EE5907R: Pattern Recognition

Lecture 3: Parameter Estimation for Supervised Learning

National University of Singapore

Dr. Sun Ying

Introduction

- When conditional densities $p(x|\omega_j)$ and a priori probabilities $P(\omega_j)$ are known
  - Bayesian Decision Theory
  - For Gaussian data
    - ✓ Quadratic classifiers (general case)
    - ✓ Linear classifiers (equal covariance)
- In most situations, the true distributions are not available. ☹
  - Estimate $p(x|\omega_j)$ and $P(\omega_j)$ from the training data
  - Two common approaches
    - ✓ Parameter Estimation (this lecture)
    - ✓ Non-parametric Density Estimation (next lecture)

Outline

- Supervised Learning
- Parameter Estimation Problem
- Maximum Likelihood Estimation (MLE)
- Bayesian Parameter Estimation (BPE)
- Numerical Examples
- Problems of Dimensionality
- Summary

Parameter Estimation: Example

- A fish expert says:
  - the length of salmon follows Gaussian distribution $N(\mu_1, \sigma_1^2)$ and the length of sea bass ~ $N(\mu_2, \sigma_2^2)$.
- Labeled training data

- Need to:
  - estimate prior probabilities
  - estimate parameters μ_1, σ_1^2, μ_2, σ_2^2 using maximum-likelihood or Bayesian parameter estimation methods

Then we can apply Bayesian decision theory!

Supervised Learning

- To design classifiers with data samples and associated labels or class categories
  - Given training samples D and associated class category j, (j =1, 2,…,c)

    - ✓ The class labels - ground truth, produced by domain experts, i.e., doctors and other specialists, might be expensive to get in many applications.
  - Learn the class conditional probabilities $p(x|\omega_j)$ and prior probabilities $P(\omega_j)$
    - ✓ Prior probabilities $P(\omega_j)$ are easy to estimate if we have enough samples:

$$\mathrm{P}(\omega_j) = |D_j| \Big/ \sum_{k=1}^{C} |D_k|$$

A Parameter Estimation Problem

- Design a classifier from training samples
  - Apply Bayes rule if we knew:
    - ✓ Priors: $P(\omega_i)$
    - ✓ Class-conditional densities: $p(x \mid \omega_i)$
  - Estimate unknown probabilities and probability densities from training samples
    - ✓ No problem with prior estimation
    - ✓ Samples are often too small for estimation of class-conditional densities
  - Parameter estimation
    - ✓ Assume a particular form for the density (e.g., Gaussian) so that only the parameters (e.g., mean and variance) need to be determined

Probability Density Estimation

- Parameter Estimation:
  - –Assume known distribution models with unknown parameters, e.g., Gaussian
  - –Estimate these unknown parameters (e.g., mean and variance) from training samples
  - –From estimating an unknown *function* to estimating a limited number of unknown *parameters*

$$\mathbf{x} = [l, \quad w, \quad c, \ldots]^T$$

$\mathbf{x} \sim N(\mu_1, \Sigma_1)$

$\mathbf{x} \sim N(\mu_2, \Sigma_2)$

- Two Common Methods:
  - –Maximum-Likelihood Estimation
  - –Bayesian Parameter Estimation

Maximum-likelihood vs. Bayesian Parameter Estimation

- Maximum likelihood Estimation
  - Parameters viewed as *fixed but unknown* quantities!
  - Estimate the value of the unknown quantities; the best estimate maximizes the probability of obtaining the samples observed.
  - Easier to understand and interpret
  - Requires high confidence on assumed distribution models $p(\mathbf{x}|\omega_j)$
- Bayesian Parameter Estimation
  - Parameters viewed as *random variables* with known prior distribution
  - Estimate the distribution of the values of the random variables
  - Represented as a weighted average of models (parameters), hard to understand
  - Requires no assumption on distribution model for $p(\mathbf{x}|\omega_j, D)$
- In either approach, use $P(\omega_i|x)$ for classification

Maximum-Likelihood Estimation

- Assumptions:
  - –We have c classes with samples in class j having been drawn independently according to $p(\mathbf{x} \mid \omega_j)$
    - ✓ For instance, $p(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$
  - –$p(x \mid \omega_j)$ has a known parametric form, determined by θ_j
    - ✓ $p(x \mid \omega_j) \equiv p(x \mid \omega_j, \theta_j)$
  - –We want to estimate θ_j for each category $j = 1, 2, \ldots, c$
  - –Parameters for different classes are functionally independent
    - ✓ Samples in D_i give no information about θ_j if $i \neq j$
- Solve c separate problems (drop class distinction)
  - –Parameter estimation is an identical procedure for all classes

$\mathbf{x} \sim N(\mu_1, \Sigma_1)$

$\mathbf{x} \sim N(\mu_2, \Sigma_2)$

Maximum-Likelihood Estimation

- Given a set $D = (x_1, x_2, \ldots, x_n)$ where the n samples are drawn *independently* from *identical* distribution $p(\mathbf{x} \mid \theta)$, estimate parameter θ
- ML estimate of θ maximizes $p(D|\boldsymbol{\theta})$

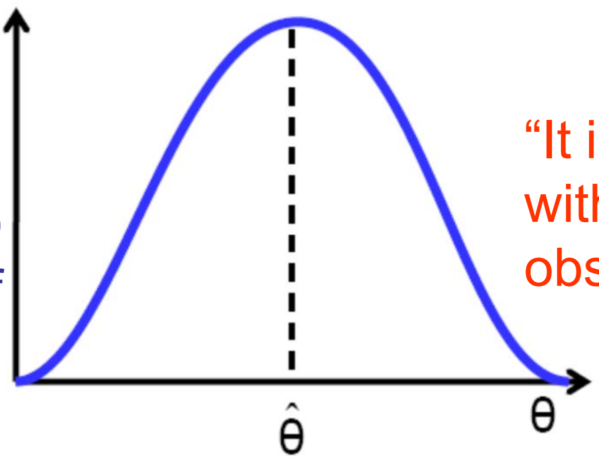
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}), \qquad p(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

arg max: argument of the maximum

D is an i.i.d. set

$p(D|\boldsymbol{\theta})$

Likelihood of θ w.r.t. the set of samples *D*

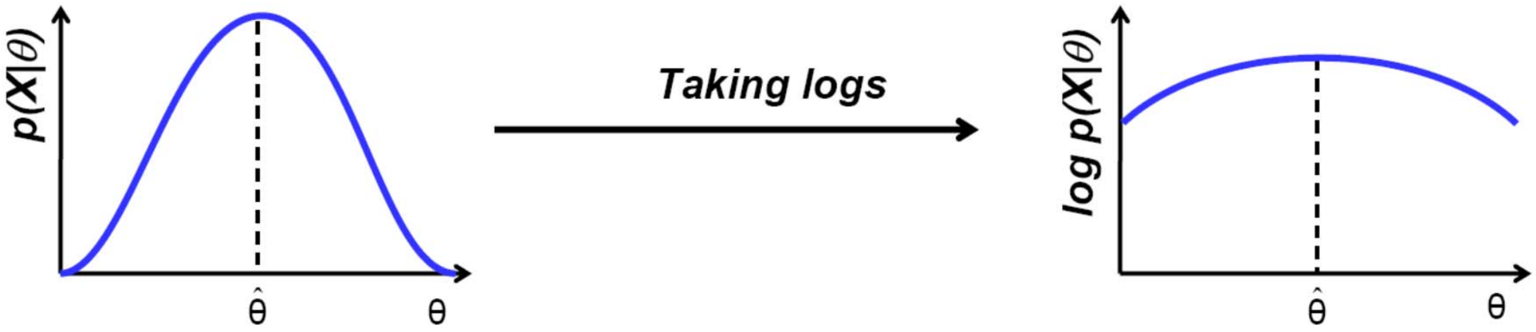


"It is the value of θ that best agrees with or supports the actually observed training samples"

Log-Likelihood Function

- Logarithm is monotonically increasing
  - –Maximizing log-likelihood ↔ maximizing likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left[p(D \mid \boldsymbol{\theta}) \right] = \arg\max_{\boldsymbol{\theta}} \left[\ln p(D \mid \boldsymbol{\theta}) \right]$$



- Define log-likelihood function

$$l(\boldsymbol{\theta}) \equiv \ln p(D \mid \boldsymbol{\theta})$$

Maximum Log-Likelihood

- The ML estimate can be written as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \; l(\boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \ln p(D \mid \boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \ln \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

– Easier to maximize a sum of terms than a product!

