Density Estimation: ML, MAP, Bayesian estimation

CE-725: Statistical Pattern Recognition
Sharif University of Technology
Spring 2013

Soleymani

Outline

- Introduction
- Maximum-Likelihood Estimation
- Maximum A Posteriori Estimation
- Bayesian Estimation

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 for density function
    - The form of the density function is determined entirely by the data

Class-Conditional Densities

- We usually do not know the class-conditional densities $p(\boldsymbol{x}|\omega_i)$.
  - However, we might have prior knowledge about:
    - Functional forms of these densities
    - Ranges for the values of their unknown parameters

- We can separate training data of different classes and use the set $\mathcal{D}_i$ containing training samples of class ω_i to estimate $p(\boldsymbol{x}|\omega_i)$
  - $\hat{p}(\boldsymbol{x}|\omega_i) = p(\boldsymbol{x}|\omega_i, \mathcal{D}_i)$
  - Estimating $p(\boldsymbol{x}|\omega_i)$ from $\mathcal{D}_i$ can be considered as an unsupervised density estimation problem

Parametric Density Estimation

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

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

Maximum Likelihood Estimation (MLE)

- 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 (statistically independent, identically distributed samples)

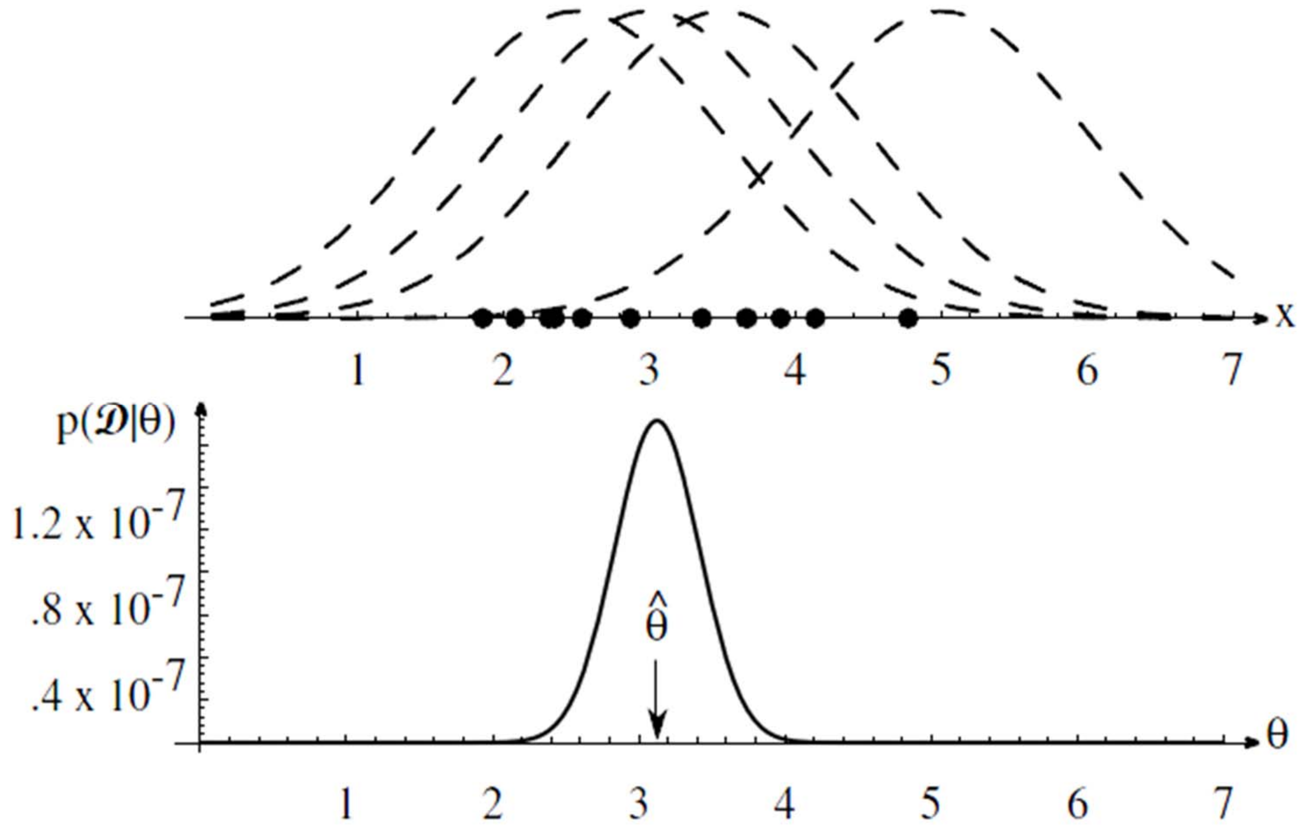
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

