

# CS 581 – ADVANCED ARTIFICIAL INTELLIGENCE

## TOPIC: MLE & BE



**Mustafa Bilgic**



<http://www.cs.iit.edu/~mbilgic>



<https://twitter.com/bilgicm>

# PARAMETER ESTIMATION

- Imagine we have a thumbtack
- Flip it, and it comes as heads or tails



- $P(\text{Heads}) = \theta$ ,  $P(\text{Tails}) = 1 - \theta$
- Assume we flip it  $(a + b)$  times and it comes head  $a$  times. What is  $\theta$  if
  - $a = 4, b = 6$
  - $a = 42, b = 58$
  - $a = 407, b = 593$
- Can you prove your answers?
- Can you associate a confidence score with your estimates?

# WE WILL SEE TWO APPROACHES

1. Maximum likelihood estimation
2. Bayesian estimation

# MAXIMUM LIKELIHOOD ESTIMATION

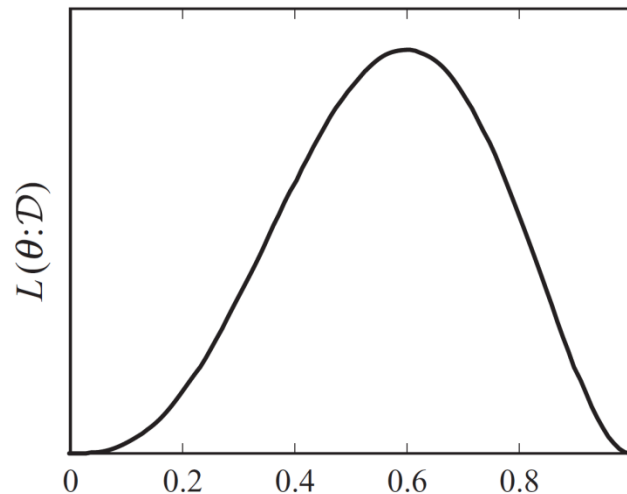
# MAXIMUM LIKELIHOOD ESTIMATION (MLE)

## – FOR THE THUMBSTACK EXAMPLE

- Assume we have a set of thumbstack tosses
  - $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$  where  $a$  are Heads and  $b$  are Tails
- Hypothesis space:  $[0, 1]$
- Find a “good” scoring  $\theta \in [0, 1]$
- Let  $f(\theta: \mathcal{D})$  be the score of  $\theta$  given  $\mathcal{D}$ , where a high score is a “good” score
- Learning:  $\operatorname{argmax}_{\theta \in [0, 1]} f(\theta: \mathcal{D})$
- What is  $f(\theta: \mathcal{D})$ ?
- The space  $[0, 1]$  is infinite. How do we search this space efficiently?

# LIKELIHOOD

- What is the probability, or *likelihood*, of seeing the sequence H, T, T, H, H?
  - $\theta * (1 - \theta) * (1 - \theta) * \theta * \theta = \theta^3(1 - \theta)^2$



When is  $L(\theta; \mathcal{D})$  maximum?

# LIKELIHOOD/LOG-LIKELIHOOD

- Number of heads =  $\alpha$ , number of tails =  $b$
- Likelihood:  $L(\theta:\mathcal{D}) = \theta^\alpha(1-\theta)^b$
- Log-likelihood:  $l(\theta:\mathcal{D}) = \alpha\log\theta + b\log(1-\theta)$
- Note that  $L(\theta:\mathcal{D})$  achieves its maximum for  $\theta$  that maximizes  $l(\theta:\mathcal{D})$
- Find  $\theta$  that maximizes the log-likelihood
- Take derivate of  $l(\theta:\mathcal{D})$  w.r.t.  $\theta$  and set it to zero

# BAYESIAN ESTIMATION



# BAYESIAN ESTIMATION

- MLE gives the same estimate of  $\theta = 0.4$ , if we have 4 H and 6 T, as well as 4M H and 6M T
- In Bayesian estimation, rather than a single  $\theta$ ,
  - We will have a prior belief about  $\theta$ :  $p(\theta)$
  - The posterior distribution over  $\theta$ :  $p(\theta | \mathcal{D})$
  - The probability distribution for the next toss:  
 $P(d_{m+1} | \mathcal{D})$

## POSTERIOR: $p(\theta \mid \mathcal{D})$

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta)P(\mathcal{D} \mid \theta)}{P(\mathcal{D})}$$

