MLE, MAP Estimation

Machine Learning

Hamid R Rabiee – Zahra Dehghanian

Spring 2025

Sharif University of Technology

Outline

- Introduction
- Maximum-Likelihood (ML) estimation
- Maximum A Posteriori (MAP) estimation

Relation of learning & statistics

- Target model in the learning problems can be considered as a statistical model

- For a fixed set of data and underlying target (statistical model), the estimation methods try to estimate the target from the available data

Density estimation

- Estimating the probability density function $p(\boldsymbol{x})$, given a set of data points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ drawn from it.

- Main approaches of density estimation:
  - Parametric: assuming a parameterized model for density function
    - A number of parameters are optimized by fitting the model to the data set
  - Nonparametric (Instance-based): No specific parametric model is assumed
    - The form of the density function is determined entirely by the data

Parametric density estimation

- Estimating the probability density function $p(\boldsymbol{x})$, given a set of data points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ drawn from it.

- Assume that $p(\boldsymbol{x})$ in terms of a specific functional form which has a number of adjustable parameters.

- Methods for parameter estimation
  - Maximum likelihood estimation
  - Maximum A Posteriori (MAP) estimation

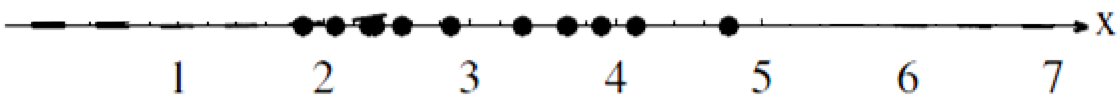
Parametric density estimation

- Goal: estimate parameters of a distribution from a dataset $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$
  - $\mathcal{D}$ contains N independent, identically distributed (i.i.d.) training samples.

- We need to determine $\boldsymbol{\theta}$ given $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$
  - How to represent $\boldsymbol{\theta}$?
    - $\boldsymbol{\theta}^*$ or $p(\boldsymbol{\theta})$?

