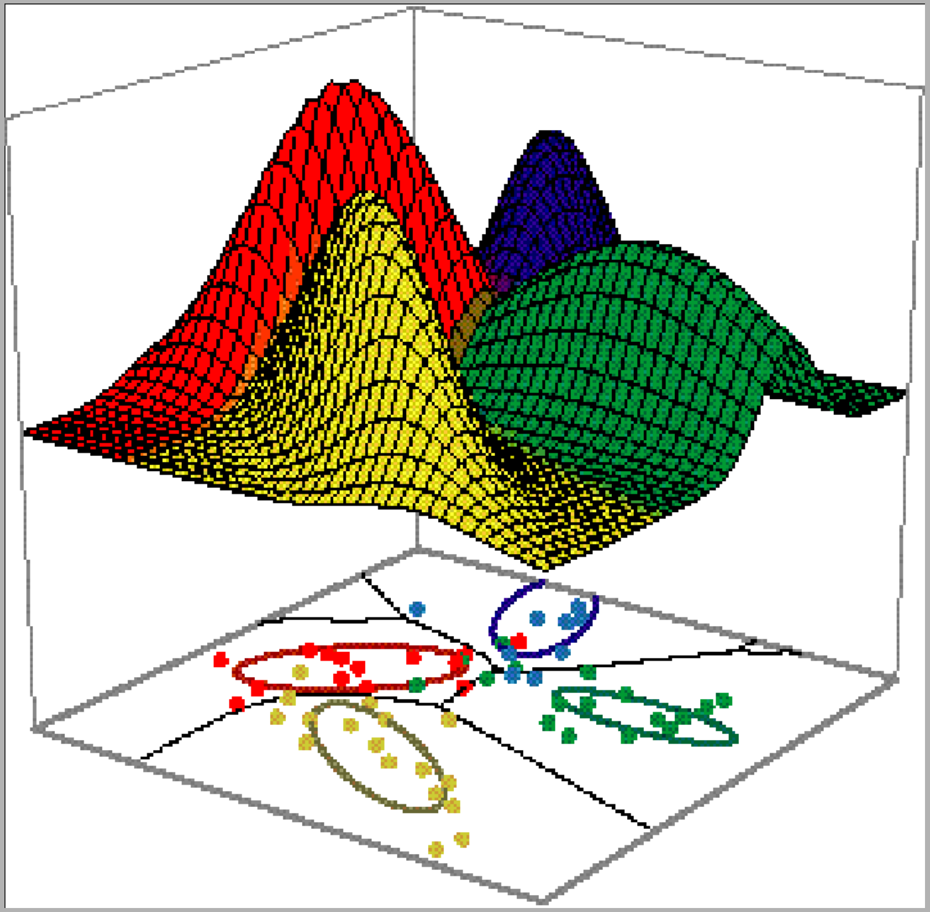


Pattern Recognition (Αναγνώριση Προτύπων)

Parameter Estimation

(Εκτίμηση Παραμέτρων)

Panos Trahanias

UNIVERSITY OF CRETE
DEPARTMENT of COMPUTER SCIENCE

Parameter Estimation

- Bayesian classifier cannot be employed when the probability density functions and prior probabilities are not known, i.e. $p(\mathbf{x}/\omega_i)$ & $P(\omega_i)$.
- Distributions can be estimated when appropriate data are available ➡ Hard Task!
- If the distribution shape is known, e.g. normal/gaussian, but not its parameters, e.g. mean and variance, then the problem is formulated as one of parameter estimation.

Parameter Estimation

- Two basic approaches:
  a. Maximum Likelihood Estimation (Εκτίμηση Μέγιστης Πιθανοφάνειας)
  b. Bayesian Parameter Estimation (Εκτίμηση παραμέτρων κατά Bayes)

Maximum Likelihood Estimation

- Assume that we take random samples from a given distribution with unknown parameters. Let us use vector ***θ*** to denote the set of parameters**.**

- If for example we know that the distribution is gaussian but we do not know its mean and variance, then:

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mu_1, \ldots, \mu_d, \sigma_1^2, \ldots, \sigma_d^2, \mathrm{cov}(x_m, x_n); m, n = 1, \ldots d; m \succ n)$$

$d + \frac{d(d+1)}{2}$ parameters

MLE Task

- For each class, estimate vector $\boldsymbol{\theta}$ using a training set $D^n=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ that includes *n* independent samples (i.i.d):

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

Likelihood of θ with respect of the sample set D^n

- **Maximum Likelihood Estimation** of $\boldsymbol{\theta}$ corresponds to that value that maximizes the above function.

- **Intuitively**, it corresponds to the value of $\boldsymbol{\theta}$ that «agrees/interprets» in the best possible way with the samples.