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)



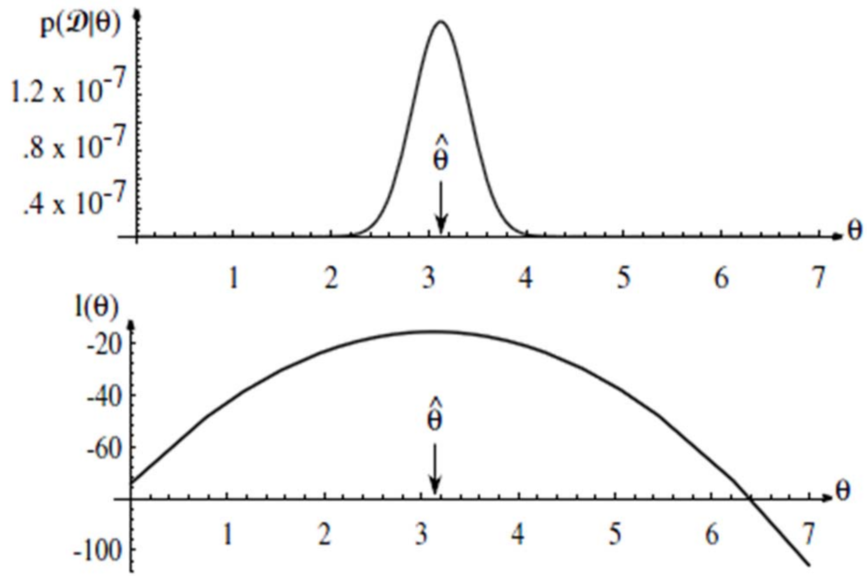
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)

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

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

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



MLE
Gaussian: Unknown $\boldsymbol{\mu}$

$$\ln p(\boldsymbol{x}^{(i)}|\boldsymbol{\mu})$$
$$= -\frac{1}{2}\ln\left\{(2\pi)^{d/2}|\boldsymbol{\Sigma}|\right\} - \frac{1}{2}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right)$$

$$\nabla_{\boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}) = \mathbf{0} \Rightarrow \nabla_{\boldsymbol{\mu}}\left(\sum_{i=1}^{N}\ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\mu}\right)\right) = \mathbf{0}$$
$$\Rightarrow \sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}\right) = \mathbf{0} \Rightarrow \widehat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}^{(i)}$$

MLE
Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}(\overbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}^{\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]}) = \mathbf{0} \Rightarrow \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu} \right) = \mathbf{0}$$

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

$$\nabla_{\boldsymbol{\Sigma}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{0}$$

$$\Rightarrow \widehat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_{ML} \right) \left(\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_{ML} \right)^{T}$$

Maximum A Posteriori (MAP) Estimation

- MAP estimation

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

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

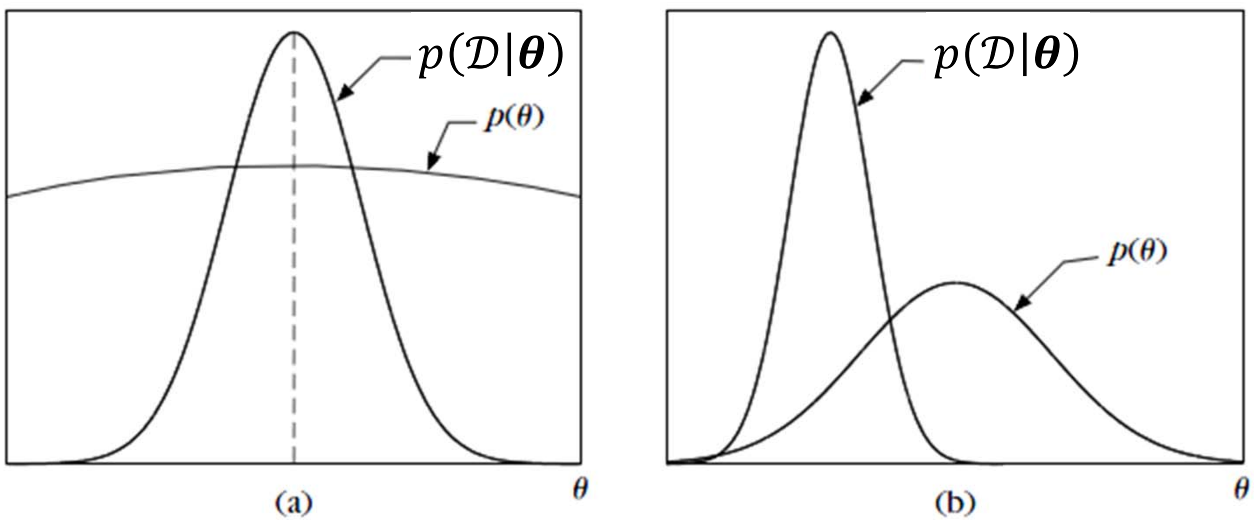
$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- Example of prior distribution:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \alpha^2 \boldsymbol{I})$$

