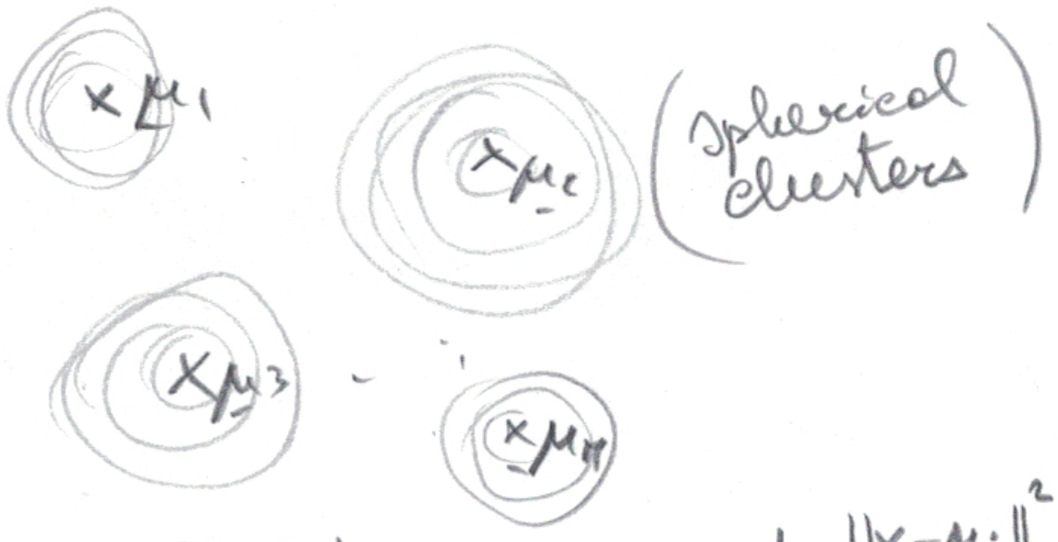
We consider now the most typical case of GAUSSIAN LIKELIHOODS WITH SPHERICAL SHAPES (G3)

$\underline{\Sigma}_i = \underline{\Sigma} = \sigma^2 \underline{I}_N$ equal covariances and all the coordinates with the same variance

$\Downarrow$

the dispersion around the means is spherical and the spread is the same for all classes.



$$f_{\underline{X}|S}(\underline{x}|a_i) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}(\underline{x}-\underline{\mu}_i)^T(\underline{x}-\underline{\mu}_i)} = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_i\|^2}$$

The MAP decision rule is

$$\hat{s} = a_i \text{ if } f_{\underline{X}|S}(\underline{x}|a_i)\pi(a_i) > f_{\underline{X}|S}(\underline{x}|a_j)\pi(a_j) \quad \forall j \neq i$$

$$\frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_i\|^2}\pi(a_i) > \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_j\|^2}\pi(a_j)$$

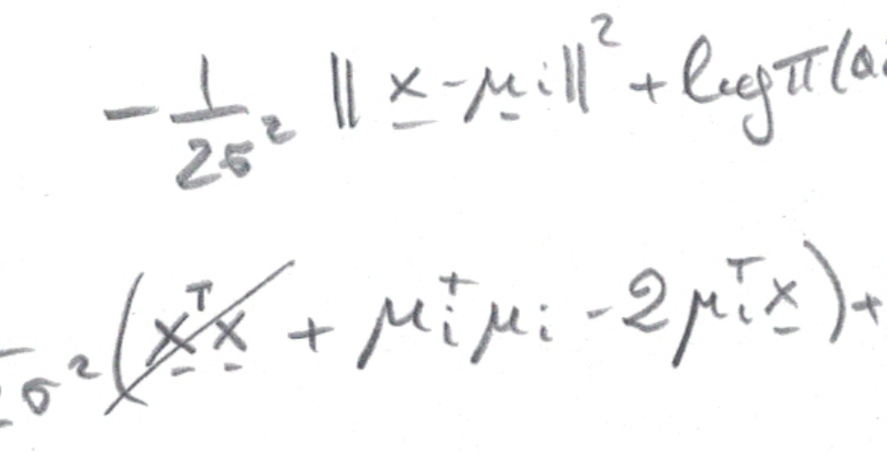
$\Downarrow$ log

$$-\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_i\|^2 + \log\pi(a_i) > -\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_j\|^2 + \log\pi(a_j) \quad (*)$$

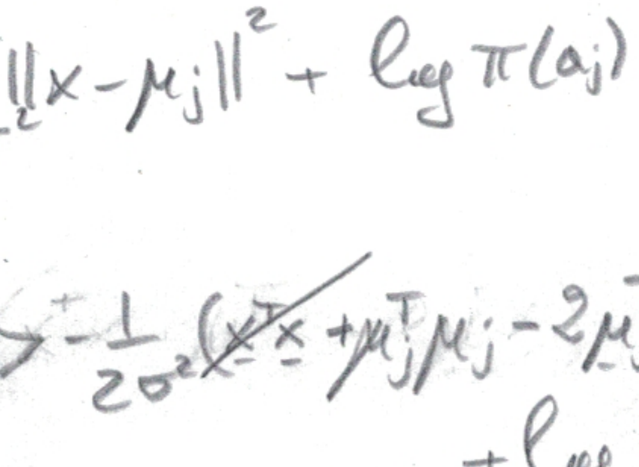
$$-\frac{1}{2\sigma^2}(\cancel{\underline{x}^T\underline{x}} + \underline{\mu}_i^T\underline{\mu}_i - 2\underline{\mu}_i^T\underline{x}) + \log\pi(a_i) > -\frac{1}{2\sigma^2}(\cancel{\underline{x}^T\underline{x}} + \underline{\mu}_j^T\underline{\mu}_j - 2\underline{\mu}_j^T\underline{x}) + \log\pi(a_j)$$

$$\underbrace{\underline{\mu}_i^T\underline{x} - \frac{1}{2\sigma^2}\underline{\mu}_i^T\underline{\mu}_i + \log\pi(a_i)}_{b_i} > \underbrace{\underline{\mu}_j^T\underline{x} - \frac{1}{2\sigma^2}\underline{\mu}_j^T\underline{\mu}_j + \log\pi(a_j)}_{b_j}$$

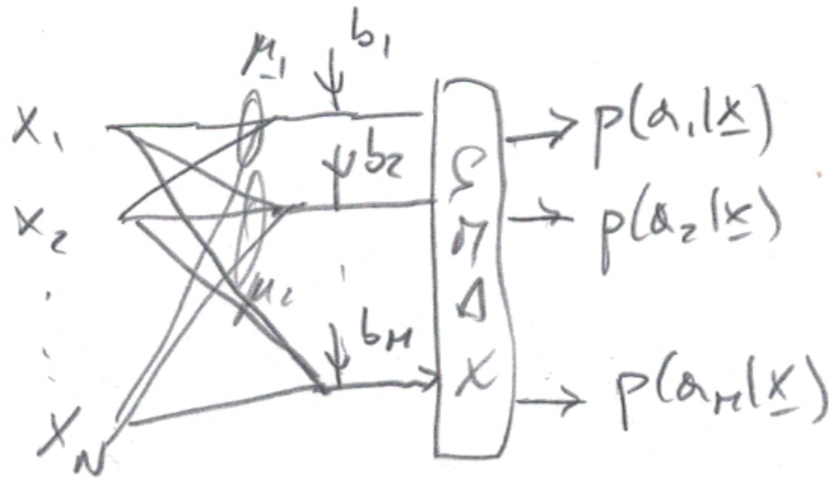
the linear classifier simplifies to



or



The potentials are obviously bounded with



Quite ill-conditioning is the case with <u>equal priors</u> (ML)

$$\pi(a_i) = \frac{1}{M} \quad i = 1, \dots, M \qquad \left(b_i = -\frac{1}{2\sigma^2}\underline{\mu}_i^T\underline{\mu}_i\right)$$

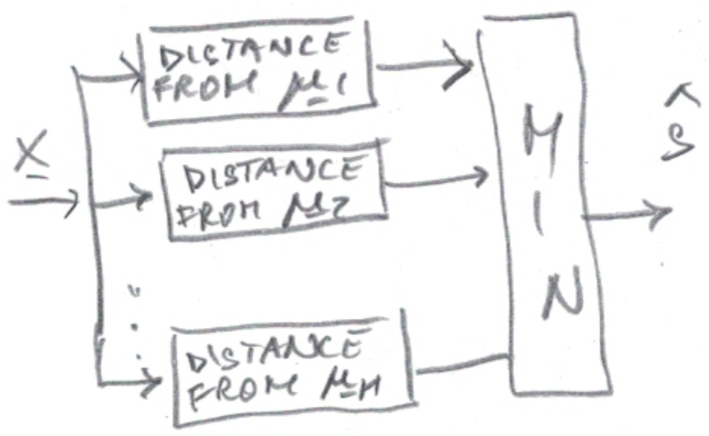
From (*)

$$\hat{s} = a_i \quad \text{if} \quad -\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_i\|^2 > -\frac{1}{2\sigma^2}\|\underline{x}-\underline{\mu}_j\|^2 \quad \forall j \neq i$$

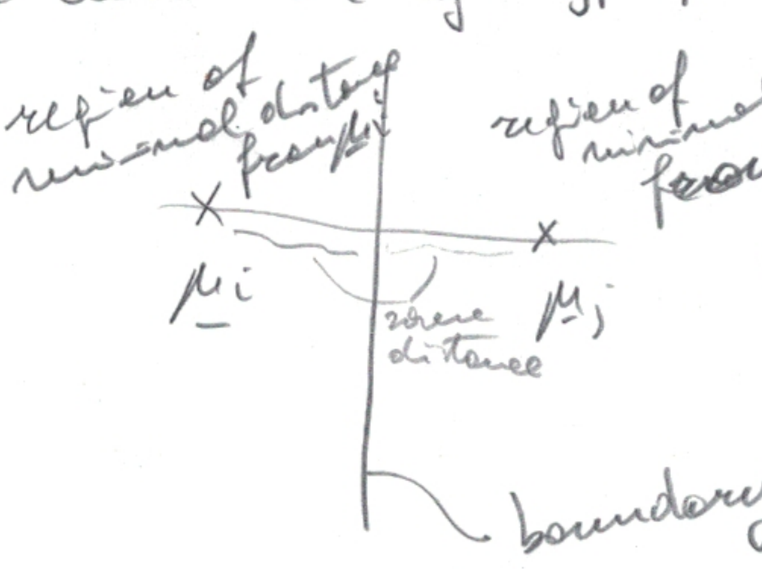
$$\underbrace{\|\underline{x}-\underline{\mu}_i\|^2}_{\text{Quadratic distance}} < \|\underline{x}-\underline{\mu}_j\|^2$$