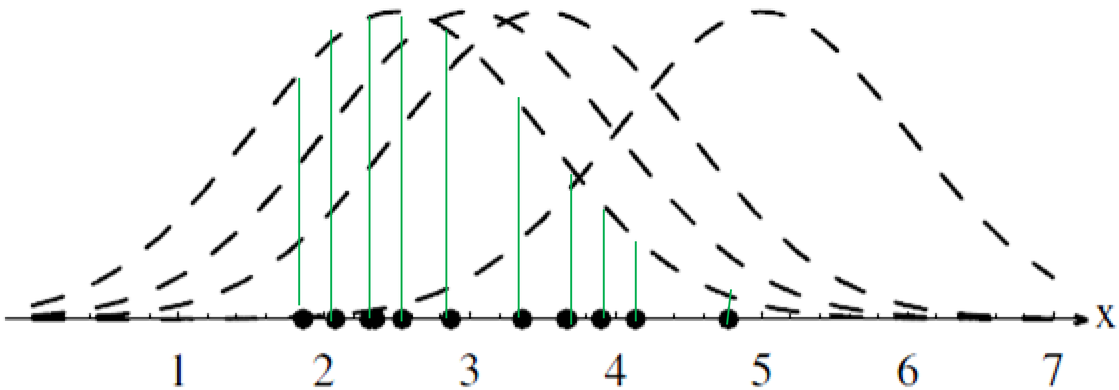
Example

-

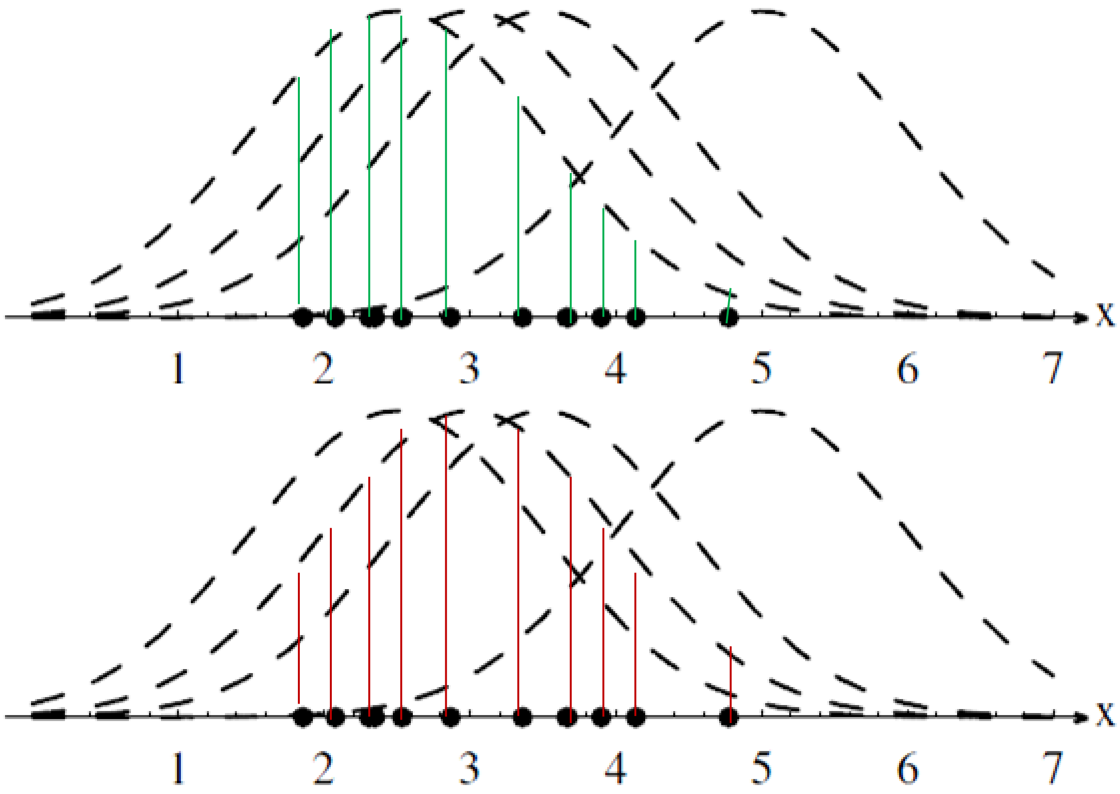


$$P(x|\mu) = N(x|\mu, 1)$$

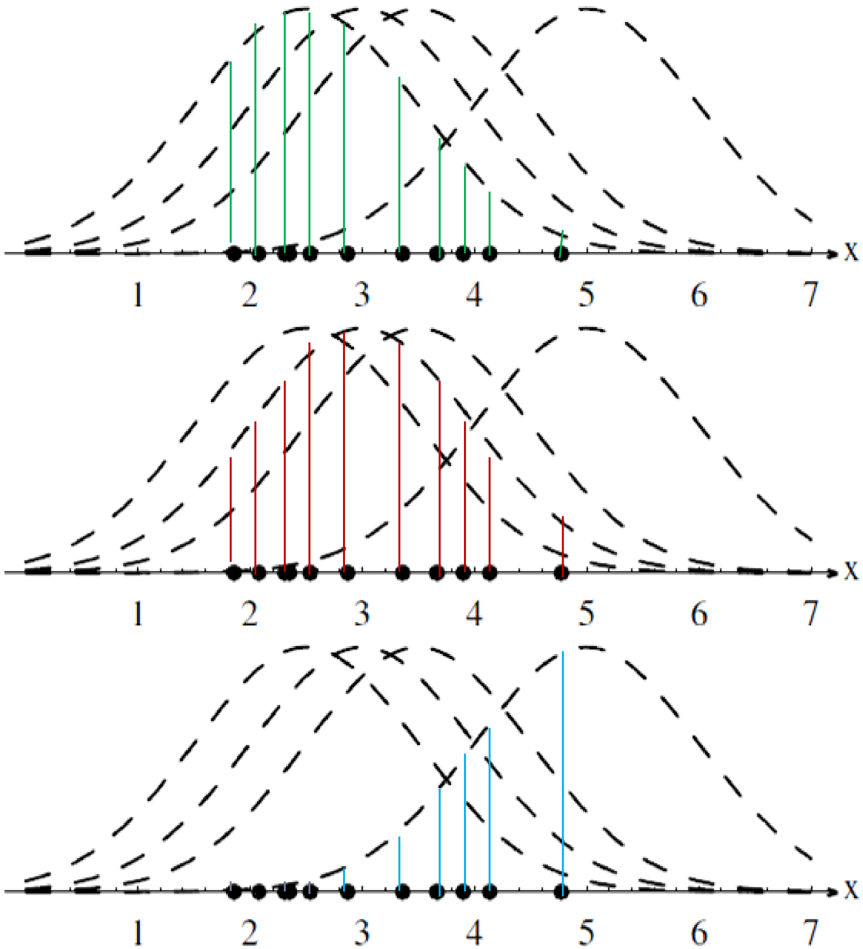
Example



Example



Example



Maximum Likelihood Estimation (MLE)

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given data.

Maximum Likelihood Estimation (MLE)

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given data.

- Likelihood is the conditional probability of observations $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\}$ given the value of parameters $\boldsymbol{\theta}$
  - Assuming i.i.d. observations:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

$p(\mathcal{D}|\boldsymbol{\theta})$ → likelihood of $\boldsymbol{\theta}$ w.r.t. the samples

Maximum Likelihood Estimation (MLE)

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given data.

- Likelihood is the conditional probability of observations $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\}$ given the value of parameters $\boldsymbol{\theta}$
  - Assuming i.i.d. observations:

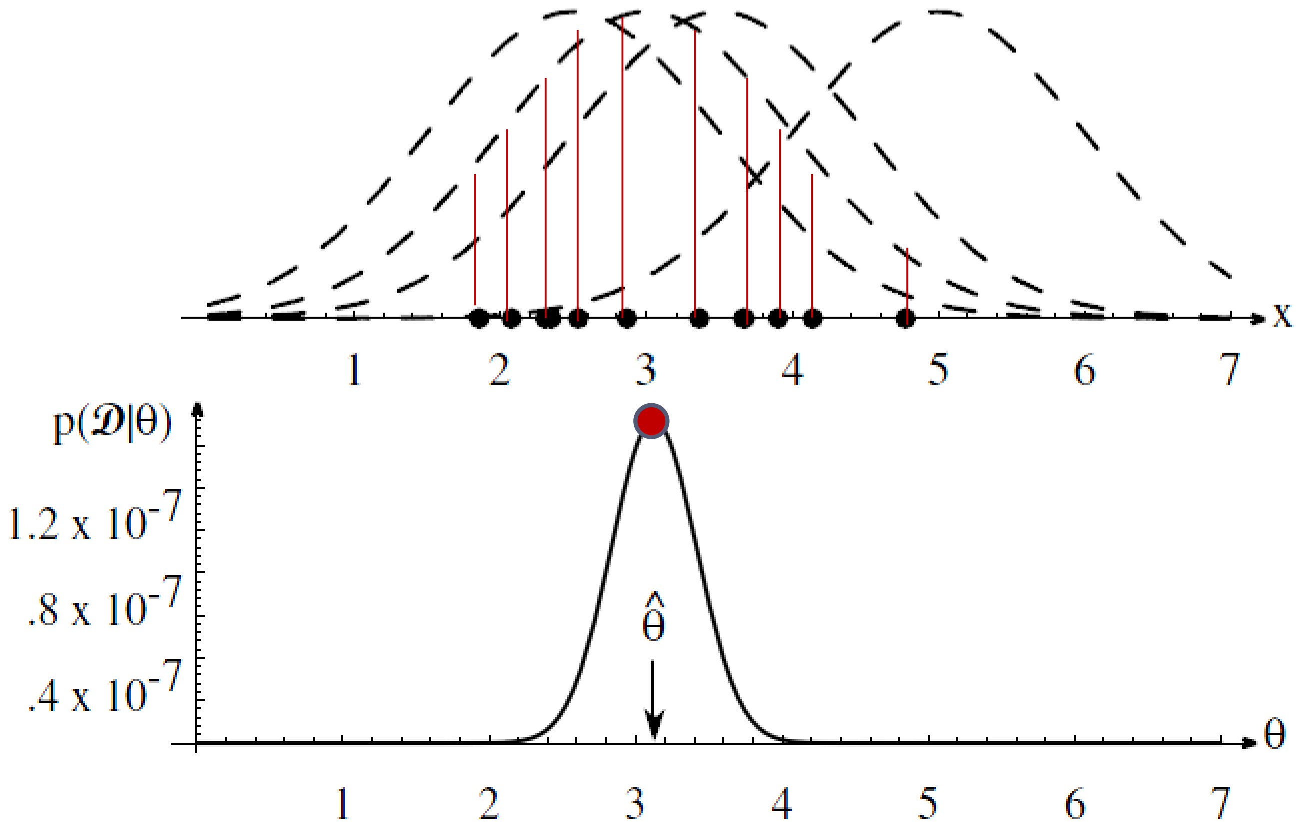
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

