ML, MAP Estimation and Bayesian

CE-717: Machine Learning
Sharif University of Technology
Fall 2019

Soleymani

Outline

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

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.

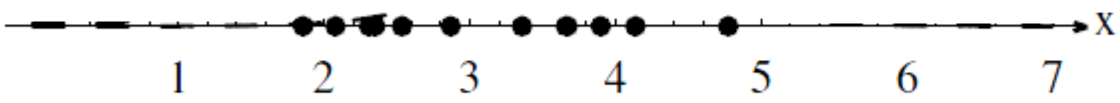
- Methods for parameter estimation
  - Maximum likelihood estimation
  - Maximum A Posteriori (MAP) estimation

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.

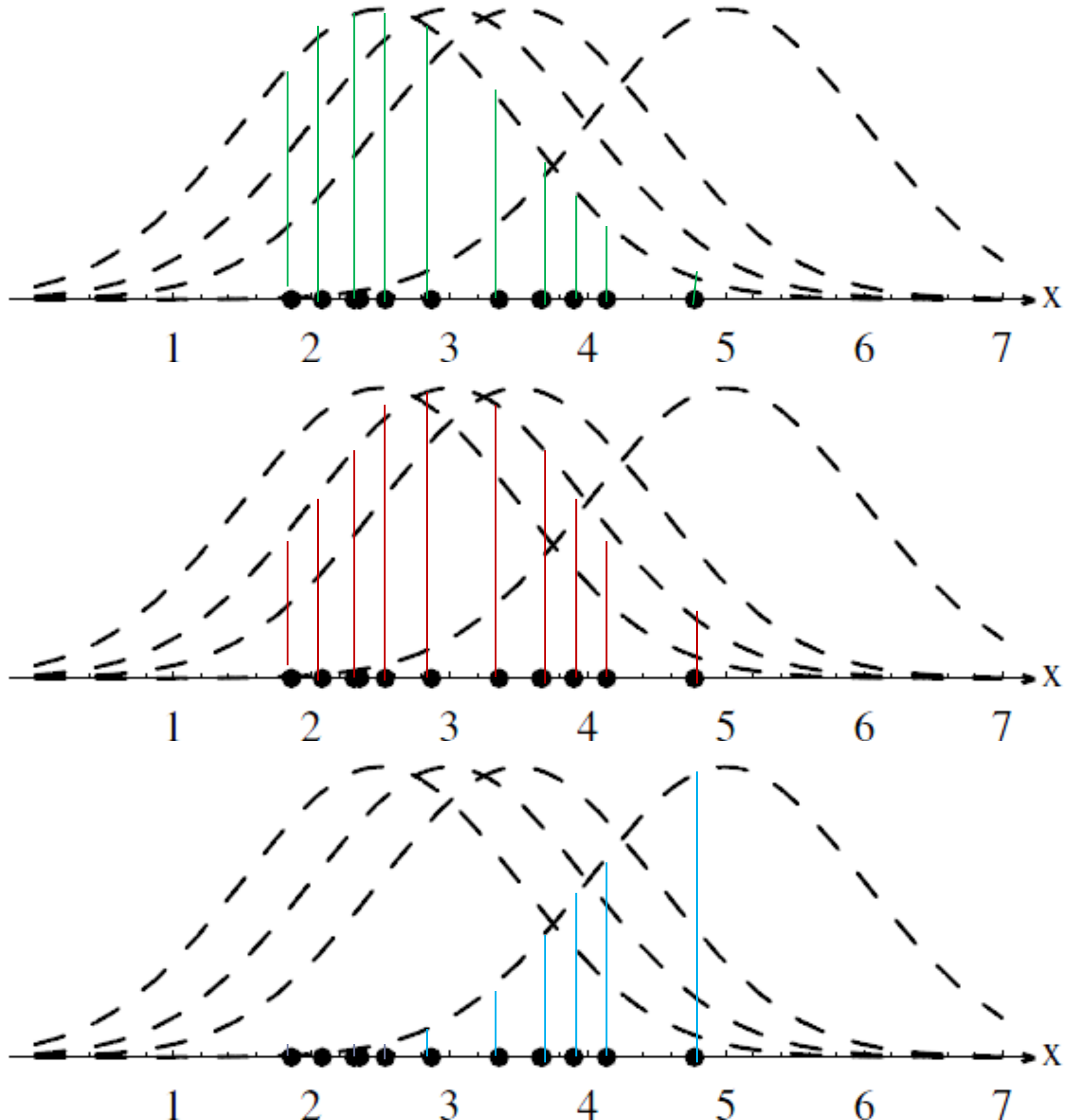
- We need to determine $\boldsymbol{\theta}$ given $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$
  - How to represent $\boldsymbol{\theta}$?
    - $\boldsymbol{\theta}^*$ or $p(\boldsymbol{\theta})$?