$\hat{s} = a_i$ if $\underline{x}$ is at minimum distance from $\underline{\mu}_i$

<u>MINIMUM DISTANCE CLASSIFIER</u>



with reference to a pair i and j the decision regions are delimited by hyperplanes placed in the middle



extending to each pair, intersect the regions ⇒



<u>VORONOI TESSELLATION</u>

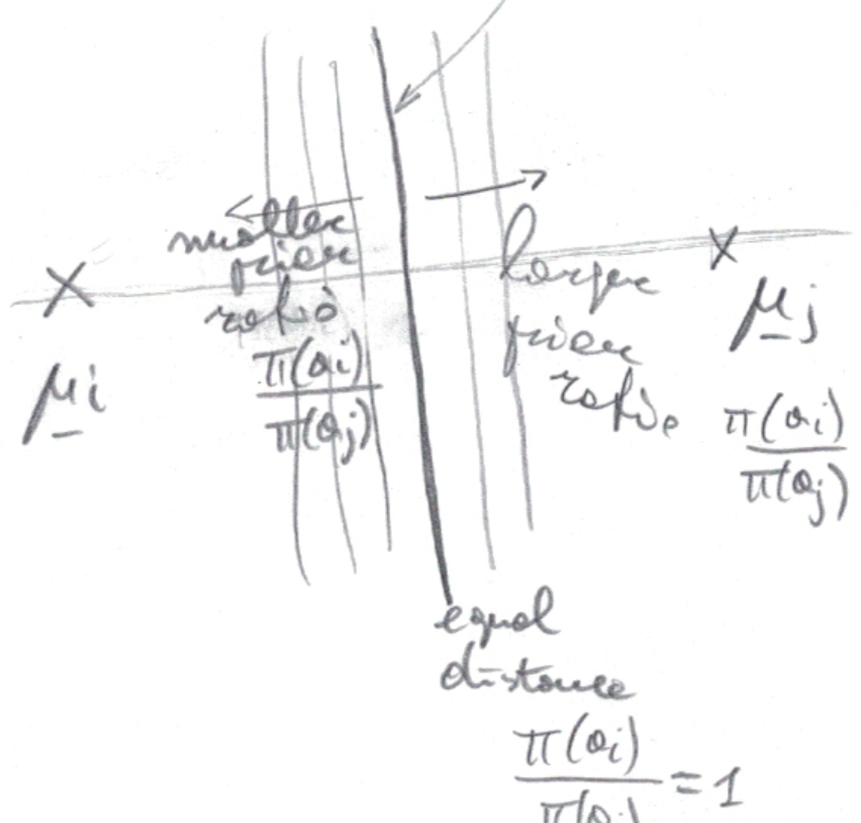
Reconsidering the MAP case for classes i and j
the set corresponds to the hyperplane

$$\underline{\mu}_i^T \underline{x} - \frac{1}{2\sigma^2}\underline{\mu}_i^T\underline{\mu}_i + \log \pi(a_i) = \underline{\mu}_j^T \underline{x} - \frac{1}{2\sigma^2}\underline{\mu}_j^T\underline{\mu}_j + \log \pi(a_j)$$

$$(\underline{\mu}_i^T - \underline{\mu}_j^T)\underline{x} - \frac{1}{2\sigma^2}(\underline{\mu}_i^T\underline{\mu}_i - \underline{\mu}_j^T\underline{\mu}_j) + \log \frac{\pi(a_j)}{\pi(a_i)} = 0$$

this hyperplane is parallel to the equal-distance hyperplane

$$(\underline{\mu}_i^T - \underline{\mu}_j)\underline{x} - \frac{1}{2\sigma^2}(\underline{\mu}_i^T\underline{\mu}_i - \underline{\mu}_j^T\underline{\mu}_j) = 0$$



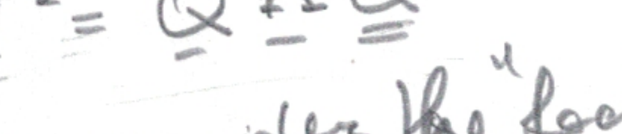
The classes with the larger priors take more space.

$$\frac{\pi(a_i)}{\pi(a_j)} = 1$$

let us now reconsider the gaussian case with equal but arbitrary covariances (G2), because it can be reduced to the spherical case (G3).

Recall that a covariance matrix $\underline{\underline{\Sigma}}$ can be factorized as

$$\underline{\underline{\Sigma}} = \underline{\underline{Q}}\,\underline{\underline{\Lambda}}\,\underline{\underline{Q}}^T$$



where $\underline{\underline{Q}} = [\underline{q}_1\, \underline{q}_2 \cdots \underline{q}_N]$ contains the N eigenvectors of $\underline{\underline{\Sigma}}$ and

$\underline{\underline{\Lambda}} = \text{diag}(\lambda_1 \lambda_2 \cdots \lambda_N)$ the eigenvalues.

Clearly $\underline{\underline{Q}}$ is unitary, ie $\underline{\underline{Q}}^T\underline{\underline{Q}} = \underline{\underline{I}}_N$, and therefore

$$\underline{\underline{\Sigma}}^{-1} = \underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1}\underline{\underline{Q}}^T$$

Now consider the "factor" $\underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1/2}$ for projecting $\underline{x}$ into a new space

$$\underline{y} = \underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T\underline{x} \qquad \underline{\underline{W}}^T$$

The means for the classes will be

$$\underline{m}_i = \underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T\underline{\mu}_i$$

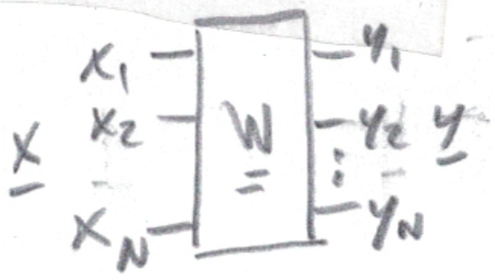
and the covariances become diagonal with all unit variances.

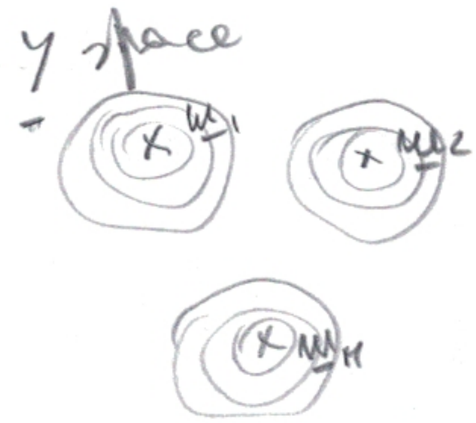
$$\underline{\underline{S}}_i = E[(\underline{y} - \underline{m}_i)(\underline{y} - \underline{m}_i)^T] = E[\underline{y}\underline{y}^T] - \underline{m}_i\underline{m}_i^T$$

$$= E\left[\underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T \underline{x}\,\underline{x}^T \underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1/2}\right] - \underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T\underline{\mu}_i\underline{\mu}_i^T\underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1/2}$$

$$= \underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T\left(\underbrace{E[\underline{x}\,\underline{x}^T] - \underline{\mu}_i\underline{\mu}_i^T}_{\underline{\underline{\Sigma}}_i}\right)\underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1/2}$$

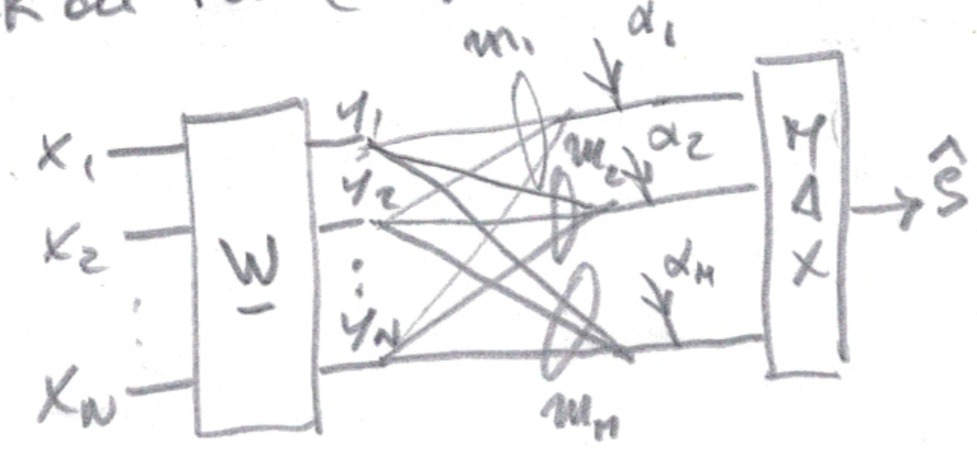
$$= \underline{\underline{\Lambda}}^{-1/2}\underline{\underline{Q}}^T\,\underline{\underline{Q}}\,\underline{\underline{\Lambda}}\,\underline{\underline{Q}}^T\underline{\underline{Q}}\,\underline{\underline{\Lambda}}^{-1/2} = \underline{\underline{I}}_N$$