likelihood of $\boldsymbol{\theta}$ w.r.t. the samples

- Maximum Likelihood estimation

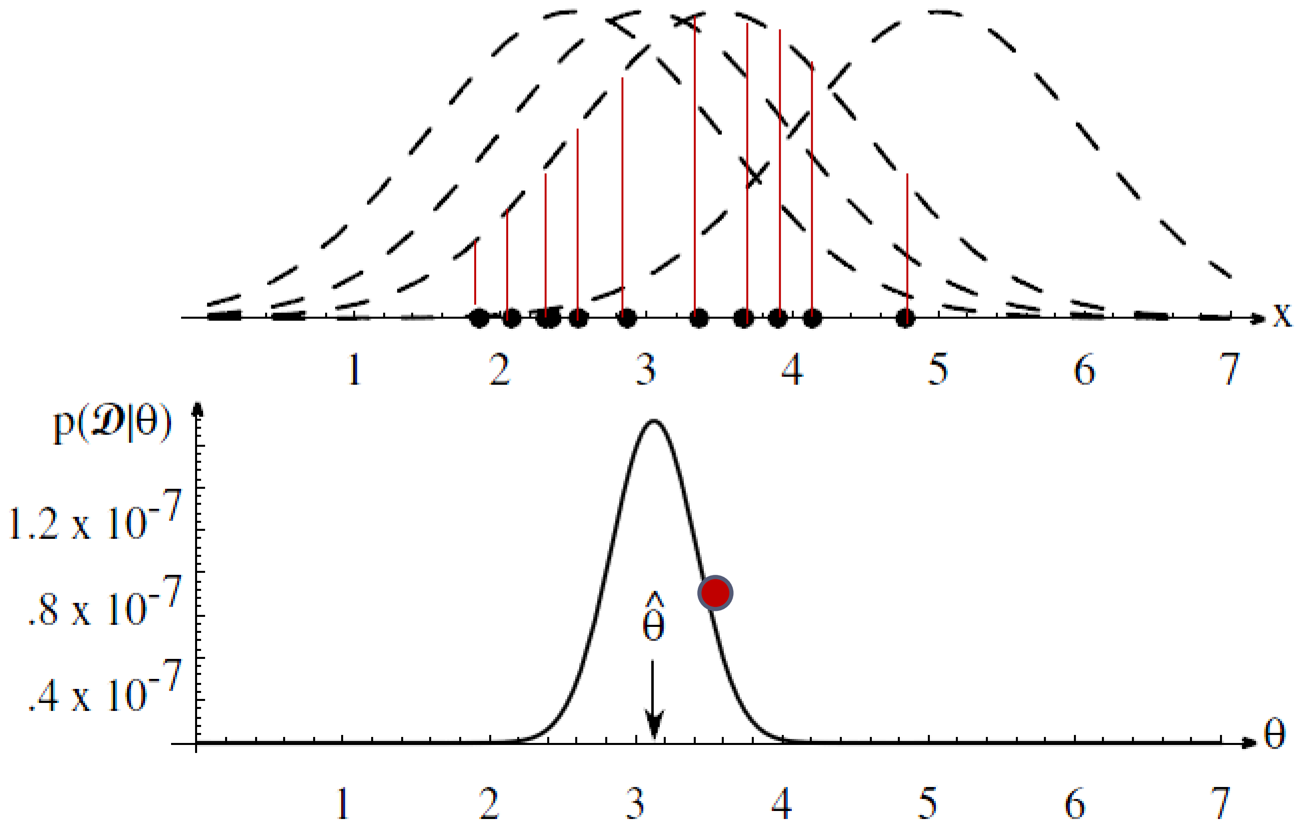
$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)



