Chapter 3 (part 2):
Maximum-Likelihood and Bayesian Parameter Estimation

- Bayesian Estimation (BE)
- Bayesian Parameter Estimation: Gaussian Case
- Bayesian Parameter Estimation: General Estimation
- Sufficient Statistics

- 3.3 **Bayesian Estimation** (Bayesian learning to pattern classification problems)
  - In MLE **θ** was supposed fixed
  - In BE **θ** is a random variable
  - The computation of posterior probabilities $P(\omega_i|\mathbf{x})$ lies at the heart of Bayesian classification
  - But what If the priors and class-conditional densities are unknown?
  - Goal: compute $P(\omega_i|\mathbf{x})$ using all of the information at our disposal such as $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$ → $P(\omega_i|\mathbf{x}, \mathcal{D})$.

    Given the sample $\mathcal{D}$, Bayes formula can be written

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

- Supervised case → we separate the training samples by class into *c* subsets $\mathcal{D}_1, \ldots, \mathcal{D}_c$, with the samples in $\mathcal{D}_i$ belonging to ω_i.
- In most cases of interest the samples in $\mathcal{D}_i$ have no influence on $p(\mathbf{x}|\omega_j, \mathcal{D})$ if $i \neq j$.
- Suppose priors are known; $P(\omega_i)=P(\omega_i|\mathcal{D})$

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

The Parameter Distribution

- The desired probability density $p(\mathbf{x})$ is unknown. We assume that it has a known parametric form (vector $\boldsymbol{\theta}$).
- So, the function $p(\mathbf{x}|\boldsymbol{\theta})$ is completely known.
- Any information we might have about $\boldsymbol{\theta}$ prior to observing the samples is assumed to be contained in a known prior density $p(\boldsymbol{\theta})$.
- Observation of the samples converts this to a posterior density $p(\boldsymbol{\theta}/\mathcal{D})$**,** which, we hope, is sharply peaked about the true value of $\boldsymbol{\theta}$.

- Note that we are changing our supervised learning problem (pdf) into an unsupervised density estimation problem (parameter vector).
- Our basic goal is to compute $p(\mathbf{x}|\mathcal{D})$, which is as close as we can come to obtaining the unknown $p(\mathbf{x})$.
- We do this by integrating the joint density $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ over $\boldsymbol{\theta}$.

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

- We can write $p(\mathbf{x}, \boldsymbol{\theta}/\mathcal{D})$ as the product $p(\mathbf{x}/\boldsymbol{\theta}, \mathcal{D})\ p(\boldsymbol{\theta}/\mathcal{D})$. Since the selection of **x** and that of the training samples in $\mathcal{D}$ is done independently, $p(\mathbf{x}/\boldsymbol{\theta}, \mathcal{D})$ is merely $p(\mathbf{x}/\boldsymbol{\theta})$.

$$\boxed{p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}} \qquad (25)$$

$$\boxed{p(\mathbf{x}\,|\,\mathcal{D}) = \int p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D})\,d\boldsymbol{\theta}} \tag{25}$$

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

$$p(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k\,|\boldsymbol{\theta})$$

Looking more carefully at (25) and assuming that $p(\boldsymbol{\theta}/\mathcal{D})$ is known, then $p(\mathbf{x}/\mathcal{D})$ is nothing but the average of $p(\mathbf{x}/\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, that is, $p(\mathbf{x}\,|\,\mathcal{D}) = E_{\boldsymbol{\theta}}\left[p(\mathbf{x}\,|\,\boldsymbol{\theta})\right]$

If we assume that a large enough number of samples $\boldsymbol{\theta}_{\mathrm{i}}$, $i = 1, 2 \ldots , L$, of the random vector $\boldsymbol{\theta}$ are available, one can compute the corresponding values $p(\mathbf{x}/\boldsymbol{\theta}_{\mathrm{i}})$ and then approximate the expectation as the mean value

$$p(\mathbf{x}\,|\,\mathcal{D}) = \frac{1}{L}\sum_{i=1}^{L} p(\mathbf{x}|\boldsymbol{\theta}_i)$$

- Bayesian estimation approach estimates a distribution for $p(\mathbf{x}/\mathcal{D})$ rather than making point estimates like ML.
- This key equation links the desired class-conditional density $p(\mathbf{x}/\mathcal{D})$ to the posterior density $p(\boldsymbol{\theta}/\mathcal{D})$ for the unknown parameter vector.
- If $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply about some value $\hat{\boldsymbol{\theta}}$, we obtain $p(\mathbf{x}|\mathcal{D}) \approx p(\mathbf{x}|\,\hat{\boldsymbol{\theta}})$, i.e., the result we would obtain by substituting the estimate $\hat{\boldsymbol{\theta}}$ for the true parameter $\boldsymbol{\theta}$.
- When the unknown densities have a known parametric form, the samples exert their influence on $p(\mathbf{x}|\mathcal{D})$ through the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$.
- We should also point out that in practice, the integration in Eq. 25 is often performed numerically, for instance by Monte-Carlo simulation.

3.4 Bayesian Parameter Estimation:

Gaussian Case

Goal: use Bayesian estimation techniques to calculate the a posteriori density $p(\theta/\mathcal{D})$ and the desired probability density $p(\mathbf{x}/\mathcal{D})$ for the case where $p(\mathbf{x} \mid \boldsymbol{\mu}) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- **The univariate case**: $p(\mu/\mathcal{D})$

  μ is the only unknown parameter

$$p(\mathrm{x} \mid \mu) \sim N(\mu, \sigma^2)$$

$$\mathrm{p}(\mu) \sim N(\mu_0, \sigma_0^2)$$

μ_0 and σ_0 are known!

- Roughly speaking, μ_0 represents our best a priori guess for μ, and σ_0^2 measures our uncertainty about this guess.
- Imagine that a value is drawn for μ from a population governed by the probability law *p*(μ). Once this value is drawn, it becomes the true value of μ and completely determines the density for *x*.
- Suppose now that *n* samples $x_1, \ldots, x_n$ are independently drawn from the resulting population. Letting $\mathcal{D} = \{x_1, \ldots, x_n\}$, we use Bayes' formula to obtain

$$p(\mu \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mu).p(\mu)}{\int p(\mathcal{D} \mid \mu).p(\mu)d\mu} = \alpha \prod_{k=1}^{k=n} p(x_k \mid \mu).p(\mu)$$

where α is a normalization factor that depends on $\mathcal{D}$ but is independent of μ.

This equation shows how the observation of a set of training samples affects our ideas about the true value of μ; it relates the prior density *p*(μ) to an a posteriori density *p*(μ/$\mathcal{D}$).

Because $p(x_k|\mu) \sim N(\mu, \sigma^2)$ and $p(\mu) \sim N(\mu_0, \sigma_0^2)$, we have

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

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

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

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

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

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

$$= \alpha'' \exp -\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)\mu^2 - 2(\frac{1}{\sigma^2}(n\hat{\mu}_n) + \frac{\mu_0}{\sigma_0^2})\mu \right] \tag{1}$$

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

- Where factors that do not depend on μ have been absorbed into the constants α, α′, and α″. Thus, $p(\mu|\mathcal{D})$ is an exponential function of a quadratic function of μ, i.e., is again a normal density.
- Since this is true for any number of training samples, $p(\mu|\mathcal{D})$ remains normal as the number n of samples is increased, and $p(\mu|\mathcal{D})$ is said to be a reproducing density and $p(\mu)$ is said to be a conjugate prior.