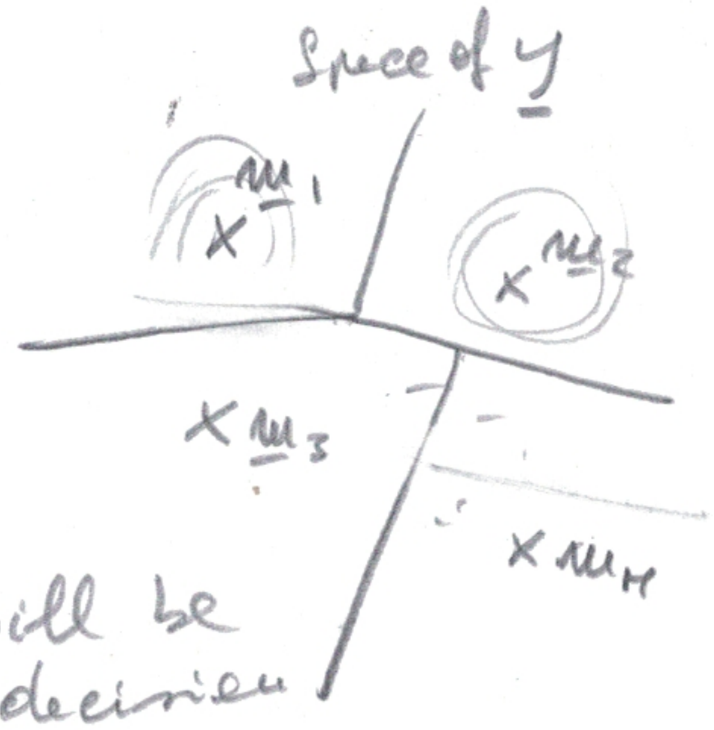




$$f_{\underline{Y}|S}(\underline{y}|\Omega_i) = N(\underline{y}; \underline{m}_i, \underline{\underline{I}}_N)$$

Now the classification problem can start from $\underline{y}$ and fall back on the (B3) case in which





$$\alpha_i = -\frac{1}{2}\underline{m}_i^T\underline{m}_i + \log \pi(\Omega_i)$$

In the space of $\underline{y}$ all likelihoods will be spherical with unit variance and the decision regions will be delimited by hyperplanes.

If $\pi(\Omega_i) = \frac{1}{M}$ $i = 1 \dots M$, the decision regions will constitute a minimum distance VORONOI TESSELLATION.

This also clarifies the geometric interpretation

THE RANK-DEFICIENT CASE

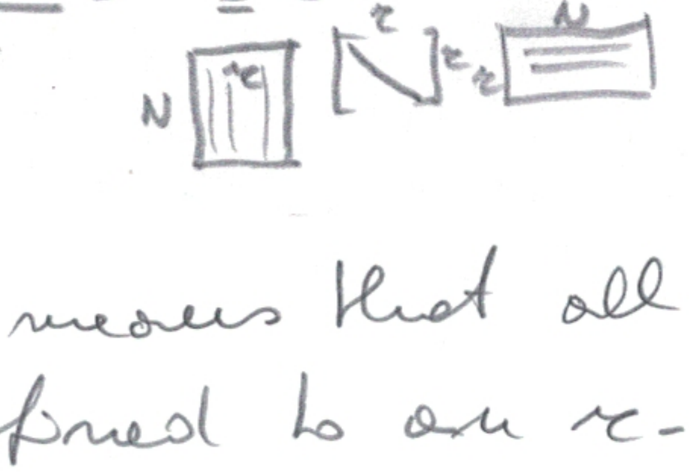
In the factorization of $\underline{\underline{\Sigma}}$ it may happen that only r eigenvalues are non null. With reference to a partitioned factorization

$$\underline{\underline{\Sigma}} = [\underline{\underline{Q}}_r \; \underline{\underline{Q}}_{N-r}] \begin{bmatrix} \underline{\underline{\Lambda}}_r & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{\Lambda}}_{N-r} \end{bmatrix} \begin{bmatrix} \underline{\underline{Q}}_r^T \\ \underline{\underline{Q}}_{N-r}^T \end{bmatrix}$$

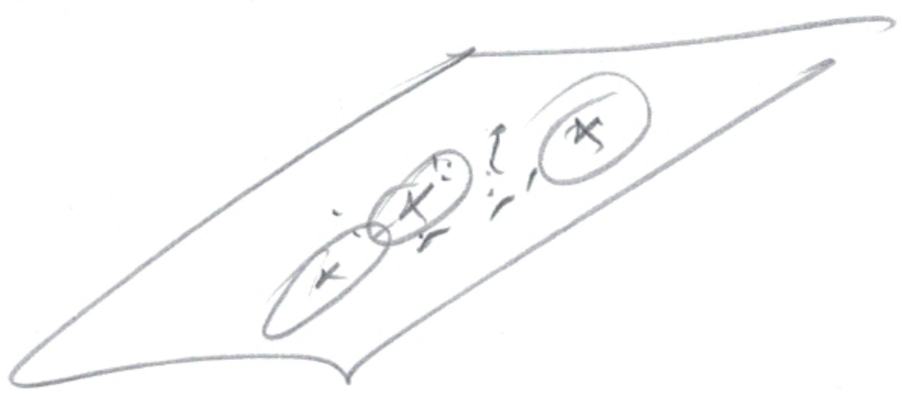


if $\underline{\underline{\Lambda}}_{N-r} = \underline{\underline{0}}$

$$\underline{\underline{\Sigma}} = \underline{\underline{Q}}_r \underline{\underline{\Lambda}}_r \underline{\underline{Q}}_r^T$$

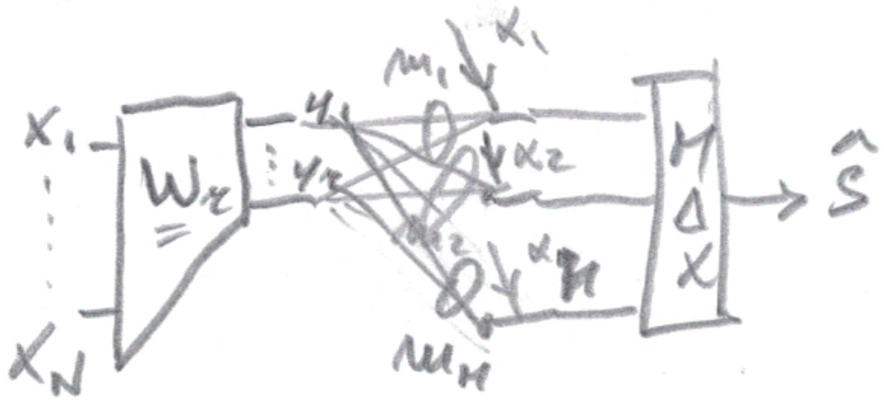


this means that all the distribution is confined to an r-dimensional linear space



The transformation becomes rectangular

$$W_{\underline{\underline{r}}} = Q_{\underline{\underline{r}}} \Lambda^{-1/2} \qquad \underline{y} = \Lambda^{-1/2} Q_{\underline{\underline{r}}}^T \underline{x}$$



ALMOST RANK-DEFICIENT COVARIANCES:

When the eigenvalues $\{\lambda_{r+1}, \ldots, \lambda_N\}$ are very small, we may consider to ignore them, as in the KLT (Karhunen-Loève Transform) also named PCA (Principal Components Analysis). Clearly there would be a loss of information (just as in lossy coding), but a classifier can be built as well accepting small approximation. Recall that in an almost rank-deficient case the data points are in the close neighbourhood of an hyperplane.



Projecting everything on that linear subspace may be just fine.

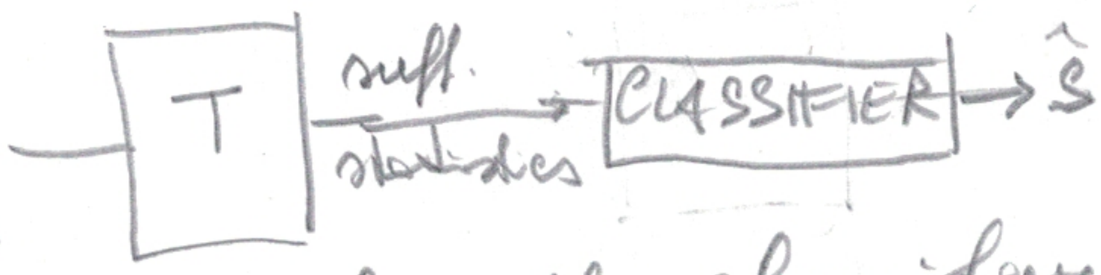
OTHER DISTRIBUTIONS

There are many other parametric distributions that can be used to design the MAP classifier that needs

$$f_{\underline{X}|S}(\underline{x}|a_i)\pi(a_i) \quad \text{or} \quad \log f_{\underline{X}|S}(\underline{x}|s) + \log \pi(a_i) \qquad i=1,\dots,M.$$

The likelihoods $f_{\underline{X}|S}(\underline{x}|a_i)$, or the log-likelihoods $\log f_{\underline{X}|S}(\underline{x}|s)$, may be very complex functions that depend on a large number of parameters. For example, neural networks, that will be discussed in the following, could be used, and their parameters be "learned" from data.

In estimation theory a transformation on the data that preserves information for the estimation of a variable (in this case the class) produces what are called "sufficient statistics". Therefore a classifier may take the form of the cascade



where perhaps the classifier has a simpler (maybe linear) structure.

However, there exist some classical distributions that can be used as likelihoods and that lead to simple classifiers.

THE EXPONENTIAL FAMILY

A very popular family of distributions functions that share with the gaussian distribution many properties, is the exponential family

$$f_{\underline{x}|a_i}(\underline{x}|a_i) = h(\underline{x})\, e^{\underline{\eta}_i^T \underline{T}(\underline{x}) - c_i}$$

where $h(\underline{x})$ is a positive function that does not depend on i, $\underline{T}(\underline{x})$ is vector function that also does not depend on i, and $\underline{\eta}_i$ and c_i are a vector and a scalar respectively, that are associated to a_i.

Evidently for this function to be a valid pdf it must happen that

$$\int_{\mathcal{X}} f_{\underline{x}|a_i}(\underline{x}|a_i)\, d\underline{x} = 1 \quad \forall i \qquad \text{i.e.}$$

$$\int_{\mathcal{X}} h(\underline{x})\, e^{\underline{\eta}_i^T \underline{T}(\underline{x}) - c_i}\, d\underline{x} = 1$$

$$\int_{\mathcal{X}} h(\underline{x})\, e^{\underline{\eta}_i^T \underline{T}(\underline{x})}\, d\underline{x} = e^{c_i}$$

$$c_i = \log \int_{\mathcal{X}} h(\underline{x})\, e^{\underline{\eta}_i^T \underline{T}(\underline{x})}\, d\underline{x}$$

c_i is a normalizing parameter, also known in physics, as the "log-partition function".