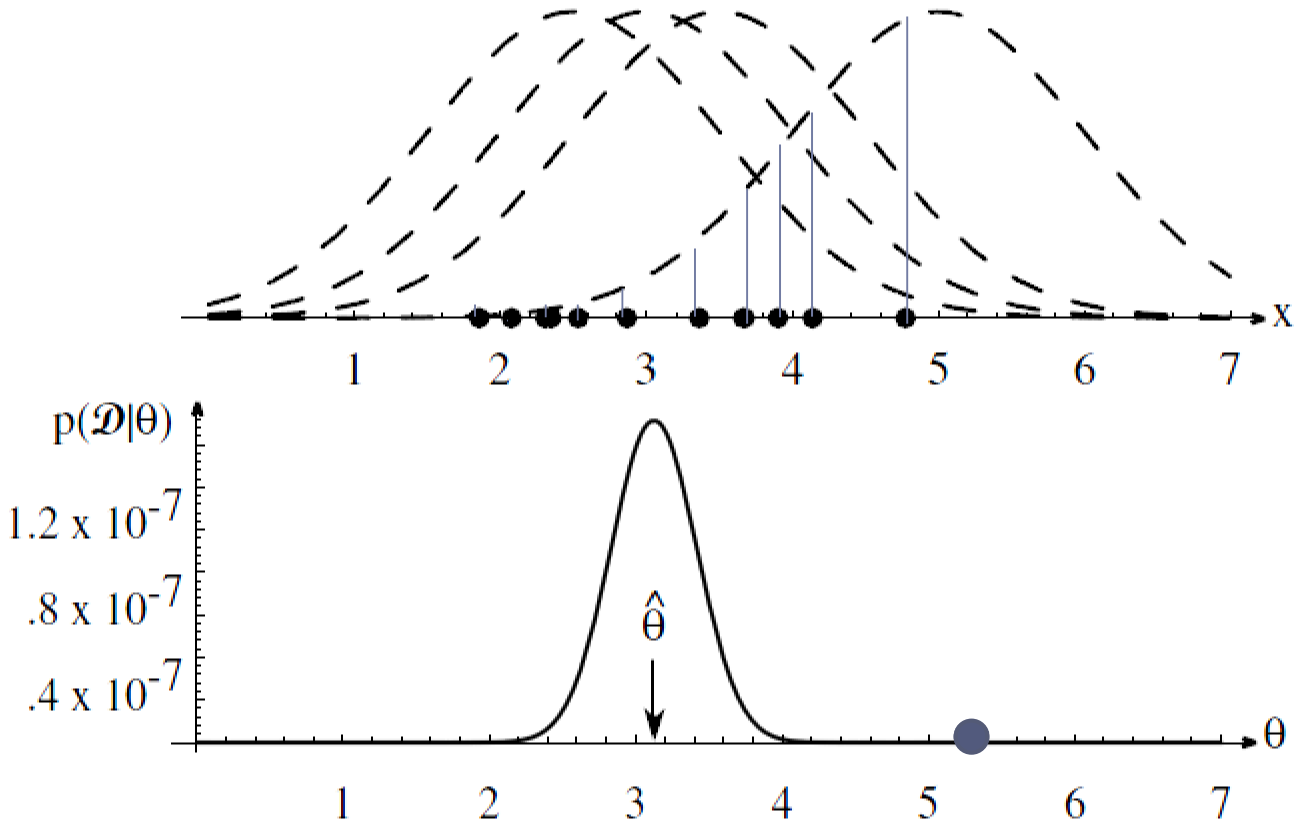
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



$\hat{\theta}$ best agrees with the observed samples

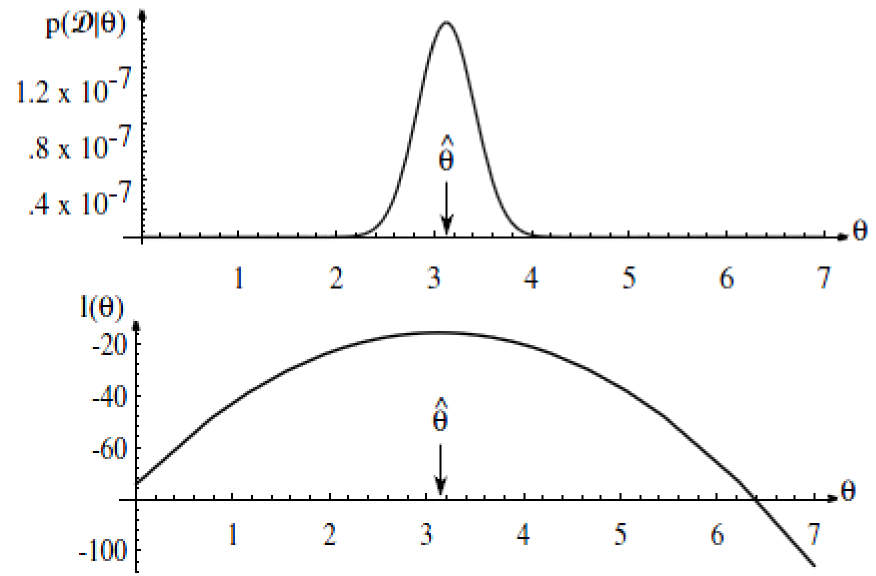
Maximum Likelihood Estimation (MLE)



$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)