Example



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

Example



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:

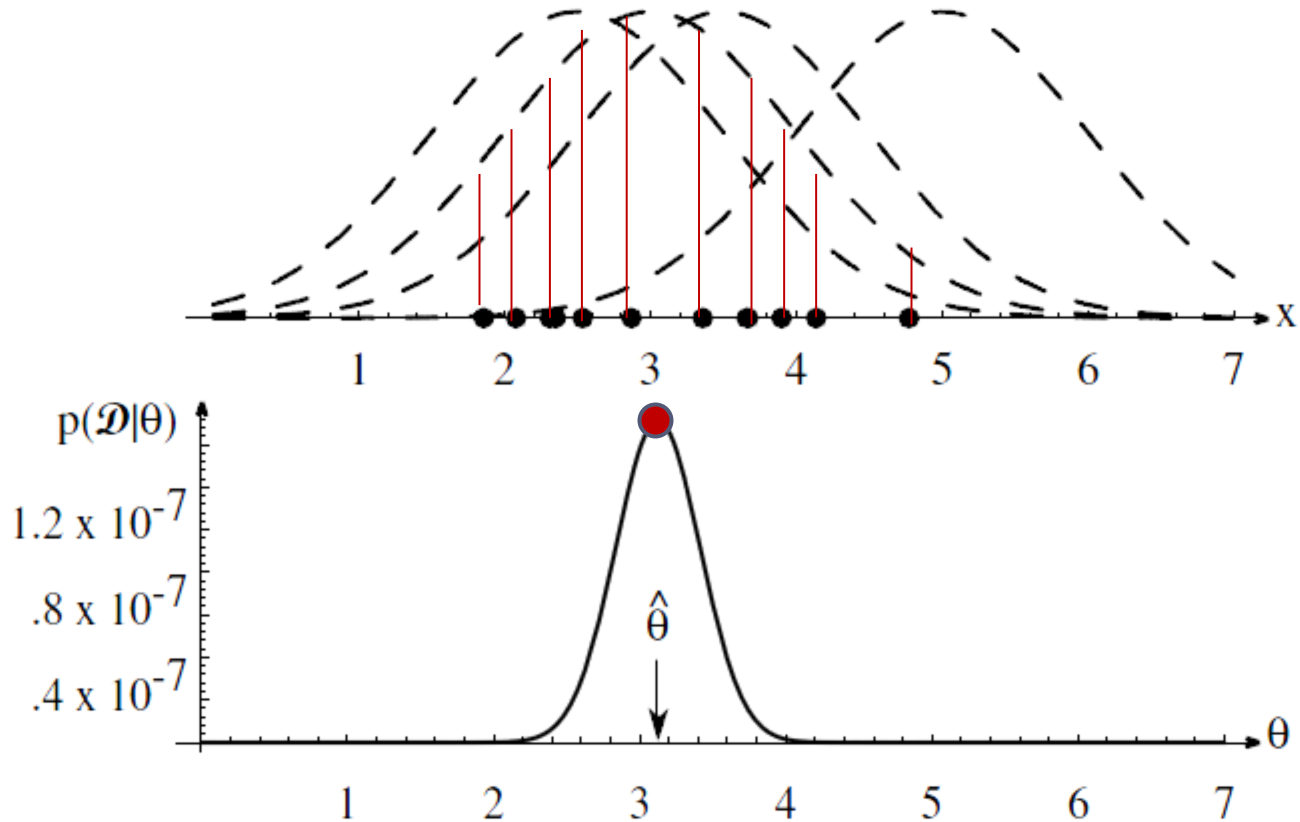
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