– An added advantage for Gaussian distributions

Optimal Estimation

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

- Determine θ that maximizes the log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} [l(\boldsymbol{\theta})]$$

Optimal Estimation (Cont'd)

- Recall that

$$l(\boldsymbol{\theta}) = \ln p(D \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

Necessary conditions for an optimum:

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}$$

Need to identify the true global maximum

MLE: Example

- Suppose that *N* samples $x_1, x_2 \ldots, x_N$ are drawn independently according to the following probability density function:

$$p(x \mid \theta) = \begin{cases} \theta^2 x e^{-\theta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Derive the maximum likelihood estimate of the parameter θ

Solution

- Find the likelihood $p(D\,|\,\theta) = p(x_1, x_2, \ldots, x_N\,|\,\theta)$

$$p(x_1, x_2, \ldots, x_n\,|\,\theta) = \prod_{k=1}^{N} p(x_k\,|\,\theta)$$

$$p(x\,|\,\theta) = \begin{cases} \theta^2 x e^{-\theta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Log-likelihood

$$l(\theta) = \sum_{k=1}^{N} \ln p(x_k\,|\,\theta)$$

$$l(\theta) = \sum_{k=1}^{N} \left(2\ln\theta + \ln x_k - \theta x_k\right)$$

Solution (Cont'd)

- Log-likelihood $l(\theta) = \sum_{k=1}^{N} \left(2\ln\theta + \ln x_k - \theta x_k\right)$

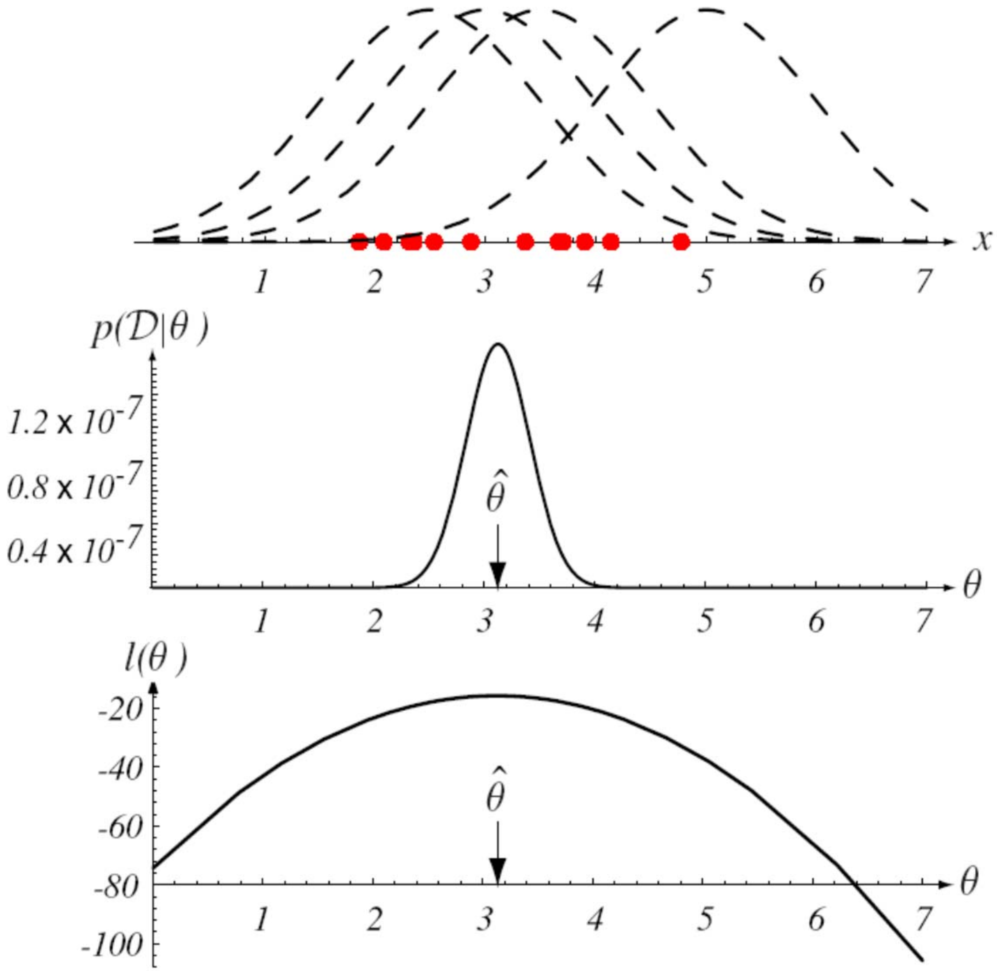
- Taking derivative

$$\nabla_\theta l(\theta) = \sum_{k=1}^{N} \left(\frac{2}{\theta} - x_k\right)$$

$$= \frac{2N}{\theta} - \sum_{k=1}^{N} x_k$$

Let $\nabla_\theta l(\theta) = 0$ $\Rightarrow$ $\frac{2N}{\theta} = \sum_{k=1}^{N} x_k$ $\Rightarrow$ $\hat{\theta} = \frac{2N}{\sum_{k=1}^{N} x_k}$

A Gaussian Example



Training points drawn from a Gaussian of a particular variance, but unknown mean

Likelihood $p(D|\theta)$ as a function of the mean

Log-likelihood $l(\theta)$

From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*

$p(x|\theta)$ vs. $p(D|\theta)$

Gaussian Case 1: Unknown μ

-Univariate

- Given a data set $D = (x_1, x_2, \ldots, x_n)$ where the *n* samples are drawn *independently* from *identical* distribution $N(\mu,\sigma^2)$, where σ^2 is known
- What is the ML estimate of the mean μ?

$$\hat{\theta} = \arg\max_{\theta} \sum_{k=1}^{n} \ln p(x_k \mid \theta) \qquad p(x_k \mid \theta) \sim N(\mu, \sigma^2); \ \theta = \mu$$

$$= \arg\max_{\theta} \sum_{k=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_k - \theta)^2}{2\sigma^2}\right)\right]$$

$$= \arg\max_{\theta} \sum_{k=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_k - \theta)^2}{2\sigma^2}\right]$$

Gaussian Case 1: Unknown μ

-Univariate (Cont'd)

- The maxima (or minima) of a function are defined by the zeros of its derivative:

$$\hat{\theta} = \arg\max_{\theta} \sum_{k=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{(x_k - \theta)^2}{2\sigma^2} \right]$$

$$\frac{\partial \sum_{k=1}^{n} \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{(x_k - \theta)^2}{2\sigma^2} \right]}{\partial \theta} = 0$$

$$\sum_{k=1}^{n} (x_k - \theta) = 0 \Rightarrow \hat{\mu} = \hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

A very intuitive result: the ML estimate of the mean is the average value of the training data

Gaussian Case 1: Unknown μ

-Multivariate

- For each sample vector:

$$p(\mathbf{x}_k \mid \boldsymbol{\theta}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}); \quad \boldsymbol{\theta} = \boldsymbol{\mu}$$

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\boxed{\frac{\partial}{\partial \mathbf{x}}\left[\mathbf{x}^t \mathbf{A}\mathbf{x}\right] = 2\mathbf{A}\mathbf{x}}$$

- Summing over all sample vectors:

$$\nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu}) = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

Gaussian Case 1: Unknown μ

-Multivariate (Cont'd)

The ML estimate for mean vector μ must satisfy:

$$\sum_{k=1}^{n} \mathbf{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0$$

$$\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0$$

Sample mean - arithmetic average of the training samples

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

The ML estimate of the mean for the multivariate Gaussian is the sample mean vector.

Gaussian Case 2: Unknown μ and Σ -Univariate

$$p(x_k \mid \boldsymbol{\theta}) \sim N\left(\mu, \sigma^2\right) \qquad \boldsymbol{\theta} = [\theta_1, \theta_2]^t = \left[\mu, \sigma^2\right]^t$$

$$l = \ln p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

Derivative of the log-likelihood of a single point:

$$\nabla_{\boldsymbol{\theta}} l = \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \ln p(x_k \mid \boldsymbol{\theta}) \\ \dfrac{\partial}{\partial \theta_2} \ln p(x_k \mid \boldsymbol{\theta}) \end{bmatrix}$$

$$\nabla_{\boldsymbol{\theta}} l = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

Gaussian Case 2: Unknown μ and Σ -Univariate (Cont'd)

For the full log-likelihood

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \end{cases}$$

By substituting $\hat{\mu} = \hat{\theta}_1; \quad \hat{\sigma}^2 = \hat{\theta}_2$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k; \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

The ML estimate of the variance is the sample variance of the training data set

Gaussian Case 2: Unknown μ and Σ -Multivariate

- For each sample vector:

$$p(\mathbf{x}_k \mid \boldsymbol{\theta}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

Derivative of the log-likelihood w.r.t. μ:

$$\boxed{\frac{\partial}{\partial \mathbf{x}}\left[\mathbf{x}^t \mathbf{A}\mathbf{x}\right] = 2\mathbf{A}\mathbf{x}}$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

- Summing over all sample vectors:

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = \mathbf{0} \quad \Longrightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

Sample mean vector

Gaussian Case 2: Unknown μ and Σ -Multivariate (Cont'd)

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\ln\left[(2\pi)^d |\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$= -\frac{1}{2}\ln\left[(2\pi)^d\right] - \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

Derivative of the log-likelihood w.r.t. Σ

$$\nabla_{\boldsymbol{\Sigma}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\left[\frac{1}{|\boldsymbol{\Sigma}|}|\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}\right]$$

Matrix derivatives

$$\boxed{\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|\left(\mathbf{X}^{-1}\right)^t} \qquad \boxed{\frac{\partial \mathbf{a}^t \mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-t}\mathbf{a}\mathbf{b}^t\mathbf{X}^{-t}}$$