$P(\mathcal{D})$  does not depend on  $\theta$ . Hence, it can be treated as a constant from the perspective of  $\theta$ .

$$p(\theta \mid \mathcal{D}) \propto p(\theta)P(\mathcal{D} \mid \theta)$$

Next, assume each data point is independent given  $\theta$ :  $d_i \perp d_j \mid \theta$

$$P(\mathcal{D} \mid \theta) = P(d_1 \mid \theta)P(d_2 \mid \theta)P(d_3 \mid \theta) \cdots P(d_m \mid \theta) = \prod_{i=1}^m P(d_i \mid \theta)$$

Hence, the posterior becomes

$$p(\theta \mid \mathcal{D}) \propto p(\theta) \prod_{i=1}^m P(d_i \mid \theta)$$

## PREDICTION: $P(d_{m+1} \mid \mathcal{D})$

$$P(d_{m+1} \mid \mathcal{D}) = \int_0^1 P(d_{m+1} \mid \theta, \mathcal{D}) p(\theta \mid \mathcal{D}) d\theta$$

Assuming  $d_i \perp d_j \mid \theta$ :

$$P(d_{m+1} \mid \mathcal{D}) = \int_0^1 P(d_{m+1} \mid \theta) p(\theta \mid \mathcal{D}) d\theta$$

Using the posterior equation from the previous slide:

$$P(d_{m+1} \mid \mathcal{D}) \propto \int_0^1 P(d_{m+1} \mid \theta) p(\theta) \prod_{i=1}^m P(d_i \mid \theta) d\theta$$

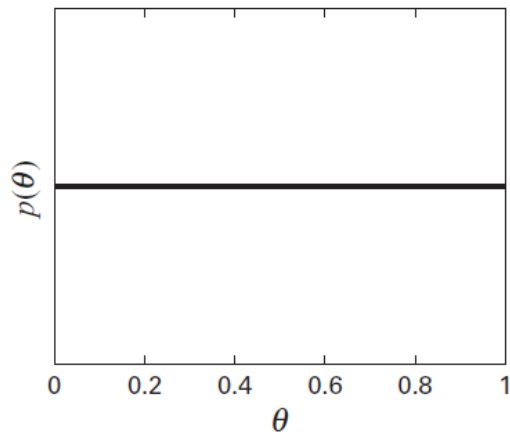
# UNIFORM PRIOR

- Assume  $a$  Heads and  $b$  Tails
- Assume a uniform prior over  $\theta$ . That is,  $p(\theta) = 1$
- What is  $P(d_{m+1} = \text{Heads} \mid \mathcal{D})$ ?
  - $(a+1)/(a+b+2)$
- What is  $p(\theta \mid \mathcal{D})$ ?
  - $\text{Beta}(a+1, b+1)$

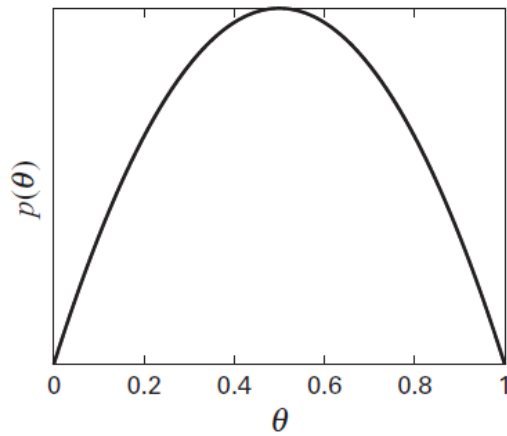
# BETA DISTRIBUTION

- $\theta \sim \text{Beta}(\alpha, \beta)$  if  $p(\theta) = \gamma \theta^{\alpha-1} (1-\theta)^{\beta-1}$  where  $\gamma$  is a normalizing constant
- Mean:  $\alpha/(\alpha+\beta)$
- Mode:  $(\alpha-1)/(\alpha+\beta-2)$
- Note that the mode is closer to the mean when  $\alpha$  and  $\beta$  are large
- Read more at
  - [https://en.wikipedia.org/wiki/Beta\\_distribution](https://en.wikipedia.org/wiki/Beta_distribution)