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

- Maximum Likelihood estimation

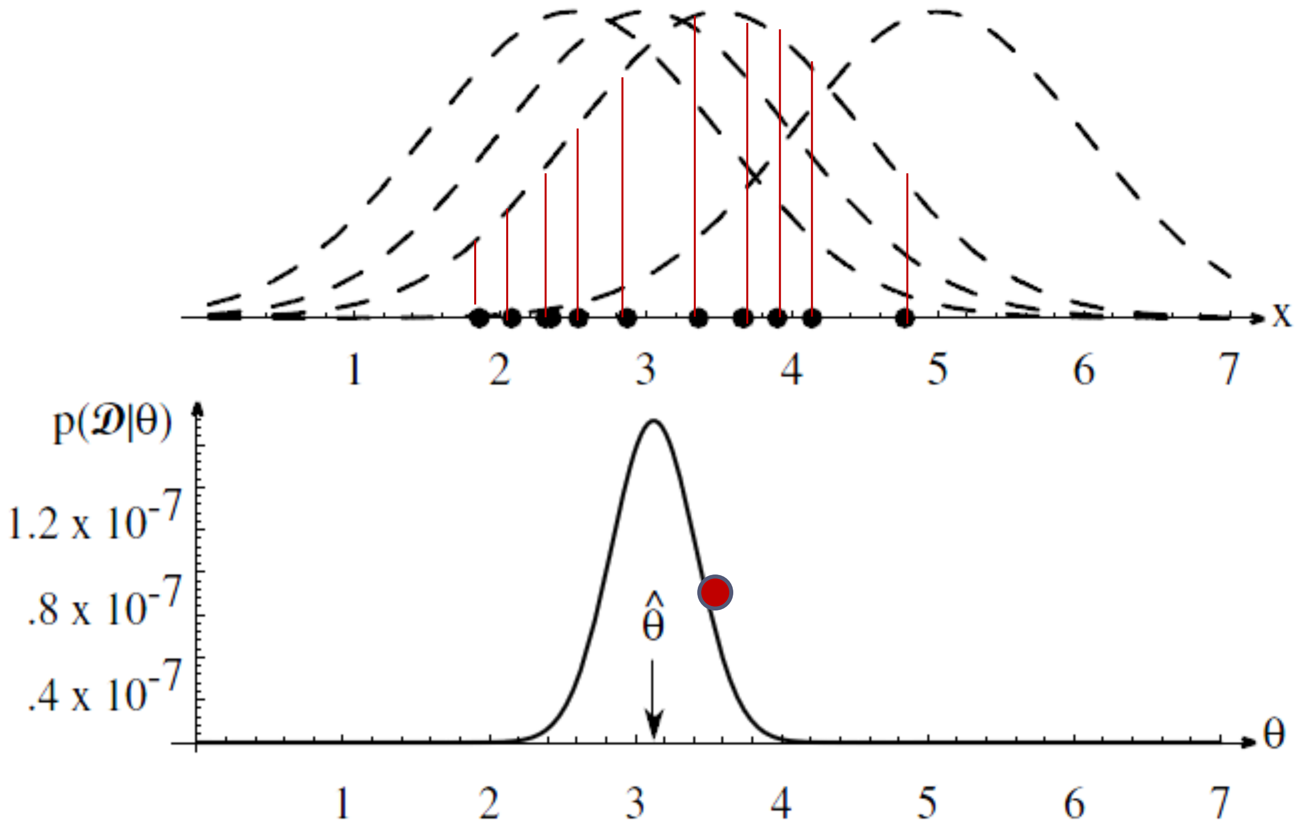
$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)