Maximum A Posteriori (MAP) Estimation

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



$\widehat{\boldsymbol{\theta}}_{MAP} \cong \widehat{\boldsymbol{\theta}}_{ML}$ $\qquad$ $\widehat{\boldsymbol{\theta}}_{MAP} > \widehat{\boldsymbol{\theta}}_{ML}$

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

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

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

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

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

Bayesian Estimation

- **Given**: samples $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$, a priori information about pdf $p(\boldsymbol{\theta})$, the form of the density $p(\boldsymbol{x}|\boldsymbol{\theta})$

- **Goal**: compute the conditional pdf $p(\boldsymbol{x}|\mathcal{D})$ as an estimate of $p(\boldsymbol{x})$.

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

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

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

Bayesian Estimation

- Parameters are considered as a vector $\boldsymbol{\theta}$ of 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 $p(\boldsymbol{\theta}|\mathcal{D})$
  - To find the conditional pdf $p(\boldsymbol{x}|\mathcal{D})$, we must first specify $p(\boldsymbol{\theta}|\mathcal{D})$
  - Then, $p(\boldsymbol{x}|\mathcal{D})$ is found as $p(\boldsymbol{x}|\mathcal{D}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$

Bayesian Estimation
Gaussian: Unknown μ (known σ)

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

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu} \propto \prod_{i=1}^{N} p(x^{(i)}|\mu)p(\mu)$$

$$p(\mu|\mathcal{D}) = \alpha \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2\right\} \times \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

$$= \alpha' \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{N}\left(\frac{x^{(i)}-\mu}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\}$$

$$= \alpha'' \exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\sum_{i=1}^{N} x^{(i)}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

Bayesian Estimation
Gaussian: Unknown μ (known σ)

$$\Rightarrow p(\mu|\mathcal{D}) \sim N(\mu_N, \sigma_N^2) \longrightarrow$$ $p(\mu)$: conjugate prior

$$\mu_N = \left(\frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\right)\bar{x}_N + \left(\frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2}\right)\mu_0$$

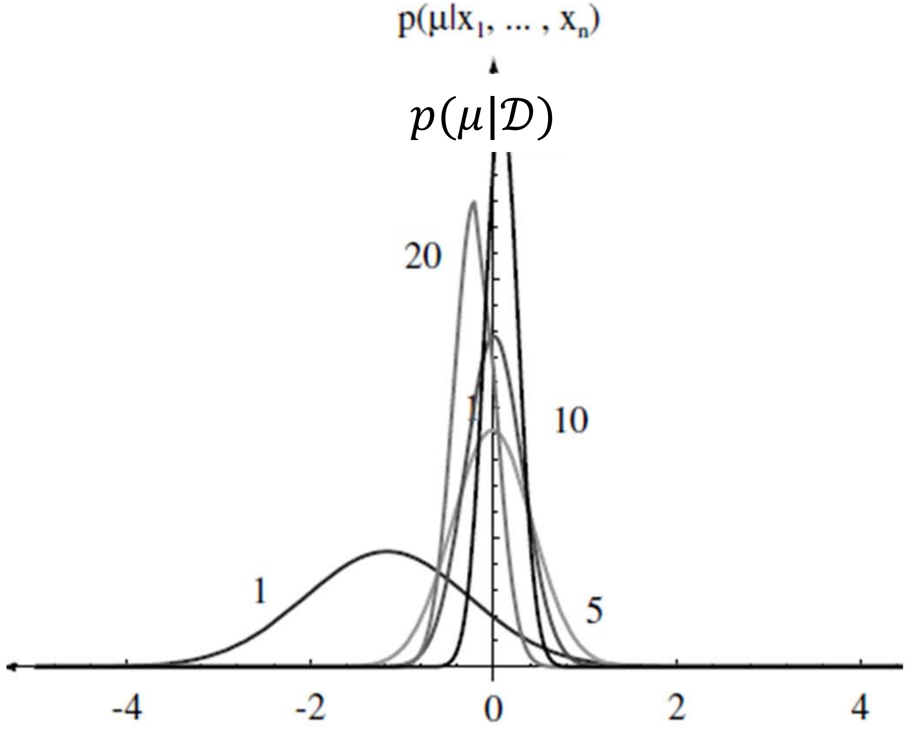
$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2}$$