If we write $p(\mu \mid \mathcal{D}) \sim N(\mu_n, \sigma_n^2)$ (2)

Identifying (1) and (2) yields:

$$\begin{cases} \dfrac{1}{\sigma_n^2} = \dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2} \\[2ex] \dfrac{\mu_n}{\sigma_n^2} = \dfrac{n}{\sigma_2}\,\bar{x}_n + \dfrac{\mu_0}{\sigma_0^2} \end{cases}$$

where $\bar{x}_n$ is the sample mean

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

We solve explicitly for μ_n and $\sigma^2{}_n$ and obtain

$$\boxed{\begin{aligned} \mu_n &= \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0 \\ \sigma_n^2 &= \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \end{aligned}}$$

● μ_n represents our best guess for μ after observing *n* samples, and $\sigma^2{}_n$ measures our uncertainty about this guess.

● $\sigma^2{}_n$ decreases monotonically with *n. E*ach additional observation decreases our uncertainty about the true value of μ.

● The relative balance between prior knowledge and empirical data is set by the ratio of σ^2 to $\sigma^2{}_0$, which is sometimes called the *dogmatism*.

- The posterior, p(μ|D), becomes more sharply peaked as *n* grows large. This is known as Bayesian learning.

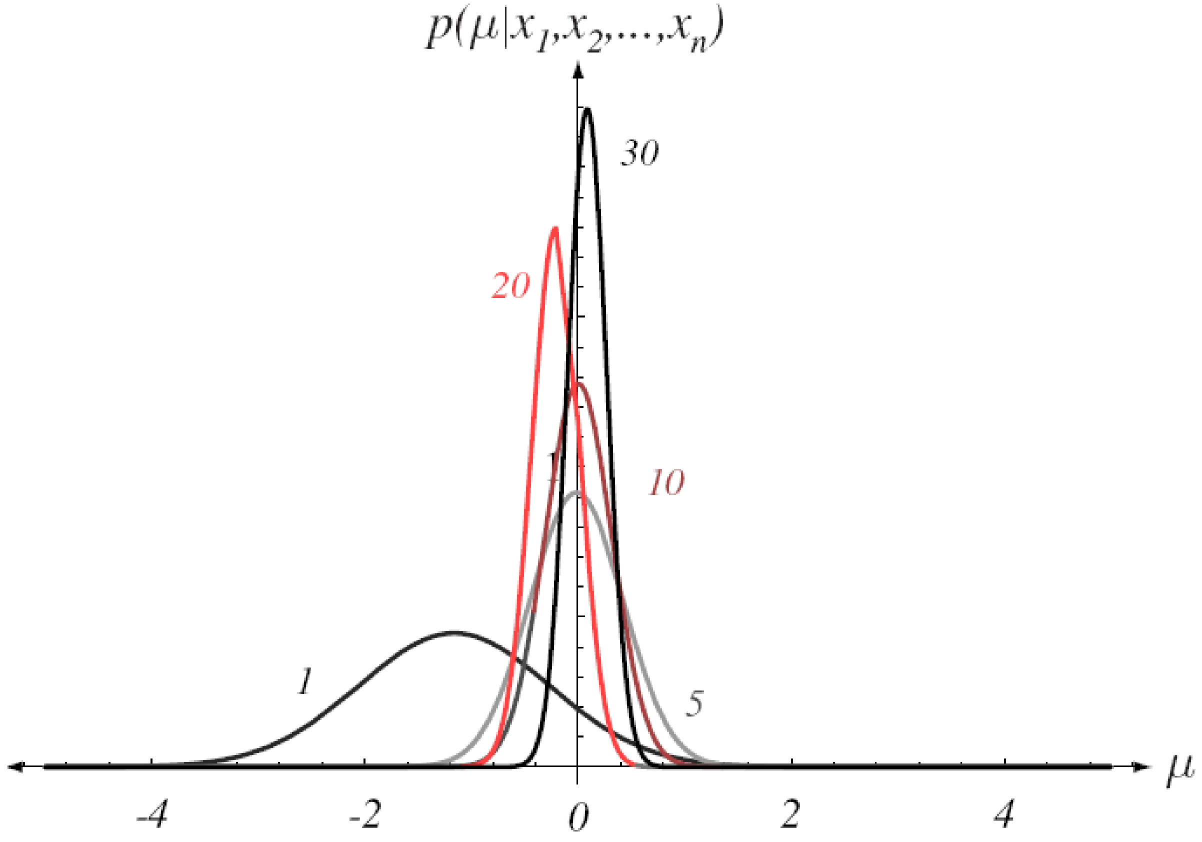


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one dimension. The posterior distribution estimates are labeled by the number of training samples used in the estimation.

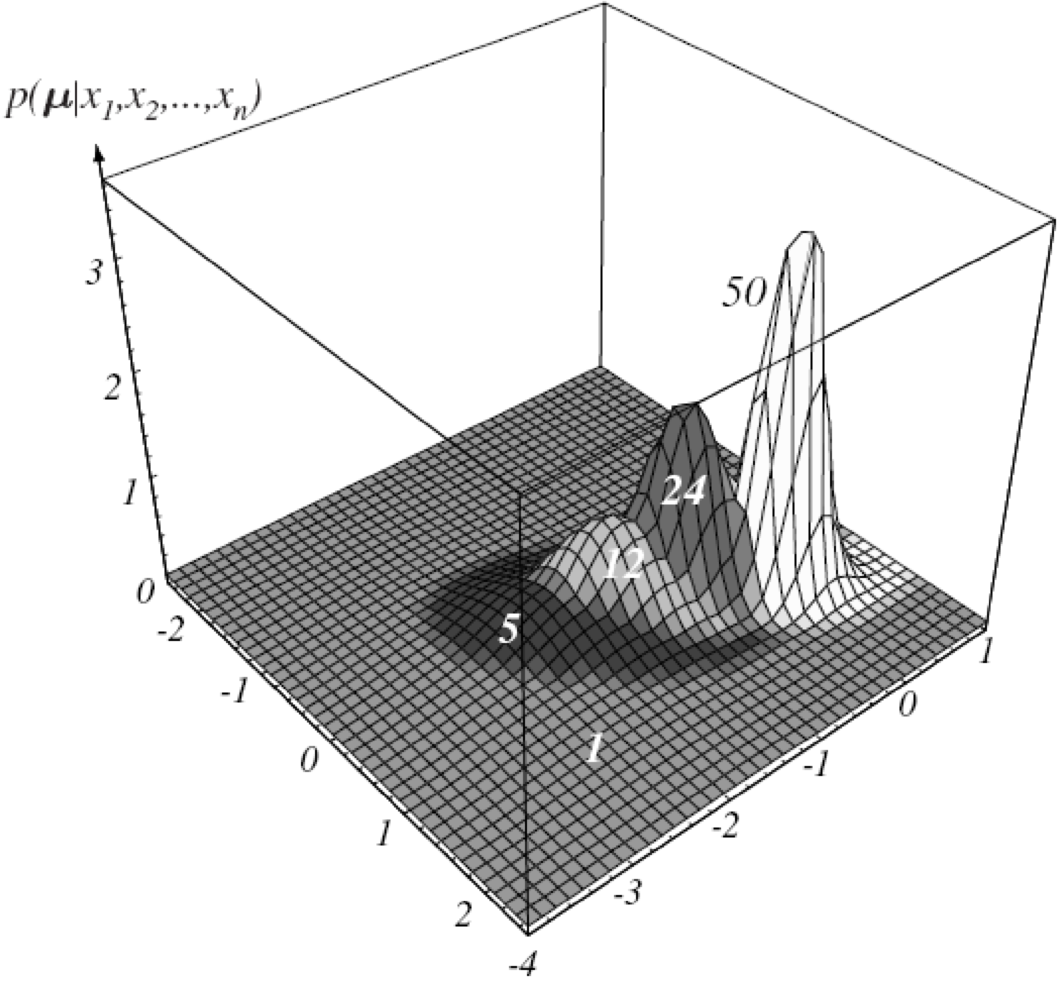


FIGURE 3.2. Bayesian learning of the mean of normal distributions in two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation.