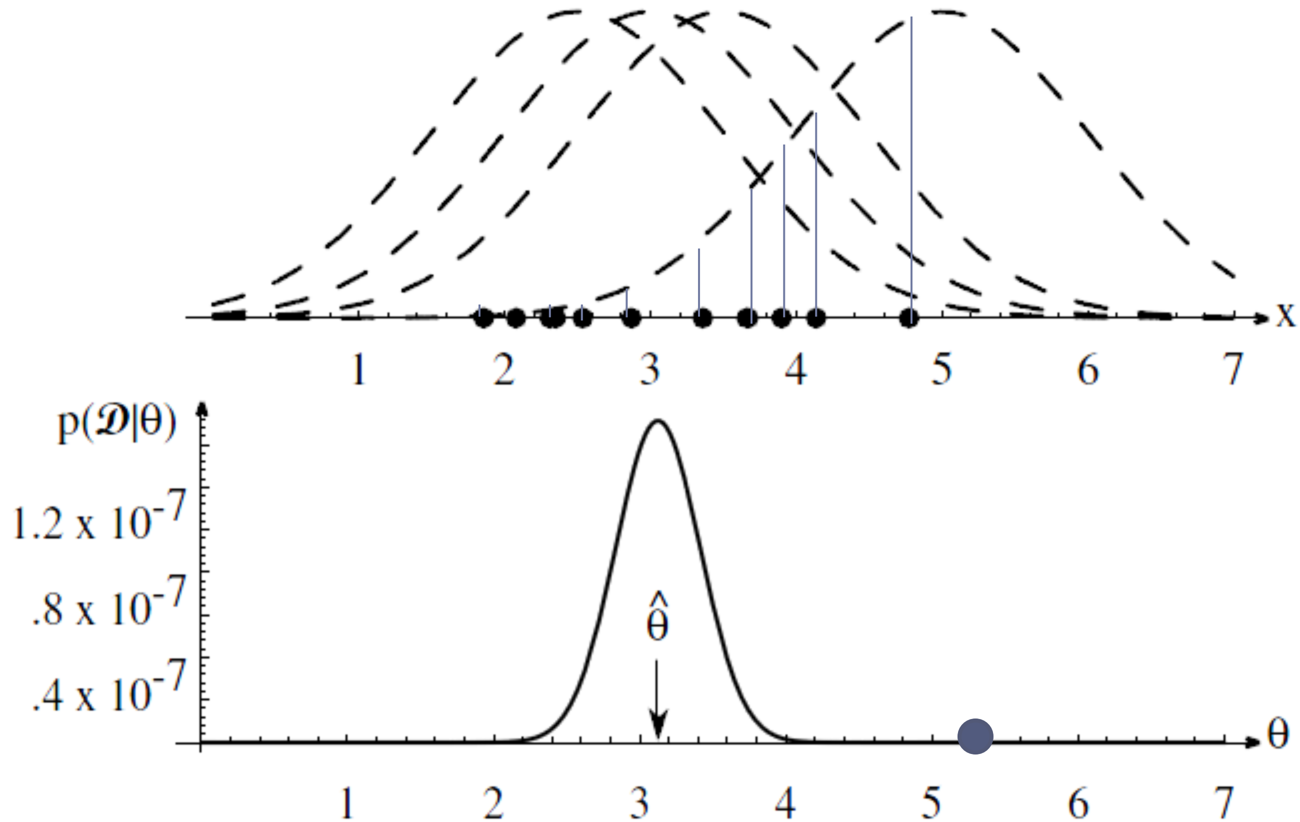
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



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

Maximum Likelihood Estimation (MLE)



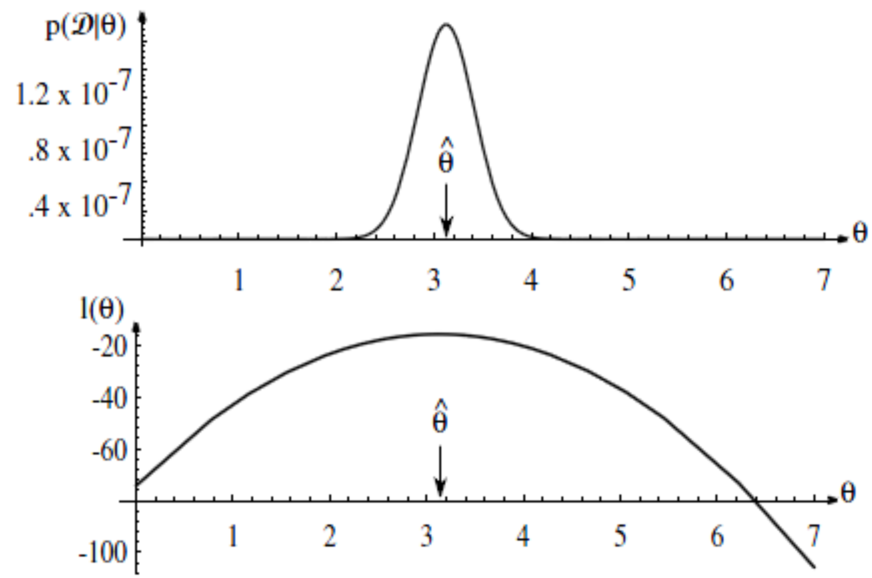
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)

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

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

- 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)

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

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

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

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

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: Multinomial distribution

- Multinomial distribution (on variable with K state):