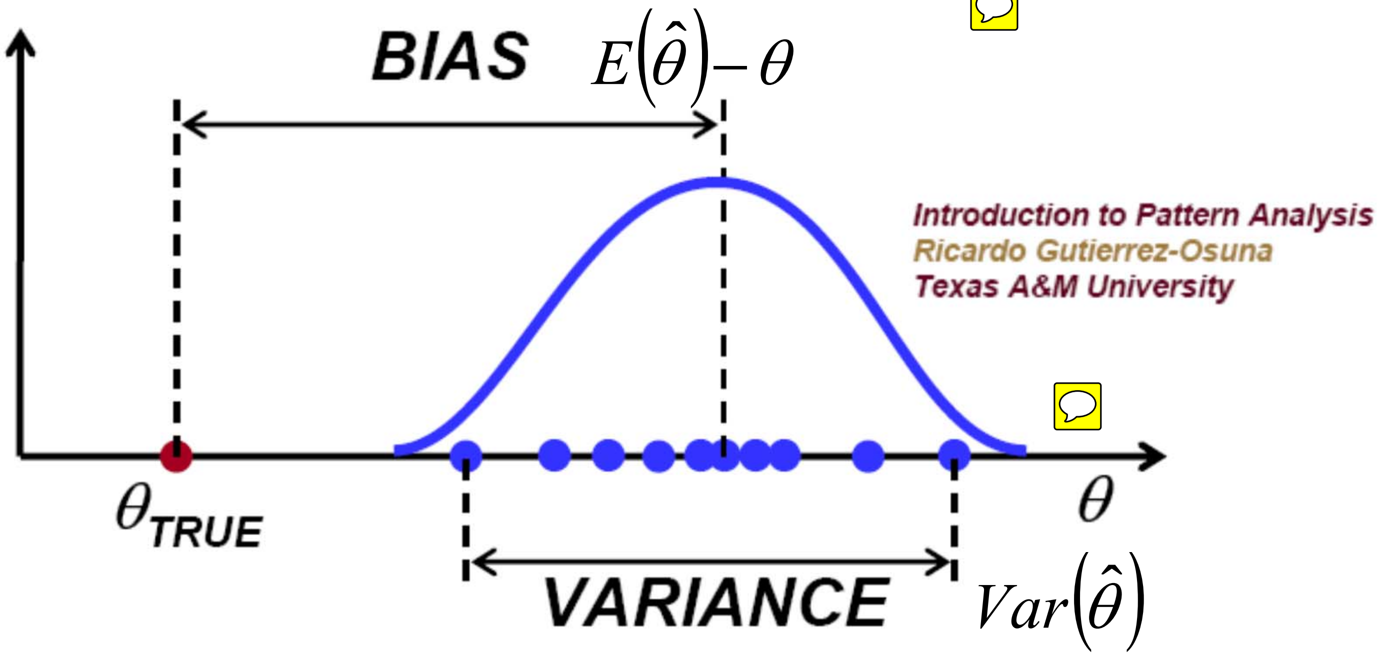
$$\sum_{k=1}^{n}\left[\boldsymbol{\Sigma} - (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right] = \mathbf{0} \Longrightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

Sample covariance matrix

Bias and Variance

- Bias and variance are used to measure how good an estimate is.

How close is the estimate to the true value?



How much does the estimate change for different runs?

Bias

- ML estimate for μ is *unbiased*

$$E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \mu$$

- ML estimate for σ^2 is *biased*

$$E\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Hint: $x_i - \bar{x} = x_i - \mu + \mu - \bar{x} = (x_i - \mu) - \sum_{j=1}^{n} (x_j - \mu) \Big/ n$

– For $n \to \infty$ the bias becomes zero asymptotically

– The bias is only noticeable when we have few samples

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Bias

- An elementary unbiased estimator for σ^2

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{n}{n-1}\frac{n-1}{n}\sigma^2=\sigma^2$$

- An elementary unbiased estimator for Σ:

$$\mathbf{C}=\frac{1}{n\text{-}1}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t$$

Sample covariance matrix
Absolutely unbiased

$$\hat{\boldsymbol{\Sigma}}=\frac{n\text{-}1}{n}\mathbf{C};\quad \lim_{n\to\infty}\hat{\boldsymbol{\Sigma}}=\mathbf{C}$$

Asymptotically unbiased

Bias vs. Variance
- How to generalize better for test data

- Two components of Mean Square Error (MSE)

$$E\left[\left(\theta-\hat{\theta}\right)^2\right]=Var\left(\hat{\theta}\right)+\left[E\left(\hat{\theta}\right)-\theta\right]^2$$

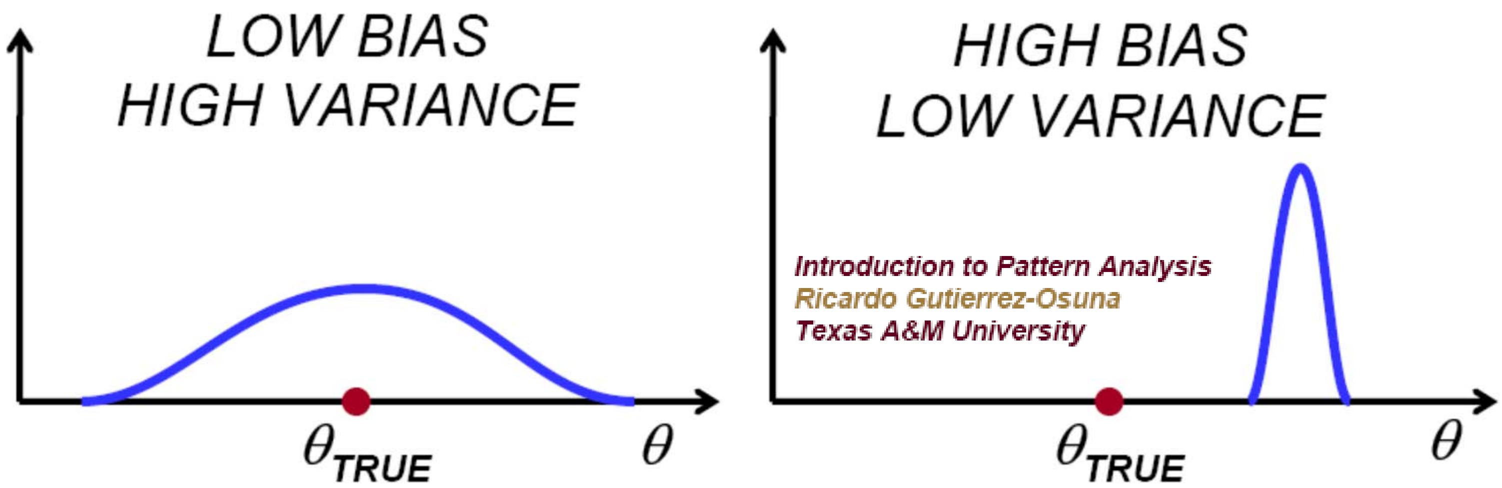
Var($\boldsymbol{X}$)=E($\boldsymbol{X}^2$)-(E($\boldsymbol{X}$))^2

MSE = variance + square of bias

- Bias – systematic, measures the accuracy or quality of match.
- Variance - sensitivity to variability in the data, measures the precision or specificity of the match.

The Bias-Variance Tradeoff

- In most cases, you can only decrease one of them at the expense of the other
  - More complex models have lower bias, but higher variance.
  - More training data → the estimation variance decreases → you can use a more complex model.

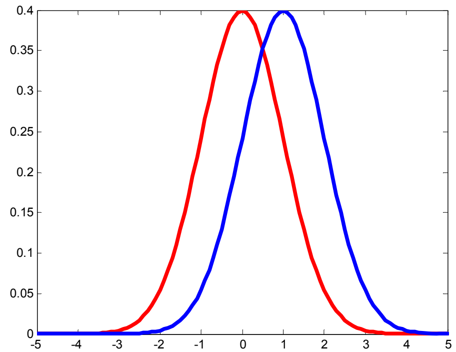


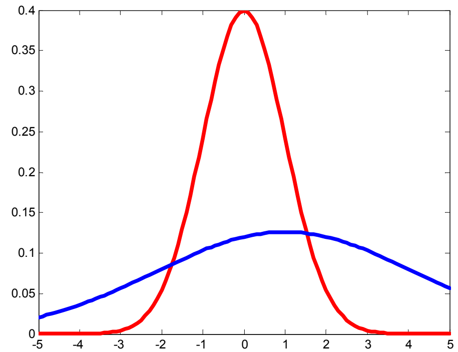
Problems with MLE

- What if our assumption about the model is wrong?
  - For instance, we may assume a distribution comes from $N(\mu, 1)$ but it actually comes from $N(\mu, 10)$
  - Leads to large *model error*

N(0, 1)

N(1, 1)





N(0, 1)

N(1, 10)

- Need reliable information concerning the models
  - Use MLE when the underlying distribution model is known; only the parameter values are to be estimated

Bayesian Estimation

- θ is a random variable ~ $p(\theta)$
- $p(\mathrm{x})$ is unknown but with a known parametric form
- Goal: compute posterior probabilities $P(\omega_i \mid \mathrm{x}, D)$, where $D = \{D_1, \ldots, D_c\}$

Given the sample *D*, Bayes formula becomes

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D) P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D) P(\omega_j \mid D)}$$

Bayesian Parameter Estimation

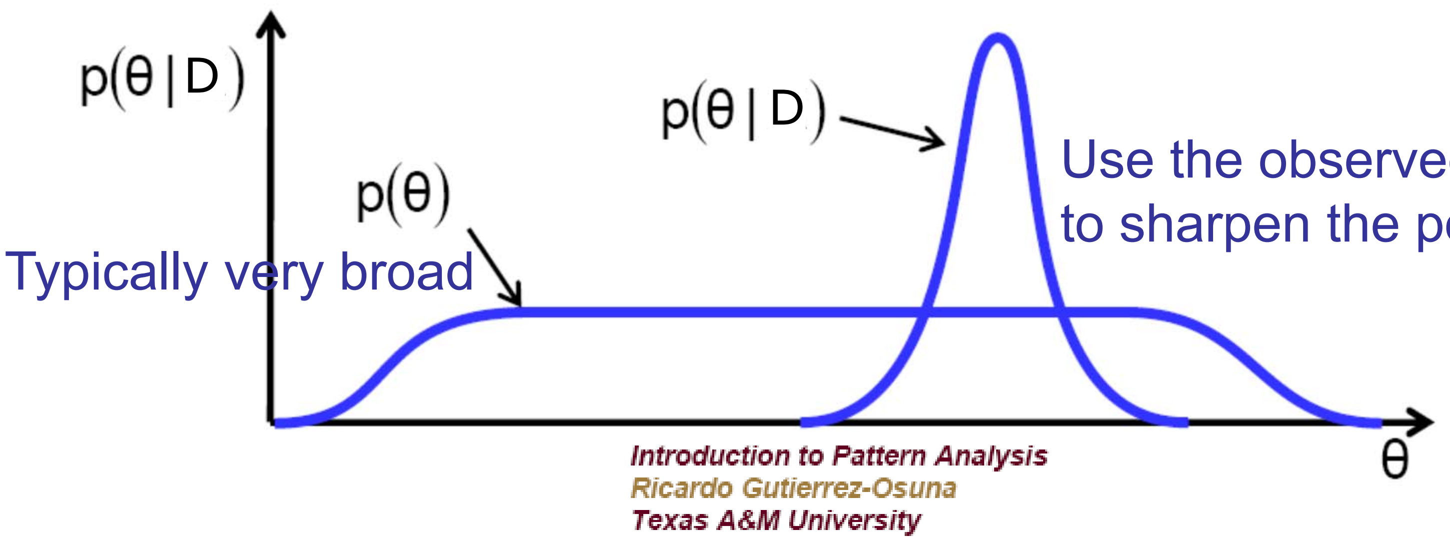
- Assumptions:
  - –Known $P(\omega_i) = P(\omega_i|D)$
  - –Samples in class D_i have no influence on $p(\mathbf{x}|\omega_j, D)$ if $i \neq j$
- Bayes Formula

$$P(\omega_i \mid \mathbf{x}, D) = \frac{\boxed{p(\mathbf{x} \mid \omega_i, D_i)} P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D_j) P(\omega_j)}$$