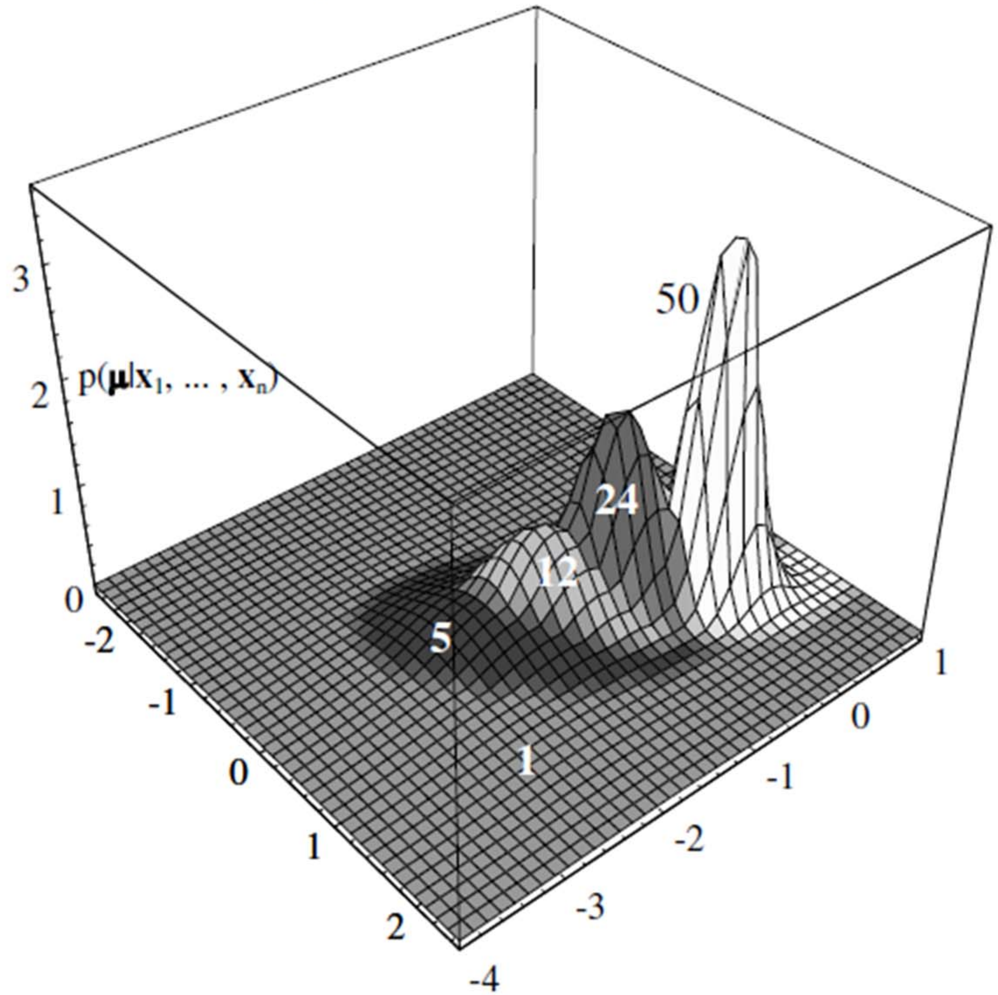
$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_N}\exp\left\{-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2 + \sigma_N^2}\right\} f(\sigma, \sigma_N)$$

$$\Rightarrow p(x|\mathcal{D}) \sim N(\mu_N, \sigma^2 + \sigma_N^2)$$

Bayesian Estimation
Gaussian: Unknown μ (known σ)





More samples $\Longrightarrow$ sharper $p(\mu|\mathcal{D})$

Some Related Definitions

- 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.

