theta.$$

- MAP estimate: use $\max_\theta f(\theta|D)$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, f(\theta|D)$$

Overall pipeline for MLE Vs Bayesian



Example:
Bayesian estimation of
Bernoulli/Binomial parameter p

Bayesian estimation of Bernoulli parameter

$p \in [0 \dots 1]$

- Choose a prior distribution over parameter θ or p of Bernoulli
  - $\mathrm{f}_{\mathrm{pr}}(\theta) \sim U(0,1)$

- Data D has n are ones and remaining N-n = m are 0s.

$D \equiv \{n, m\}$ (n: #1s, m: #0s)

$D \equiv \{x_1, x_2 \dots x_N\}$ eg: $\{0, 1, 1, \dots 0 \dots 0\}$

$$f(D|\theta) \equiv \prod_{i=1}^{n} f(x_i|\theta) = \theta^n (1-\theta)^{\underbrace{N-n}_{=m}}$$

- Posterior distribution is:

$$f_{po}(\theta|D) = \frac{f_{pr}(\theta)\, f(D|\theta)}{\int_{\theta'} f_{pr}(\theta')\, f(D|\theta')} = \frac{1 \cdot \theta^n (1-\theta)^m}{\int_{\theta'} \theta'^n (1-\theta')^m}$$

$$f_{po}(\theta|D) = \boxed{(1-\theta)^m (\theta)^n} \qquad 0 \le \theta \le 1$$

$Z \leftarrow$ normalizer.

mode of $f_{po}(\theta|D)$

$$\max_{\theta} \; (1-\theta)^m \, \theta^n$$

$$\Rightarrow \frac{\partial}{\partial \theta}\left[(1-\theta)^m \theta^n\right] = -m(1-\theta)^{m-1}\theta^n + n\,\theta^{n-1}(1-\theta)^m = 0$$

$$\Rightarrow -m\theta + n(1-\theta) = 0$$

$$\theta(n+m) = n \Rightarrow \theta = \frac{n}{m+n}$$

Beta Random Variable (Generic defn. of Beta)

X is a **Beta Random Variable**: X ~ Beta(*a*, *b*)

- Probability Density Function (PDF): (where *a*, *b* > 0)

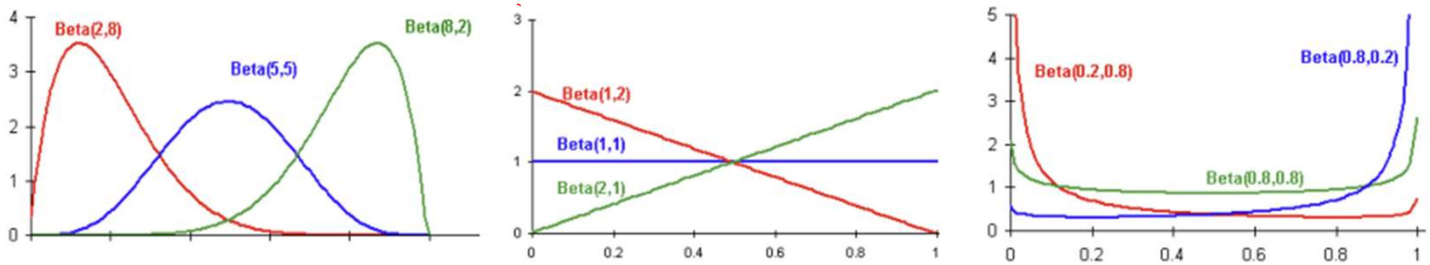
$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