- $$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

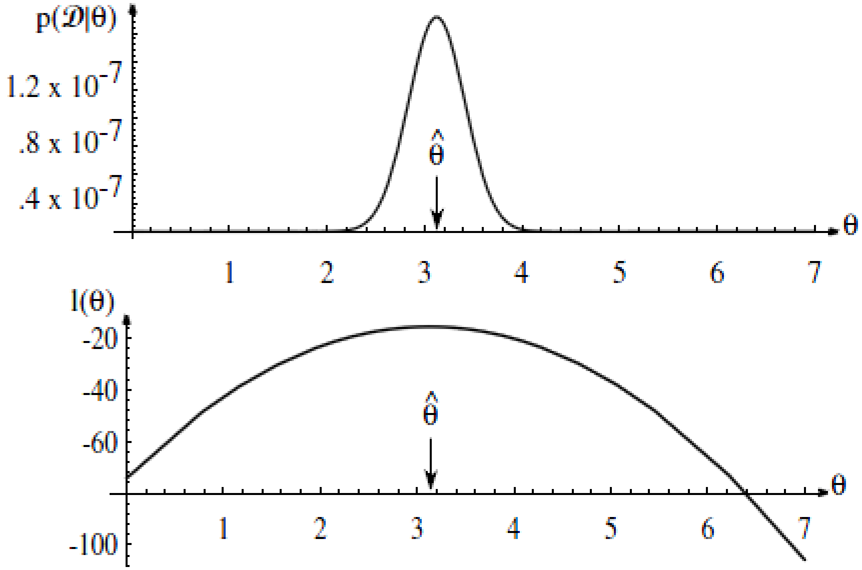


Maximum Likelihood Estimation (MLE)

- $$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right)$$

$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right)$$

- Thus, we solve $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$ to find global optimum



MLE Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^{x}(1-\theta)^{1-x}$$

MLE Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}}$$

MLE Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}}$$

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta) = \sum_{i=1}^{N} \{x^{(i)} \ln \theta + (1 - x^{(i)}) \ln(1-\theta)\}$$

MLE Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}}$$

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta) = \sum_{i=1}^{N} \{x^{(i)} \ln\theta + (1 - x^{(i)}) \ln(1-\theta)\}$$

$$\frac{\partial \ln p(\mathcal{D}|\theta)}{\partial \theta} = 0 \Rightarrow \theta_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N} = \frac{m}{N}$$

MLE Bernoulli: example

- Example: $\mathcal{D} = \{1,1,1\}, \hat{\theta}_{ML} = \frac{3}{3} = 1$
  - Prediction: all future tosses will land heads up

- Overfitting to $\mathcal{D}$

MLE Gaussian: unknown μ

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\ln p(x^{(i)}|\mu) = -\ln\{\sqrt{2\pi}\sigma\} - \frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2$$

$$\frac{\partial \mathcal{L}(\mu)}{\partial \mu} = 0 \Rightarrow \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{N} \ln p\left(x^{(i)}\middle|\mu\right)\right) = 0 \Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2}\left(x^{(i)} - \mu\right)$$

$$= 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

MLE corresponds to many well-known estimation methods.

MLE Gaussian: unknown μ and σ

$$\boldsymbol{\theta} = [\mu, \sigma]$$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma} = 0 \Rightarrow \hat{\sigma}^{2}{}_{ML} = \frac{1}{N} \sum_{i=1}^{N} \left(x^{(i)} - \hat{\mu}_{ML}\right)^{2}$$

Maximum A Posteriori (MAP) estimation

- MAP estimation

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

Maximum A Posteriori (MAP) estimation

- MAP estimation

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

- Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

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

Maximum A Posteriori (MAP) estimation

- MAP estimation

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

- Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

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

- Example of prior distribution:

$$p(\theta) = \mathcal{N}(\theta_0, \sigma^2)$$

MAP estimation Gaussian: unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$ μ is the only unknown parameter

$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$ μ_0 and σ_0 are known

MAP estimation Gaussian: unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$
$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$

μ is the only unknown parameter
μ_0 and σ_0 are known

$$\frac{d}{d\mu} \ln\left(p(\mu) \prod_{i=1}^{N} p(x^{(i)}|\mu)\right) = 0$$

MAP estimation Gaussian: unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$
$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$

μ is the only unknown parameter
μ_0 and σ_0 are known

$$\frac{d}{d\mu}\ln\left(p(\mu)\prod_{i=1}^{N} p(x^{(i)}|\mu)\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) - \frac{1}{\sigma_0^2}(\mu - \mu_0) = 0$$

MAP estimation
Gaussian: unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$ $\quad$ μ is the only unknown parameter

$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$ $\quad$ μ_0 and σ_0 are known

$$\frac{d}{d\mu} \ln\left(p(\mu) \prod_{i=1}^{N} p\left(x^{(i)}|\mu\right)\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) - \frac{1}{\sigma_0^2}(\mu - \mu_0) = 0$$

$$\Rightarrow \hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2} \sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}$$

MAP estimation
Gaussian: unknown μ

$p(x|\mu) \sim N(\mu, \sigma^2)$ μ is the only unknown parameter

$p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2)$ μ_0 and σ_0 are known

$$\frac{d}{d\mu} \ln\left(p(\mu) \prod_{i=1}^{N} p(x^{(i)}|\mu)\right) = 0$$

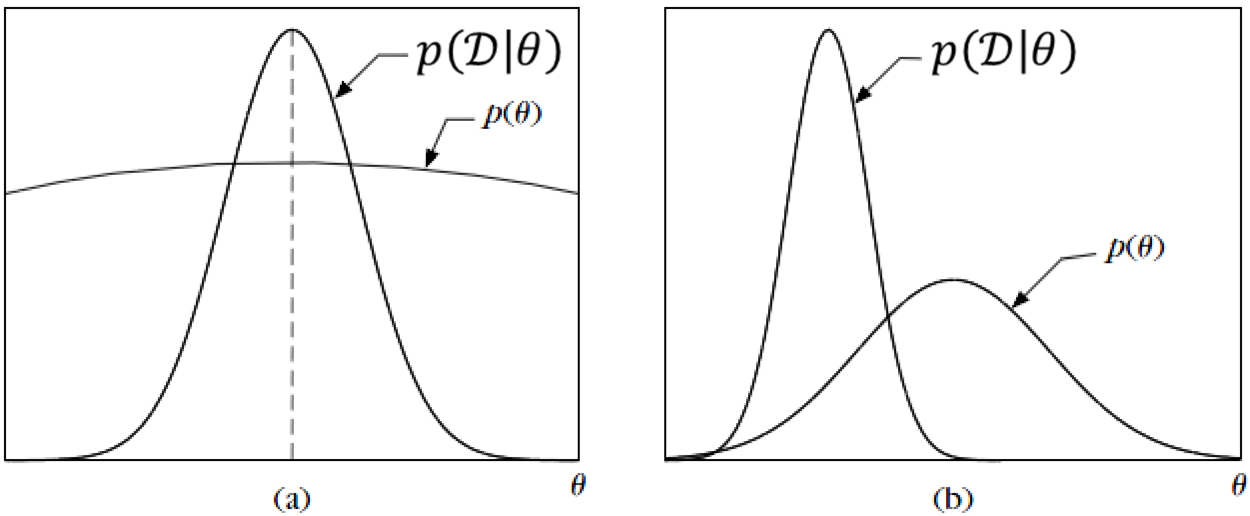
$$\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2}(x^{(i)} - \mu) - \frac{1}{\sigma_0^2}(\mu - \mu_0) = 0$$