- **The univariate case:** $p(x/\mathcal{D})$
  - $p(\mu/\mathcal{D})$ computed
  - $p(x/\mathcal{D})$ remains to be computed! ($p(x/\mathcal{D})$ is really $p(x/\omega_i, \mathcal{D}_i)$.)
  - From

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

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

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp[-\frac{1}{2}(\frac{\mu-\mu_n}{\sigma_n})^2]$$

We have

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

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

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

where

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

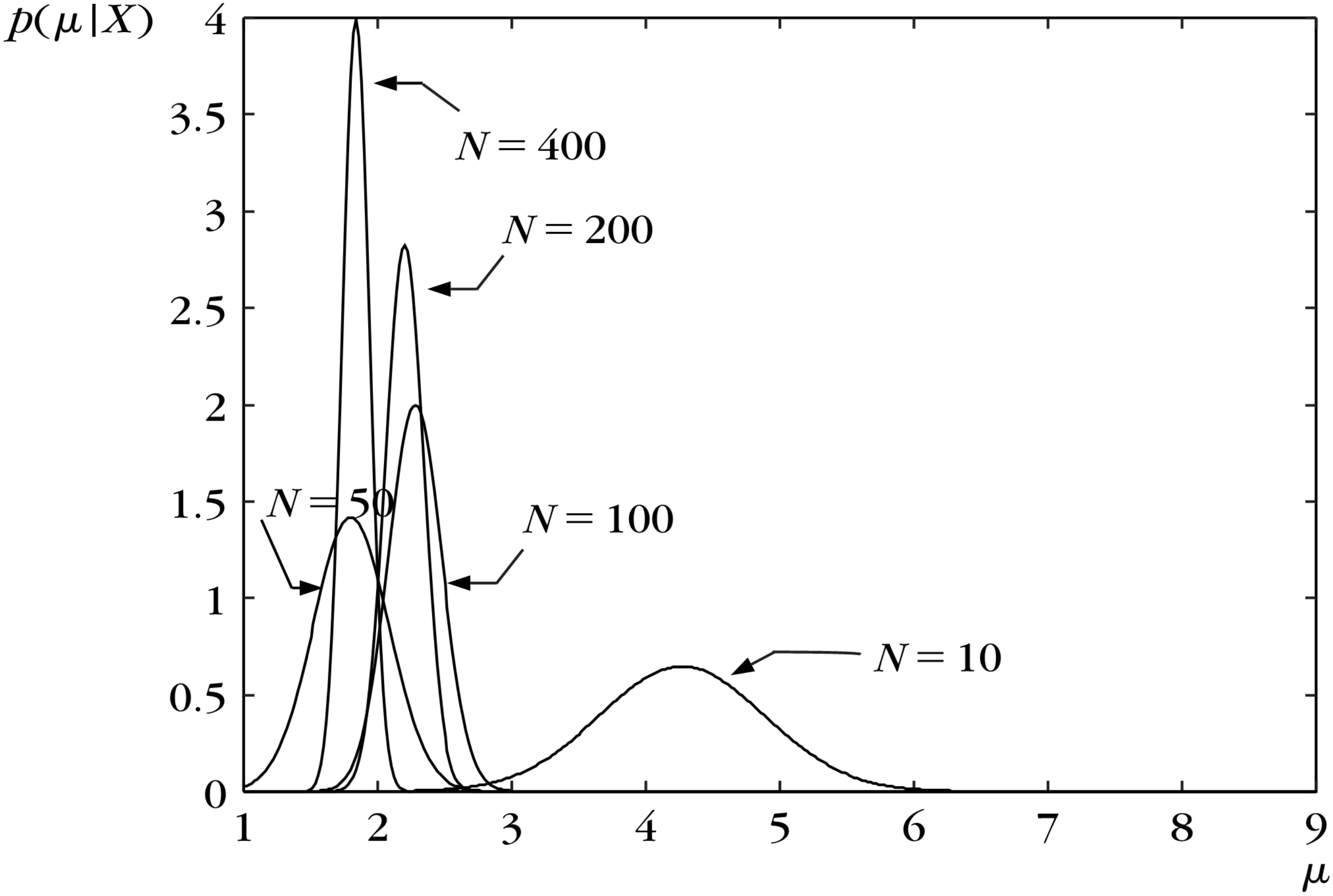
That is, as a function of x, $p(x/\mathcal{D})$ is proportional to $\exp[-(1/2)(x\text{-}\mu_n)^2/(\sigma^2+\sigma^2_n)]$, and hence $p(x/\mathcal{D})$ is normally distributed with mean μ_n and variance $\sigma^2+\sigma^2_n$:

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

In other words, to obtain the class-conditional density $p(x/\mathcal{D})$, whose parametric form is known to be $p(x/\mu) \sim N(\mu, \sigma^2)$, we merely replace μ by μ_n and σ^2 by $\sigma^2+\sigma^2_n$.

(Desired class-conditional density $p(x/D_j, \omega_j)$) Therefore: $p(x/D_j, \omega_j)$ together with $P(\omega_j)$ and using Bayes formula, we obtain the Bayesian classification rule:

$$\max_{\omega_j}\left[P(\omega_j \mid x,\mathcal{D})\right] \equiv \max_{\omega_j}\left[p(x \mid \omega_j,\mathcal{D}_j).P(\omega_j)\right]$$



A sequence of the posterior pdf estimates. As the number of training points increases, the posterior pdf becomes more spiky (the ambiguity decreases) and its center moves toward the true mean value of the data [Theo09].

The Multivariate Case

- The treatment of the multivariate case in which **Σ** is known but **μ** is not, is a direct generalization of the univariate case.

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

- where **Σ**, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\mu}_0$ are assumed to be known.
- After observing a set D of n independent samples $\mathbf{x}_1$, ..., $\mathbf{x}_n$, we use Bayes' formula to obtain

$$p(\boldsymbol{\mu}\,|\,D) = \alpha \prod_{k=1}^{n} p(\mathbf{x}_k\,|\,\boldsymbol{\mu})\,p(\boldsymbol{\mu})$$

$$= \alpha' \exp\left[-\frac{1}{2}(\boldsymbol{\mu}^{\mathrm{t}}(n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathrm{t}}(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}_k + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0))\right]$$

which has the form

$$p(\boldsymbol{\mu} \mid \mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^{\mathrm{t}} \Sigma_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

Thus, $p(\boldsymbol{\mu}/\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, and once again we have a reproducing density.

$$\begin{cases} \Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1} \\ \Sigma_n^{-1} \boldsymbol{\mu}_n = n\Sigma^{-1} \hat{\boldsymbol{\mu}}_n + \Sigma_0^{-1} \mu_0 \end{cases}$$