Parameter space: $\boldsymbol{\theta} = [\theta_1, \dots, \theta_K]$

$\theta_i \in [0,1]$

$$\sum_{k=1}^{K} \theta_k = 1$$

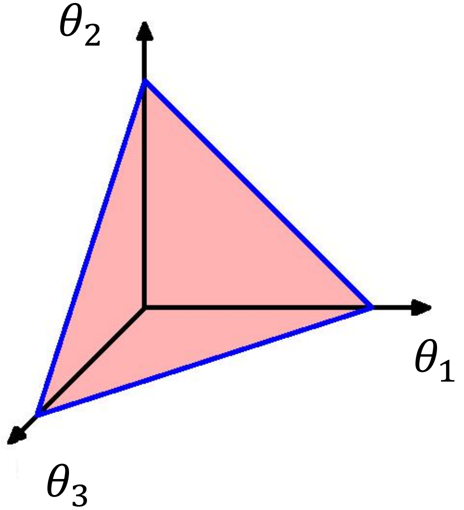
$\boldsymbol{x} = [x_1, \dots, x_K]$

$x_k \in \{0,1\}$

$$\sum_{k=1}^{K} x_k = 1$$

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}$$

$P(x_k = 1) = \theta_k$



MLE: Multinomial distribution

$$\mathcal{D} = \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\right\}$$

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}} = \prod_{k=1}^{K} \theta_k^{\boxed{\sum_{i=1}^{N} x_k^{(i)}}}$$

$$N_k = \sum_{i=1}^{N} x_k^{(i)}$$

$$\sum_{k=1}^{K} N_k = N$$

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \ln p(\mathcal{D}|\boldsymbol{\theta}) + \lambda(1 - \sum_{k=1}^{K} \theta_k)$$

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} x_k^{(i)}}{N} = \frac{N_k}{N}$$

MLE
Gaussian: unknown μ

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

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

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

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

MLE corresponds to many well-known estimation methods.

MLE
Gaussian: unknown μ and σ

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

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

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

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

Maximum A Posteriori (MAP) estimation

- MAP estimation

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

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

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

- 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

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

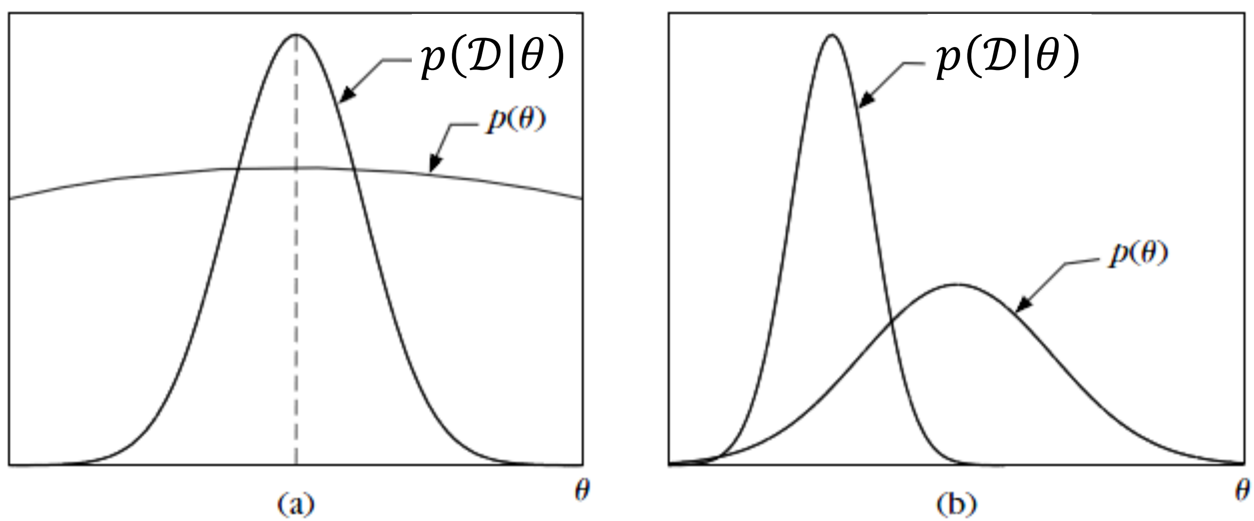
$$\Rightarrow \sum_{i=1}^{N}\frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) - \frac{1}{\sigma_0^2}(\mu - \mu_0) = 0$$

