CS 581 – Advanced Artificial Intelligence

Topic: Parameter Estimation

Mustafa Bilgic

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

Parameter Estimation

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

- P(Heads) = θ, P(Tails) = 1- θ
- Assume we flip it $(a + b)$ times and it comes head a times. What is θ 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

PROBABILITY OF DATA

- Experiment with thumbtack tosses
- Data is H, T, H, H, T, T, T, H, T, T
  - 4 H, 6 T
- What is P(H, T, H, H, T, T, T, H, T, T)?
- If $P(H) = 0.3$
  - $0.3 \times 0.7 \times 0.3 \times 0.3 \times 0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7$
  - $0.3^4 \times 0.7^6$
  - 9.53×10^{-4}

Likelihood of D

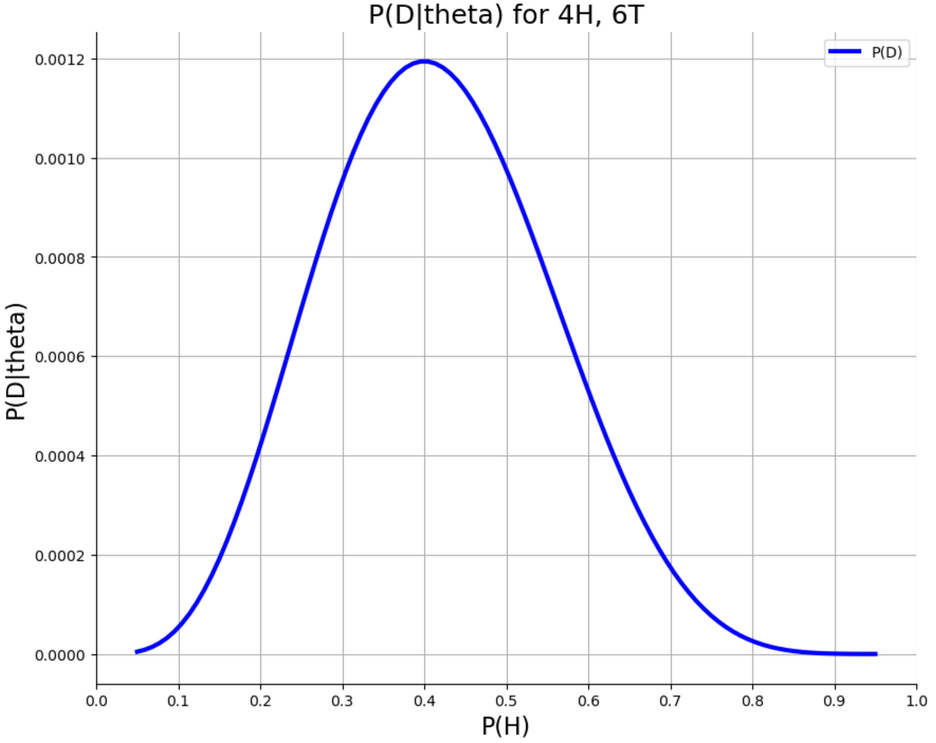
PROBABILITY OF DATA, GIVEN θ

$P(H); \boldsymbol{\theta}$	$\boldsymbol{\theta^4 \times (1-\theta)^6}$	$P(D \mid \theta)$
0.0	$0^4 \times 1^6$	0.0
0.1	$0.1^4 \times 0.9^6$	0.53×10^{-4}
0.2	$0.2^4 \times 0.8^6$	4.19×10^{-4}
0.3	$0.3^4 \times 0.7^6$	9.53×10^{-4}
0.4	$0.4^4 \times 0.6^6$	$\mathbf{11.94 \times 10^{-4}}$
0.5	$0.5^4 \times 0.5^6$	9.77×10^{-4}
0.6	$0.6^4 \times 0.4^6$	5.31×10^{-4}
0.7	$0.7^4 \times 0.3^6$	1.75×10^{-4}
0.8	$0.8^4 \times 0.2^6$	0.26×10^{-4}
0.9	$0.9^4 \times 0.1^6$	0.01×10^{-4}
1.0	$1^4 \times 0^6$	0.0