$$\mathrm{B}(z_1, z_2) = \frac{\Gamma(z_1)\,\Gamma(z_2)}{\Gamma(z_1 + z_2)}$$

$z_1, z_2 \equiv$ integers

$$\frac{(z_1-1)!\,(z_2-1)!}{(z_1+z_2-1)!}$$

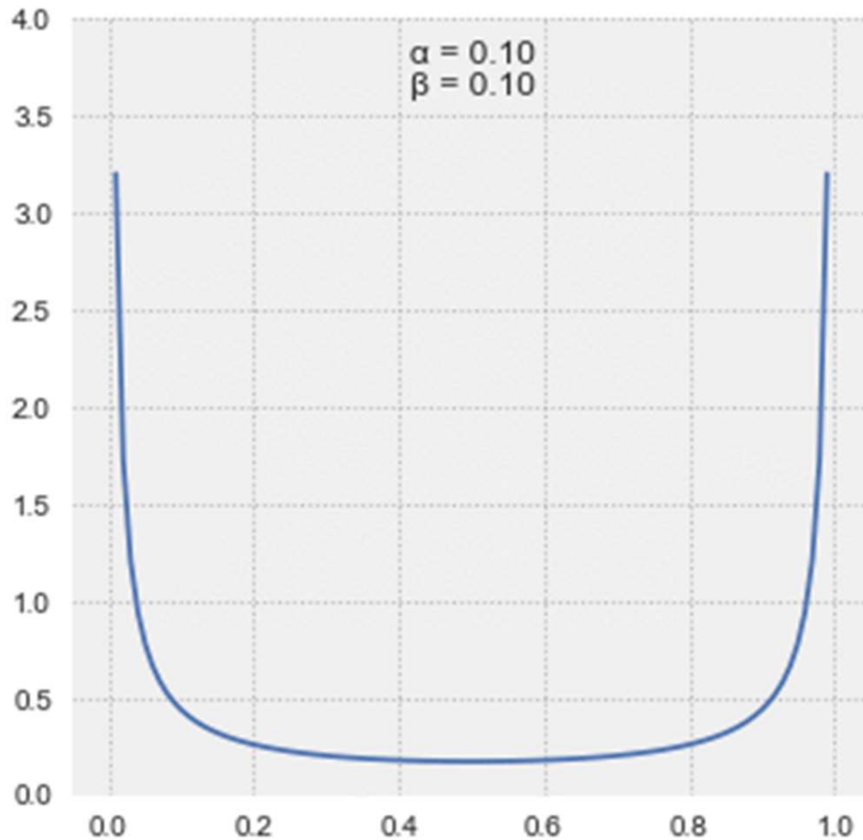


- Symmetric when *a* = *b*

$$\boxed{E[X] = \frac{a}{a+b}} \qquad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Mode : $\frac{a-1}{a+b-1}$

The shapes of the Beta distribution



Beta distribution is the distribution of probabilities

More properties of Beta distributions

- Uniform distribution U(0,1) = B(1,1)

$$f(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} = \frac{1\cdot 1}{B(a,b)} = 1$$

- Relationship between Beta and Gamma distribution

$$f(x|\alpha,\lambda) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$$

  - Let Y = G(a,1) and W= G(b,1)

$$f(y|a,1) = \frac{e^{-y} y^{a-1}}{\Gamma(a)} \qquad f(w|b,1) = \frac{e^{-w}\cdot w^{b-1}}{\Gamma(b)}$$

  - The X = Y/(Y+W) follows a Beta distribution B(a,b)

$$X = \frac{Y}{Y+W} \quad \text{then} \quad X \sim B(a,b)$$

Expected value of the posterior of Binomial

$$f_{po}(\theta \mid D) = \frac{\theta^n (1-\theta)^m}{Z} = B(a = n+1,\ b = m+1)$$

$D = \{n, m\}$

n: #1s, m: #0s

$$\hat{\theta}_{marg} = \frac{n+1}{n+m+2}$$

Laplace smoothing..

Contrast with MLE

$$\hat{\theta}_{MLE} = \frac{n}{m+n}$$

Bass guitar example: $\hat{\theta}_{marg} = \frac{1}{9}$

$$\hat{\theta}_{MLE} = \frac{0}{7}$$