- Solve c separate problems (drop class distinction)
  - –Use a set D of samples drawn independently according to the fixed but unknown $p(\mathbf{x})$ to determine $p(\mathbf{x}|D)$.

Bayesian Parameter Estimation

- The parameters are assumed to be random variables with some (assumed) known a priori distribution $p(\theta)$
- Bayesian method seeks to estimate the posterior $p(\theta|D)$



The Parameter Distribution

- Assumptions:
  - –$p(\mathbf{x})$ is unknown but with a known parametric form
  - –$p(\mathbf{x}|\theta)$ is completely known
  - –prior density $p(\theta)$ is known

Integrating the joint density over θ:

$$P(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

$$p(\mathbf{x} \mid \boldsymbol{\theta}, D) = p(\mathbf{x} \mid \boldsymbol{\theta})$$

$$P(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

- Links the desired $p(\mathbf{x}|D)$ to $p(\theta|D)$
  - –Average $p(x|\theta)$ over the possible values of θ
  - –Integration can be performed numerically

Bayesian Parameter Estimation: Procedures

- Use Bayes rule to calculate:
  - *a posteriori* density

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}$$

  - By assuming i.i.d.

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

  - Class-conditional density

$$P(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

Integration can be difficult!

The Univariate Case: $p(\mu \mid D)$

- μ is the only unknown parameter

$$p(x \mid \mu) \sim N\left(\mu, \sigma^2\right)$$

- *known* prior density - μ_0 and σ_0 are known!

$$p(\mu) \sim N\left(\mu_0, \sigma_0^2\right)$$

The crucial assumption is that the prior distribution for μ is known, not necessarily normal.

The Univariate Case: $p(\mu \mid D)$ (Cont'd)

Let $D = \{x_1, \ldots, x_n\}$, where $x_1, \ldots, x_n$ are independently drawn, then:

$$p(\mu \mid D) = \frac{p(D \mid \mu) p(\mu)}{\int p(D \mid \mu) p(\mu) d\mu}$$

Relates $p(\mu)$ to $p(\mu|D)$

$$= \alpha \prod_{k=1}^{n} p(x_k \mid \mu) p(\mu)$$

$$= \alpha p(\mu) \prod_{k=1}^{n} p(x_k \mid \mu)$$

$$p(\mu \mid D)$$

$$= \alpha \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \prod_{k=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_k - \mu)^2}{2\sigma^2}\right)\right]$$

The Univariate Case: $p(\mu \mid D)$ (Cont'd)

$$p(\mu \mid D) = \alpha' \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\prod_{k=1}^{n}\left[\exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right)\right]$$

$$p(\mu \mid D) = \alpha' \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\sigma_0^2}+\sum_{k=1}^{n}\frac{(x_k-\mu)^2}{\sigma^2}\right)\right]$$

$p(\mu \mid D)$ is a normal density and it remains normal as the number of training samples is increased.

Reproducing density: $p(\mu \mid D) \sim N(\mu_n, \sigma_n^2)$

Conjugate prior: $p(\mu) \sim N(\mu_0, \sigma_0^2)$

The Univariate Case: $p(\mu \mid D)$ (Cont'd)

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu + \frac{\mu_n^2}{\sigma_n^2}\right)\right]$$

$$p(\mu \mid D) = \alpha' \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\sigma_0^2} + \sum_{k=1}^{n}\frac{(x_k-\mu)^2}{\sigma^2}\right)\right]$$

The coefficient of μ^2: $-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)$

$$\boxed{\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k}$$

The coefficient of μ: $-\frac{1}{2}\left(\frac{-2\mu_0}{\sigma_0^2} + \frac{-2\sum_{k=1}^{n} x_k}{\sigma^2}\right) = -\frac{1}{2}\left(\frac{-2\mu_0}{\sigma_0^2} + \frac{-2n\hat{\mu}_n}{\sigma^2}\right)$

The Univariate Case: $p(\mu \mid D)$ (Cont'd)

Find μ_n and σ_n^2 by equating coefficients:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} . \mu_0$$

Linear combination of $\hat{\mu}_n$ and μ_0, where $\hat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$

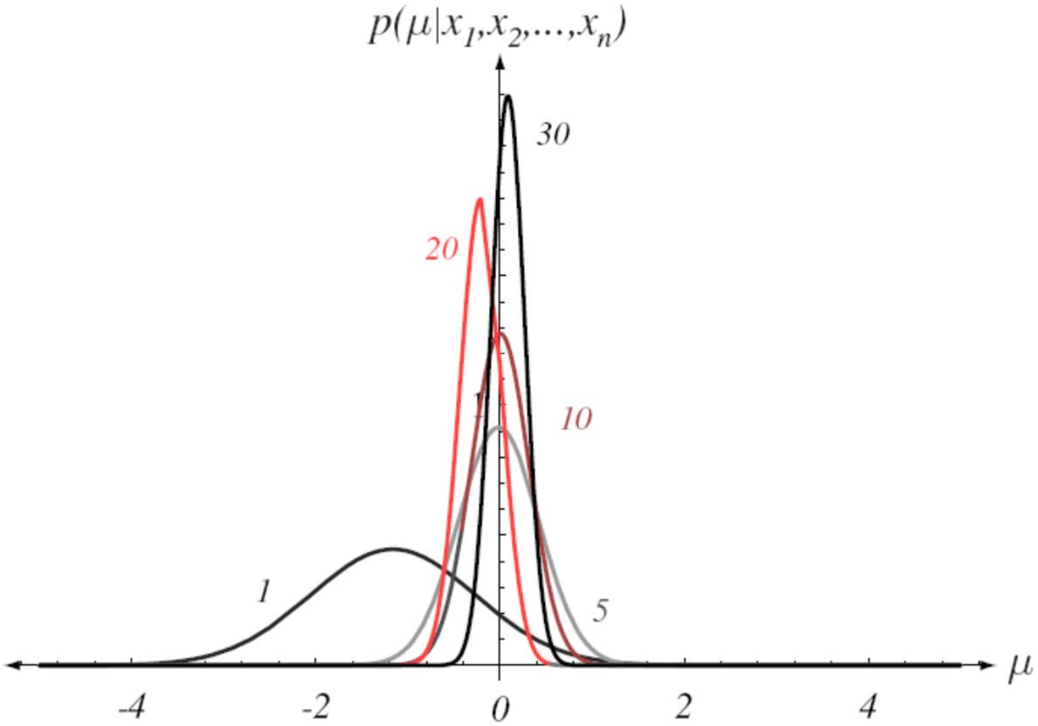
$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

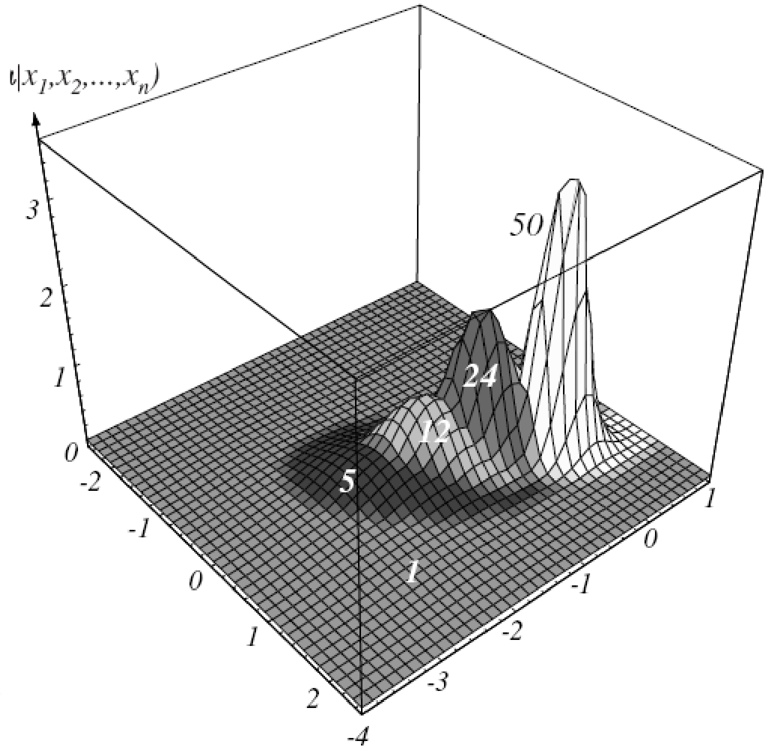
Decreases monotonically with *n* – each additional observation decreases our uncertainty about *μ*.

