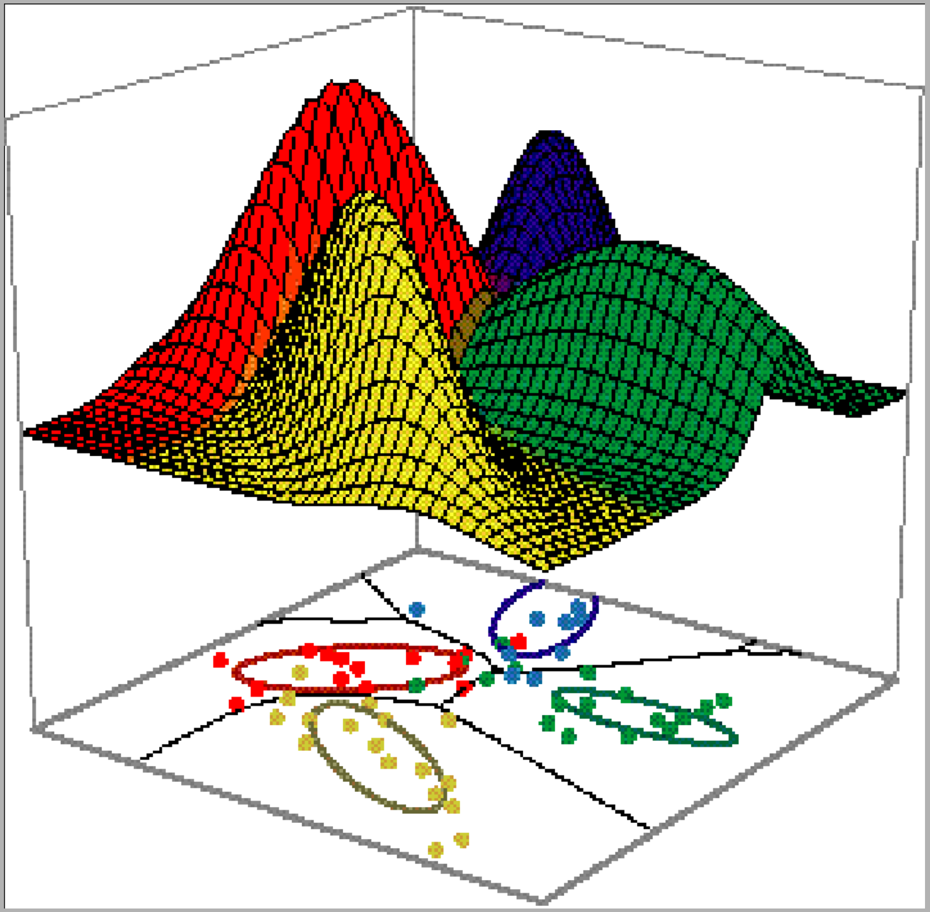


# Pattern Recognition (Αναγνώριση Προτύπων)

## Bayesian Decision Theory
(Μπεϋζιανή Θεωρία Αποφάσεων)


Panos Trahanias

UNIVERSITY OF CRETE
DEPARTMENT of COMPUTER SCIENCE

# *Bayes Decision theory*

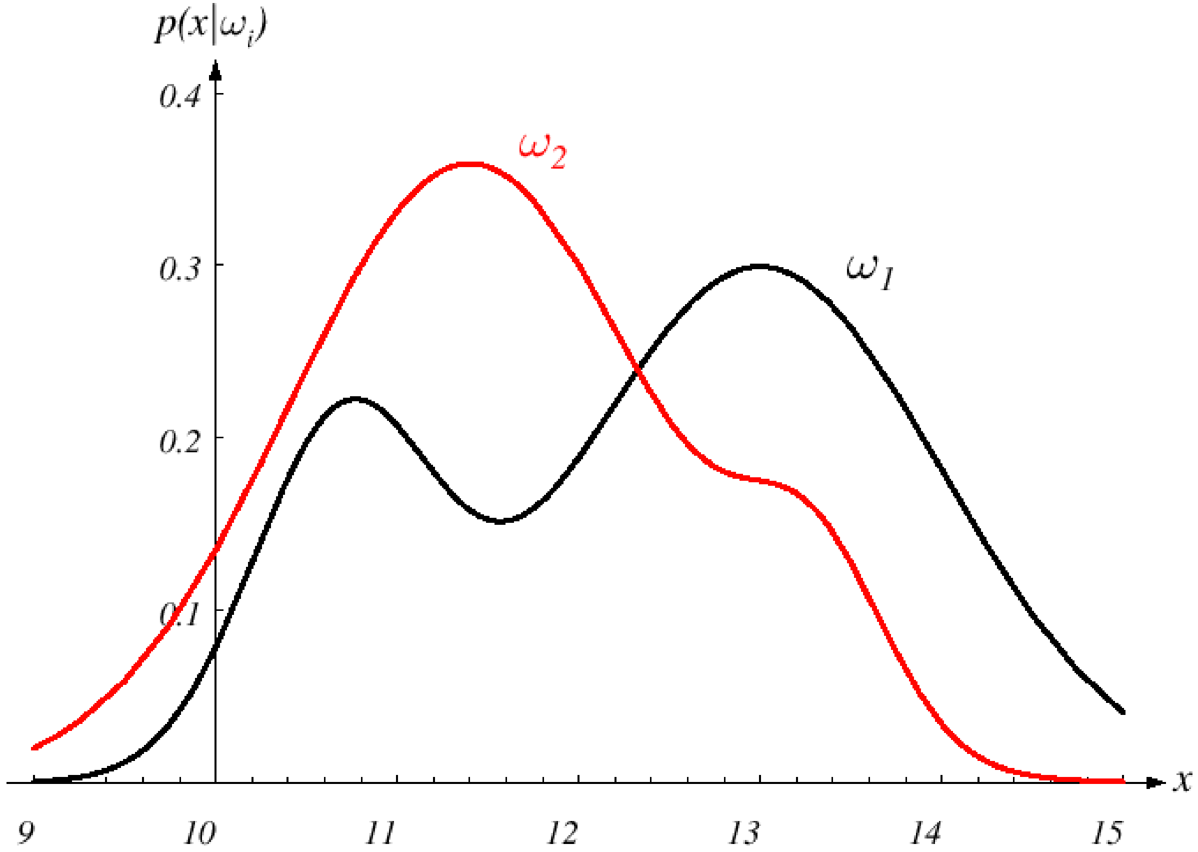
- Statistically optimal classification.
- Based on the probabilistic description of the classification problem/task.
- It assumes:
  - Classification problem can be probabilistically stated
  - Relevant values and probability functions are known (not valid in practice).