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

The solution of these equations for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_n$ is simplified by knowledge of the matrix identity

$$(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A},$$

After a little manipulation

$$\begin{cases} \boldsymbol{\mu}_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\hat{\boldsymbol{\mu}}_n + \frac{1}{n}\Sigma\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\boldsymbol{\mu}_0 \\ \Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\frac{1}{n}\Sigma \end{cases}$$

Linear combination of $\hat{\boldsymbol{\mu}}_n$ and $\boldsymbol{\mu}_0$

The proof that $p(\mathbf{x}/\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}+\boldsymbol{\Sigma}_n)$ can be obtained as before by performing the integration

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

Or, if $\mathbf{x}=\boldsymbol{\mu}+\mathbf{y}$ and $p(\boldsymbol{\mu}/\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ and $p(\mathbf{y}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Since the sum of two independent, normally distributed vectors is again a normally distributed vector whose mean is the sum of the means and whose covariance matrix is the sum of the covariance matrices we have

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

- Bayesian Parameter Estimation: General Theory

  - The Bayesian approach has been applied to compute $p(\mathbf{x}|\mathcal{D})$. It can be applied to any situation in which the unknown density can be parameterized:

    The basic assumptions are:

    - The form of $p(\mathbf{x}|\theta)$ 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 about θ is contained in a set $\mathcal{D}$ of n random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ that follows $p(\mathbf{x})$

The basic problem is:

“Compute the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$”

then “Derive $p(\mathbf{x}|\mathcal{D})$”

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

Using Bayes’ formula, we have:

$$p(\boldsymbol{\theta}\,|\,\mathcal{D}) = \frac{p(\mathcal{D}\,|\,\boldsymbol{\theta}).p(\boldsymbol{\theta})}{\int p(\mathcal{D}\,|\,\boldsymbol{\theta}).p(\boldsymbol{\theta})d\boldsymbol{\theta}} \qquad (50)$$

And by independence assumption:

$$p(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{k=n} p(\mathbf{x}_k\,|\,\boldsymbol{\theta})$$

Suppose that $p(\mathcal{D}/\boldsymbol{\theta})$ reaches a sharp peak at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$ If the prior density $p(\boldsymbol{\theta})$ is not zero at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$ and does not change much in the surrounding neighborhood, then $p(\boldsymbol{\theta}/\mathcal{D})$ also peaks at that point.

From (25) $$\boxed{p(\mathbf{x}\,|\,\mathcal{D})=\int p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D})d\boldsymbol{\theta}}$$

$p(\mathbf{x}/\mathcal{D})$ will be approx. $p(\mathbf{x}\,|\,\hat{\boldsymbol{\theta}})$, which is the ML result.

If the peak of $p(\mathcal{D}/\boldsymbol{\theta})$ is very sharp, then the influence of prior information on the uncertainty in the true value of $\boldsymbol{\theta}$ can be ignored.

However, the Bayes solution tells us how to use all of the available information to compute the desired density $p(\mathbf{x}|\mathcal{D})$.

Two questions remain: the difficulty of carrying out these Computations and the convergence of $p(\mathbf{x}/\mathcal{D})$ to $p(\mathbf{x})$.

Convergence: $\mathcal{D}^{\mathrm{n}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\mathrm{n}}\}.$

$$p(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{k=n} p(\mathbf{x}_k\,|\,\boldsymbol{\theta}) \longrightarrow p(\mathcal{D}^n\,|\,\boldsymbol{\theta}) = p(\mathbf{x}_n\,|\,\boldsymbol{\theta})\,p(\mathcal{D}^{n-1}\,|\,\boldsymbol{\theta})$$

$$\longrightarrow \quad p(\boldsymbol{\theta}\,|\,\mathcal{D}^n) = \frac{p(\mathbf{x}_n\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D}^{n-1})}{\int p(\mathbf{x}_n\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathcal{D}^{n-1})\,d\boldsymbol{\theta}} \qquad (53)$$

$$p(\boldsymbol{\theta}\,|\,\mathcal{D}^0) = p(\boldsymbol{\theta}) \longrightarrow p(\boldsymbol{\theta}),\; p(\boldsymbol{\theta}\,|\,\mathbf{x}_1),\; p(\boldsymbol{\theta}\,|\,\mathbf{x}_1, \mathbf{x}_2),\; \ldots$$

● This is called the Recursive Bayes Incremental or on-line Learning because we have a method for incrementally updating our estimates.

…Sufficient statistics

$$p(\boldsymbol{\theta} \mid \mathcal{D}^{n}) = \frac{p(\mathcal{D}^{n} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D}^{n})} = \frac{p(\mathbf{x}_{n} \mid \boldsymbol{\theta})\, p(\mathcal{D}^{n-1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D}^{n})}$$

$$= \frac{p(\mathbf{x}_{n} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})\, p(\mathcal{D}^{n-1})}{p(\mathcal{D}^{n})} = \frac{p(\mathbf{x}_{n} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{p(\mathbf{x}_{n})}$$

$$= \frac{p(\mathbf{x}_{n} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{\int p(\mathbf{x}_{n} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})\, d\boldsymbol{\theta}}$$

Example 1: Recursive Bayes learning

Suppose we believe our one-dimensional samples come from a uniform distribution

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

We know only that our parameter is bounded, $0 < \theta \leq 10$

$\mathcal{D} = \{4, 7, 2, 8\}$ selected randomly from the underlying distribution.

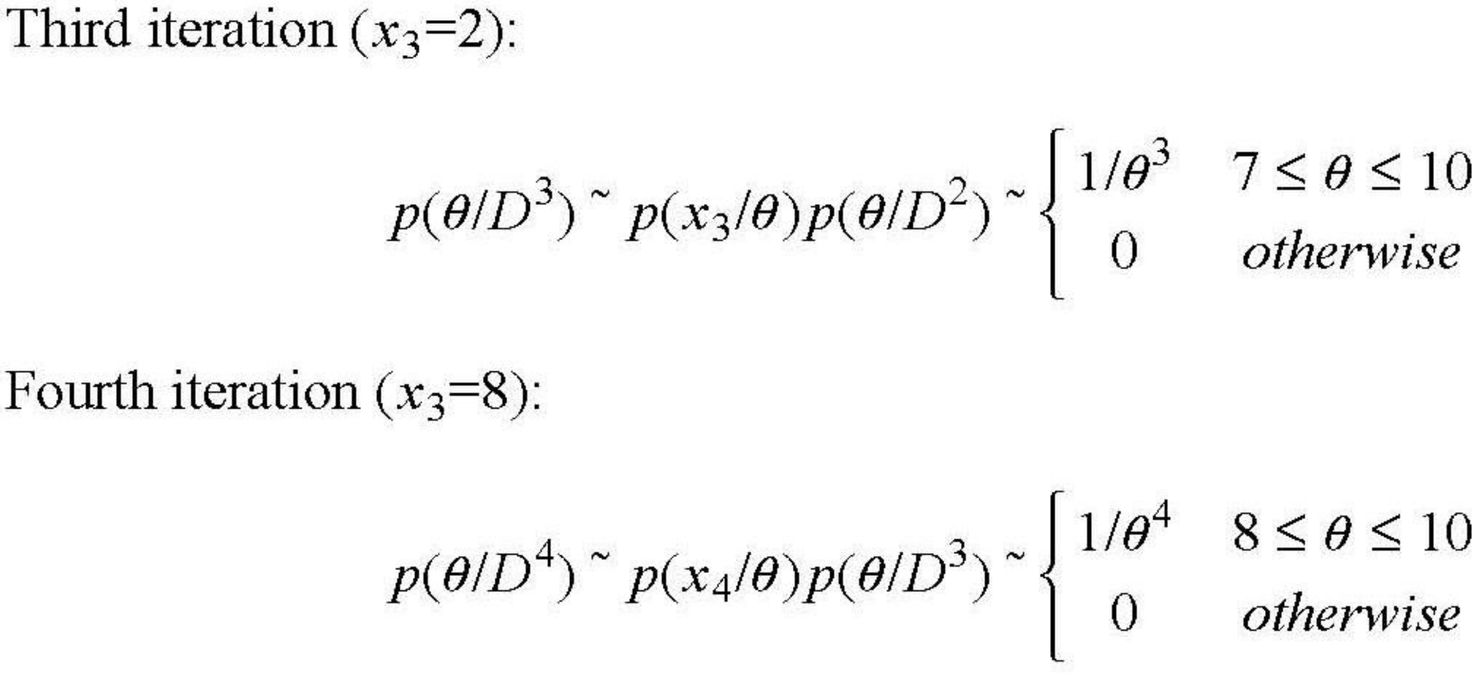
We have $p(\theta/D^0) = p(\theta) = U(0, 10)$.