The effects of σ_0^2 ?

Bayesian Learning

As n approaches infinity, $p(\mu \mid D)$ approaches a Dirac delta function.





From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*

The Univariate Case: $p(x \mid D)$

Having obtained: $p(\mu \mid D) \sim N(\mu_n, \sigma_n^2)$

$$p(x \mid D) = \int p(x \mid \mu) p(\mu \mid D) d\mu$$ Class-conditional density

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma, \sigma_n)$$

Scaling factor

$$p(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$ Normally distributed

The Univariate Case: $p(x \mid D)$ (Cont'd)

$$p(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- Observations:
  - The conditional mean μ_n is treated as if it were the true mean.
  - The known variance is increased to account for the additional uncertainty in x resulting from lack of exact knowledge of μ.
- Remarks:
  - $p(x \mid D)$ is the desired class-conditional density $p(x \mid D_j, \omega_j)$;
  - Together with prior probabilities, we now have needed information to apply the Bayesian classification rule:

$$\max_{\omega_j} P(\omega_j \mid x, D) \equiv \max_{\omega_j} p(x \mid \omega_j, D_j) P(\omega_j)$$

The Multivariate Case

Assume:

$$p(\mathbf{x} \mid \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{and} \quad p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

Similar to the univariate case,

$$p(\boldsymbol{\mu} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n),$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma}$$

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

Sample mean

The Multivariate Case (Cont'd)

Performing the integration:

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid D) d\boldsymbol{\mu}$$

$$p(\mathbf{x} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

Alternatively, view **x** as the sum of two independent variables:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{y}$$

$$p(\boldsymbol{\mu} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \qquad p(\mathbf{y}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$E(\mathbf{x}) = E(\boldsymbol{\mu} + \mathbf{y}) = E(\boldsymbol{\mu}) + E(\mathbf{y}) = \boldsymbol{\mu}_n$$

$$E(\mathbf{x} - \boldsymbol{\mu}_n)(\mathbf{x} - \boldsymbol{\mu}_n)^t = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n$$

Bayesian Parameter Estimation: General Theory

- $p(\mathrm{x} \mid \mathrm{D})$ computation can be applied to any situation in which the unknown density can be parameterized.
- The basic assumptions are:
  - The form of $p(\mathrm{x} \mid \theta)$ is assumed known, but the value of θ is not known exactly.
  - Our knowledge about θ is assumed to be contained in a known prior density $p(\theta)$.
  - The rest of our knowledge of θ is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows $p(\mathrm{x})$.

The Basic Problem

Compute the posterior density $p(\theta \mid D)$, then compute $p(\mathbf{x} \mid D)$

By Bayes formula: $$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

By independence assumption: $$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

By integration: $$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

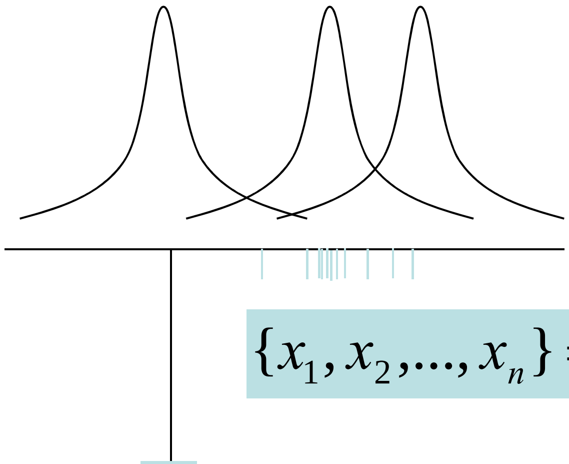
Bayesian solution tells us how to use *all* the available information to compute the desired density $p(\mathbf{x} \mid D)$.

ML Vs. Bayesian

goal: Find $p(x \mid \omega_i)$ from D

$p(x \mid \omega_i) \sim N(\mu, \sigma^2)$

μ is unknown

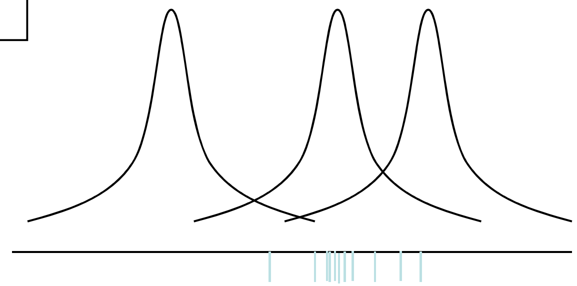


$\{x_1, x_2, \ldots, x_n\} = D$

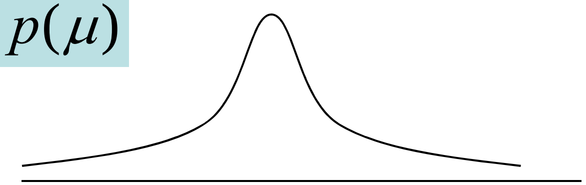
μ

sub-goal: Find μ
which maximizes $p(D \mid \mu)$

goal: Find $p(x \mid \omega_i, D)$



$p(\mu)$



sub-goal: Find $p(\mu \mid D)$
and then use

$$p(x \mid D) = \int p(x \mid \mu, D)\, p(\mu \mid D)\, d\mu$$

ML Vs. Bayesian (Cont'd)

- In both cases, parameters defining the underlying distribution are estimated.
- Both methods assume the form of the density is known, but the value of the parameter is unknown.
- ML methods estimate the point value of a parameter (a fixed point).
- Bayesian methods estimate the distribution of a parameter (a random variable).

ML Vs. Bayesian (Cont'd)

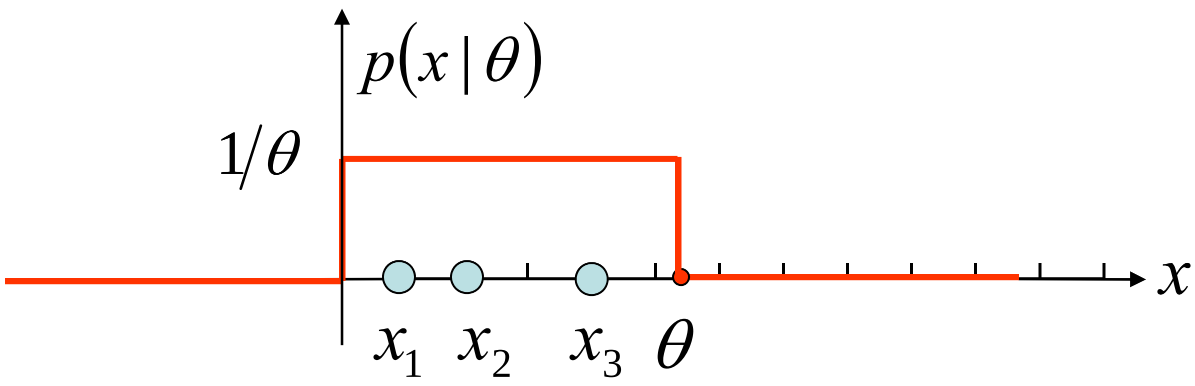
- Computational complexity:
  - –Bayesian methods are in nature more complex than ML methods
  - –Complex multidimensional integration vs. differential calculus techniques or gradient search
- Interpretability:
  - –ML solution is easy to interpret and understand.
- Information used:
  - –Bayesian methods use more information about the problem than do ML methods (prior information).
- The Bayesian estimate will approach the ML solution as $n \rightarrow \infty$

MLE: Numerical Example

- X is uniformly distributed with parameter θ

$$p(x \mid \theta) = \begin{cases} 1/\theta & x \in [0, \theta] \\ 0 & \text{otherwise} \end{cases}$$

- We are given 3 samples $D=\{1,2,4\}$ that are independently drawn from $p(x|\theta)$.



- What is the maximum likelihood estimate of θ?

MLE: Solution

- Find the likelihood of observing D={1,2,4}

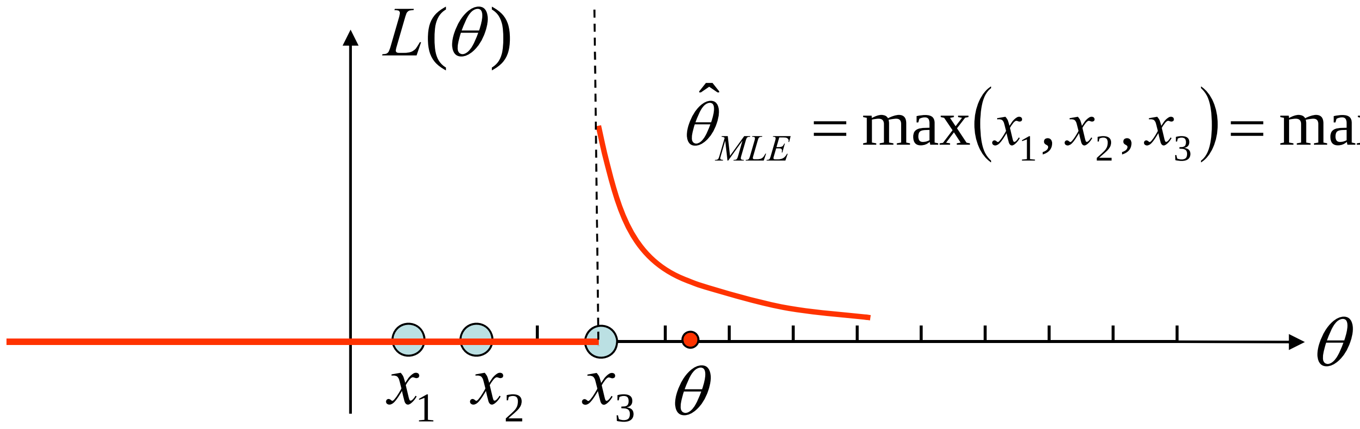
$$\mathbf{x} = [x_1 \quad x_2 \quad x_3]^t = [1 \quad 2 \quad 4]^t$$

$$L(\theta) = p(\mathbf{x} \mid \theta) = p(x_1 \mid \theta) p(x_2 \mid \theta) p(x_3 \mid \theta)$$

$$L(\theta) = \begin{cases} 1/\theta^3 & \text{if } \theta \geq \max(x_1, x_2, x_3) \\ 0 & \text{if } \theta < \max(x_1, x_2, x_3) \end{cases}$$