A MAP classifier with an exponential likelihood has to compute

$$\int_{\underline{x}/s} (\underline{x}|a_i)\pi(a_i) = h(\underline{x})\, e^{\underline{\eta}_i^T \underline{T}(\underline{x}) - c_i}\, \pi(a_i)$$

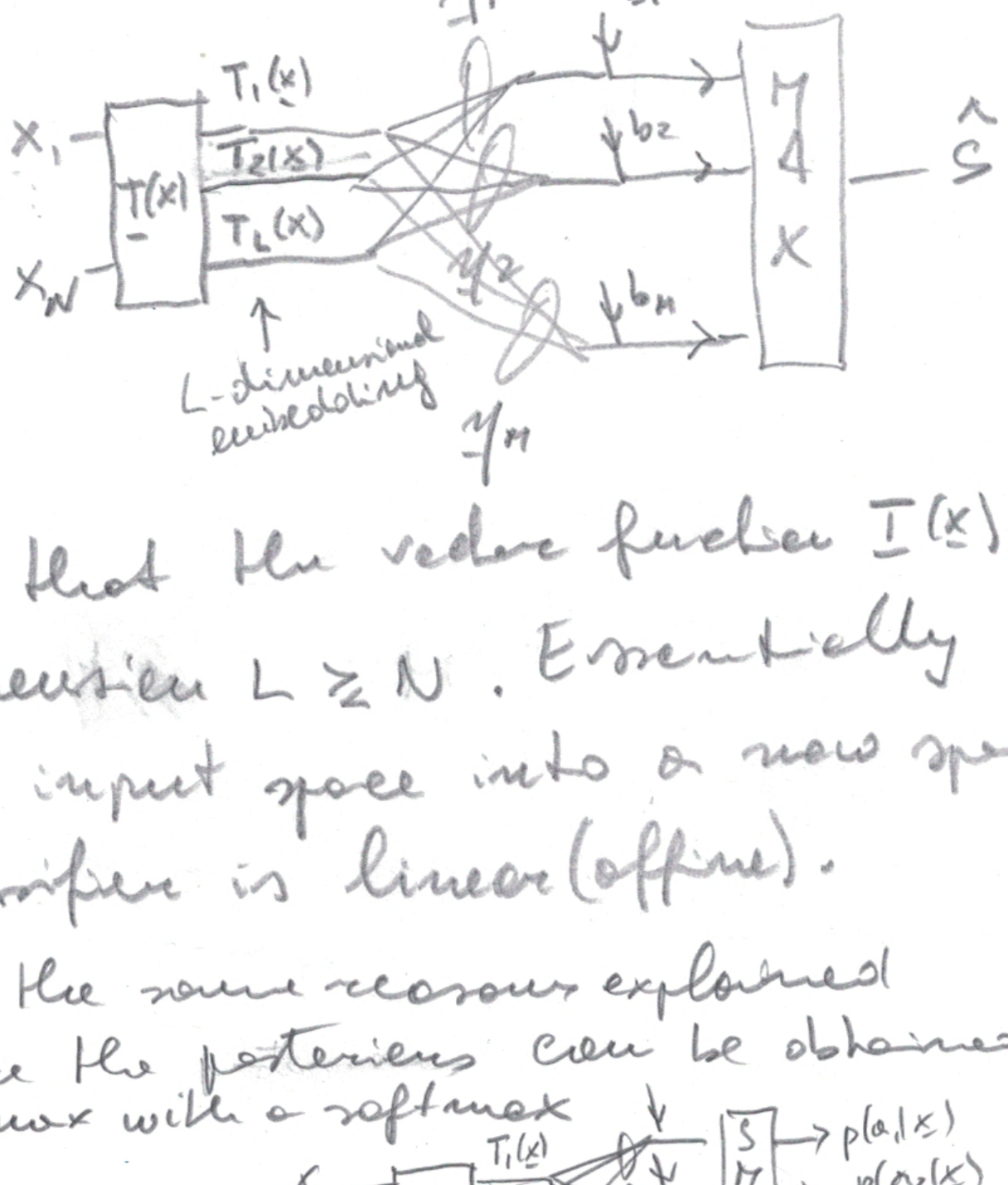
or equivalently the log

$$\log f_{\underline{x}|s}(\underline{x}|a_i) + \log \pi(a_i)$$

$$= \log h(\underline{x}) + \underline{\eta}_i^T \underline{T}(\underline{x}) - c_i + \log \pi(a_i)$$

Since $\log h(\underline{x})$ does not depend on i, the classifier gives

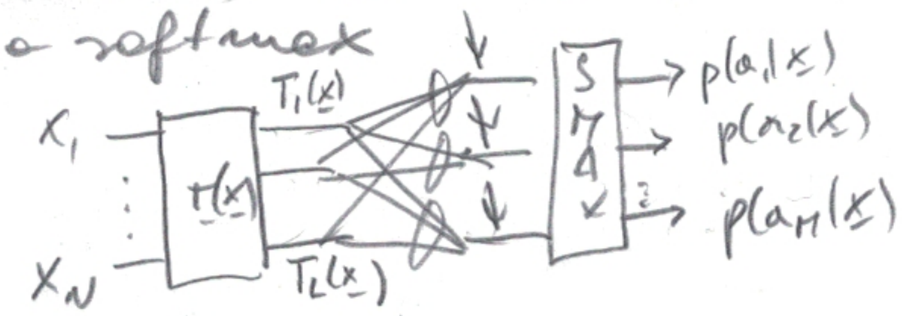
$$\hat{s} = a_i \text{ if } \underline{\eta}_i^T \underline{T}(\underline{x}) - \underbrace{c_i + \log \pi(a_i)}_{b_i} > \underline{\eta}_j^T \underline{T}(\underline{x}) - \underbrace{c_j + \log \pi(a_j)}_{b_j} \quad \forall j \neq i$$



Note that the vector function $\underline{T}(\underline{x})$ may have a dimension $L \geq N$. Essentially $\underline{T}(\underline{x})$ maps the input space into a new space where the classifier is linear (affine).

L-dimensional embedding

For the same reasons explained before the posteriors can be obtained simply by replacing the max with a softmax



It is easy to see that the gaussian pdfs with equal covariances ($\Sigma_i = \Sigma \ \forall i$) belongs to the exponential family:

$$f_{\underline{x}|s}(\underline{x}|a_i) = \frac{1}{(2\pi)^{N/2}|\underline{\underline{\Sigma}}|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_i)}$$

$$= \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\underline{x}^T\underline{\underline{\Sigma}}^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_i^T\underline{\underline{\Sigma}}^{-1}\underline{\mu}_i + \underline{\mu}_i^T\underline{\underline{\Sigma}}^{-1}\underline{x}}$$

$$= \underbrace{\frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\underline{x}^T\underline{\underline{\Sigma}}^{-1}\underline{x}}}_{h(\underline{x})} e^{\underbrace{\underline{\mu}_i^T\underline{\underline{\Sigma}}^{-1}}_{\eta_i^T}\underbrace{\underline{x}}_{T(\underline{x})} - \underbrace{\frac{1}{2}\underline{\mu}_i^T\underline{\underline{\Sigma}}^{-1}\underline{\mu}_i}_{c_i}}$$

$$\underline{T}(\underline{x}) = \underline{x}$$

or

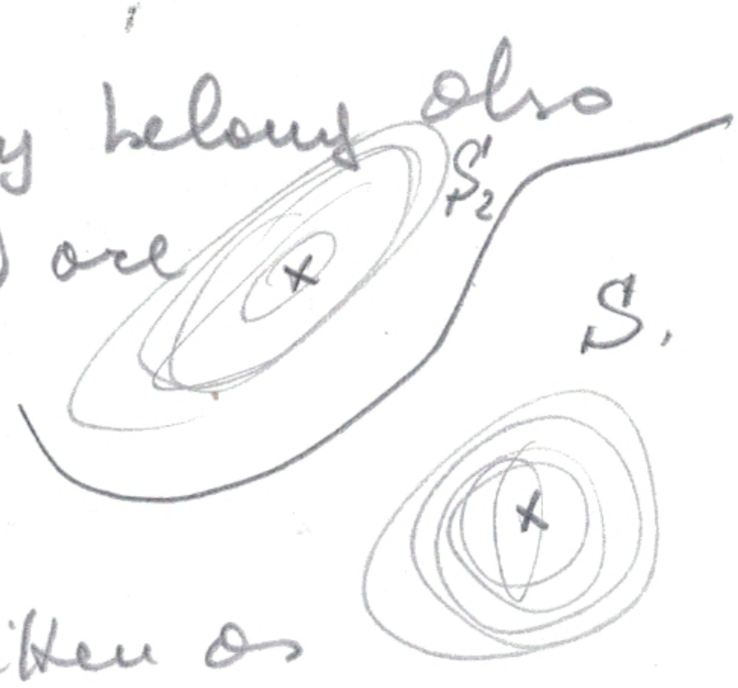
$$T(\underline{x}) = \underline{\underline{\Sigma}}^{-1}\underline{x}$$

THE BINARY CASE

The case in which there are only two classes $\Lambda = \{a_1, a_2\}$, deserves special attention for the very many applications areas in which it emerges.

For example: RADAR DETECTION (Target present/not present) COMMUNICATIONS (bit 0 or 1); ALARM SYSTEM (ON/OFF); COVID TESTING (POSITIVE/NEGATIVE), etc....