Probability of Data, Given θ

$\theta^4(1-\theta)^6$

$P(H); \boldsymbol{\theta}$	$\boldsymbol{\theta^4 \times (1-\theta)^6}$	
0.0	$0^4 \times 1^6$	0.0
0.1	$0.1^4 \times 0.9^6$	0.53×10^{-4}
0.2	$0.2^4 \times 0.8^6$	4.19×10^{-4}
0.3	$0.3^4 \times 0.7^6$	9.53×10^{-4}
0.4	$0.4^4 \times 0.6^6$	$\mathbf{11.94 \times 10^{-4}}$
0.5	$0.5^4 \times 0.5^6$	9.77×10^{-4}
0.6	$0.6^4 \times 0.4^6$	5.31×10^{-4}
0.7	$0.7^4 \times 0.3^6$	1.75×10^{-4}
0.8	$0.8^4 \times 0.2^6$	0.26×10^{-4}
0.9	$0.9^4 \times 0.1^6$	0.01×10^{-4}
1.0	$1^4 \times 0^6$	0.0



P(Data | θ)

- Data is 4000 H, 6000 T
- What is $P(D \mid \theta)$ if P(H) is θ?
  - $\theta^{4000} \times (1-\theta)^{6000}$
- If $\theta = 0.4$
  - $0.4^{4000} \times 0.6^{6000} = 0$ (underflow!)

Log Probability of Data, Given θ

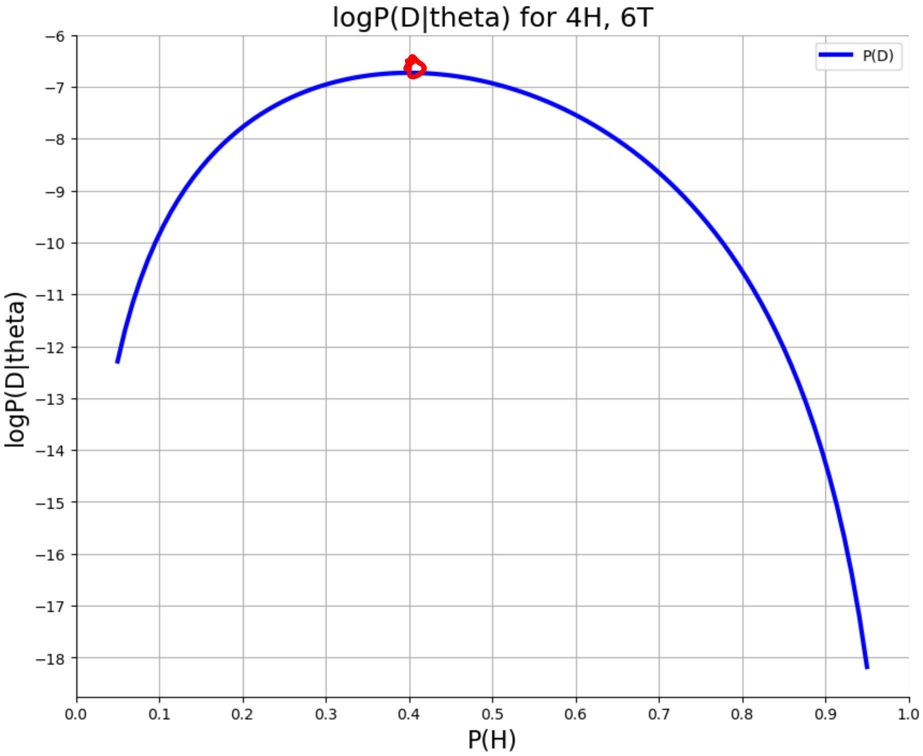
- Experiment with thumbtack tosses
- Data is H, T, H, H, T, T, T, H, T, T
  - 4 H, 6 T
- What is $\ln(P(\text{H, T, H, H, T, T, T, H, T, T}))$?
- If $P(H) = 0.3$
  - $\ln(0.3^4 \times 0.7^6)$
  - $4 * \ln(0.3) + 6 * \ln(0.7)$
  - -6.956