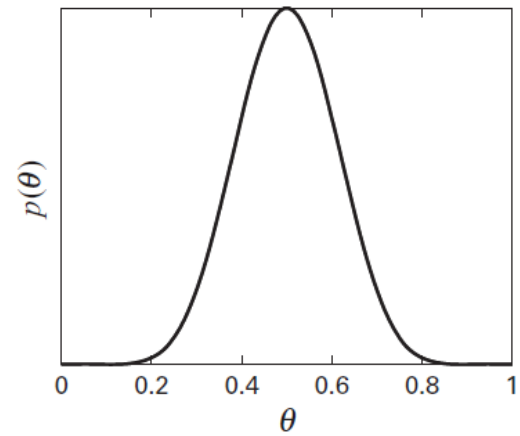
# BETA DISTRIBUTION



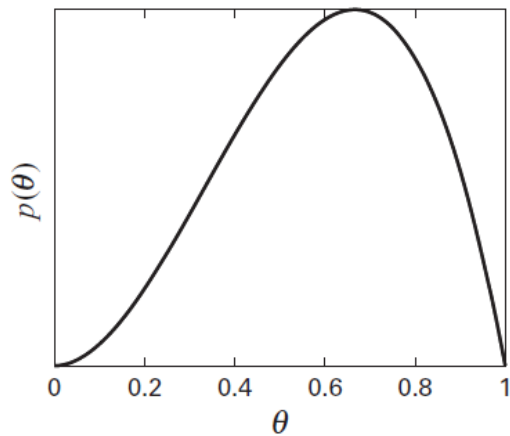
*Beta(1,1)*



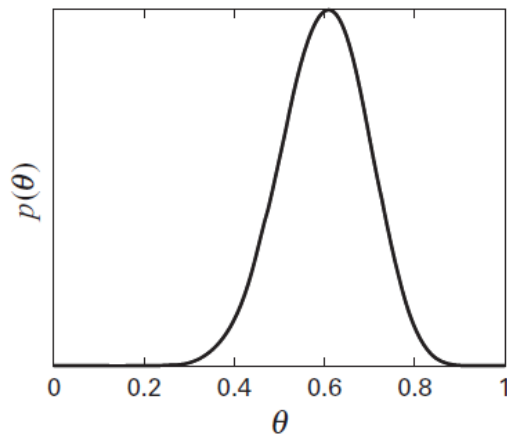
*Beta(2,2)*



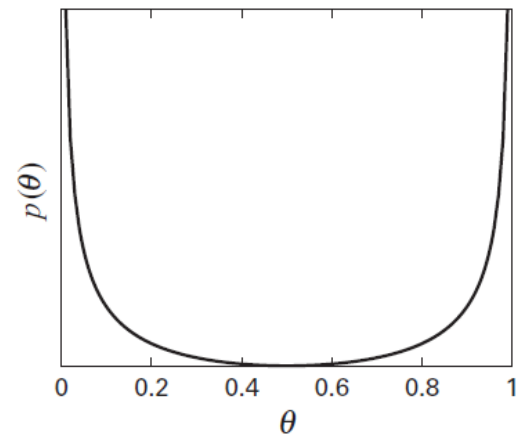
*Beta(10,10)*



*Beta(3,2)*



*Beta(15,10)*



*Beta(0.5,0.5)*

# BETA DISTRIBUTION

- What is  $P(d_{M+1}=True \mid d_1, \dots, d_M)$  if the prior is  $Beta(\alpha, \beta)$ ?
  - $P(X[M+1]=True \mid D) = (a + \alpha) / (a + b + \alpha + \beta)$
- What is the posterior,  $P(\theta \mid D)$ , if the prior is  $Beta(\alpha, \beta)$ ?
  - $P(\theta \mid D) = Beta(a + \alpha, b + \beta)$
- $\alpha$  and  $\beta$  work like pseudo-counts for the positive and negative cases respectively
- What values to choose for  $\alpha$  and  $\beta$ ?
  - It depends on our belief and the strength of our belief

# DIRICHLET PRIORS

- Generalizes the Beta distribution for multinomials

$$\theta \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K) \text{ if } P(\theta) \propto \prod_{k=1}^K \theta_k^{\alpha_k - 1}$$

- What is  $P(d_{M+1}=v_i | D)$  if the prior is Dirichlet?
  - $P(d_{M+1}=v_i | D) = (n_i + \alpha_i) / (|D| + \alpha)$  where  $n_i$  is the number of times the  $i^{\text{th}}$  case appears in  $D$  and  $\alpha = \alpha_1 + \alpha_2 + \dots + \alpha_K$
- What is the posterior,  $P(\theta | D)$ , if the prior is Dirichlet?
  - $P(\theta | D) = \text{Dirichlet}(n_1 + \alpha_1, n_2 + \alpha_2, \dots, n_K + \alpha_K)$