The 1st data x_1 = 4 arrives

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

The 2nd data x_2 = 7 arrives

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



Third iteration (x_3=2):

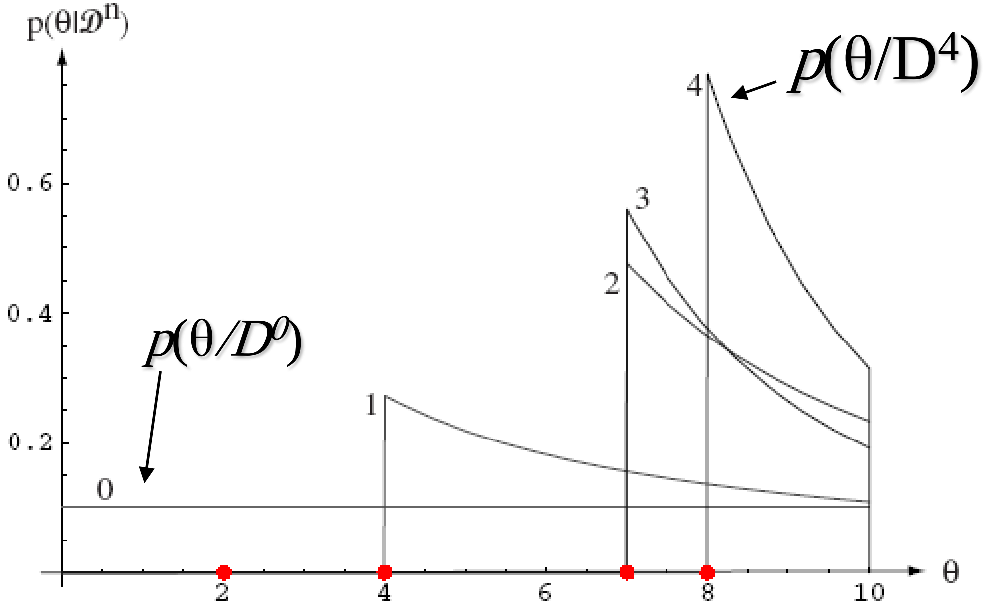
$$p(\theta/D^3) \sim p(x_3/\theta)p(\theta/D^2) \sim \begin{cases} 1/\theta^3 & 7 \leq \theta \leq 10 \\ 0 & otherwise \end{cases}$$

Fourth iteration (x_3=8):

$$p(\theta/D^4) \sim p(x_4/\theta)p(\theta/D^3) \sim \begin{cases} 1/\theta^4 & 8 \leq \theta \leq 10 \\ 0 & otherwise \end{cases}$$

So we have

$$p(\theta / D^n) \propto \frac{1}{\theta^n}, \quad for \quad \max_x [D^n] \leq \theta \leq 10$$

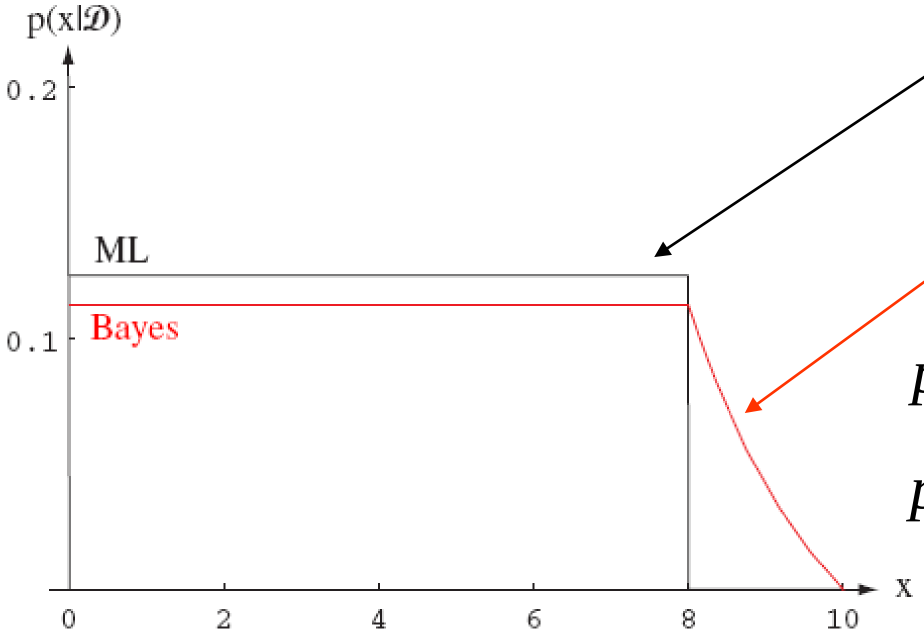


$p(\theta/D^4)$ peaks at $\hat{\theta} = 8$

$p(\theta/D^0)$

Given our full data set, the maximum likelihood solution here is clearly $\hat{\boldsymbol{\theta}} = 8$, and this implies a uniform $p(x/\mathcal{D}) \sim U(0, 8)$.

According to our Bayesian methodology, which requires the integration in Eq. 49, the density is uniform up to x=8, but has a tail at higher values — an indication that the influence of our prior $p(\theta)$ has not yet been swamped by the information in the training data.



ML estimate:

$$p(x/\hat{\theta}) \sim U(0,8)$$

Bayesian estimate:

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

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

Whereas the maximum likelihood approach estimates a *point* in $\boldsymbol{\theta}$ space, the Bayesian approach instead estimates a *distribution*. Technically speaking, then, we cannot directly compare these estimates. It is only when the second stage of inference is done — that is, we compute the distributions $p(x/\mathcal{D})$, as shown in the above figure — that the comparison is fair.

Identifiablity

- For most of the typically encountered probability densities $p(\mathbf{x}/\boldsymbol{\theta})$, the sequence of posterior densities does indeed converge to a delta function. This implies that with a large number of samples there is only one value for $\boldsymbol{\theta}$ that causes $p(\mathbf{x}/\boldsymbol{\theta})$ to fit the data, i.e., that $\boldsymbol{\theta}$ can be determined uniquely from $p(\mathbf{x}/\boldsymbol{\theta})$.
- When this is the case, $p(\mathbf{x}/\boldsymbol{\theta})$ is said to be *identifiable*.