LogProbability of Data, Given θ

$P(H); \boldsymbol{\theta}$	$\mathbf{4 \times ln(\theta) + 6 \times ln(1-\theta)}$	$ln\ P(D\mid\theta)$
0.0	$4 \times ln(0) + 6 \times ln(1)$	$-\infty$
0.1	$4 \times ln(0.1) + 6 \times ln(0.9)$	-9.84
0.2	$4 \times ln(0.2) + 6 \times ln(0.8)$	-7.78
0.3	$4 \times ln(0.3) + 6 \times ln(0.7)$	-6.96
0.4	$4 \times ln(0.4) + 6 \times ln(0.6)$	$\boxed{\mathbf{-6.73}}$
0.5	$4 \times ln(0.5) + 6 \times ln(0.5)$	-6.93
0.6	$4 \times ln(0.6) + 6 \times ln(0.4)$	-7.54
0.7	$4 \times ln(0.7) + 6 \times ln(0.3)$	-8.65
0.8	$4 \times ln(0.8) + 6 \times ln(0.2)$	-10.55
0.9	$4 \times ln(0.9) + 6 \times ln(0.1)$	-14.24
1.0	$4 \times ln(1) + 6 \times ln(0)$	$-\infty$

LogProbability of Data, Given θ

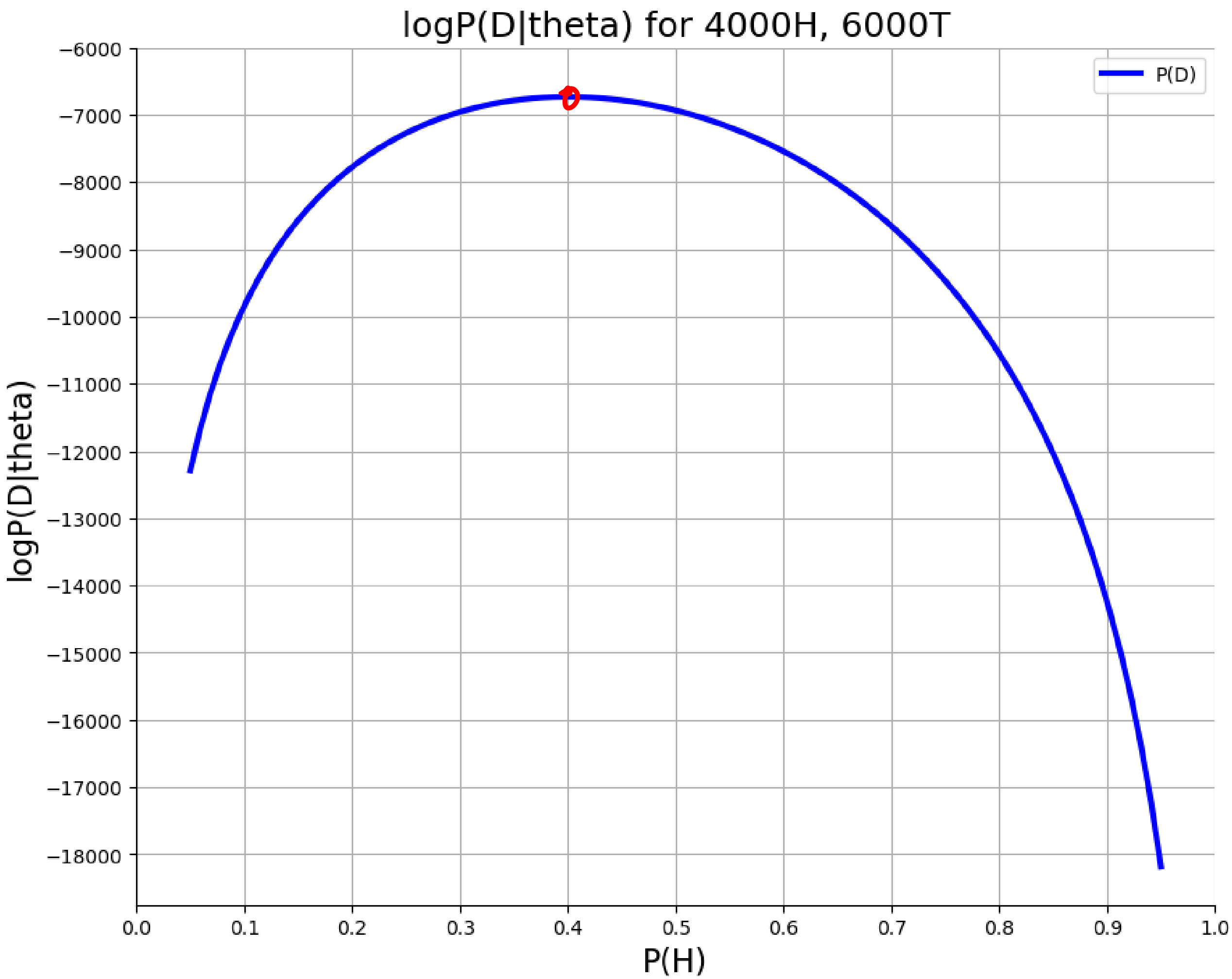
$P(H); \boldsymbol{\theta}$	$\boldsymbol{4 \times ln(\theta) + 6 \times ln(1-\theta)}$	
0.0	$4 \times ln(0) + 6 \times ln(1)$	$-\infty$
0.1	$4 \times ln(0.1) + 6 \times ln(0.9)$	-9.84
0.2	$4 \times ln(0.2) + 6 \times ln(0.8)$	-7.78
0.3	$4 \times ln(0.3) + 6 \times ln(0.7)$	-6.96
0.4	$4 \times ln(0.4) + 6 \times ln(0.6)$	$\mathbf{-6.73}$
0.5	$4 \times ln(0.5) + 6 \times ln(0.5)$	-6.93
0.6	$4 \times ln(0.6) + 6 \times ln(0.4)$	-7.54
0.7	$4 \times ln(0.7) + 6 \times ln(0.3)$	-8.65
0.8	$4 \times ln(0.8) + 6 \times ln(0.2)$	-10.55
0.9	$4 \times ln(0.9) + 6 \times ln(0.1)$	-14.24
1.0	$4 \times ln(1) + 6 \times ln(0)$	$-\infty$



