CS 1675 Introduction to Machine Learning
Lecture 7

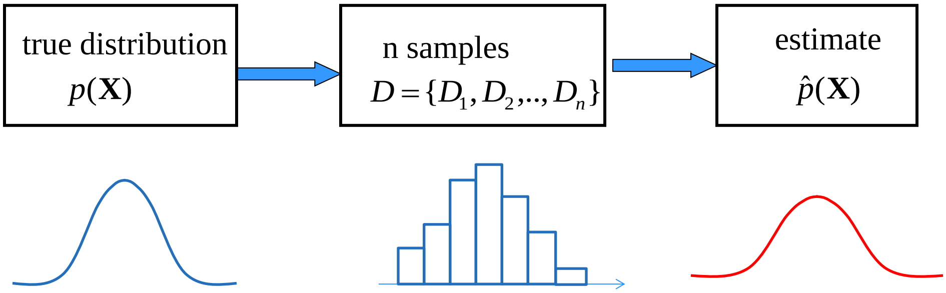
Density estimation II

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$
$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D



true distribution $p(\mathbf{X})$

n samples $D = \{D_1, D_2, .., D_n\}$

estimate $\hat{p}(\mathbf{X})$

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML) parameter estimation:**
  - **maximizes the data likelihood**

$$\boxed{\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi)}$$

Log-likelihood $\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$

Maximization of the data likelihood = maximization of the data log-likelihood

$$\boxed{\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)}$$

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta, \xi)$$

Optimize log-likelihood (the same as maximizing likelihood)

$$l(D, \theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} =$$

$$\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) = \log\theta \underline{\sum_{i=1}^{n} x_i} + \log(1-\theta) \underline{\sum_{i=1}^{n} (1-x_i)}$$

N_1 - number of heads seen N_2 - number of tails seen

Maximum likelihood (ML) estimate.

Optimize log-likelihood

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Set derivative to zero

$$\frac{\partial l(D,\theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{(1-\theta)} = 0$$

Solving $\theta = \dfrac{N_1}{N_1 + N_2}$

ML Solution: $\theta_{ML} = \dfrac{N_1}{N} = \dfrac{N_1}{N_1 + N_2}$

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Head: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

Tail: $$(1-\theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

Bayesian parameter estimation

The ML estimate picks just one value of the parameter

- **Problem:** if there are two different parameter values that are close in terms of the likelihood, using only one of them may introduce a strong bias, if we use it, for example, for predictions.

Bayesian parameter estimation

- Remedies the limitation of one choice
- Uses the posterior distribution for parameters Θ
- Posterior 'covers' all possible parameter values (and their "weights")

Parameter posterior

Data Likelihood

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)} \longleftarrow \text{Parameter prior}$$

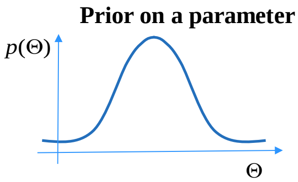
Bayesian parameter estimation

What does it do?

- Prior and Posterior 'covers' all possible parameter values (and their "weights")

Assume: we have a model of $p(x \mid \Theta)$ with a parameter Θ

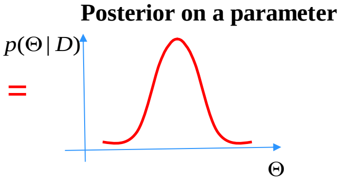
- **Bayesian parameter estimation:**



+

Data +
$p(D \mid \Theta)$

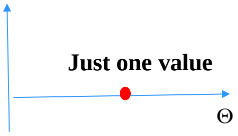
=



- **ML parameter estimate**

Data +
$p(D \mid \Theta)$

=



Bayesian parameter estimation

Bayesian parameter estimation

- Uses the posterior distribution for parameters
- Posterior 'covers' all possible parameter values (and their "weights")

$$\underset{\text{Parameter posterior}}{p(\Theta \mid D, \xi)} = \frac{\overset{\text{Data Likelihood}}{p(D \mid \Theta, \xi)}\, p(\Theta \mid \xi) \leftarrow \text{Parameter prior}}{p(D \mid \xi)}$$

- **How to use the posterior for modeling** p(X)?