Likelihood Function

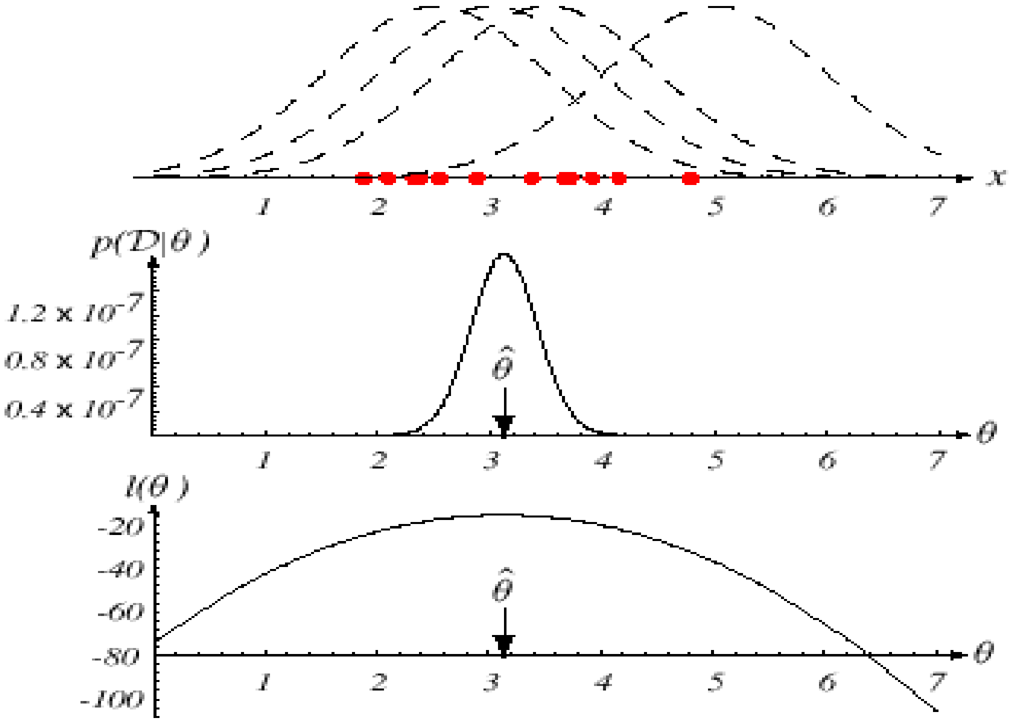


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Log-Likelihood

- ***θ*** that maximizes the likelihood function, maximizes also its log, that is more convenient to use:

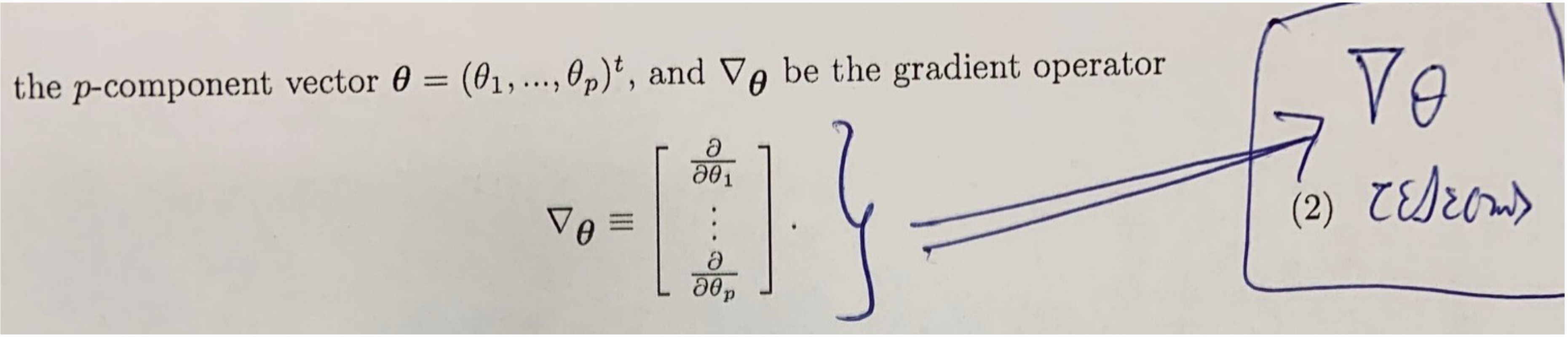
$$l(\boldsymbol{\theta}) \equiv \ln p(D^n \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

θ that maximizes the above function can be obtained by setting its derivative over ***θ*** equal to zero, and solving for ***θ***.