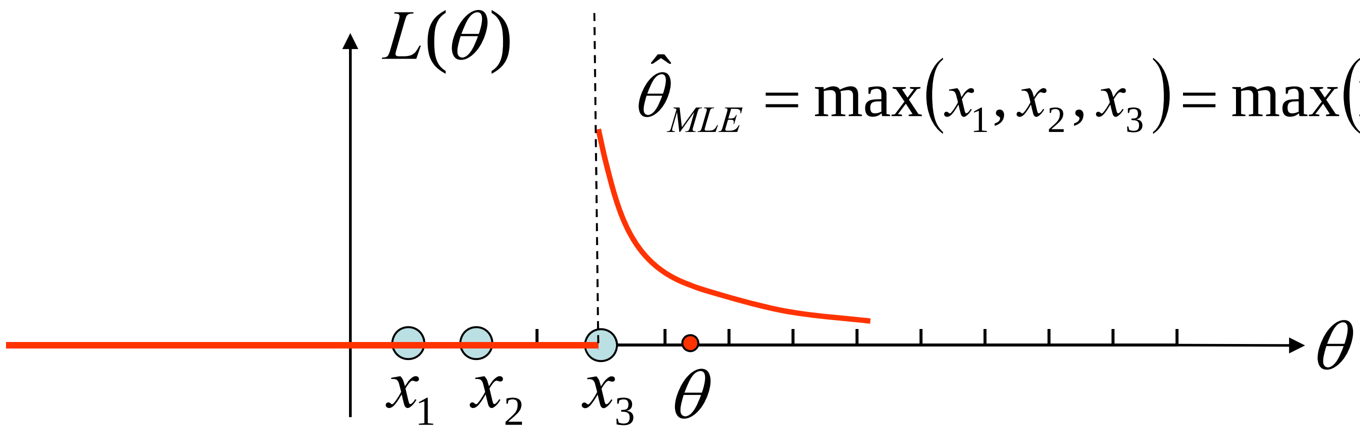
No sample is greater than θ

Any sample is greater than θ



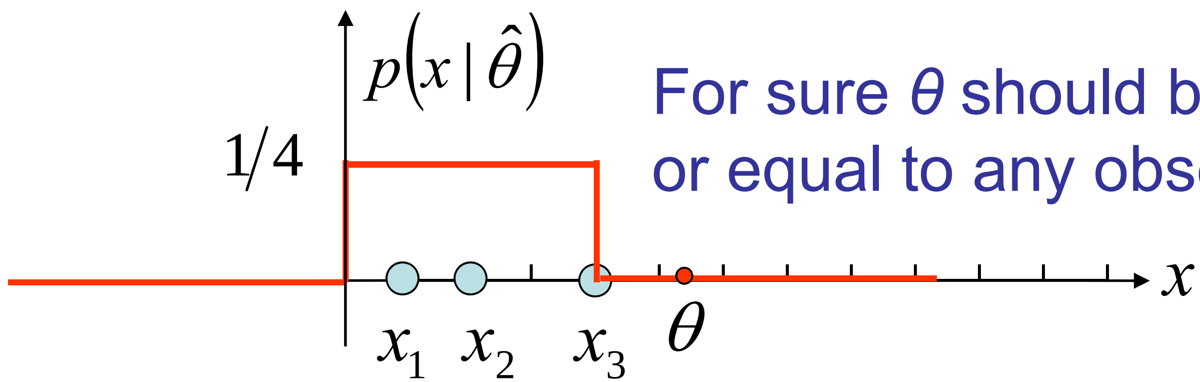
$$\hat{\theta}_{MLE} = \max(x_1, x_2, x_3) = \max(1,2,4) = 4$$

MLE: Solution (Cont'd)



$$\hat{\theta}_{MLE} = \max(x_1, x_2, x_3) = \max(1,2,4) = 4$$

The ML estimate of the density function

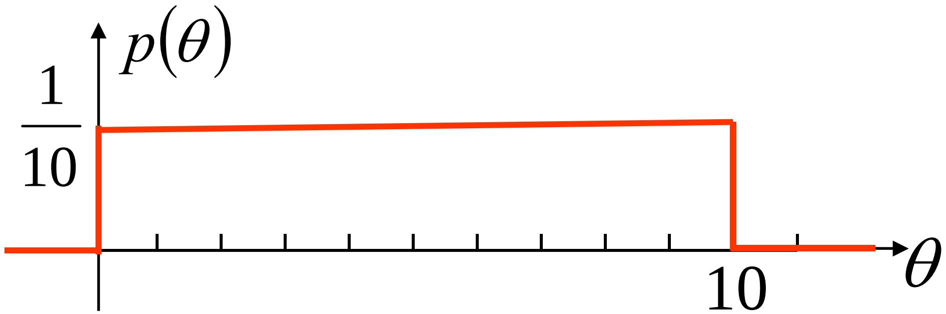


For sure θ should be greater than or equal to any observed x!

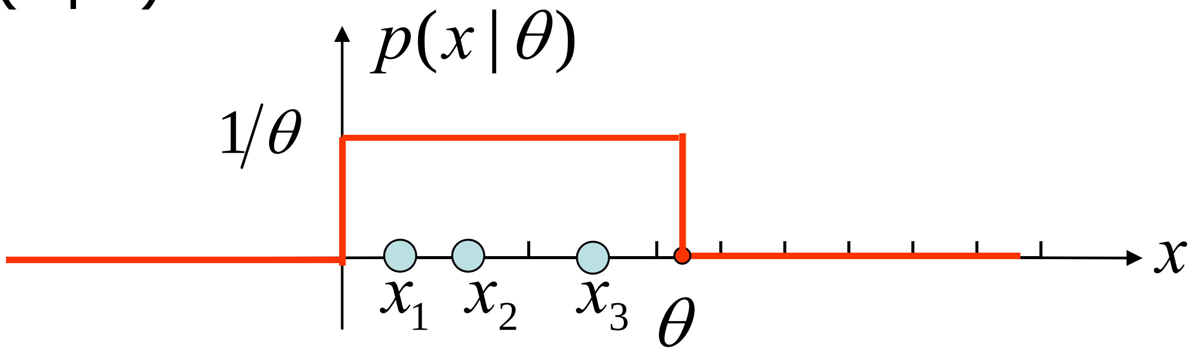
BPE: Numerical Example

- X is uniformly distributed with parameter θ
- And we assume a uniform prior for θ

$$p(\theta) = \begin{cases} 1/10 & \theta \in [0,10] \\ 0 & \text{otherwise} \end{cases}$$



- Given 3 independently drawn samples D={1,2,4}, what is the Bayesian estimate of the density function $p(x|D)$?

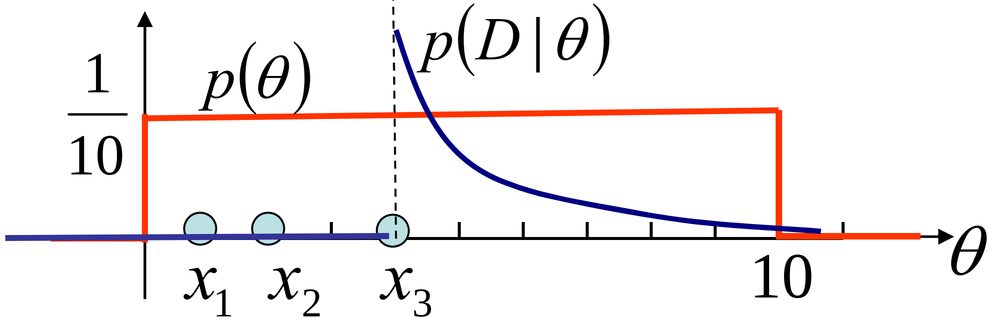


BPE: Solution

- In order to estimate $p(x|D)$, we first compute $p(\theta|D)$

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \qquad p(D|\theta)=\prod_{k=1}^{n}p(x_k|\theta)$$

Recall that $$p(D|\theta)=\begin{cases}1/\theta^3 & \text{if } \theta\geq\max(x_1,x_2,x_3)\\ 0 & \text{if } \theta<\max(x_1,x_2,x_3)\end{cases}$$



Thus $$p(\theta|D)=\begin{cases}c\dfrac{1}{\theta^3} & \text{if } \max(x_1,x_2,x_3)\leq\theta\leq 10\\ 0 & \text{otherwise}\end{cases}$$

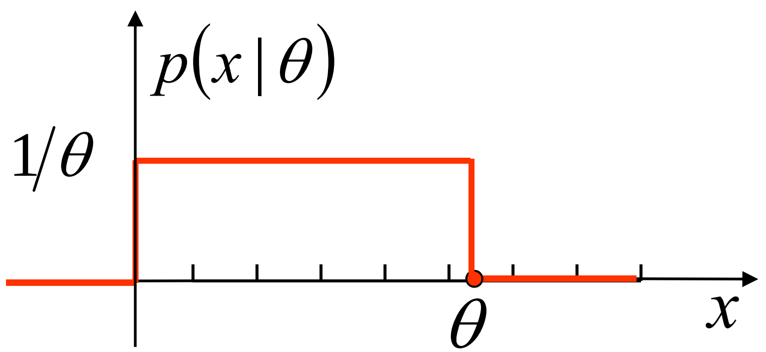
BPE: Solution (Cont'd)

- Having obtained $p(\theta|D)$, now we compute $p(x|D)$

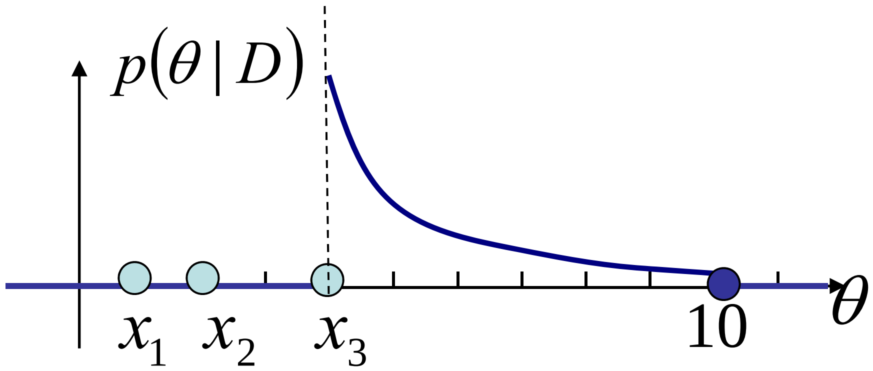
$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

Note that the integration is over θ.