$$\Rightarrow \boxed{\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2} \sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}}$$

$\frac{\sigma_0^2}{\sigma^2} \gg 1$ or $N \rightarrow \infty \Rightarrow \hat{\mu}_{MAP} = \hat{\mu}_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}$

Maximum A Posteriori (MAP) estimation

- Given a set of observations $\mathcal{D}$ and a prior distribution $p(\boldsymbol{\theta})$ on parameters, the parameter vector that maximizes $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is found.

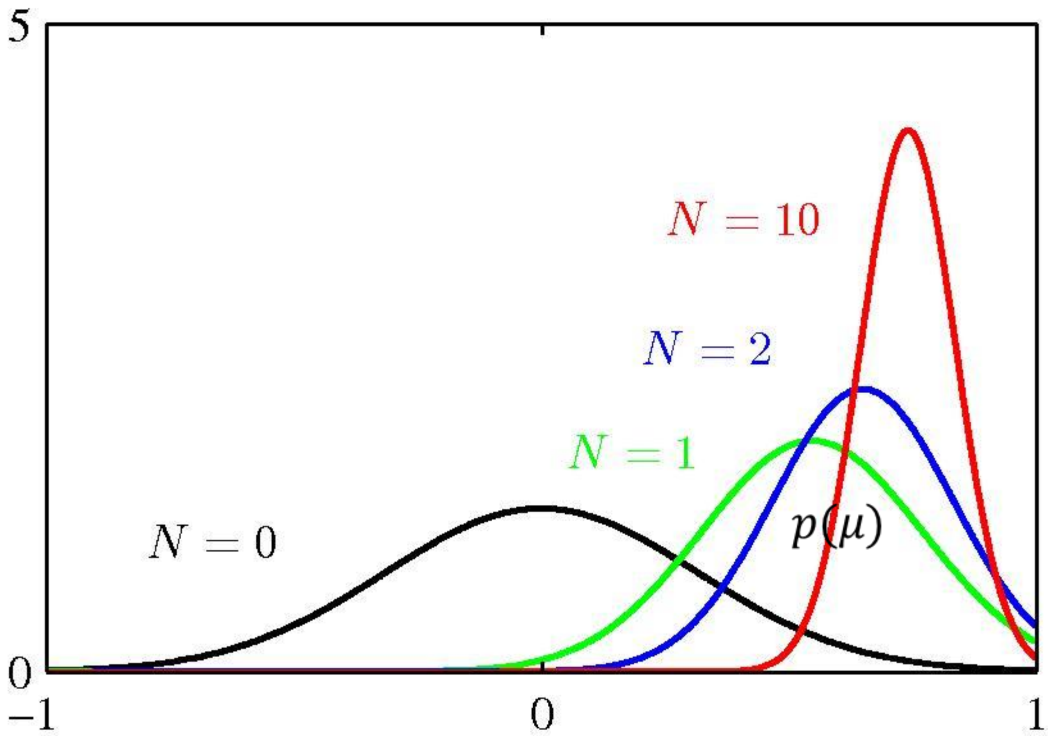


$\hat{\theta}_{MAP} \cong \hat{\theta}_{ML}$

$\hat{\theta}_{MAP} > \hat{\theta}_{ML}$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

MAP estimation
Gaussian: unknown μ (known σ)



[Bishop]

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

$$p(\mu|\mathcal{D}) = N(\mu|\mu_N, \sigma_N)$$

$$\mu_N = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2}\sum_{i=1}^N x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2}N}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

More samples $\Longrightarrow$ sharper $p(\mu|\mathcal{D})$
Higher confidence in estimation

Conjugate Priors

- We consider a form of prior distribution that has a simple interpretation as well as some useful analytical properties

- Choosing a prior such that the posterior distribution that is proportional to $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ will have the same functional form as the prior.

$$\forall \boldsymbol{\alpha}, \mathcal{D} \;\; \exists \boldsymbol{\alpha}' \quad P(\boldsymbol{\theta}|\boldsymbol{\alpha}') \propto P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\alpha})$$

Having the same functional form

Prior for Bernoulli Likelihood

- **Beta distribution** over $\theta \in [0,1]$:

$$\mathrm{Beta}(\theta|\alpha_1, \alpha_0) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

$$\mathrm{Beta}(\theta|\alpha_1, \alpha_0) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