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

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = \mathbf{0}$$

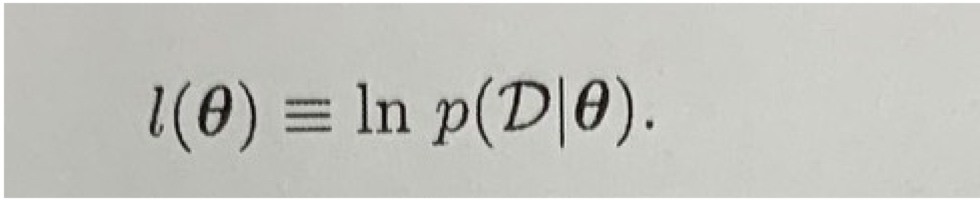
Log-Likelihood



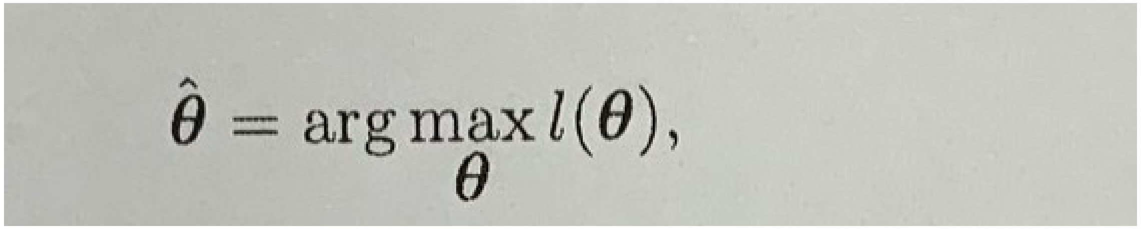
the p-component vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$, and $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

$$\left.\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}. \right\} \longrightarrow \nabla\theta$$

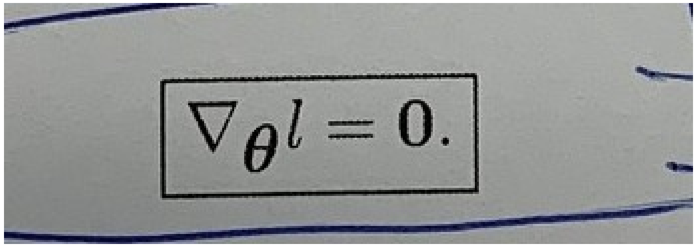
(2) τελεστής



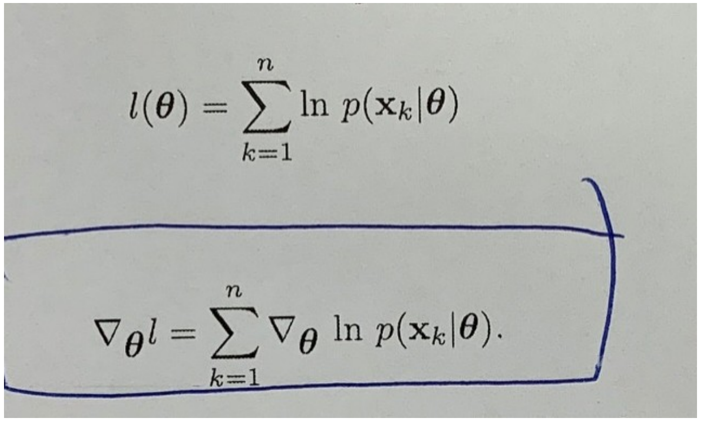
$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta}).$$



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



$$\boxed{\nabla_{\boldsymbol{\theta}} l = 0.}$$



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

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

Special Cases – Normal Distribution

- Normal distribution with unknown mean **μ**
  - ○ $\mathrm{p}(\mathbf{x}_{\kappa} \mid \boldsymbol{\mu}) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

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

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

  - ○ Maximum likelihood estimator of $\boldsymbol{\mu}$ satisfies:

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

  - ○ Left multiplying by $\boldsymbol{\Sigma}$ we obtain:

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

Accordingly, it is the arithmetic mean of the training samples!

Special Cases – Normal Distribution

- Normal distribution with unknown mean *μ* and variance *σ*²:
  - $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$

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

$$\nabla_{\boldsymbol{\theta}} l = \begin{pmatrix} \dfrac{\partial}{\partial\theta_1}(\ln P(x_k \mid \boldsymbol{\theta})) \\ \dfrac{\partial}{\partial\theta_2}(\ln P(x_k \mid \boldsymbol{\theta})) \end{pmatrix} = \mathbf{0}$$

$$\begin{cases} \dfrac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

Special Cases – Normal Distribution

- ➢ Taking into consideration all samples:

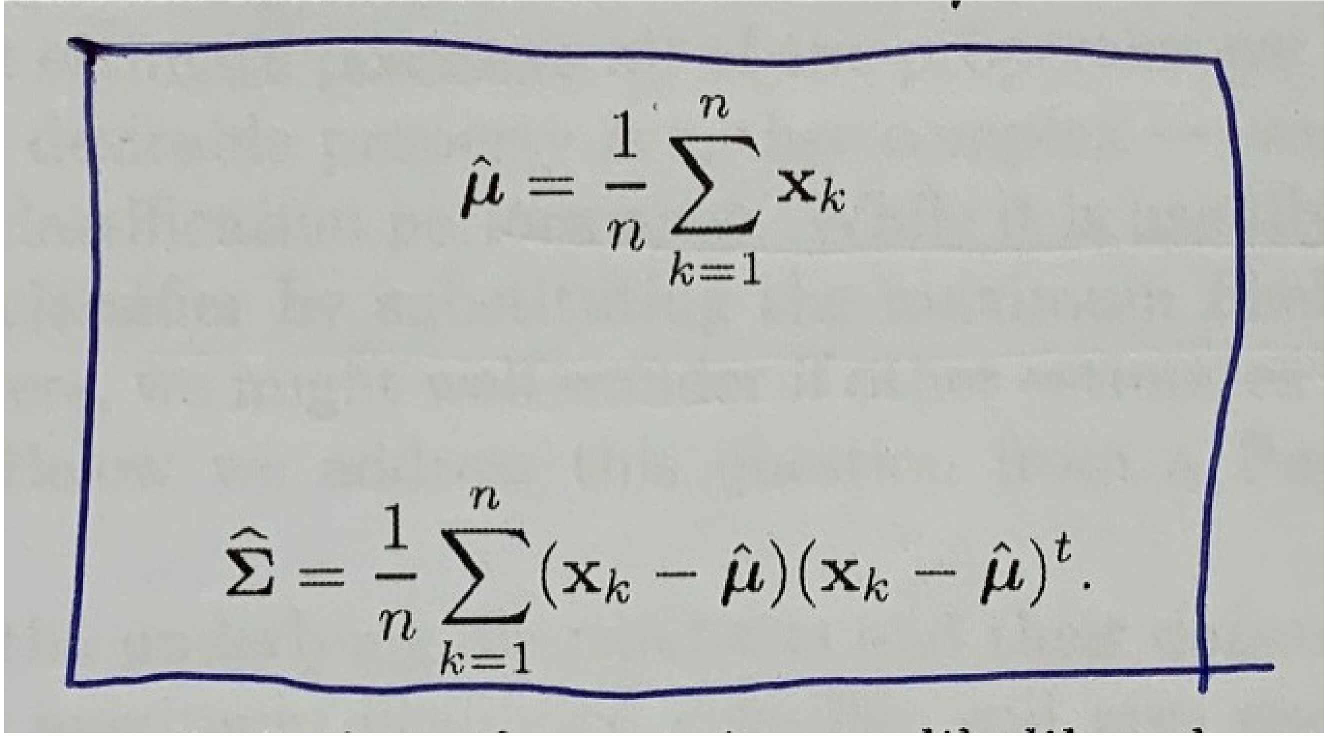
$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

- ➢ From (1) and (2), we obtain:

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

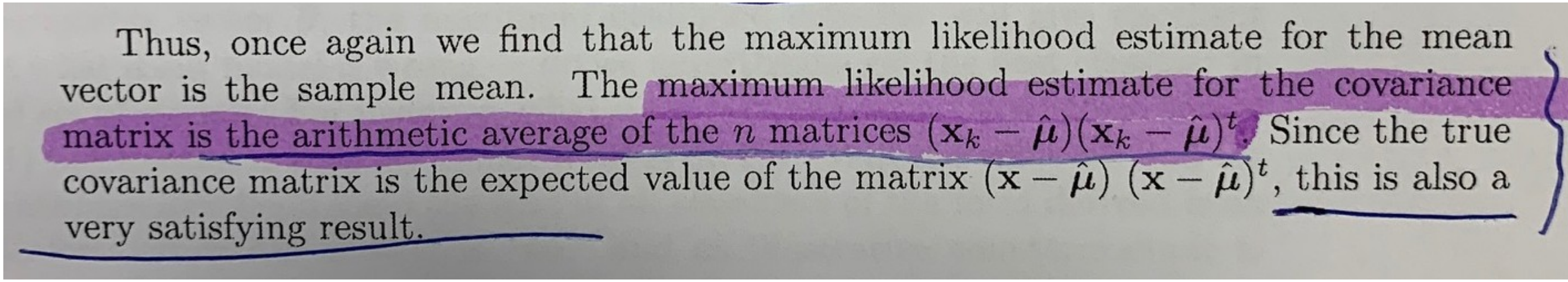
Special Cases – Normal Distribution

The case of multidimensional gaussian distribution:



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

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



Thus, once again we find that the maximum likelihood estimate for the mean vector is the sample mean. The maximum likelihood estimate for the covariance matrix is the arithmetic average of the n matrices $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$. Since the true covariance matrix is the expected value of the matrix $(\mathbf{x} - \hat{\boldsymbol{\mu}})\,(\mathbf{x} - \hat{\boldsymbol{\mu}})^t$, this is also a very satisfying result.

Bayesian Estimator

- In contrast to MLE, where we assumed that the unknown parameters have constant values, the Bayesian estimator (BE) assumes that the unknown parameters are random variables and follow an priori known p.d.f
- Therefore, BE estimates a distribution of the values of $\boldsymbol{\theta}$ and not the values themselves. BE provides more information, but is often difficult to calculate.
- The existence of training data allows the conversion of prior information to posterior p.d.f. ➡ phenomenon of learning (Bayesian learning) where each new observation refines the posterior probability.