The two likelihoods (that may belong also to different parametric classes) are



$$f_{\underline{X}|S}(\underline{x}|a_1) \qquad f_{\underline{X}|S}(\underline{x}|a_2)$$

The MAP criterion can be written as

$$f_{\underline{X}|S}(\underline{x}|a_1)\pi(a_1) \underset{a_2}{\overset{a_1}{\gtrless}} f_{\underline{X}|S}(\underline{x}|a_2)\pi(a_2)$$

Recall also that $\pi(a_2) = 1 - \pi(a_1)$, i.e. we need only one prior. The test can be re-written in many equivalent ways

likelihood ratio $L(\underline{x})$

$$\frac{f_{\underline{X}|S}(\underline{x}|a_1)}{f_{\underline{X}|S}(\underline{x}|a_2)} \underset{a_2}{\overset{a_1}{\gtrless}} \underbrace{\frac{\pi(a_2)}{\pi(a_1)}}_{\lambda}$$

$\Downarrow \log$

log-likelihood ratio $l(\underline{x})$

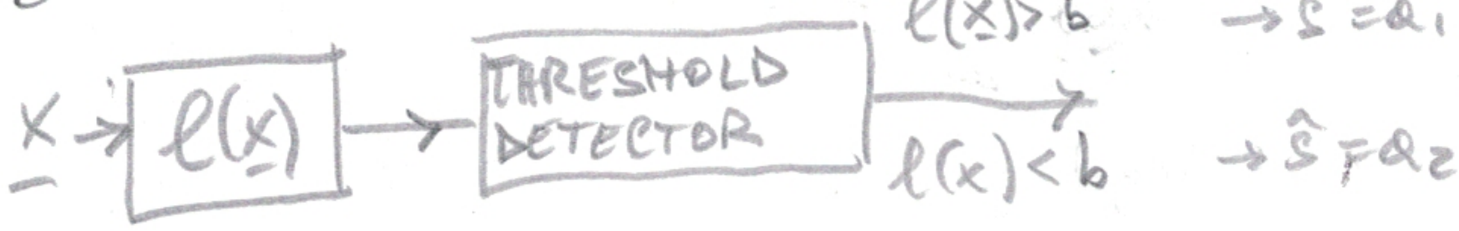
$$\log \frac{f_{\underline{X}|S}(\underline{x}|a_1)}{f_{\underline{X}|S}(\underline{x}|a_2)} \underset{a_2}{\overset{a_1}{\gtrless}} \underbrace{\log \frac{\pi(a_2)}{\pi(a_1)}}_{\log \lambda = b}$$

$$\log f_{\underline{X}|S}(\underline{x}|a_1) - \log f_{\underline{X}|S}(\underline{x}|a_2) \underset{a_2}{\overset{a_1}{\gtrless}} \log \pi(a_2) - \log \pi(a_1)$$

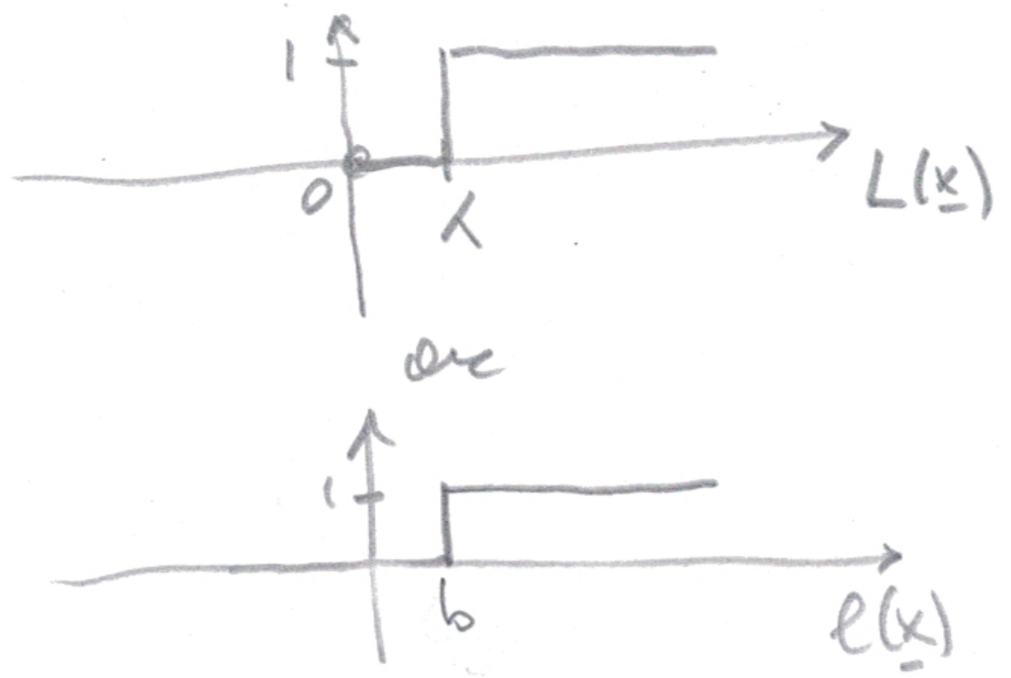
The classifier can be drawn as



or



If we code a_1 as 1 and a_2 as 0, the threshold detector becomes a step function



Since these functions are complementary to one another

$$P(S=a_1|\underline{x}) \cdot P(S=a_2|\underline{x}) = 1 - P(S=a_1|\underline{x}),$$

we may just consider one of them, say

$$P(S=a_1|\underline{x}) = \frac{f_{\underline{X}|S}(\underline{x}|a_1)\,\pi(a_1)}{f_{\underline{X}}(\underline{x})} \propto f_{\underline{X}|S}(\underline{x}|a_1)\,\pi(a_1)$$

Rewriting

$$P(a_1|\underline{x}) = \frac{f_{\underline{x}|S}(\underline{x}|a_1)\,\pi(a_1)}{f_{\underline{x}|S}(\underline{x}|a_1)\,\pi(a_1) + f_{\underline{x}|S}(\underline{x}|a_2)\,\pi(a_2)}$$

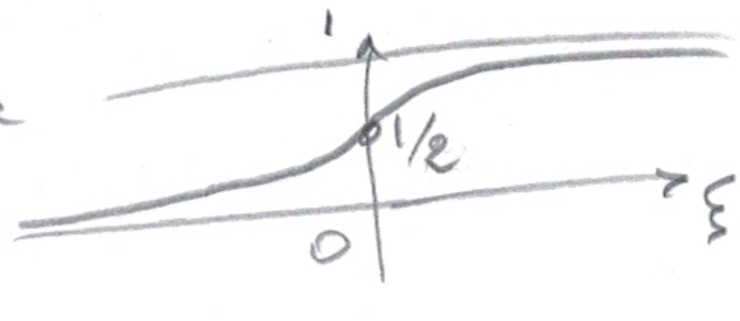
$$= \frac{1}{1 + \dfrac{f_{\underline{x}|S}(\underline{x}|a_2)\,\pi(a_2)}{f_{\underline{x}|S}(\underline{x}|a_1)\,\pi(a_1)}} = \frac{1}{1 + e^{\log \frac{f_{\underline{x}|S}(\underline{x}|a_2)}{f_{\underline{x}|S}(\underline{x}|a_1)} + \log \frac{\pi(a_2)}{\pi(a_1)}}},$$

$$= \frac{1}{1 + e^{-\left[\underbrace{\log \frac{f_{\underline{x}|S}(\underline{x}|a_1)}{f_{\underline{x}|S}(\underline{x}|a_2)}}_{\ell(\underline{x})} + \underbrace{\log \frac{\pi(a_1)}{\pi(a_2)}}_{-b}\right]}} = \frac{1}{1+e^{-[\ell(\underline{x})-b]}}$$

---

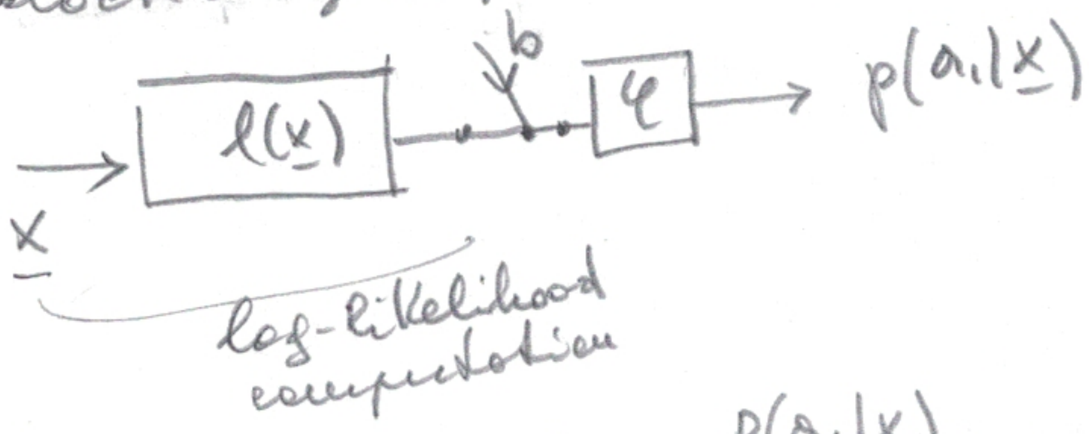
the function $\varphi(\xi) = \dfrac{1}{1+e^{-\xi}}$ is the <u>logistic function</u>

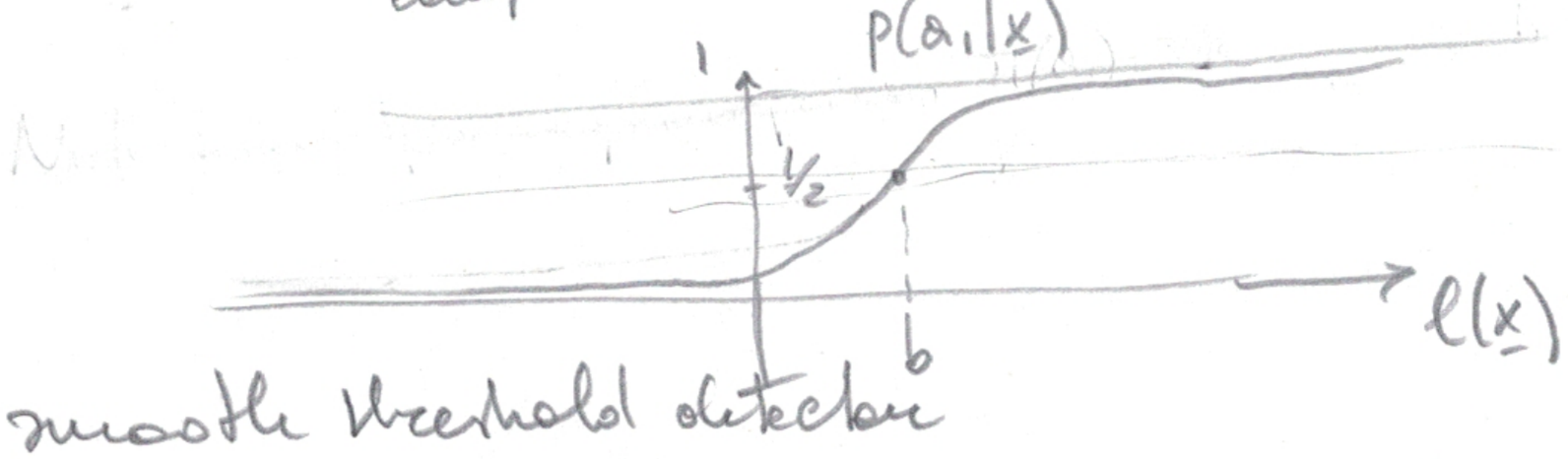
that has the <u>sigmoidal</u> shape



---

the block diagram for the receiver becomes





smooth threshold detector

MAP DECISION:

if $p(a_1|\underline{x}) > \frac{1}{2} \Rightarrow \hat{s} = a_1$

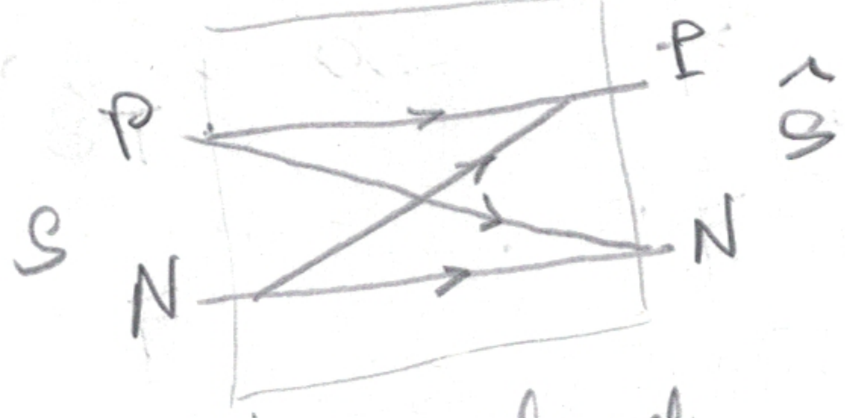
else $\hat{s} = a_2$

Note that for uniform priors $\pi(a_1) = \pi(a_2) = \frac{1}{2}$, $b = 0$

The binary classification problem deserves further attention because when it is applied to a DETECTION PROBLEM, with an alphabet $\mathcal{A} = \{\text{Positive}, \text{Negative}\} \equiv \{P, N\}$ we may be interested in the probabilities

$P_r\{\hat{S}=P \mid S=N\}$ FALSE POSITIVE (FALSE ALARM)

$P_r\{\hat{S}=N \mid S=P\}$ FALSE NEGATIVE (MISSED DETECTION)



binary channel

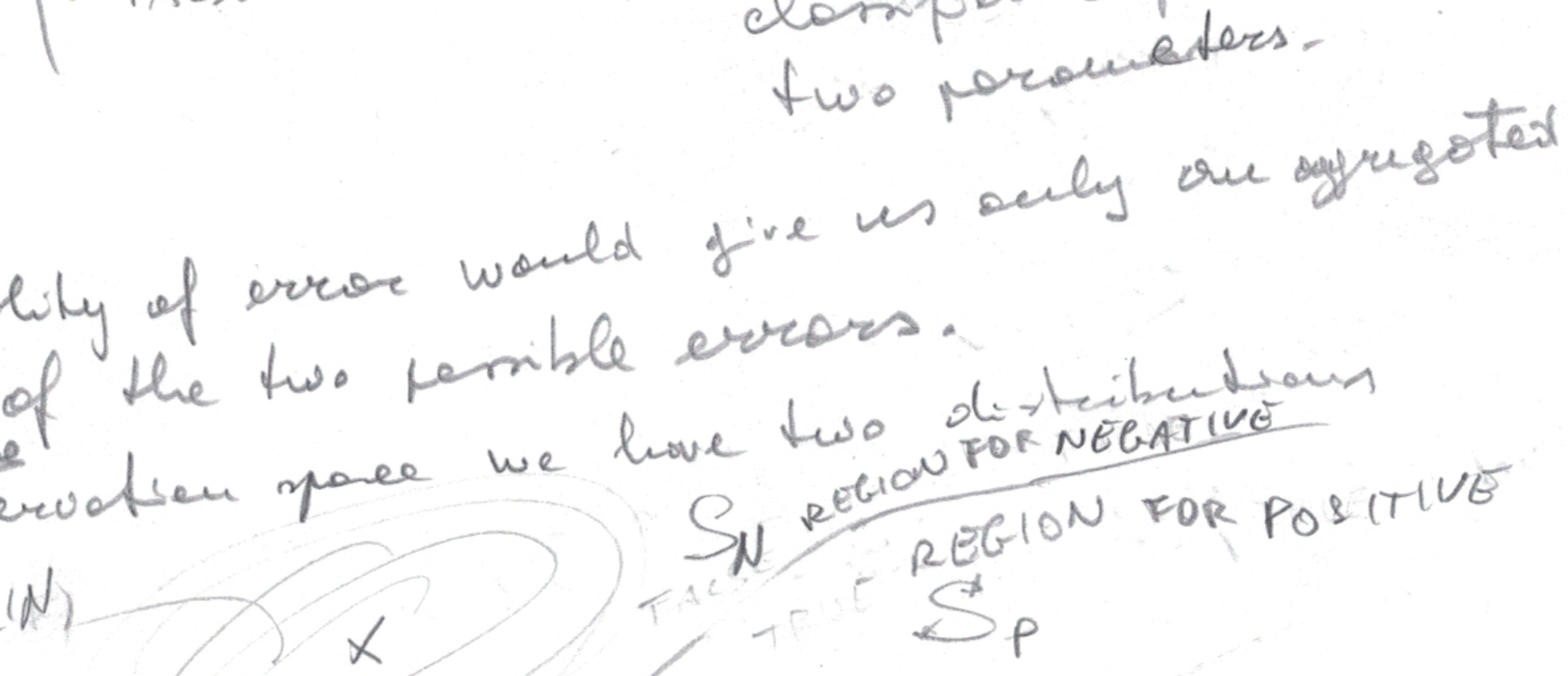
$$\underline{\underline{P}}_e = \sum_{\hat{S}} \begin{pmatrix} P_r\{\hat{S}=P \mid S=P\} & P(\hat{S}=N \mid S=P) \\ P_r\{\hat{S}=P \mid S=N\} & P_r\{\hat{S}=N \mid S=N\} \end{pmatrix}$$

$\hat{S}$; TRUE POSITIVE; FALSE NEGATIVE; FALSE POSITIVE; TRUE NEGATIVE

The matrix is row-stochastic, i.e.
$P(\hat{S}=P \mid S=P) = 1 - P(\hat{S}=N \mid S=P)$
and
$P(\hat{S}=P \mid S=N) = 1 - P(\hat{S}=N \mid S=N)$
Therefore the whole classifier depends on only two parameters.

The Probability of error would give us only an aggregated account of the two possible errors.

Looking at the observation space we have two distributions



The false negative event happens when a sample from $f_{\underline{X}|S}(\underline{x}|P)$ falls in the S_N region. The probability of this to happen is

$$P_r\{\hat{S}=N \mid S=P\} = \int_{S_N} f_{\underline{X}|P}(\underline{x}|P)\, d\underline{x}$$

Similarly a false positive happens when a sample from $f_{\underline{x}|S}(\underline{x}|N)$ falls into the S_P region. This happens with probability

$$P_2\{\hat{S}=P|S=N\} = \int_{S_P} f_{\underline{x}|S}(\underline{x}|N)\,d\underline{x}$$

It is not always easy to compute the above probabilities because we have to integrate over a possibly large N-dimensional space.

Consider now a few special (typical) cases:

GAUSSIAN LIKELIHOODS

$$f_{\underline{x}|S}(\underline{x}|a_1) = N(\underline{x};\underline{\mu}_1,\underline{\Sigma}_1) = \frac{1}{(2\pi)^{N/2}|\Sigma_1|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)^T\underline{\Sigma}_1^{-1}(\underline{x}-\underline{\mu}_1)}$$

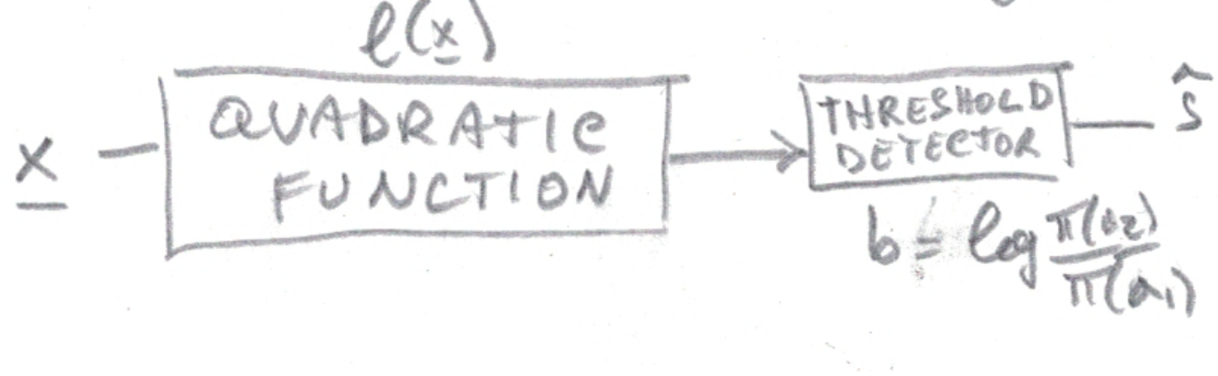
$$f_{\underline{x}|S}(\underline{x}|a_2) = N(\underline{x};\underline{\mu}_2,\underline{\Sigma}_2) = \frac{1}{(2\pi)^{N/2}|\Sigma_2|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_2)^T\underline{\Sigma}_2^{-1}(\underline{x}-\underline{\mu}_2)}$$

$$L(\underline{x}) = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} e^{-\frac{1}{2}\left[(\underline{x}-\underline{\mu}_1)^T\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)^T\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2)\right]}$$

$$\ell(\underline{x}) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\left(\underline{x}^T\Sigma_1^{-1}\underline{x} + \underline{\mu}_1^T\Sigma_1^{-1}\underline{\mu}_1 - 2\underline{\mu}_1^T\Sigma_1^{-1}\underline{x}\right.$$

$$\left. - \underline{x}^T\Sigma_2^{-1}\underline{x} - \underline{\mu}_2^T\underline{\Sigma}_2^{-1}\underline{\mu}_2 + 2\underline{\mu}_2^T\Sigma_2^{-1}\underline{x}\right)$$

$$= \underbrace{\frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}}_{\text{cst.}} - \frac{1}{2}\underbrace{\underline{x}^T\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\underline{x}}_{\text{quadratic}} + \underbrace{\left(\underline{\mu}_1^T\Sigma_1^{-1}-\underline{\mu}_2^T\Sigma_2^{-1}\right)\underline{x}}_{\text{linear}} - \underbrace{\frac{1}{2}\underline{\mu}_1^T\Sigma_1^{-1}\underline{\mu}_1 + \frac{1}{2}\underline{\mu}_2^T\Sigma_2^{-1}\underline{\mu}_2}_{\text{cst.}}$$