P(DATA | θ)

- Data is 4000 H, 6000 T
- What is $P(D \mid \theta)$ if P(H) is θ?
  - $\theta^{4000} \times (1-\theta)^{6000}$
- If $\theta = 0.4$
  - $0.4^{4000} \times 0.6^{6000} = 0$ (underflow!)
- $\log(P(D \mid \theta))$
  - $4000 \times \ln(0.4) + 6000 \times \ln(0.6) = -6730.12$

LOG(P(4000H, 6000T | θ))



LIKELIHOOD/LOG-LIKELIHOOD

- Number of heads = a, number of tails = b
- Likelihood: $L(\theta : \mathcal{D}) = \theta^a (1-\theta)^b$
- Log-likelihood: $l(\theta : \mathcal{D}) = a \log(\theta) + b \log(1-\theta)$
- Note that θ that maximizes likelihood $L(\theta : \mathcal{D})$ is the same θ that maximizes the log-likelihood $l(\theta : \mathcal{D})$
  - For non-negative x and y, $x \geq y \iff \log(x) \geq \log(y)$
- How to find θ that maximizes the log-likelihood?
  - Take derivate of $l(\theta : \mathcal{D})$ w.r.t. θ 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 θ,
  - We assume a prior belief about θ: $\boldsymbol{p(\theta)}$, and we estimate
    - The posterior distribution over θ: $\boldsymbol{p(\theta \mid \mathcal{D})}$
    - The probability distribution for the next toss: $\boldsymbol{P(d_{m+1} \mid \mathcal{D})}$

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 θ. Hence, it can be treated as a constant from the perspective of θ.