Bayesian Estimation

- Bayesian Learning for pattern classification problems.

  - The computation of the posterior p.d.f. is the basis of Bayesian classification.
  - Goal: Computation of $P(\omega_i \mid \boldsymbol{x}, D)$ given the set of training samples $D=\{D_1,\ldots,D_c\}$, where the samples in the set D_j correspond to the class j, $j=1,\ldots,c$.
  - For each new, unclassified sample, $\boldsymbol{x}$, the Bayes rule gives:

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

Bayesian Estimation

➢ We assume that

- ○ The priors $P(\omega_i)$ are known, thus $P(\omega_i/D) = P(\omega_i)$.
- ○ Only the samples of the set D_i hold information about the p.d.f. $p(\mathbf{x}/\omega_i,D_i)$ ➡ *c independent problems of estimation of* $p(\mathbf{x}/\omega_i,D_i)$ *arise, which can also be written as* $p(\mathbf{x}/D_i)$.

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

We want to estimate this function (also written as $p(\mathbf{x}/D_i)$)

General Methodology of Bayesian Estimation

- For each class, we know the form of the p.d.f. $p(\boldsymbol{x}|\boldsymbol{\theta})$, but the value of the parameter vector $\boldsymbol{\theta}$ is unknown.
- We have some initial knowledge of $\boldsymbol{\theta}$ in the form of a priori p.d.f. $p(\boldsymbol{\theta})$.
- For each class, we have a set $D^n=\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from n statistically independent samples. Then:

$$p(\mathbf{x}/D)=\int_{\boldsymbol{\theta}} p\left(\mathbf{x}/\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}/D\right)d\boldsymbol{\theta}$$

Key relationship: Connects the conditional p.d.f. $p(\boldsymbol{x}/D)$ with the posterior p.d.f. $p(\boldsymbol{\theta}/D)$ of the parameter vector. It states that the $p(\boldsymbol{x}/D)$ is a linear combination of $p(\boldsymbol{x}/\boldsymbol{\theta})$ with weights $p(\boldsymbol{\theta}/D)$.

If $p(\boldsymbol{\theta}/D)$ has a steep unique maximum at $\boldsymbol{\theta}^*$ then:

$$p(\mathbf{x}/D)\approx p(\mathbf{x}/\boldsymbol{\theta}^*)$$

General Methodology of Bayesian Estimation

- For the computation of $p(\boldsymbol{\theta}/D)$, we have:

$$p(\boldsymbol{\theta}/D)=\frac{p(D/\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p\left(D/\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}}$$

- Due to the independence of the training samples, $p(D/\boldsymbol{\theta})=\prod_{k=1}^{n} p(\mathbf{x}_k/\boldsymbol{\theta})$

- If $p(D/\boldsymbol{\theta})$ is centered around $\boldsymbol{\theta}^*$ with a large peak at this point, and if $p(\boldsymbol{\theta}^*)$ is not 0, then $p(\boldsymbol{\theta}/D)$ also has a large peak at $\boldsymbol{\theta}^*$, and therefore will be

$$p(\mathbf{x}/D)\approx p(\mathbf{x}/\boldsymbol{\theta}^*)$$

- But the point $\boldsymbol{\theta}^*$, as described above, is the MLE estimator of $\boldsymbol{\theta}$!!

Bayesian Learning

- An issue of interest is that of the computation and the convergence of the sequence of p.d.f. *p(**θ**/D^n^)*, where we restored the index n of the number of training samples in the set *D^n^*.

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

- Therefore,

$$p(\boldsymbol{\theta}/D^n) = \frac{p(\mathbf{x}_n/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D^{n-1})}{\int_{\boldsymbol{\theta}} p(\mathbf{x}_n/\boldsymbol{\theta})\,p(\boldsymbol{\theta}/D^{n-1})\,d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta}/D^0) = p(\boldsymbol{\theta})$$

- The above relationship creates a sequence of p.d.f. $p(\boldsymbol{\theta}/\mathbf{x}_1)$, $p(\boldsymbol{\theta}/\mathbf{x}_1, \mathbf{x}_2)$,…, $p(\boldsymbol{\theta}/\mathbf{x}_1,\ldots,\mathbf{x}_n)$ ➡ Bayesian recursive (or incremental) learning.