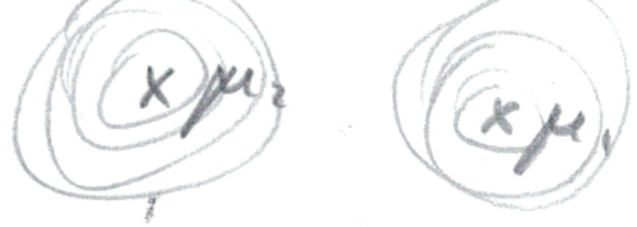


The computation of the probabilities of false positive and false negative are challenging. Therefore consider a much simpler case with equal spherical likelihoods with $\Sigma_1 = \Sigma_2 = \sigma^2 I_N$

The log-likelihood becomes linear

$$l(\underline{x}) = -\frac{1}{2\sigma^2}(\underline{\mu}_1^T\underline{\mu}_1 - \underline{\mu}_2^T\underline{\mu}_2) + \frac{1}{\sigma^2}(\underline{\mu}_1 - \underline{\mu}_2)^T\underline{x}$$





and must be compared to a threshold $b = \log\frac{\pi(a_2)}{\pi(a_1)}$

We have already shown that when $\pi(a_2) = \pi(a_1) = \frac{1}{2}$ (Uniform priors) the decision is a minimum distance classifier.



This is easily seen by adding and subtracting to $l(\underline{x})$, $\frac{1}{2\sigma^2}\underline{x}^T\underline{x}$

$$l(\underline{x}) = -\frac{1}{2\sigma^2}\underline{x}^T\underline{x} - \frac{1}{2\sigma^2}\underline{\mu}_1^T\underline{\mu}_1 + \frac{1}{\sigma^2}\underline{\mu}_1^T\underline{x} + \frac{1}{2\sigma^2}\underline{x}^T\underline{x} + \frac{1}{2\sigma^2}\underline{\mu}_2^T\underline{\mu}_2 - \frac{1}{\sigma^2}\underline{\mu}_2^T\underline{x}$$

$$= -\frac{1}{2\sigma^2}(\underline{x}-\underline{\mu}_1)^T(\underline{x}-\underline{\mu}_1) + \frac{1}{2\sigma^2}(\underline{x}-\underline{\mu}_2)^T(\underline{x}-\underline{\mu}_2)$$

$$= -\frac{1}{2\sigma^2}\left(d^2(\underline{x},\underline{\mu}_1) - d^2(\underline{x},\underline{\mu}_2)\right)$$

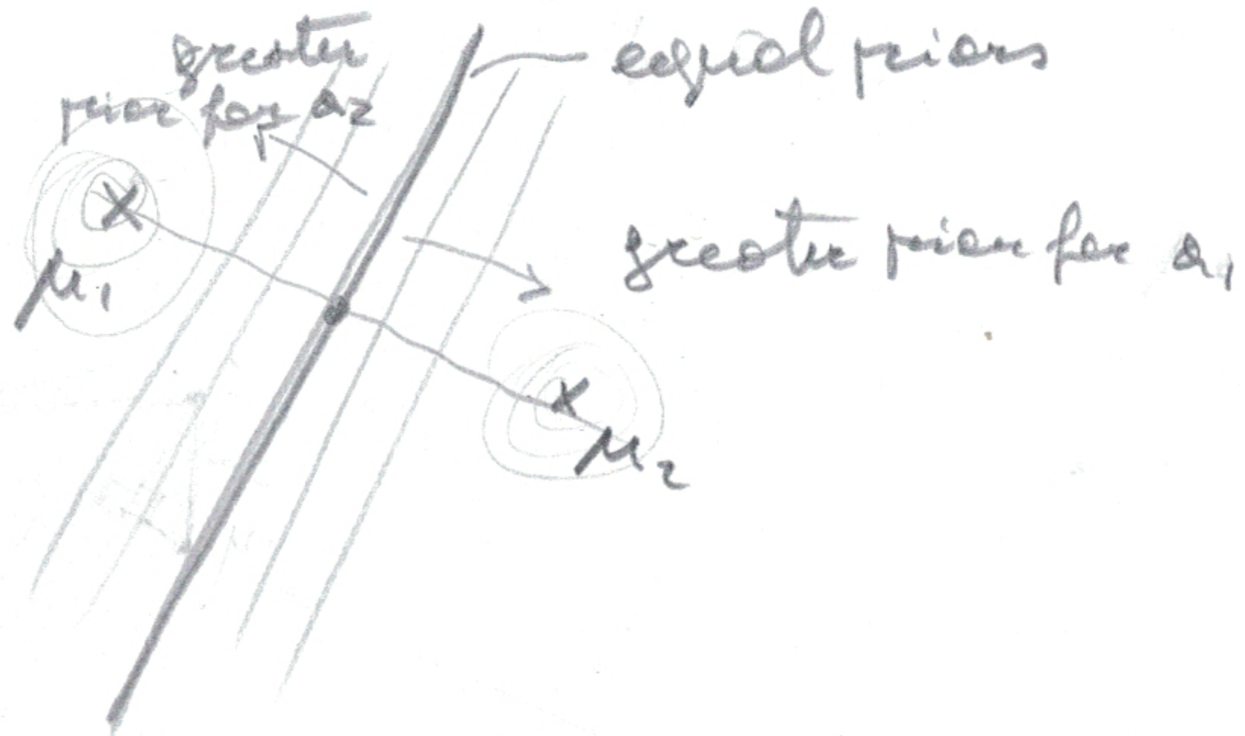
$$-\frac{1}{2\sigma^2}\left(d^2(\underline{x},\underline{\mu}_1) - d^2(\underline{x},\underline{\mu}_2)\right) \underset{a_2}{\overset{a_1}{\gtrless}} \log\frac{\pi(a_2)}{\pi(a_1)}$$

$$d^2(\underline{x},\underline{\mu}_1) - d^2(\underline{x},\underline{\mu}_2) \underset{a_1}{\overset{a_2}{\lessgtr}} 2\sigma^2 \log\frac{\pi(a_1)}{\pi(a_2)}$$

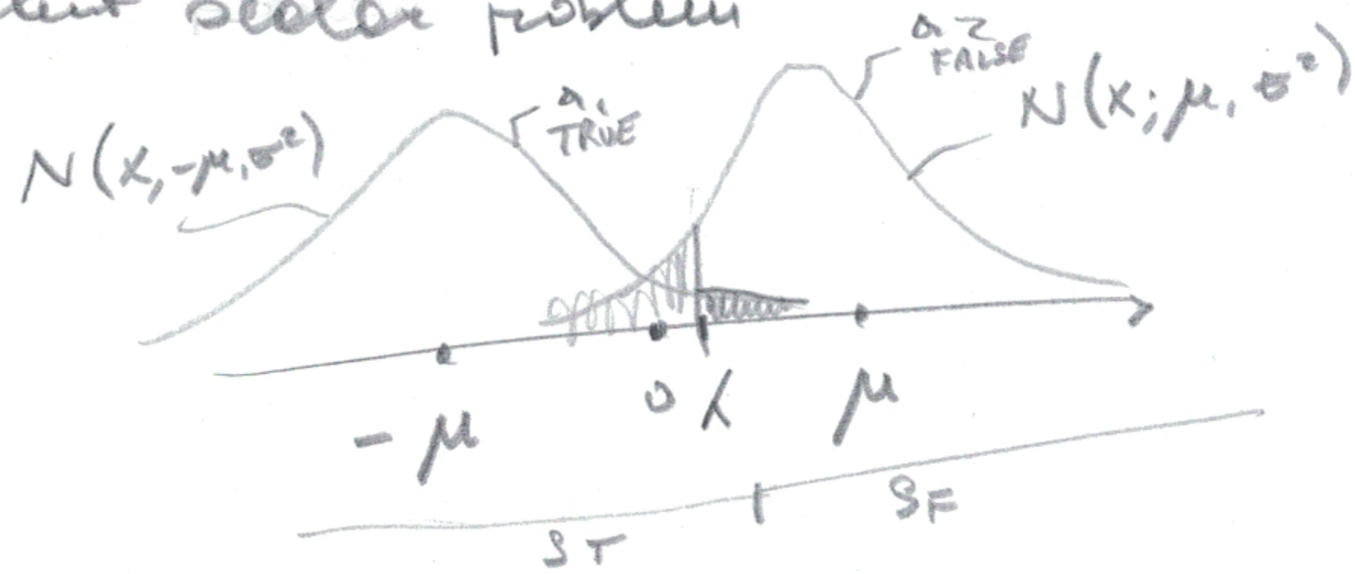
The equation of the separating hyperplane is

$$\frac{1}{\sigma^2}(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{x} - \frac{1}{2\sigma^2}(\underline{\mu}_1^T \underline{\mu}_1 - \underline{\mu}_2^T \underline{\mu}_2) + \log \frac{\pi(\Omega_1)}{\pi(\Omega_2)} = 0$$

We know that when $\log \frac{\pi(\Omega_1)}{\pi(\Omega_2)} = 0$ the separating hyperplane is crossing the line between $\underline{\mu}_1$ and $\underline{\mu}_2$ perpendicularly. Therefore for any threshold we have parallel hyperplanes