- There are occasions, when more than one value of $\boldsymbol{\theta}$ may yield the same value for $p(\mathbf{x}/\boldsymbol{\theta})$. In such cases, $\boldsymbol{\theta}$ cannot be determined uniquely from $p(\mathbf{x}/\boldsymbol{\theta})$, and $p(\mathbf{x}/\mathcal{D}_n)$ will peak near all of the values of $\boldsymbol{\theta}$ that explain the data. Fortunately, this ambiguity is erased by the integration in Eq. 26, since $p(\mathbf{x}|\boldsymbol{\theta})$ is the same for all of these values of $\boldsymbol{\theta}$. Thus, $p(\mathbf{x}/\mathcal{D}_n)$ will typically converge to $p(\mathbf{x})$ whether or not $p(\mathbf{x}/\boldsymbol{\theta})$ is identifiable.

- Given a large number of samples, $p(\boldsymbol{\theta}/D^n)$ will have a very strong peak at $\hat{\boldsymbol{\theta}}$; in this case:

$$p(\mathbf{x}/D) \cong p(\mathbf{x}/\hat{\boldsymbol{\theta}})$$

- There are cases where $p(\boldsymbol{\theta}/D^n)$ contains more than one peaks (i.e., more than one $\boldsymbol{\theta}$ explains the data); in this case, the solution $p(\mathbf{x}/\boldsymbol{\theta})$ should be obtained by integration.

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

- In general, $p(\mathbf{x}/D^n)$ converges to $p(\mathbf{x}|\,\boldsymbol{\theta})$ whether or not having one maximum.

When do Maximum Likelihood and Bayes methods differ?

- Maximum likelihood and Bayes solutions are equivalent in the asymptotic limit of infinite training data.
- **Computational complexity**: maximum Likelihood methods are often to be preferred since they require merely differential calculus techniques or gradient search for $\hat{\boldsymbol{\theta}}$, rather than a possibly complex multidimensional integration needed in Bayesian estimation.

- **Interpretability**: In many cases the maximum likelihood solution will be easier to interpret and understand since it returns the single best model from the set the designer provided.
- In contrast Bayesian methods give a weighted average of models (parameters), often leading to solutions more complicated and harder to understand than those provided by the designer. The Bayesian approach reflects the remaining uncertainty in the possible models.
- **The prior information:** such as in the form of the underlying distribution $p(\mathbf{x}|\boldsymbol{\theta})$.

- A maximum likelihood solution $p(\mathbf{x}\,|\,\hat{\boldsymbol{\theta}})$ must of course be of the assumed parametric form; not so for the Bayesian solution.
- In Example 1, the Bayes solution was not of the parametric form originally assumed, i.e., a uniform $p(\mathbf{x}|\mathcal{D})$. In general, through their use of the full $p(\boldsymbol{\theta}|\mathcal{D})$ distribution Bayesian methods use more of the information brought to the problem than do maximum likelihood methods.
- If the information is reliable, Bayes methods can be expected to give better results.
- Further, general Bayesian methods with a “flat” or uniform prior (i.e., where no prior information is explicitly imposed) are equivalent to maximum likelihood methods.

- When $p(\boldsymbol{\theta}/\mathcal{D})$ is broad, or asymmetric around $\hat{\boldsymbol{\theta}}$, the methods are quite likely to yield $p(\mathbf{x}/\mathcal{D})$ distributions that differ from one another.

Sources of classification error

- **Bayes or indistinguishability error**: the error due to overlapping densities $p(\mathbf{x}/\omega_i)$ for different values of i. This error is an inherent property of the problem and can never be eliminated.
- **Model error**: the error due to having an incorrect model. The model error in ML and Bayes methods rarely differ.
- **Estimation error**: the error arising from the fact that the parameters are estimated from a finite sample. Can be reduced by increasing the training data.

* Noninformative Priors and Invariance

- The information about the prior is based on the designer's knowledge of the problem domain.

- We expect the prior distributions to be "translation and scale invariance" – they should not depend on the actual value of the parameter.

- A prior that satisfies this property is referred to as a "noninformative prior":

  - The Bayesian approach remains applicable even when little or no prior information is available.

  - Such situations can be handled by choosing a prior density giving equal weight to all possible values of θ.

  - Priors that seemingly impart no prior preference, the so-called noninformative priors, also arise when the prior is required to be invariant under certain transformations.

  - Frequently, the desire to treat all possible values of θ equitably leads to priors with infinite mass. Such noninformative priors are called improper priors.

Example of Noninformative Priors

- For example, if we assume the prior distribution of a mean of a continuous random variable is independent of the choice of the origin, the only prior that could satisfy this is a uniform distribution (which isn’t possible).

- Fisher argued that *Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge*.

- If we have no information about θ we also have no information about for example $1/\theta$ but a uniform prior on θ does not correspond to a uniform prior for $1/\theta$.

*Sufficient Statistics

- Practically the direct computation and tabulation of $p(\mathcal{D}|\boldsymbol{\theta})$ or $p(\boldsymbol{\theta}|\mathcal{D})$ is very difficult.
- An analytic and computationally feasible maximum likelihood solution lies in being able to find a parametric form for $p(\mathbf{x}|\boldsymbol{\theta})$ that on the one hand matches the characteristics of the problem and on the other hand allows a reasonably tractable solution.
- Learning the parameters of a multivariate Gaussian density → the sample mean and sample covariance.
- What about other distributions?

- A *sufficient statistic* is a (possibly vector-valued) function $\mathbf{s}$ of the samples $\mathcal{D}$ that contains all of the information relevant to estimating some parameter $\boldsymbol{\theta}$. → $p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$. → Treating $\boldsymbol{\theta}$ as a random variable, limiting the definition to a Bayesian domain.
- To avoid such a limitation, the conventional definition is as follows: A statistic $\mathbf{s}$ is said to be *sufficient* for $\boldsymbol{\theta}$ if $p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$. If we think of $\boldsymbol{\theta}$ as a random variable, we can write

$$p(\boldsymbol{\theta}\,|\,\mathbf{s},\mathcal{D}) = \frac{p(\mathcal{D}\,|\,\mathbf{s},\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\mathbf{s})}{p(\mathcal{D}\,|\,\mathbf{s})} = p(\boldsymbol{\theta}\,|\,\mathbf{s})$$

Sufficient statistics are summary statistics of a dataset which are such that the distribution of the data is independent of the parameters of the underlying distribution when conditioned on the statistic.

- It becomes evident that $p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$ if $\mathbf{s}$ is sufficient for $\boldsymbol{\theta}$. Conversely, if $\mathbf{s}$ is a statistic for which $p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$, and if $p(\boldsymbol{\theta}|\mathbf{s})\neq 0$, it is easy to show that $p(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$.
- For a Gaussian distribution the sample mean and covariance, taken together, represent a sufficient statistic for the true mean and covariance; if these are known, all other statistics such as the mode, range, higher-order moments, number of data points, etc., are superfluous when estimating the true mean and covariance.

- Any function of the samples $\mathcal{D}$ is a *statistic*.
- A *sufficient statistic* is a function $\mathbf{s}=\boldsymbol{\varphi}(\mathcal{D})$ of the samples $\mathcal{D}$ that contains all the information necessary for estimating the parameters $\boldsymbol{\theta}$.
- Using sufficient statistics, we can make the computation of $p(\mathcal{D}|\boldsymbol{\theta})$ or $p(\boldsymbol{\theta}|\mathcal{D})$ much **less expensive**.