Bayesian Learning – Normal Distribution

- ➢ Problem: Computation of p.d.f. $p(\boldsymbol{\theta}/D^n)$ and $p(\mathbf{x}/D^n)$ when we assume that $p(\mathbf{x}/\theta)=p(\mathbf{x}/\mu)=N(\mu,\sigma^2)$ (i.e. $\theta=\mu$) and $p(\mu)=N(\mu_0,\sigma_0^2)$. μ_0 is the best a prior knowledge of μ and σ_0^2 indicates our uncertainty.
- ➢ It follows that $p(\mu/D^n) = N(\mu_n,\sigma_n^2)$, where:

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0,$$

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

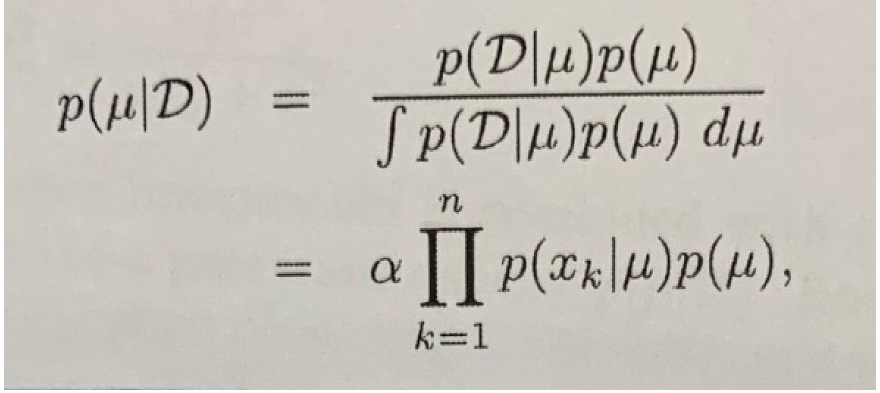
$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- ➢ μ_n represents our best knowledge of μ after observing n training samples and σ_n^2 measures our uncertainty.
- ➢ Also, $p(\mathbf{x}/D^n)= N(\mu_n, \sigma^2+\sigma_n^2)$.

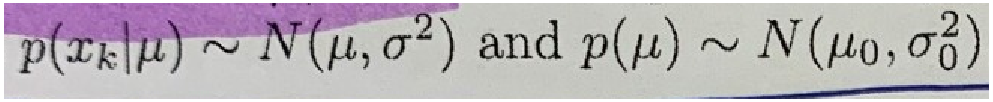
✔ *As σ_0^2→inf, we have $\mu_n = \hat{\mu}_n$ for every n, that is, we return to the estimation of maximum likelihood.*

✔ *As n→inf, we have $\sigma_n^2 \approx \sigma^2 / n$, that is, for a fairly large number of training samples, the accuracy of the estimate of μ does not depend on the uncertainty of our a priori knowledge, σ_0^2.*

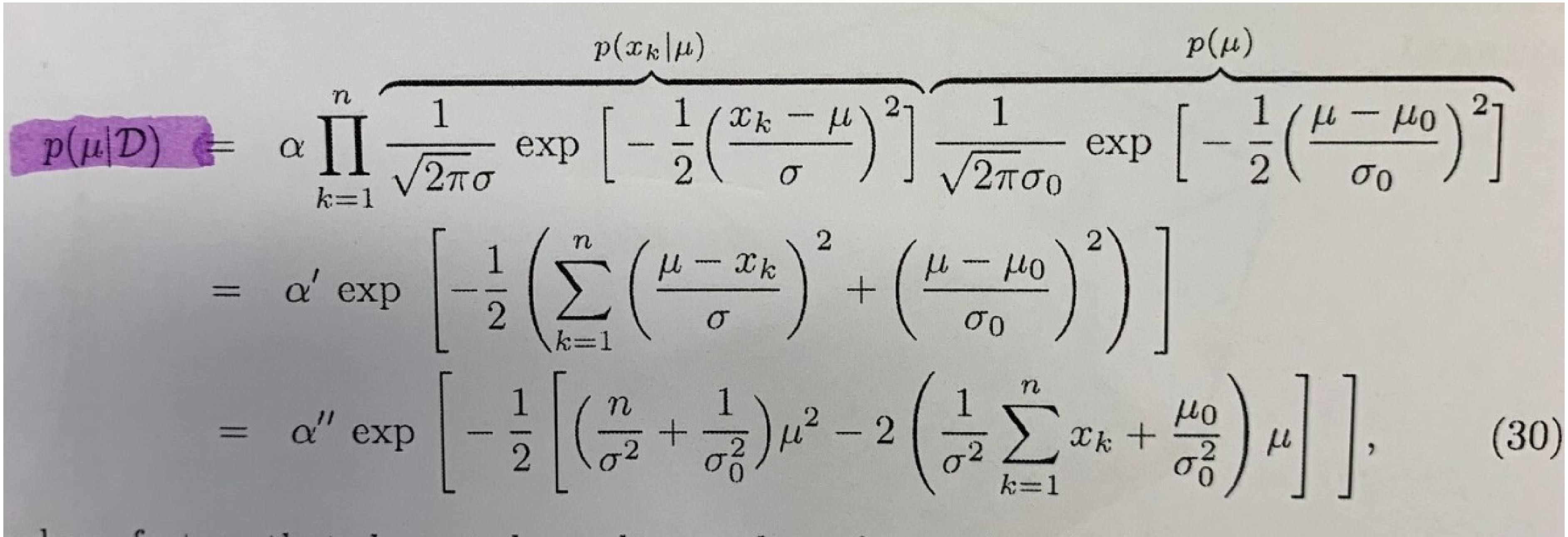
Bayesian Learning – Normal Distribution



$$\begin{aligned} p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\;d\mu} \\ &= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu), \end{aligned}$$