Recall that $$p(\theta \mid D) = \begin{cases} c\dfrac{1}{\theta^3} & \text{if } \max(x_1, x_2, x_3) \le \theta \le 10 \\ 0 & \text{otherwise} \end{cases}$$



$0 \le x \le \theta$



$\max(x_1, x_2, x_3) \le \theta \le 10$

BPE: Solution (Cont'd)

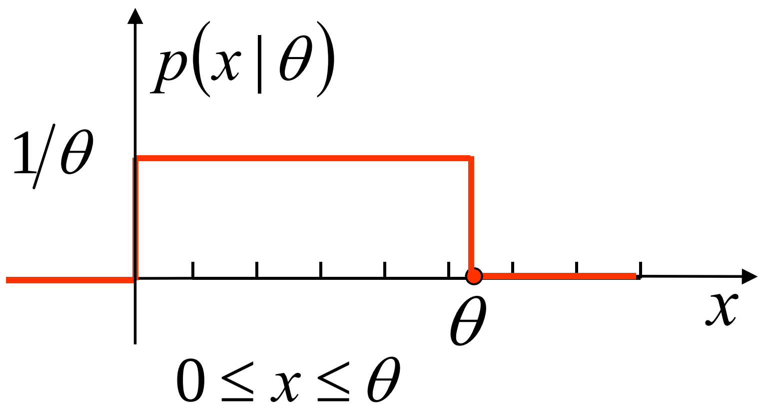
$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

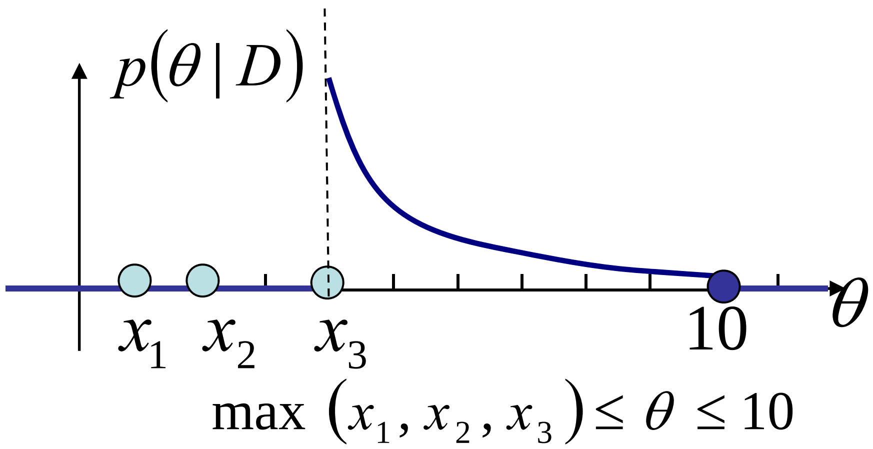
For any $x \in [0, 10]$, $p(x \mid \theta) p(\theta \mid D) \neq 0$ requires θ to satisfy:

$$x \leq \theta \quad \text{and} \quad \max(x_1, x_2, x_3) \leq \theta \leq 10$$

Therefore,

$$p(x \mid \theta) p(\theta \mid D) = \begin{cases} c\dfrac{1}{\theta^4} & \text{if } \max(x_1, x_2, x_3, x) \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$





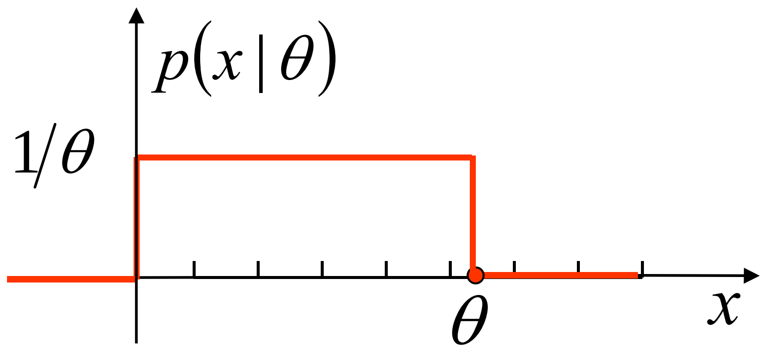
BPE: Solution (Cont'd)

- Two cases: Different ranges of integration

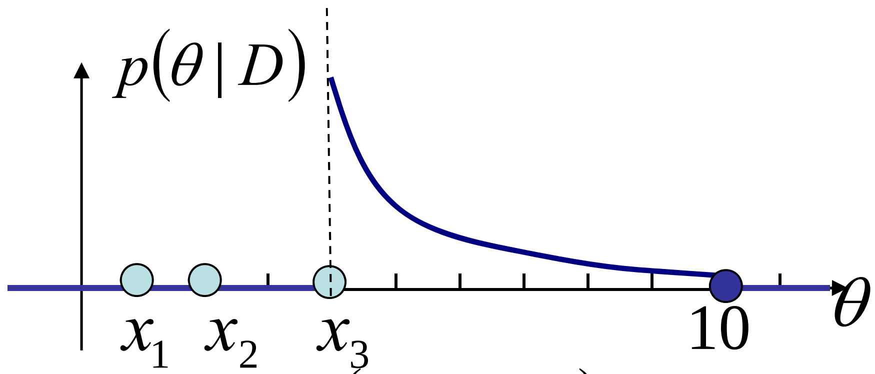
$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

If $0 \le x \le \max(x_1, x_2, x_3)$, then $\max(x_1, x_2, x_3, x) = \max(x_1, x_2, x_3)$

$$p(x \mid D) = \int_{\max(x_1, x_2, x_3)}^{10} c \frac{1}{\theta^4} d\theta = \alpha$$ Independent of x



$0 \le x \le \theta$



$\max(x_1, x_2, x_3) \le \theta \le 10$

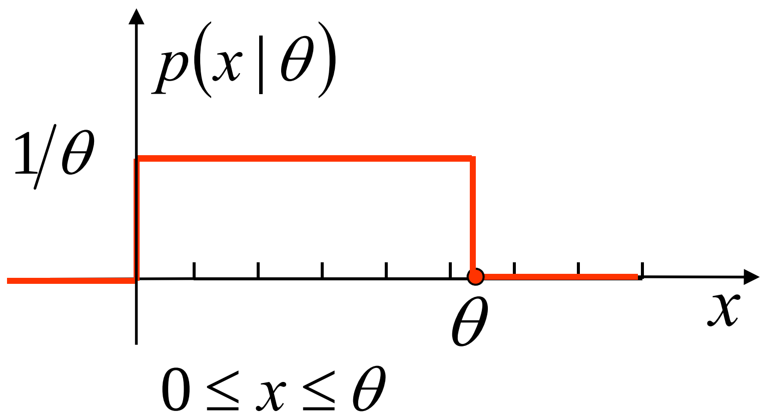
BPE: Solution (Cont'd)

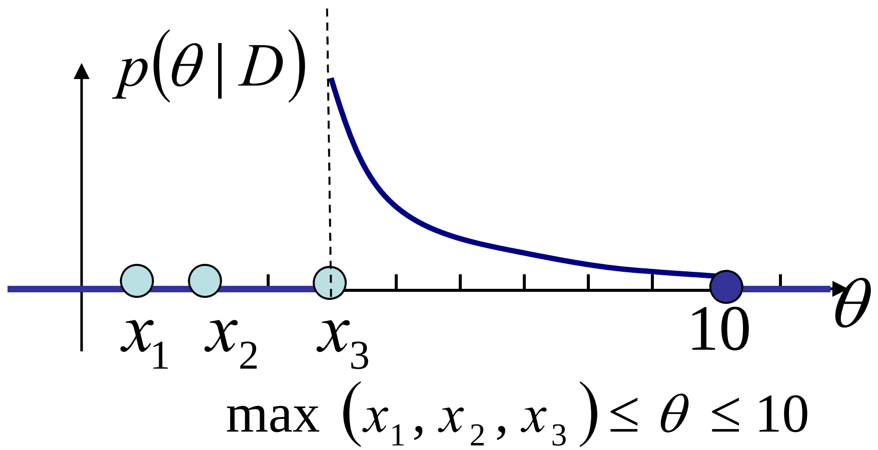
- Two cases: Different ranges of integration