- A fundamental theorem concerning sufficient statistics is the *Factorization Theorem*, which states that **s** is sufficient for **θ** if and only if $p(\mathcal{D}/\boldsymbol{\theta})$ can be factored into the product of two functions, one depending only on **s** and **θ**, and the other depending only on the training samples.
- The virtue of the Factorization Theorem is that it allows us to shift our attention from the rather complicated density $p(\mathcal{D}/\mathbf{s}, \boldsymbol{\theta})$, used to define a sufficient statistic, to the simpler function

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

Factorization

- **Theorem 3.1(Factorization)** *A statistic* $\mathbf{s}$ *is sufficient for* $\boldsymbol{\theta}$ *if and only if the probability* $P(\mathcal{D}|\boldsymbol{\theta})$ *can be written as the product*

  $$P(\mathcal{D}|\boldsymbol{\theta}) = g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D}),$$

  for some function $h(\cdot)$.
- There are trivial ways of constructing sufficient statistics. For example we can define $\mathbf{s}$ to be a vector whose components are the n samples themselves: $\mathbf{x}_1, ..., \mathbf{x}_n$. In that case $g(\mathbf{s}, \boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})$ and $h(\mathcal{D}) = 1$.
- The factoring of $p(\mathcal{D}|\boldsymbol{\theta})$ into $g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D})$ is not unique.

- If $f(\mathbf{s})$ is any function of $\mathbf{s}$, then $g'(\mathbf{s}, \boldsymbol{\theta}) = f(\mathbf{s})g(\mathbf{s}, \boldsymbol{\theta})$ and $h'(\mathcal{D}) = h(\mathcal{D})/f(\mathbf{s})$ are equivalent factors. This kind of ambiguity can be eliminated by defining the *kernel density.*

$$\bar{g}(\mathbf{s} \mid \boldsymbol{\theta}) = \frac{g(\mathbf{s}, \boldsymbol{\theta})}{\int g(\mathbf{s}, \boldsymbol{\theta}) d\boldsymbol{\theta}} \qquad (63)$$

  which is invariant to this kind of scaling.

- Significance: most practical applications of parameter estimation involve simple sufficient statistics and simple kernel densities.

- It can be shown that for any classification rule, we can find another based solely on sufficient statistics that has equal or better performance.

- So we can reduce an extremely large data set down to a few numbers — the sufficient statistics.
- In the case of maximum likelihood estimation, when searching for a value of $\boldsymbol{\theta}$ that maximizes $p(\mathcal{D}/\boldsymbol{\theta}) = g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D})$, we can restrict our attention to $g(\mathbf{s}, \boldsymbol{\theta})$.
- In ML, the normalization provided by the kernel density is of no particular value unless $\bar{g}(\mathbf{s}, \boldsymbol{\theta})$ is simpler than g($\mathbf{s}$, $\boldsymbol{\theta}$).
- The significance of the kernel density is revealed however in the Bayesian case. If we substitute $p(\mathcal{D}/\boldsymbol{\theta}) = g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D})$ in Eq. 50, we obtain

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}).p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta}).p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad \Rightarrow \quad p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{g(\mathbf{s} \mid \boldsymbol{\theta}).p(\boldsymbol{\theta})}{\int g(\mathbf{s} \mid \boldsymbol{\theta}).p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad (64)$$

- If our prior knowledge of $\boldsymbol{\theta}$ is very vague, $p(\boldsymbol{\theta})$ will tend to be uniform, or changing very slowly as a function of $\boldsymbol{\theta}$. So $p(\boldsymbol{\theta}/\mathcal{D})$ is approximately the same as the kernel density.
- Roughly speaking, the kernel density is the posterior distribution of the parameter vector when the prior distribution is uniform.
- When $p(\mathbf{x}/\boldsymbol{\theta})$ is identifiable and when the number of samples is large, $g(\mathbf{s}, \boldsymbol{\theta})$ usually peaks sharply at some value $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.
- If the a priori density $p(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and if $p(\hat{\boldsymbol{\theta}})$ is not zero, $p(\boldsymbol{\theta}/\mathcal{D})$ will approach the kernel density $\bar{g}(\mathbf{s}, \boldsymbol{\theta})$.

Sufficient Statistics and the Exponential Family

- How Factorization Theorem can be used to obtain sufficient statistics? Consider the familiar *d*-dimensional normal case with fixed covariance but unknown mean, i.e., $p(\mathbf{x}/\boldsymbol{\theta}) \sim N(\boldsymbol{\theta},\boldsymbol{\Sigma})$. Here we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\theta})^t\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\theta})\right]$$

$$= \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k + \mathbf{x}_k^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k)\right]$$

$$= \exp\left[-\frac{n}{2}\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{n}\mathbf{x}_k\right)\right]$$

$$\times\left(\frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\mathbf{x}_k^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k\right]\right)$$

- This factoring isolates the **θ** dependence of $p(\mathcal{D}|\boldsymbol{\theta})$ in the first term, and hence from the Factorization Theorem we conclude that $\sum_{k=1}^{n}\mathbf{x}_k$ is sufficient for $\boldsymbol{\theta}$.
- The sample mean $\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$ is also sufficient for $\boldsymbol{\theta}$. Using this statistic, we can write

$$g(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\theta}) = \exp\left[-\frac{n}{2}\left(\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^t\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n\right)\right].$$

- From using Eq. 63, or by completing the square, we can obtain the kernel density:

$$\bar{g}(\hat{\boldsymbol{\mu}}_n,\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}\left|\frac{1}{n}\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\frac{1}{2}\left((\boldsymbol{\theta}-\hat{\boldsymbol{\mu}}_n)^t\left(\frac{1}{n}\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\theta}-\hat{\boldsymbol{\mu}}_n)\right)\right]$$

- These results make it immediately clear that $\hat{\boldsymbol{\mu}}_n$ is the maximum likelihood estimate for $\boldsymbol{\theta}$. The Bayesian posterior density can be obtained from $\bar{g}(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\theta})$ by performing the integration indicated in Eq. 64. If the a priori density is essentially uniform, $p(\boldsymbol{\theta} \mid \mathcal{D}) = \bar{g}(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\theta})$.
- The same general approach can be used to find sufficient statistics for other density functions like exponential, Rayleigh, Poisson, and many other familiar distributions. They can all be written in the form

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \alpha(\mathbf{x}) \exp[\mathbf{a}(\theta) + \mathbf{b}(\theta)^{\mathbf{t}} \mathbf{c}(\theta)] \qquad (69)$$

- If we multiply *n* terms of the form in Eq. 69 we find

$$p(D\,|\,\boldsymbol{\theta}) = \exp\left[n\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t \sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{c}(\mathbf{x}_k)\right] \prod_{k=1}^{n} \alpha\left(\mathbf{x}_k\right) = g(\mathsf{s}, \boldsymbol{\theta}) h(D)$$

where we can take

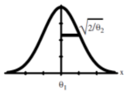
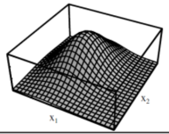
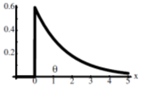
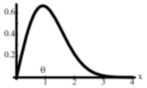
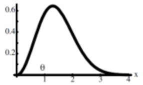
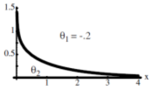
$$\boxed{\mathbf{s} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{c}(\mathbf{x}_k)}$$