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\mathbf{\Theta}} p(X \mid \mathbf{\Theta})\, p(\mathbf{\Theta} \mid D, \xi)\, d\mathbf{\Theta}$$

Bayesian parameter estimate: coin example

Calculate posterior distribution

$$p(\theta \mid D, \xi)$$

Likelihood of data

prior

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Normalizing factor

Likelihood of data for a sequence of n coin flips:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

$p(\theta \mid \xi)$ - is the prior probability on θ

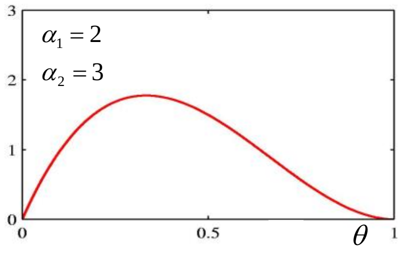
How to choose the prior probability for Bernoulli trials?

Beta distribution

Choice of prior: Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

Distribution on interval [0,1], that is: $\theta \in [0,1]$

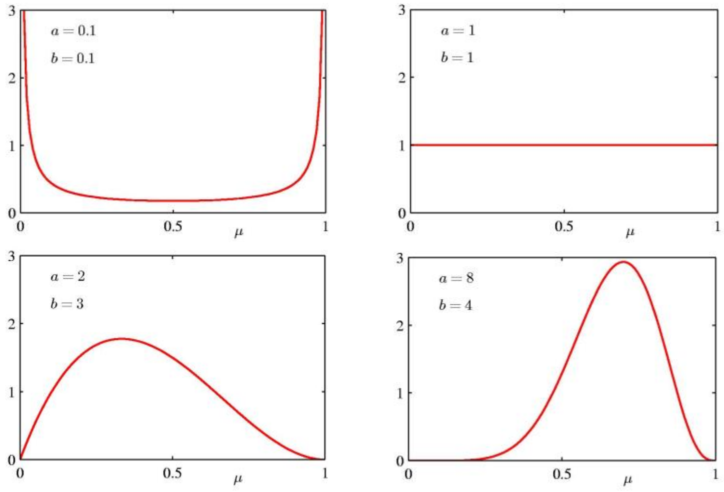


$\Gamma(x)$ - a Gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$

For integer values of x Gamma is defined by a factorial function

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

Beta distribution



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

Posterior Beta distribution

Prior Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Why to use Beta distribution?

Beta distribution "**fits**" Bernoulli trials - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution !!!!

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1+\alpha_2+N_1+N_2)}{\Gamma(\alpha_1+N_1)\Gamma(\alpha_2+N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Posterior Beta distribution

Prior Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

Why to use Beta distribution?

Beta distribution “**fits**” Bernoulli trials - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

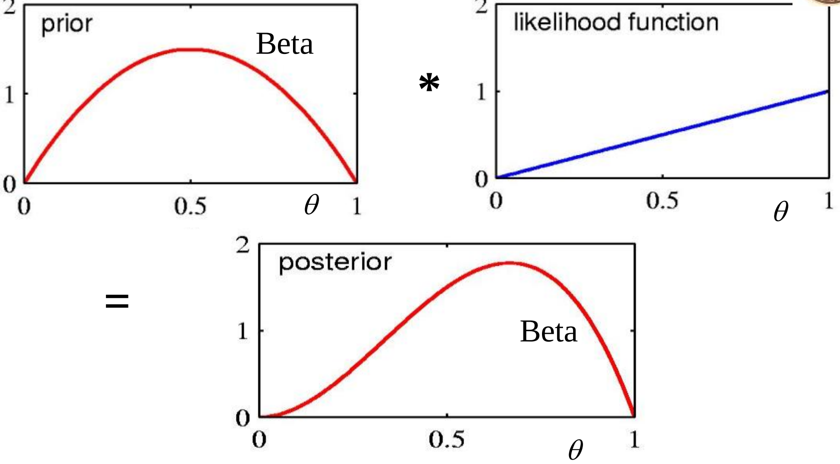
Posterior distribution is again a Beta distrib