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

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

Maximum A Posteriori (MAP) estimation

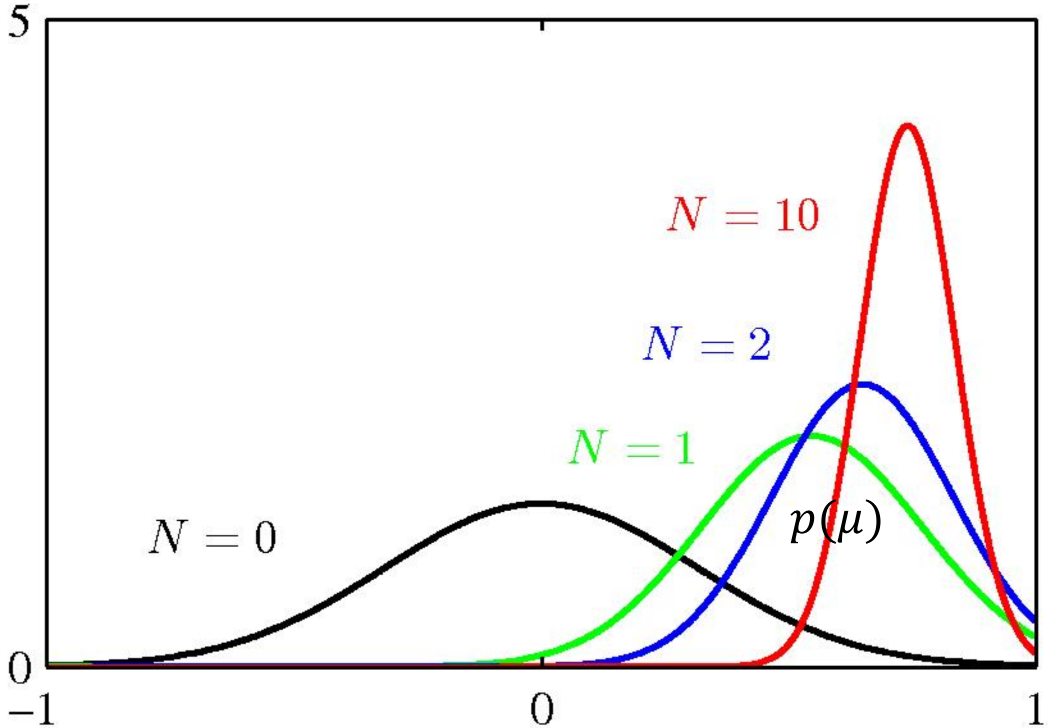
- Given a set of observations $\mathcal{D}$ and a prior distribution $p(\boldsymbol{\theta})$ on parameters, the parameter vector that maximizes $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is found.



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

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

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 <u>the same functional form</u> 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

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

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

most probable θ

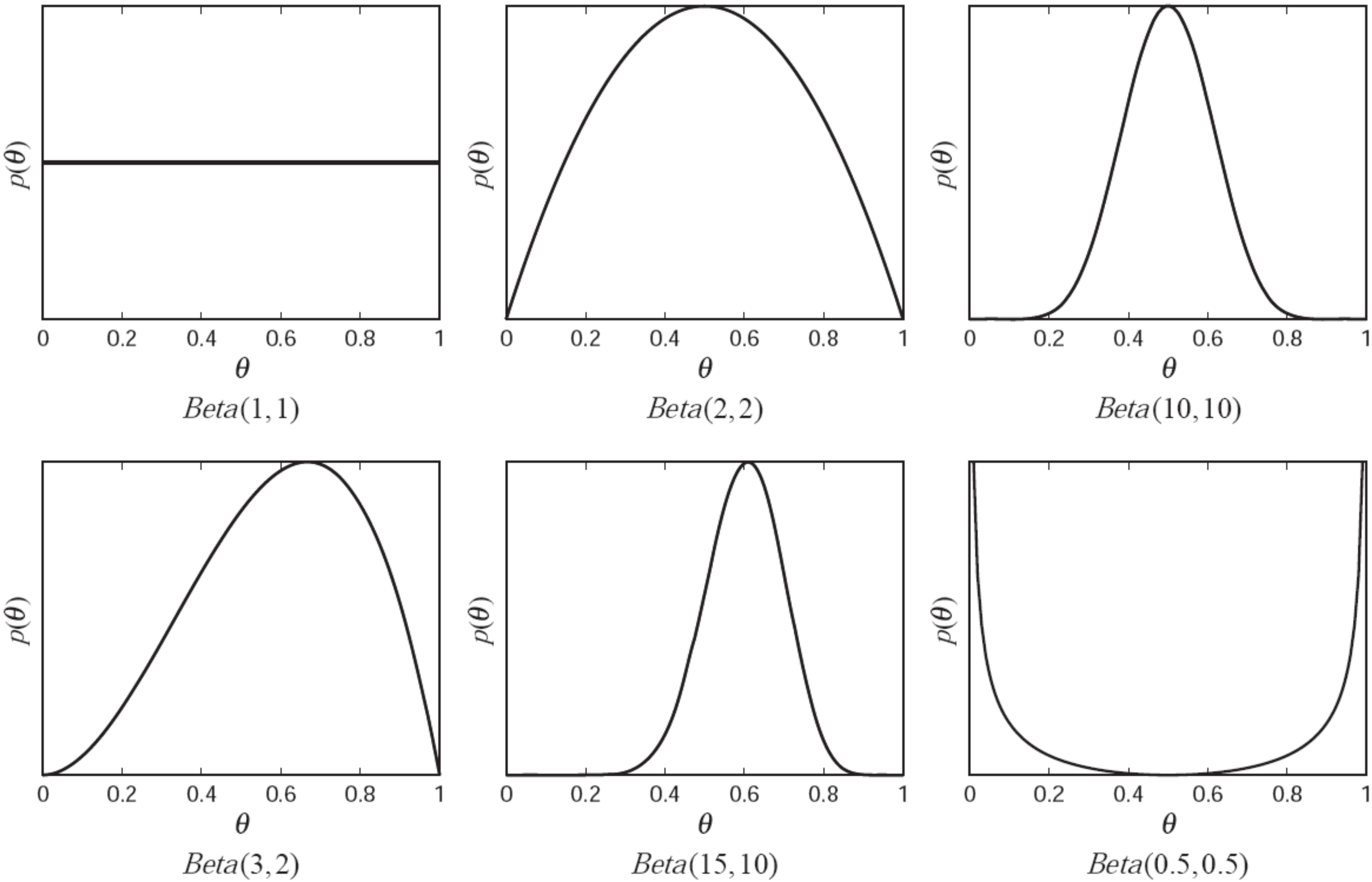
- **Beta distribution** over $\theta \in [0,1]$:

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

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

- Beta distribution is the conjugate prior of Bernoulli:

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



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

Benoulli likelihood: posterior

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

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

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

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

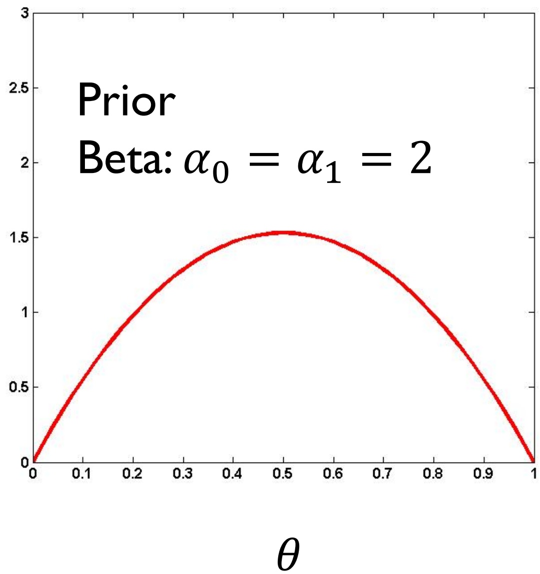
$$m = \sum_{i=1}^{N} x^{(i)}$$

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

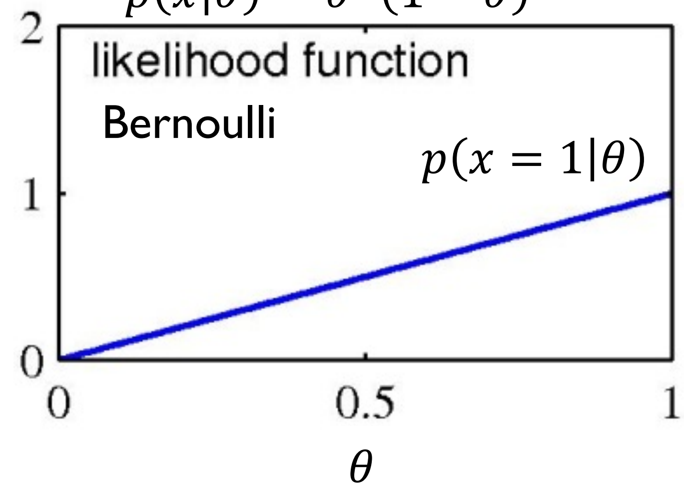
$$\alpha_1' = \alpha_1 + m$$

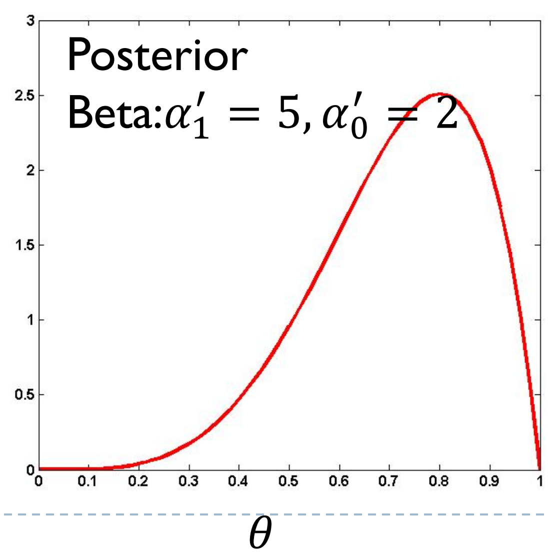
$$\alpha_0' = \alpha_0 + N - m$$

Example



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





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

$$\alpha_0 = \alpha_1 = 2$$

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

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

Toss example

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

Bayesian inference

- Parameters $\boldsymbol{\theta}$ as random variables with a priori distribution
  - Bayesian estimation utilizes the available prior information about the unknown parameter
  - As opposed to ML and MAP estimation, it does not seek a specific point estimate of the unknown parameter vector $\boldsymbol{\theta}$

- The observed samples $\mathcal{D}$ convert the prior densities $p(\boldsymbol{\theta})$ into a posterior density $p(\boldsymbol{\theta}|\mathcal{D})$
  - Keep track of beliefs about $\boldsymbol{\theta}$'s values and uses these beliefs for reaching conclusions
  - In the Bayesian approach, we first specify $p(\boldsymbol{\theta}|\mathcal{D})$ and then we compute the predictive distribution $p(\boldsymbol{x}|\mathcal{D})$

Bayesian estimation: predictive distribution

- Given a set of samples $\mathcal{D} = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$, a prior distribution on the parameters $P(\boldsymbol{\theta})$, and the form of the distribution $P(\boldsymbol{x}|\boldsymbol{\theta})$

- We find $P(\boldsymbol{\theta}|\mathcal{D})$ and then use it to specify $\hat{P}(\boldsymbol{x}) = P(\boldsymbol{x}|\mathcal{D})$ as an estimate of $P(\boldsymbol{x})$:

$$P(\boldsymbol{x}|\mathcal{D}) = \int P(\boldsymbol{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} = \int P(\boldsymbol{x}|\mathcal{D}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} = \int P(\boldsymbol{x}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

Predictive distribution

If we know the value of the parameters $\boldsymbol{\theta}$, we know exactly the distribution of $\boldsymbol{x}$

- Analytical solutions exist for very special forms of the involved functions

Benoulli likelihood: prediction

- Training samples: $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$

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

$$P(\theta|\mathcal{D}) = Beta(\theta|\alpha_1 + m, \alpha_0 + N - m)$$
$$\propto \theta^{\alpha_1 + m - 1}(1-\theta)^{\alpha_0 + (N-m) - 1}$$

$$P(x|\mathcal{D}) = \int P(x|\theta)\, P(\theta|\mathcal{D}) d\theta$$
$$= E_{P(\theta|\mathcal{D})}[P(x|\theta)]$$
$$\Rightarrow P(x=1|\mathcal{D}) = E_{P(\theta|\mathcal{D})}[\theta] = \frac{\alpha_1 + m}{\alpha_0 + \alpha_1 + N}$$

ML, MAP, and Bayesian Estimation

- If $p(\boldsymbol{\theta}|\mathcal{D})$ has a sharp peak at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ (i.e., $p(\boldsymbol{\theta}|\mathcal{D}) \approx \delta(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$), then $p(\boldsymbol{x}|\mathcal{D}) \approx p\left(\boldsymbol{x}|\widehat{\boldsymbol{\theta}}\right)$
  - In this case, the Bayesian estimation will be approximately equal to the MAP estimation.
  - If $p(\mathcal{D}|\boldsymbol{\theta})$ is concentrated around a sharp peak and $p(\boldsymbol{\theta})$ is broad enough around this peak, the ML, MAP, and Bayesian estimations yield approximately the same result.

- All three methods asymptotically $(N \to \infty)$ results in the same estimate

Summary

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

- Bayesian approach finds a predictive distribution using all the available information:
  - expected to give better results
  - needs higher computational complexity

- Bayesian methods have gained a lot of popularity over the recent decade due to the advances in computer technology.

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

Resource

- C. Bishop, "Pattern Recognition and Machine Learning", Chapter 2.