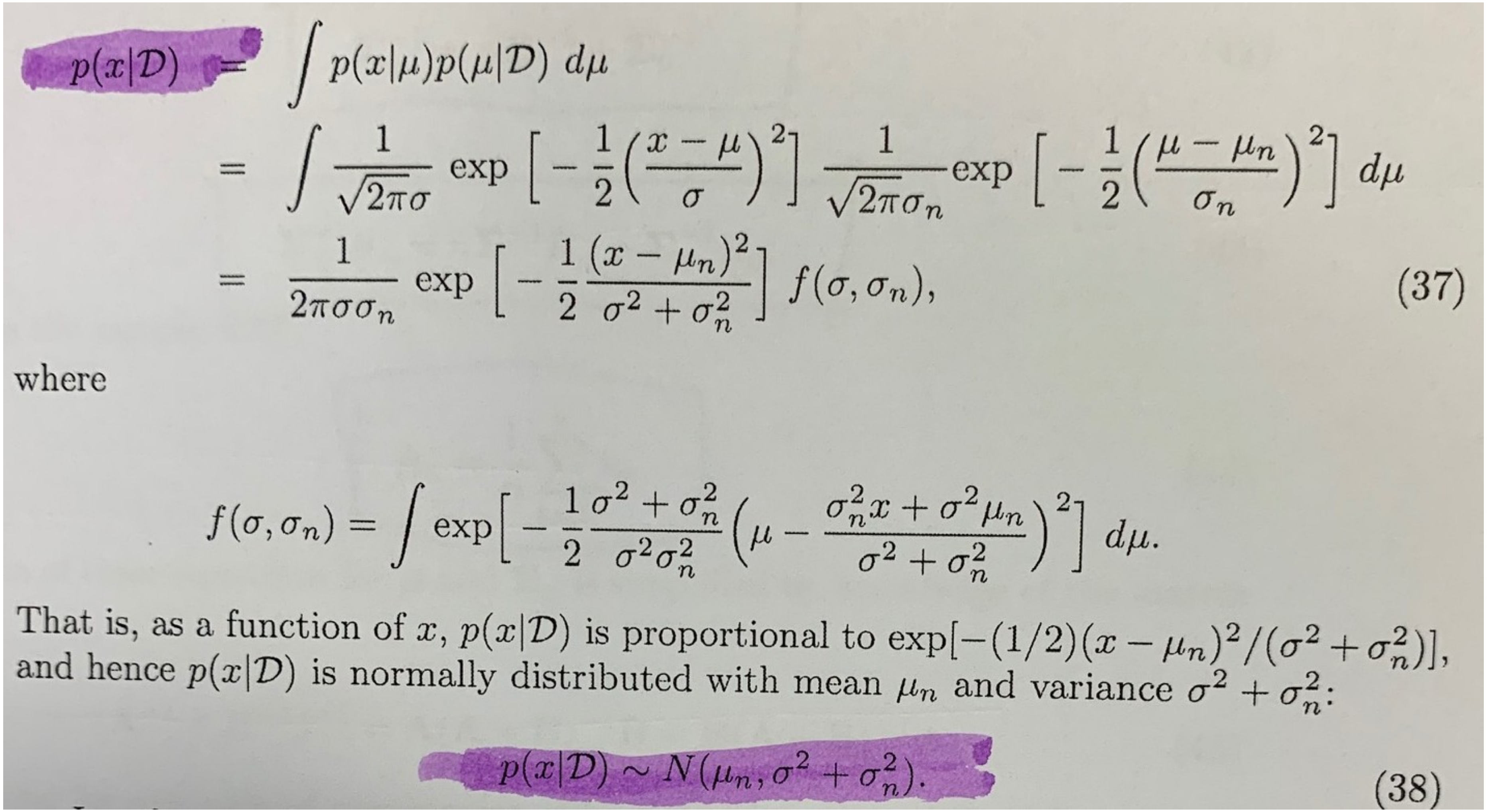


$p(x_k|\mu) \sim N(\mu, \sigma^2)$ and $p(\mu) \sim N(\mu_0, \sigma_0^2)$



$$\begin{aligned} p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\Big(\frac{x_k-\mu}{\sigma}\Big)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)^2\right]}^{p(\mu)} \\ &= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\Big(\frac{\mu-x_k}{\sigma}\Big)^2 + \Big(\frac{\mu-\mu_0}{\sigma_0}\Big)^2\right)\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}\left[\Big(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\Big)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right], \qquad (30) \end{aligned}$$

Bayesian Learning – Normal Distribution



$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\,d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n), \qquad (37)
\end{aligned}
$$