Since everything happens in one dimension (the line between $\underline{\mu}_1$ and $\underline{\mu}_2$) we can consider the equivalent scalar problem



$$\mu \triangleq \frac{d(\underline{\mu}_1, \underline{\mu}_2)}{2}$$

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

$$l(x) = -\frac{1}{2\sigma^2}(x^2 + \mu^2 + 2x\mu) + \frac{1}{2\sigma^2}(x^2 + \mu^2 - 2x\mu) \qquad \lambda$$

$$-\frac{2x\mu}{\sigma^2} \underset{\Omega_1}{\overset{\Omega_2}{\gtrless}} \log \frac{\pi(\Omega_2)}{\pi(\Omega_1)}$$

$$x \underset{\Omega_1}{\overset{\Omega_2}{\lessgtr}} \frac{\sigma^2}{2\mu} \log \frac{\pi(\Omega_1)}{\pi(\Omega_2)}$$

Now if a_1 = Positive a_2 = Negative

$$P_F\{\hat{S}=P \mid S=N\} = \int_{S_P} N(x;\mu,\sigma^2)\,dx = P_F\{x<\lambda \mid a_2\}$$

FALSE POSITIVE

P_{FA}

$$= Q\left(\frac{\lambda-\mu}{\sigma}\right)$$

$$P_F\{\hat{S}=N \mid S=P\} = \int_{S_D} N(x;-\mu,\sigma^2)\,dx = P_F\{x>\lambda \mid a_1\}$$

$P_{MD} = 1 - P_D$

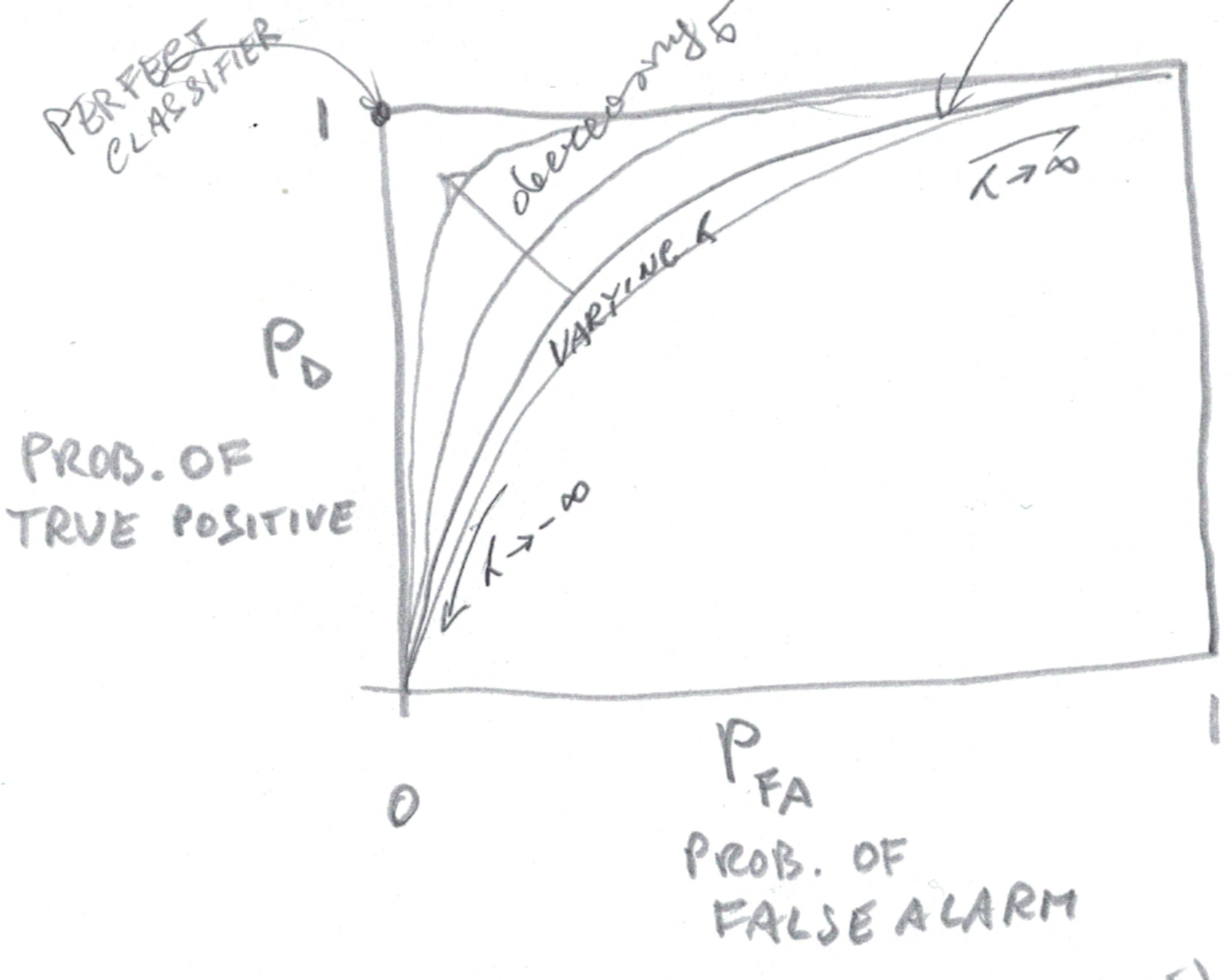
MISSED DETECTION

$$= Q\left(\frac{\lambda+\mu}{\sigma}\right)$$

We draw a diagram named ROC when λ is varied

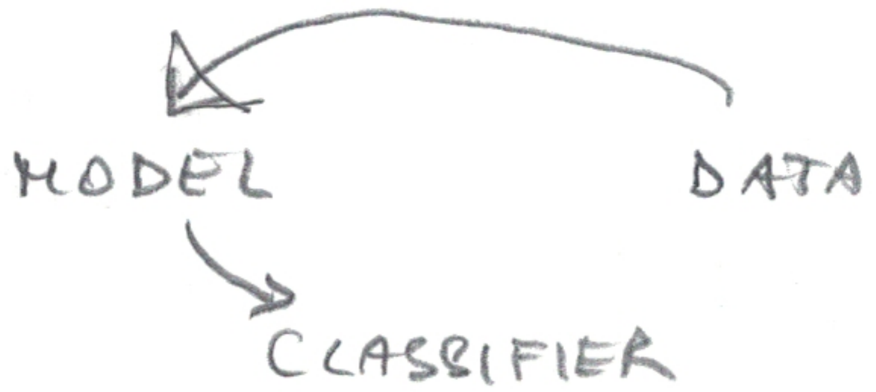
ROC

RECEIVER OPERATING CHARACTERISTIC



In practice we determine λ as a compromise between False positives and false negatives.

ESTIMATE PARAMETERS FROM DATA



In the classifier design using a model, the likelihoods and the priors need to be available. They can come from the physics of the problem, or be estimated from data.

Data is usually available as a set of pairs

$$\mathcal{L} = \{(\underline{x}[k], \underline{z}[k]), k = 1, \dots, L\}$$ TRAINING SET

Take as a main example the Gaussian mixture case. We need to estimate

$$\hat{\underline{\mu}}_1, \hat{\underline{\mu}}_2, \dots, \hat{\underline{\mu}}_M, \hat{\underline{\underline{\Sigma}}}_1, \hat{\underline{\underline{\Sigma}}}_2, \dots, \hat{\underline{\underline{\Sigma}}}_M, \hat{\pi}(a_1), \hat{\pi}(a_2), \dots \hat{\pi}(a_M)$$

The data set is partitioned in classes

$\{\underline{x}_1[k], k = 1, \dots, L_1\}$ Data vectors labeled as a_1

$\{\underline{x}_2[k], k = 1, \dots, L_2\}$ Data ... ... as a_2

$\{\underline{x}_M[k], k = 1, \dots, L_M\}$ Data ... ... as a_M

Straightforward averages give

$$\hat{\underline{\mu}}_i = \frac{1}{L_i} \sum_{k=1}^{L_i} \underline{x}_i[k] \quad ; \quad \hat{\underline{\underline{\Sigma}}}_i = \frac{1}{L_i} \sum_{k=1}^{L_i} (\underline{x}_i[k] - \hat{\underline{\mu}}_i)(\underline{x}_i[k] - \hat{\underline{\mu}}_i)^T$$

Where $i = 1, \dots, M$

COVARIANCES

(sometimes we use a normalizer $\frac{1}{L_i - 1}$, that gives more unbiased estimate ⇒ considered in this 90 by our the purpose of these no

Note that the averages (sample means) may be computed on sets of different sizes. This corresponds to different accuracies for the various likelihoods.

There is a large literature on statistics on how accurate these estimates may be if the "true" generative model is gaussian.

We prefer to leave these considerations apart of these notes and concentrate on the intuitive straightforward approach.

Estimation of the priors, may be tricky because a poor estimation may provoke a heavy effect on our classifier.

If we believe that the frequency of occurrence of the various classes in our dataset reflects the true class frequencies, then clearly

$$\hat{\pi}(a_i) = \frac{L_i}{L} \qquad i = 1, \dots, M$$

Otherwise other considerations should drive our prior choices. Common assumption is to take the priors to be uniform $\hat{\pi}(a_i) = \frac{1}{M}$ $i = 1, \dots, M$. (ML)

There is also here a large literature in what is the most appropriate (for example Bayesian inferential) prior choice. We leave this, some out too for simplicity.

NON-PARAMETRIC LIKELIHOODS

Labeled dataset vectors can be used also directly to represent likelihoods.



A vector $\underline{x}$ can be classified with minimum distance by computing the average distance from all the data points for each cluster.

$$D(\underline{x};i) = \frac{1}{L_i} \sum_{l=1}^{L_i} d(\underline{x};\underline{x}_{i[l]}) \qquad i = 1,\dots,M$$

and decide for the class that gives the smallest Distance.

The implicit assumption here is that the likelihoods are spherical and have the same variance (dispersion). Modifications can be introduced via class-normalizations and global rotation of the observation space.

The method can be computationally expensive because we may have to compute a large number of distances and memorize many patterns.

Usually it is preferred to use parametric models that correspond to simpler and clearer set of assumptions.