$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

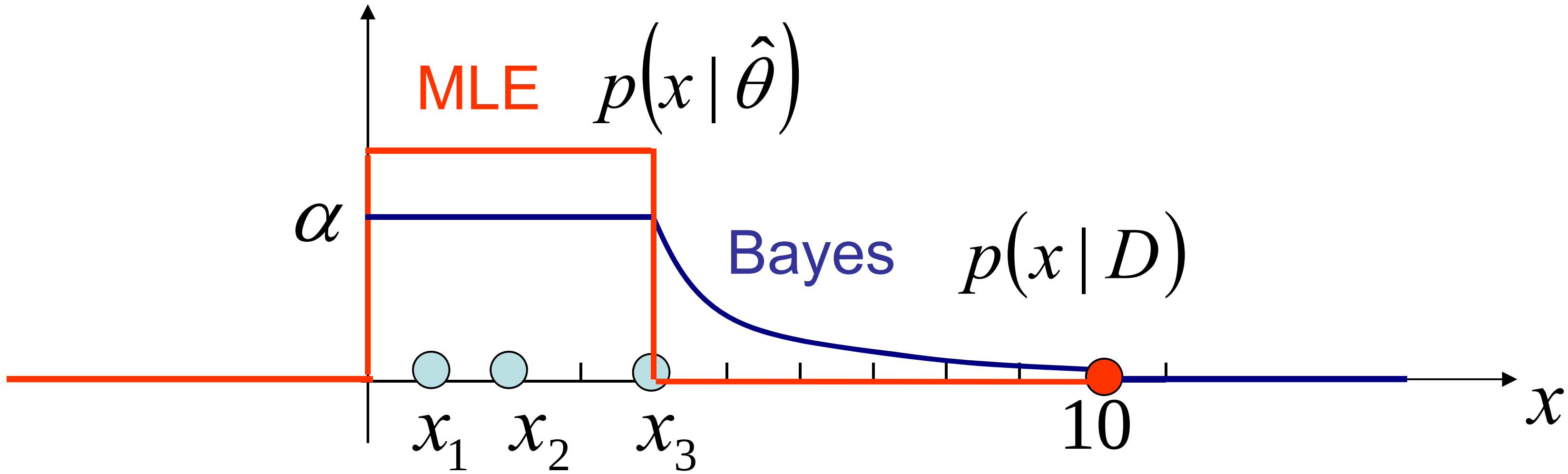
If $\max(x_1, x_2, x_3) < x \le 10$, then $\max(x_1, x_2, x_3, x) = x$

$$p(x \mid D) = \int_x^{10} c \frac{1}{\theta^4} d\theta = -\frac{c}{3}\theta^{-3}\Big|_x^{10} = -\frac{c}{3}10^{-3} + \frac{c}{3}x^{-3}$$





MLE vs. BPE



Observations:

When $x \geq \max(x_1, x_2, x_3)$ Bayes density is not zero!

Bayes density is not uniform – it does not have the functional form that we have assumed!

Sources of Classification Errors

- Bayes error
  - –Caused by overlapping conditional densities
  - –Can never be eliminated (an inherent property)
- Model error
  - –Assumption of probability density function
  - –Number of parameters
- Estimation error
  - –Parameters estimated from a *finite* sample
  - –Can be reduced by increasing the number of training samples

Problems of Dimensionality

- *"The performance of a classifier depends on the interrelationship between sample size, number of features, and classifier complexity"* from Jain's PAMI paper, pp.11
- Theoretically, one can always reduce the error rate by introducing new *independent* features.
  - Comes with increased cost and complexity
  - Additional information → improved performance
- In practice, increasing the number of features may not improve classification accuracy.
  - Wrong model assumption
  - Finite samples → inaccurate estimation of the distribution

Feature Dimensions vs. Bayes Error Rates

- Case of two-class multivariate normal with the same covariance, and equal prior probabilities:

$$P(error) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{\frac{-u^2}{2}} du \text{ , where } r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Squared Mahalanobis distance

- The probability of error decreases as r increases:

$$\lim_{r \to \infty} P(error) = 0$$

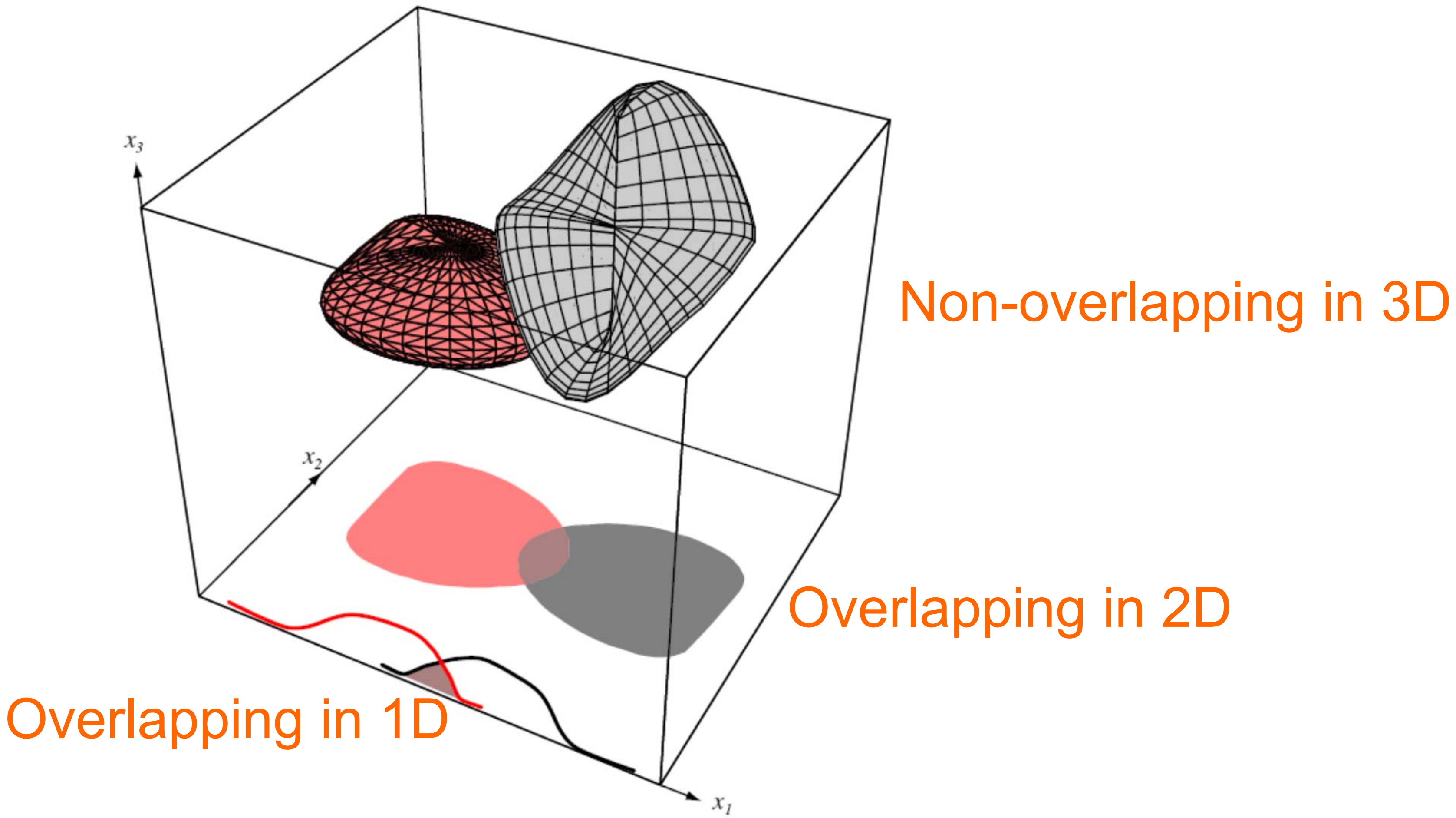
Feature Dimensions Vs. Bayes Error Rates (Cont'd)

- If features are independent:

$$\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2) \qquad r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- Most useful features are the ones for which the difference between the means is large relative to the standard deviation
- Theoretically, inclusion of additional features leads to better performance.

Bayes Errors in 3D, 2D, and 1D



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*
Copyright © 2001 by John Wiley & Sons, Inc.

Curse of Dimensionality

- If we have limited amount of training data, the accuracy of a classifier can decrease if we increase the dimensionality of the feature vector beyond a limit.
  - Higher dimensional feature space will require *huge* number of training samples.
  - In practice, one should try to limit the number of features to be used in classifier design.
  - Dimension reduction for high-dimensional data
    - ✓ In many cases, there are really few things that matter.
    - ✓ Reduce the number of dimensions by eliminating some coordinates that seem irrelevant.

Summary

- Supervised learning
  - Training samples are labeled
- Parameter estimation
  - Assume a particular form for the density (e.g., Gaussian); and estimate the parameters
- Maximum Likelihood Estimation
  - Parameters are assumed to be FIXED but unknown
  - MLE seeks the solution that "best" explains the data set, i.e., maximizing $p(D|\theta)$
- Bayesian Parameter Estimation
  - Parameters are assumed to be RANDOM variables with known prior distribution $p(\theta)$
  - BPE seeks to estimate the posterior density $p(\theta|D)$

Key Concepts

- Supervised learning
- Parameter estimation
  - –Maximum likelihood estimation
  - –Bayes parameter estimation
  - –Bias and variance
- Problems of dimensionality
  - –Relation of classification accuracy, feature dimension and training sample size

Readings

- Chapter 3, Pattern Classification by Duda, Hart, Stork, 2001, Sections 3.1 – 3.4, 3.7

Next Time: Nonparametric Techniques

- Nothing is known about the probability distribution.
- All we have is labeled training data

- Need to:
  - –estimate prior probabilities
  - –estimate the probability distribution from the labeled data using non-parametric methods

Then we can apply Bayesian decision theory!