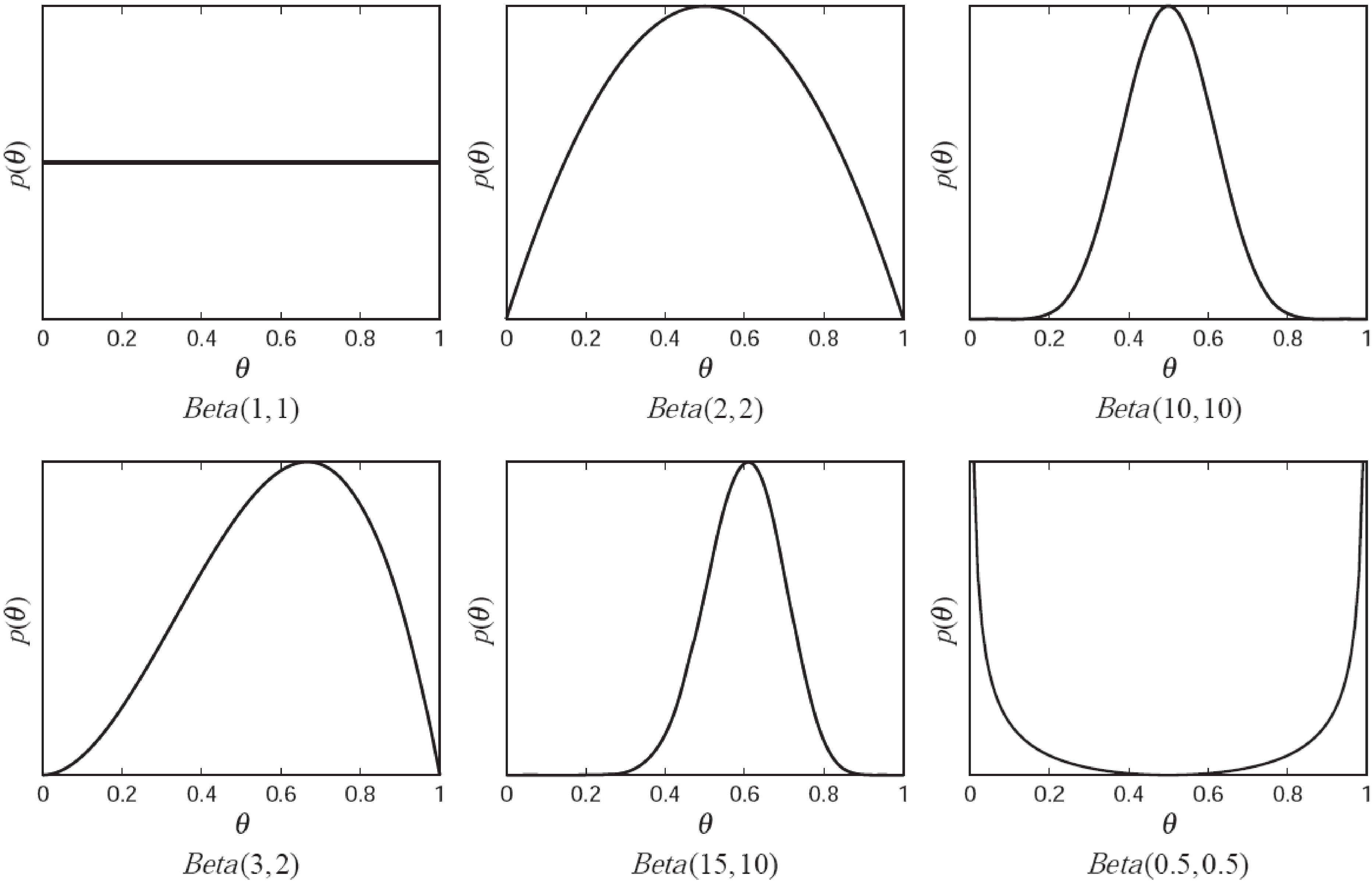
$$E[\theta] = \frac{\alpha_1}{\alpha_0 + \alpha_1}$$

$$\hat{\theta} = \frac{\alpha_1 - 1}{\alpha_0 - 1 + \alpha_1 - 1}$$

most probable θ

- Beta distribution is the conjugate prior of Bernoulli:

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


p(θ)
θ
0
0.2
0.4
0.6
0.8
1
Beta(1,1)
Beta(2,2)
Beta(10,10)
Beta(3,2)
Beta(15,10)
Beta(0.5,0.5)

Benoulli likelihood: posterior

Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$= \left(\prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{(1-x^{(i)})}\right) \underbrace{\mathrm{Beta}(\theta|\alpha_1, \alpha_0)}$$

$$\propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

Benoulli likelihood: posterior

Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$= \left(\prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{(1-x^{(i)})}\right) \underbrace{\mathrm{Beta}(\theta|\alpha_1, \alpha_0)}$$

$$\propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

$$\propto \theta^{m+\alpha_1-1}(1-\theta)^{N-m+\alpha_0-1}$$

$$m = \sum_{i=1}^{N} x^{(i)}$$

Benoulli likelihood: posterior

Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, m heads (1), $N - m$ tails (0)

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$= \left(\prod_{i=1}^{N} \theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) \mathrm{Beta}(\theta|\alpha_1, \alpha_0)$$