α_1, α_2
Are referred to as
Prior counts

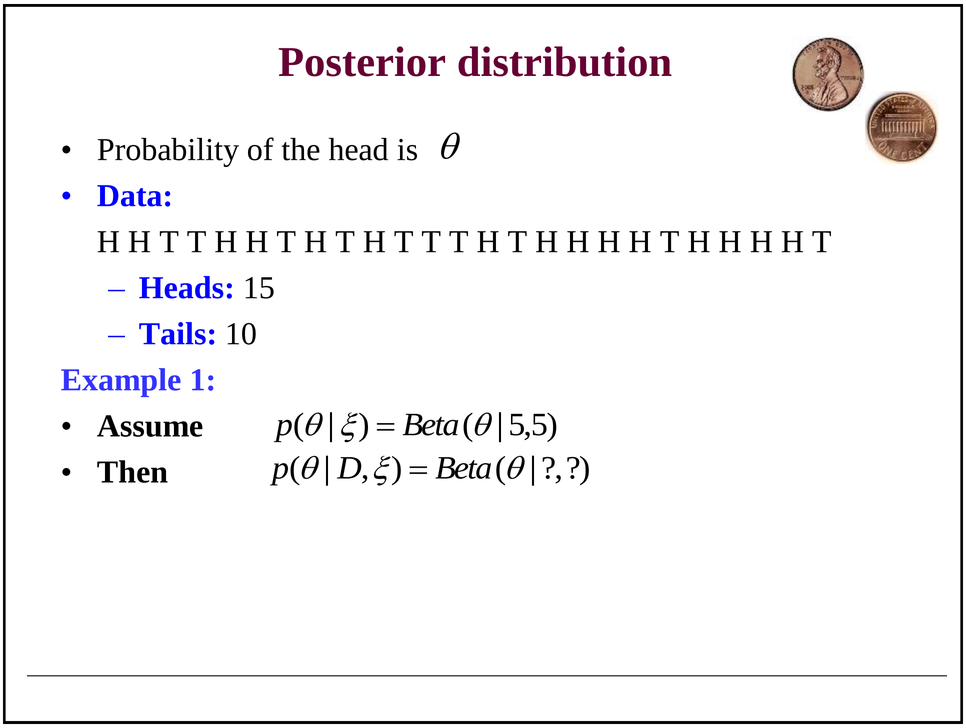
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_2 + \alpha_2 - 1}$$

Posterior distribution



$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$



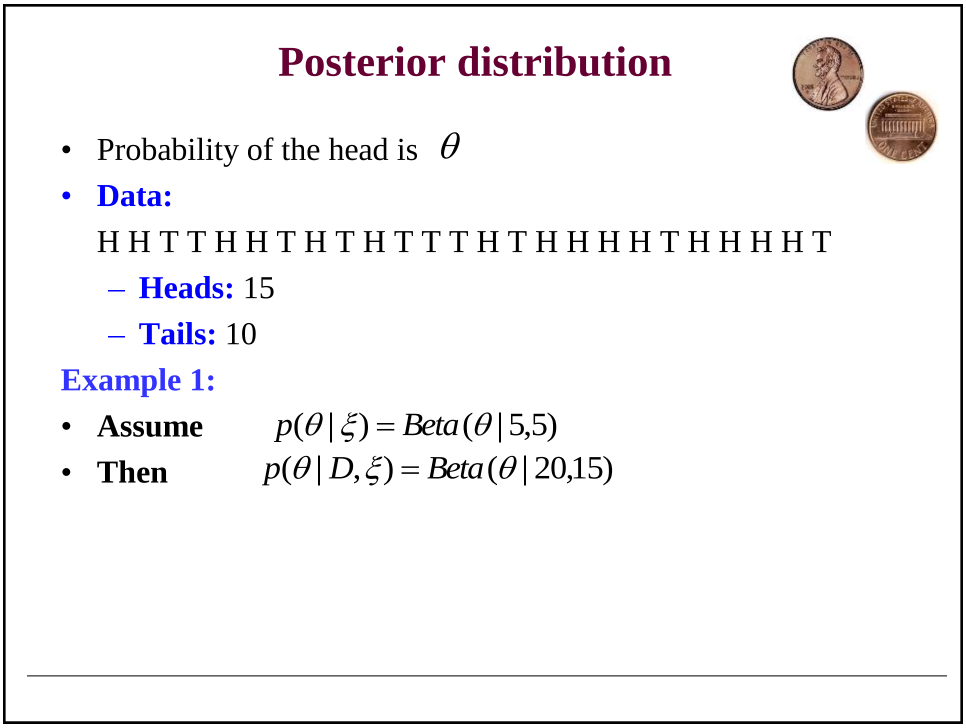
Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid ?, ?)$



Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

Example 2:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 3,1)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid ?, ?)$

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

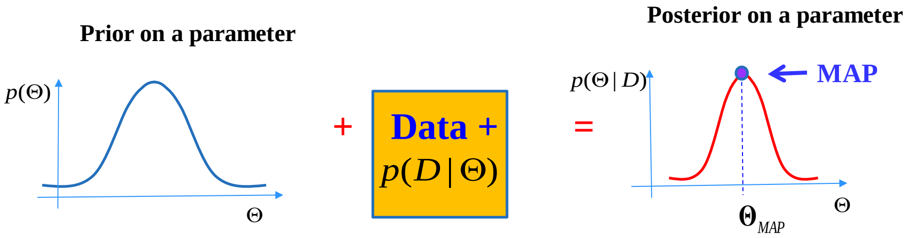
Example 2:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 3,1)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 18,11)$

Parameter estimation: MAP

- **Maximum a posteriori probability (MAP)**

  maximize $p(\mathbf{\Theta} \mid D, \xi)$



- **MAP**
  - Yields: one set of parameters $\mathbf{\Theta}_{MAP}$ (mode of the posterior)
  - Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

Maximum a posteriori estimate: MAP

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

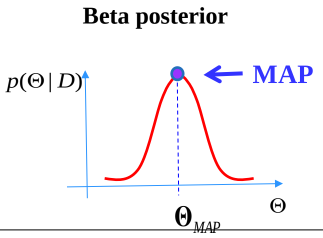
$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

$$p(\theta \mid D, \xi) = \frac{\overset{\text{Likelihood of data}}{P(D \mid \theta, \xi)}\, \overset{\text{prior}}{p(\theta \mid \xi)}}{\underset{\text{Normalizing factor}}{P(D \mid \xi)}}$$

By using Beta prior:

- **We get Beta posterior**
- **And MAP solution is:**

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$



MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Binomial distribution

Example problem: N coin flips, where each coin flip can have two results: head or tail

Outcome: N_1 - number of heads seen in N trials N_2 - number of tails seen

Model: probability of a head θ

probability of a tail $(1-\theta)$

Probability of an outcome:

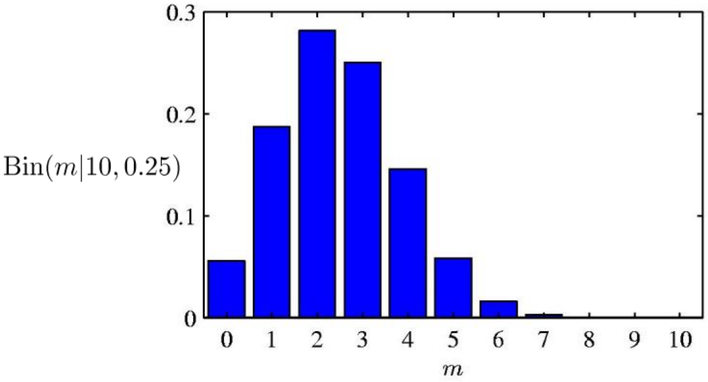
$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$ **Binomial distribution**