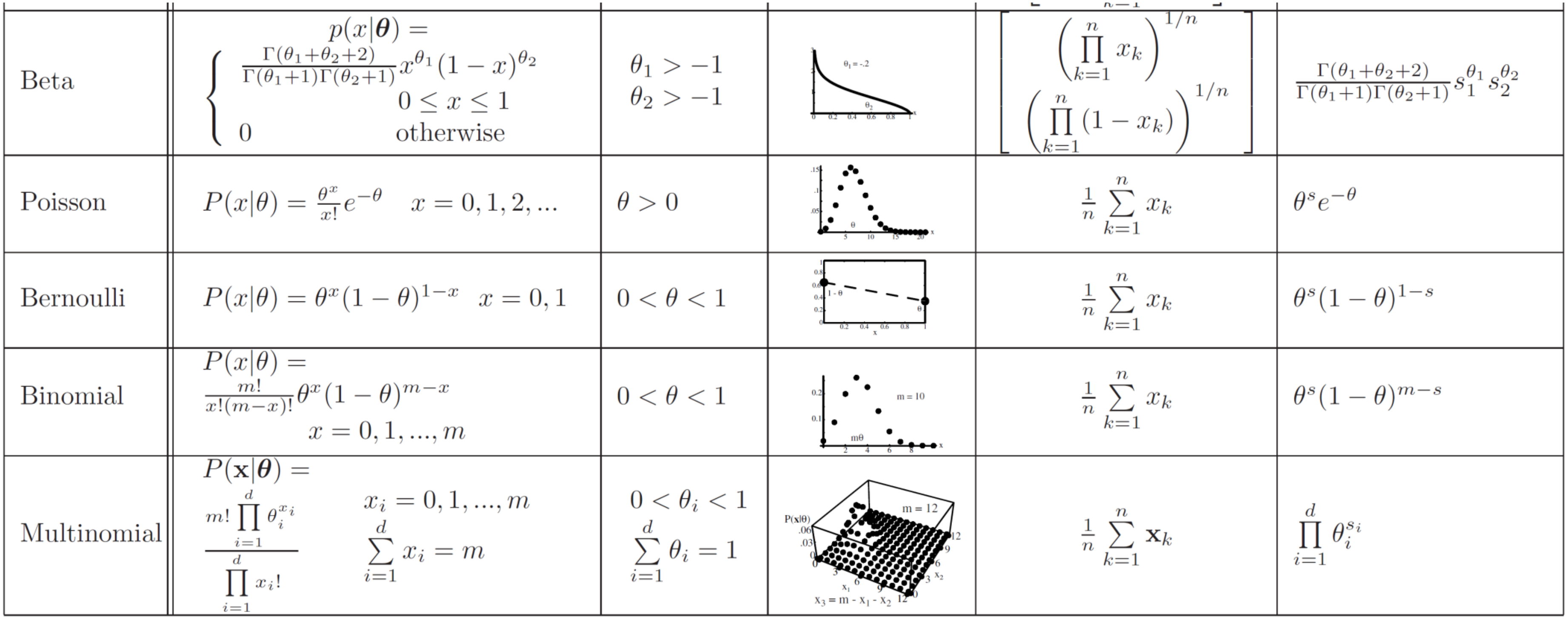
$$g(\mathbf{s}, \boldsymbol{\theta}) = \exp\left[n\ \ \mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t \mathbf{s}\ \ \right]$$

and

$$h(D) = \prod_{k=1}^{n} \alpha(\mathbf{x}_k)$$

Table 3.1:Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		$\mathbf{s}$	$[g(\mathbf{s},\boldsymbol{\theta})]^{1/n}$
Normal	$p(x\mid\boldsymbol{\theta}) =$ $\sqrt{\frac{\theta_2}{2\pi}}e^{-(1/2)\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n}\sum\limits_{k=1}^{n} x_k \\ \frac{1}{n}\sum\limits_{k=1}^{n} x_k^2 \end{bmatrix}$	$\sqrt{\theta_2}e^{-\frac{1}{2}\theta_2(s_2-2\theta_1 s_1+\theta_1^2)}$
Multi-variate Normal	$p(\mathbf{x}\mid\boldsymbol{\theta}) =$ $\frac{\lvert\boldsymbol{\Theta}_2\rvert^{1/2}}{(2\pi)^{d/2}}e^{-(1/2)(\mathbf{x}-\boldsymbol{\theta}_1)^t\boldsymbol{\Theta}_2(\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\Theta}_2$ positive definite		$\begin{bmatrix} \frac{1}{n}\sum\limits_{k=1}^{n} \mathbf{x}_k \\ \frac{1}{n}\sum\limits_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t \end{bmatrix}$	$\lvert\boldsymbol{\Theta}_2\rvert^{1/2}e^{-\frac{1}{2}[\mathrm{tr}\boldsymbol{\Theta}_2\mathbf{s}_2 -2\boldsymbol{\theta}_1^t\boldsymbol{\Theta}_2\mathbf{s}_1+\boldsymbol{\theta}_1^t\boldsymbol{\Theta}_2\boldsymbol{\theta}_1]}$
Exponential	$p(x\mid\theta) =$ $\begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k$	$\theta e^{-\theta s}$
Rayleigh	$p(x\mid\theta) =$ $\begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k^2$	$\theta e^{-\theta s}$
Maxwell	$p(x\mid\theta) =$ $\begin{cases} \frac{4}{\sqrt{\pi}}\theta^{3/2}x^2e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k^2$	$\theta^{3/2}e^{-\theta s}$
Gamma	$p(x\mid\boldsymbol{\theta}) =$ $\begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)}x^{\theta_1}e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\begin{bmatrix} \left(\prod\limits_{k=1}^{n} x_k\right)^{1/n} \\ \frac{1}{n}\sum\limits_{k=1}^{n} x_k \end{bmatrix}$	$\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)}s_1^{\theta_1}e^{-\theta_2 s_2}$



Beta	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1}(1-x)^{\theta_2} \\ \qquad 0 \le x \le 1 \\ 0 \qquad \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > -1$		$\begin{bmatrix} \left(\prod\limits_{k=1}^{n} x_k\right)^{1/n} \\ \left(\prod\limits_{k=1}^{n} (1-x_k)\right)^{1/n} \end{bmatrix}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$
Poisson	$P(x\|\theta) = \frac{\theta^x}{x!}e^{-\theta} \quad x = 0, 1, 2, \ldots$	$\theta > 0$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k$	$\theta^s e^{-\theta}$
Bernoulli	$P(x\|\theta) = \theta^x(1-\theta)^{1-x} \quad x = 0, 1$	$0 < \theta < 1$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k$	$\theta^s(1-\theta)^{1-s}$
Binomial	$P(x\|\theta) =$ $\frac{m!}{x!(m-x)!}\theta^x(1-\theta)^{m-x}$ $x = 0, 1, ..., m$	$0 < \theta < 1$		$\frac{1}{n}\sum\limits_{k=1}^{n} x_k$	$\theta^s(1-\theta)^{m-s}$
Multinomial	$P(\mathbf{x}\|\boldsymbol{\theta}) =$ $\frac{m!\prod\limits_{i=1}^{d}\theta_i^{x_i}}{\prod\limits_{i=1}^{d} x_i!}$ $x_i = 0, 1, ..., m$ $\sum\limits_{i=1}^{d} x_i = m$	$0 < \theta_i < 1$ $\sum\limits_{i=1}^{d} \theta_i = 1$		$\frac{1}{n}\sum\limits_{k=1}^{n} \mathbf{x}_k$	$\prod\limits_{i=1}^{d} \theta_i^{s_i}$

Certain probability distributions do not have sufficient statistics (e.g., Cauchy)