$$\propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

$$\propto \theta^{m+\alpha_1-1}(1-\theta)^{N-m+\alpha_0-1}$$

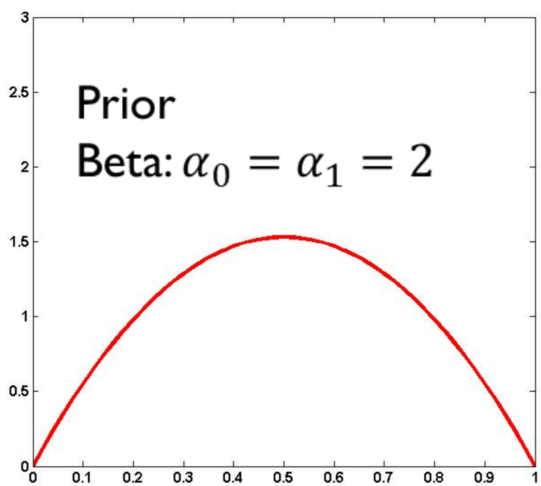
$$m = \sum_{i=1}^{N} x^{(i)}$$

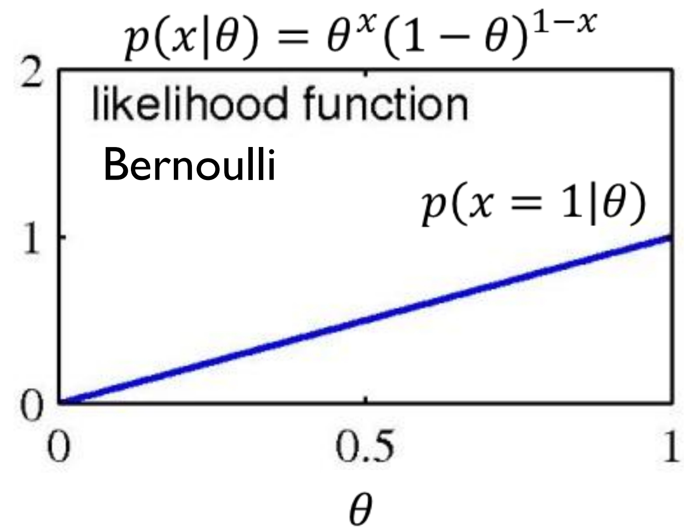
$$\Rightarrow p(\theta|\mathcal{D}) \propto Beta(\theta|\alpha_1', \alpha_0')$$

$$\alpha_1' = \alpha_1 + m$$

$$\alpha_0' = \alpha_0 + N - m$$

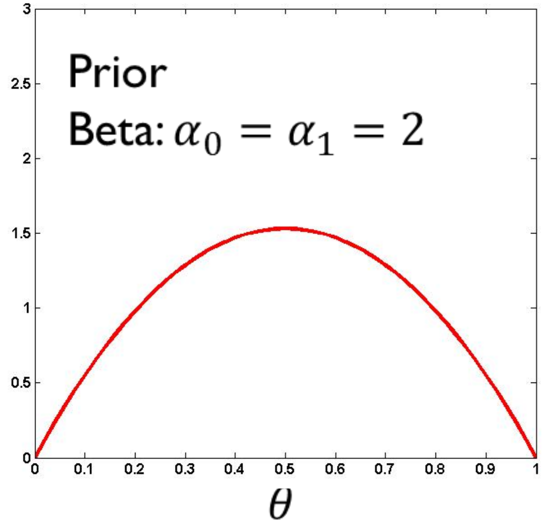
Example

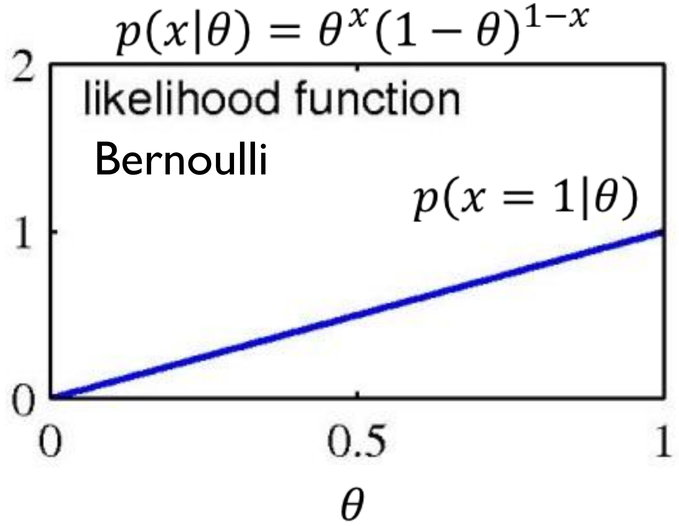


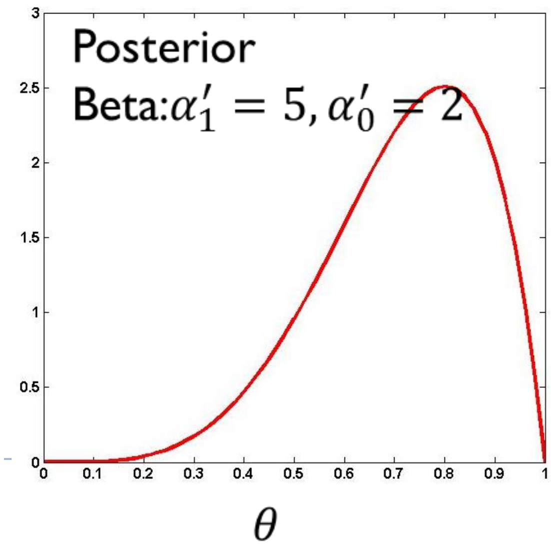


θ

Example







Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$:
m heads (1), $N-m$ tails (0)

$$\alpha_0 = \alpha_1 = 2$$

$$\mathcal{D} = \{1,1,1\} \Rightarrow \; N = 3, m = 3$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta|\mathcal{D}) = \frac{\alpha_1' - 1}{\alpha_1' - 1 + \alpha_0' - 1} = \frac{4}{5}$$

Coin toss example

- MAP estimation can avoid overfitting
  - $\mathcal{D} = \{1,1,1\}, \hat{\theta}_{ML} = 1$
  - $\hat{\theta}_{MAP} = 0.8$ (with prior $p(\theta) = \text{Beta}(\theta|2,2)$)

Summary

- ML and MAP result in a single (point) estimate of the unknown parameters vector.
  - More simple and interpretable than Bayesian estimation

- Both methods asymptotically $(N \to \infty)$ results in the same estimate.

Resources

- C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 2.
- Course CE-717, Dr. M.Soleymani