Binomial distribution:

- **models order independent sequence of Bernoulli trials**

Binomial distribution

Binomial distribution:



Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D\,|\,\theta) = \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2} = \frac{N!}{N_1!N_2!}\theta^{N_1}(1-\theta)^{N_2}$$

Log-likelihood

$$l(D,\theta) = \log\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2} = \log\underline{\frac{N!}{N_1!N_2!}} + N_1\log\theta + N_2\log(1-\theta)$$

Constant from the point of optimization !!!

ML Solution: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1+N_2}$$

The same as for a sequence of iid Bernoulli trials

Posterior density

Posterior density

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Prior choice

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Likelihood

$$P(D \mid \theta) = \frac{\Gamma(N_1+N_2)}{\Gamma(N_1)\Gamma(N_2)}\theta^{N_1}(1-\theta)^{N_2}$$

Posterior $p(\theta \mid D,\xi) = Beta(\alpha_1+N_1,\alpha_2+N_2)$

MAP estimate $\theta_{MAP} = \arg\max_{\theta}\; p(\theta \mid D,\xi)$

$$\theta_{MAP} = \frac{\alpha_1+N_1-1}{\alpha_1+\alpha_2+N_1+N_2-2}$$

Multinomial distribution

Example: **multiple rolls of a die with 6 results**

Outcome: counts of occurrences of *k* possible outcomes of *N* trials: N_i - a number of times an outcome i has been seen

$$\sum_{i=1}^{k} N_i = N$$

Model parameters: $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots\theta_k)$ **s.t.** $\sum_{i=1}^{k}\theta_i = 1$

θ_i - probability of an outcome i

Probability distribution:

$$P(N_1,N_2,\ldots N_k \mid \boldsymbol{\theta},\xi) = \frac{N!}{N_1!N_2!\ldots N_k!}\theta_1^{N_1}\theta_2^{N_2}\ldots\theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

Posterior density and MAP estimate

Choice of the prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1,..,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for the multinomial sampling

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Posterior density

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, ..\alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

MAP estimate:

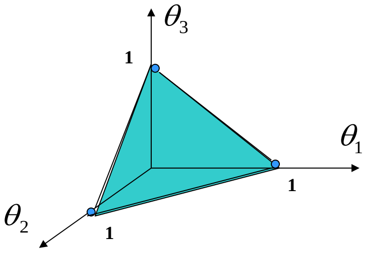
$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

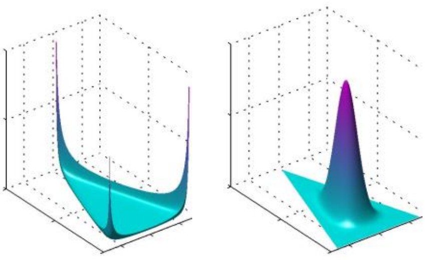
Dirichlet distribution

Dirichlet distribution:

$$Dir(\boldsymbol{\theta} \mid \alpha_1,..,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Assume: k=3





$\alpha_k = 10^{-1}$ $\alpha_k = 10^{1}$

Distribution models for random variables

Distribution models covered so far:

- **Bernoulli distribution**
  - **Model for binary random variables**

$$P(x \mid \theta) = \theta^{x}(1-\theta)^{(1-x)}$$

- **Binomial distribution**
  - **Model for order independent sets of binary outcomes**

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1}(1-\theta)^{N-N_1}$$

- **Multinomial distribution**
  - **Model for order independent sets of k-nary outcomes**

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Distribution models for random variables

Models for other types of random variables:

- **Gaussian distribution**
  - **Models of real-valued random variable**
- **Gamma distribution:**
  - **Models of random variables for positive real numbers**
- **Exponential distribution**
  - **Models of random variables for positive real numbers**
- **Poisson distribution**
  - **Models of random variables for nonegative integers**

Conjugate choices of priors for some these distributions:

- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**