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

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

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

Hence, the posterior becomes

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

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

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

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

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

Using the posterior equation from the previous slide:

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

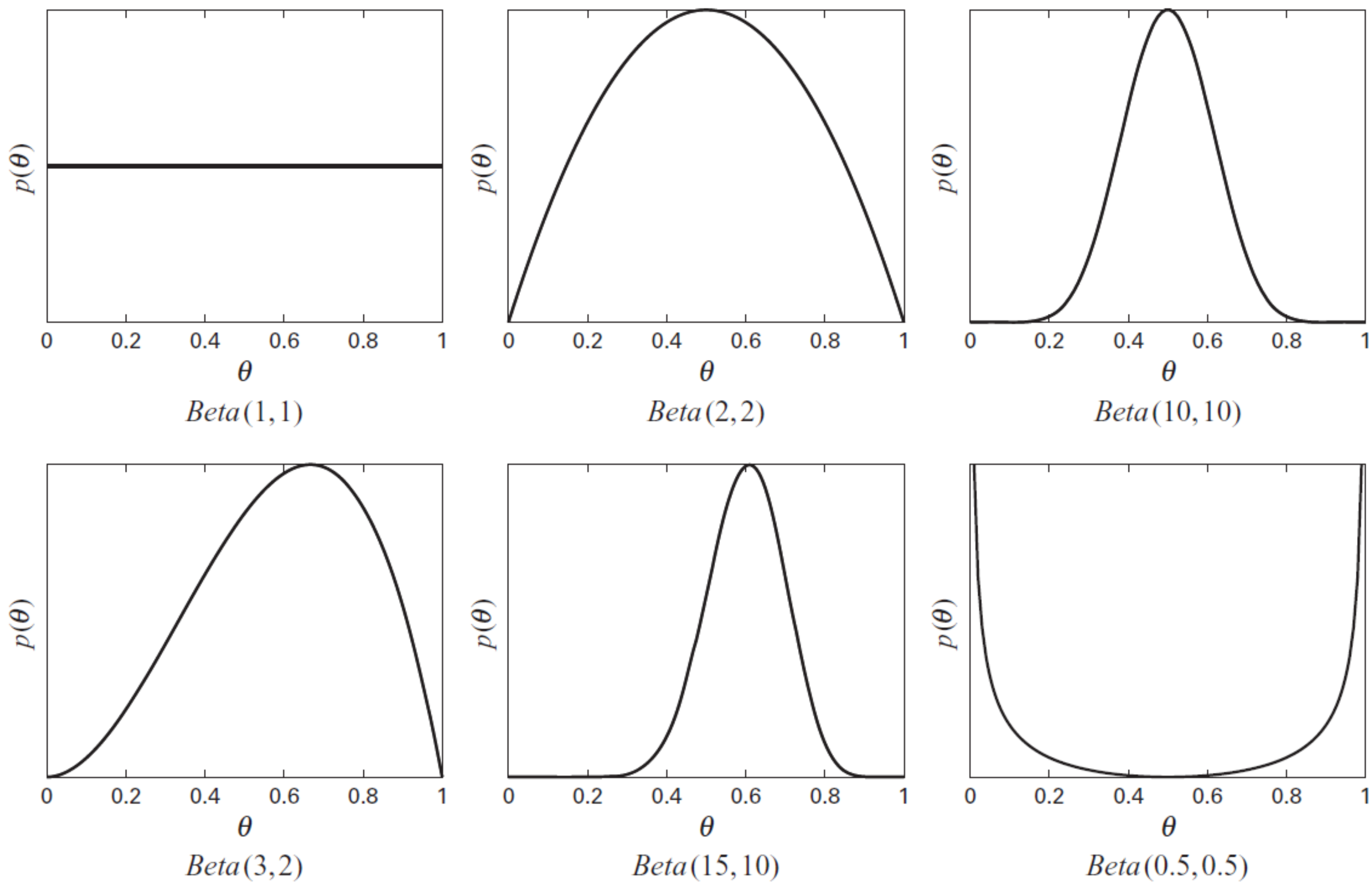
Uniform prior

- Assume a Heads and b Tails
- Assume a uniform prior over θ. That is, $p(\theta) = 1$
- What is $P(d_{m+1} = Heads \mid \mathcal{D})$?
  - (a+1)/(a+b+2)
- What is $p(\theta \mid \mathcal{D})$?
  - 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 γ is a normalizing constant
- Mean: $\alpha/(\alpha+\beta)$
- Mode: $(\alpha-1)/(\alpha+\beta-2)$
- Note that the mode is closer to the mean when α and β are large
- Read more at
  - https://en.wikipedia.org/wiki/Beta_distribution

Beta Distribution



Beta distribution

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

DIRICHLET PRIORS

- Generalizes the Beta distribution for multinomials

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

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