- Bayesian learning
  - When densities converges to a Dirac delta function centered about the true parameter value

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

Bayesian Estimation: Example

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise,} \end{cases}$$

$$0 < \theta \le 10 \quad \Rightarrow \quad p(\theta|\mathcal{D}^0) = p(\theta) = U(0,10)$$

$$\mathcal{D} = \{4, 7, 2, 8\} \qquad \mathcal{D}^i = \{x^{(1)}, \dots, x^{(i)}\}$$

$x = 4$

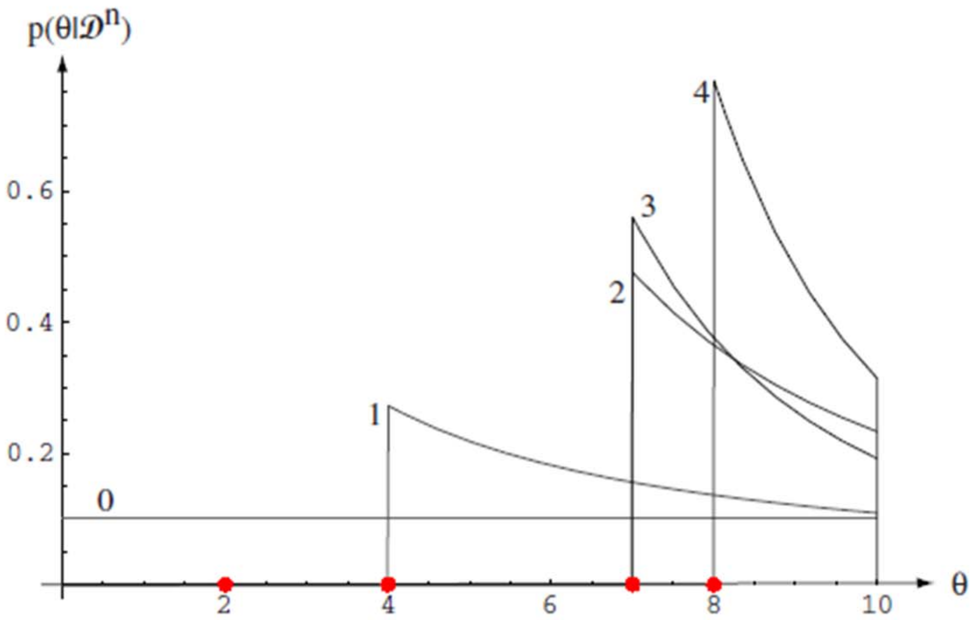
$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & \text{for } 4 \le \theta \le 10 \\ 0 & \text{otherwise,} \end{cases}$$

$x = 7$

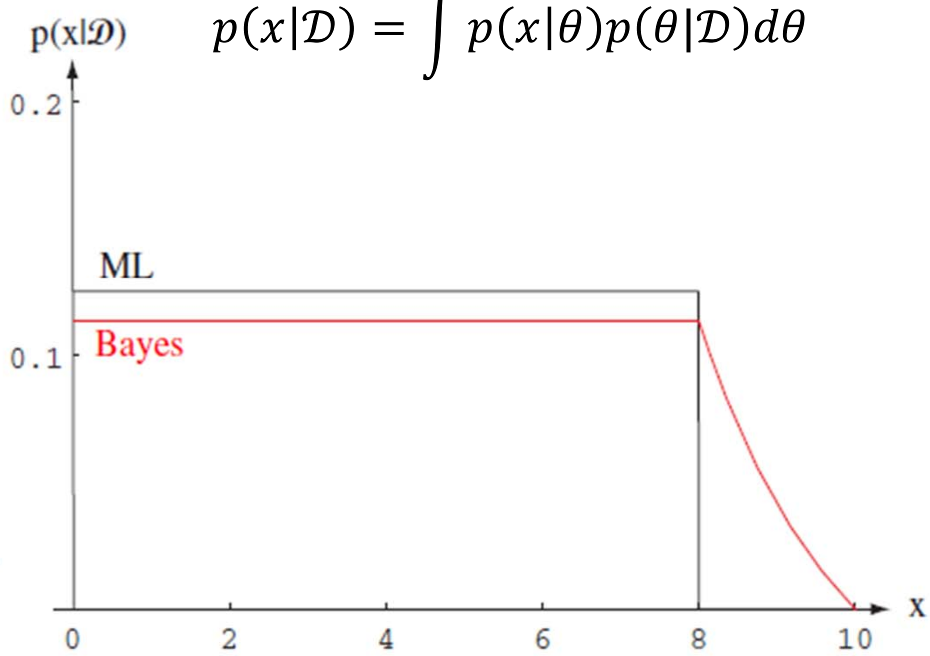
$$p(\theta|\mathcal{D}^2) \propto p(x|\theta)p(\theta|\mathcal{D}^1) = \begin{cases} 1/\theta^2 & \text{for } 7 \le \theta \le 10 \\ 0 & \text{otherwise,} \end{cases}$$

Bayesian Estimation: Example

$$p(\theta|\mathcal{D}^n) \propto 1/\theta^n \text{ for } \max_x[\mathcal{D}^n] \le \theta \le 10$$



$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$



Summary

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