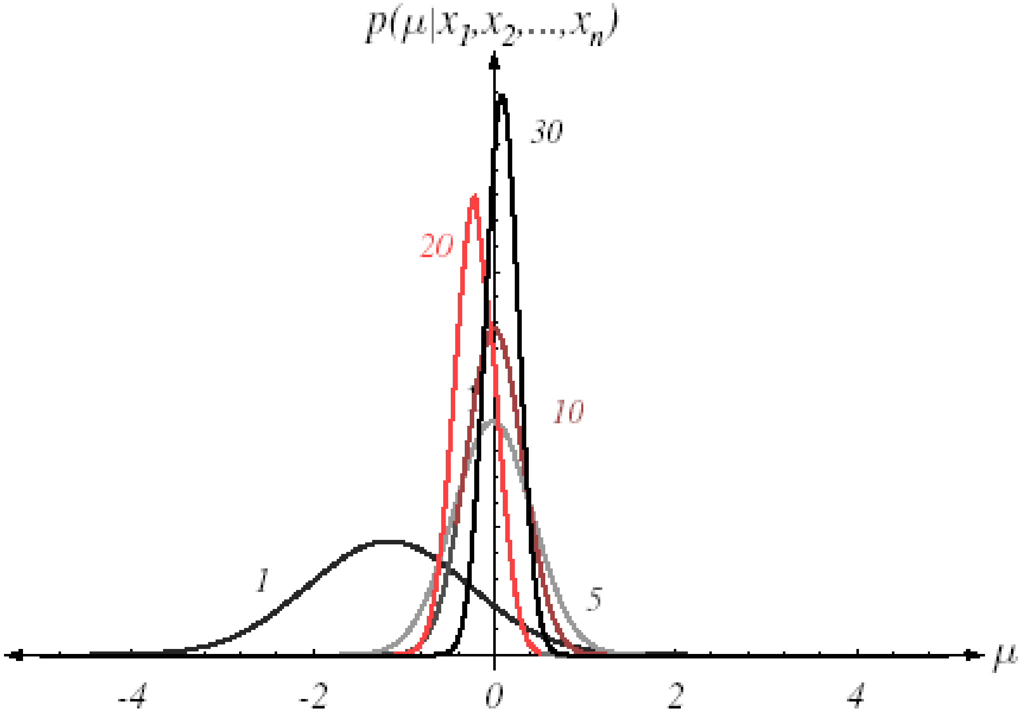
where

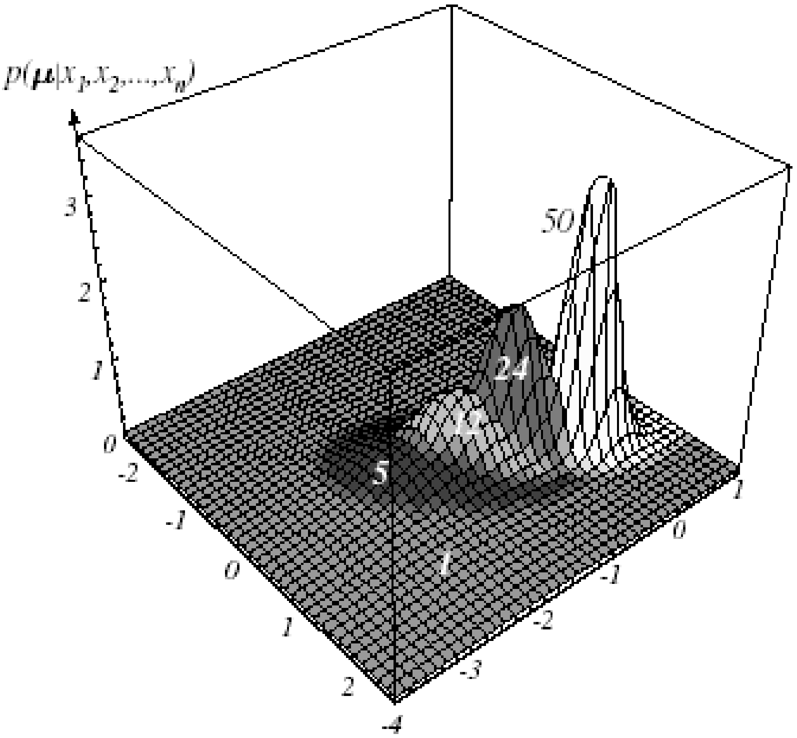
$$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu.$$

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

$$p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2). \qquad (38)$$

Bayesian Learning – Normal Distribution





As $n \to inf$, the posterior p.d.f. $p(\mu|D^n)$ becomes more and more concentrated/peaked around its middle. The phenomenon is called *Bayesian learning*.

MLE vs Bayesian Estimation

- In virtually every case, MLE & Bayesian are equivalent in case of infinite training data.
- Criteria for choosing:
  - Computational complexity ➡ MLE
  - Interpretability ➡ MLE
  - Confidence in prior info ➡ Bayesian / Bayesian with flat / uniform prior == equivalent to MLE