# Class Conditional Probabilities

$\omega_1$: Sea bass

$\omega_2$: Salmon

- $p(x \mid \omega_2)$: Conditional Probability Density Function (PDF) of the variable x given the state of nature.
- Likelihood: Given that salmon is observed, what is the probability that its weight is between 11 and 12?




weight

# Definitions and Bayes Decision Rule

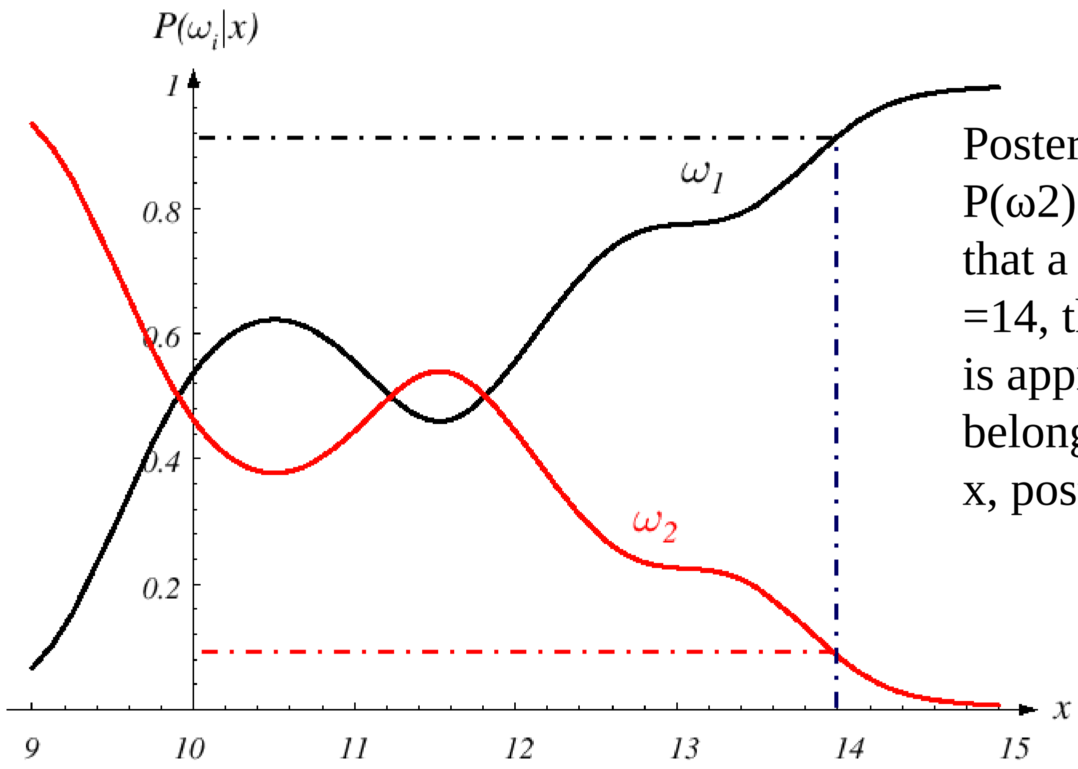
- State of nature (Κατάσταση της φύσης)
- Prior (Εκ των προτέρων πιθανότητα)
- Posterior (Εκ των υστέρων πιθανότητα)
- Likelihood (Πιθανοφάνεια)
- Evidence

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) \cdot P(\omega_j)}{\sum_{x \in X} p(x \mid \omega_j) \cdot P(\omega_j)}$$

$$= \frac{p(x \mid \omega_j) \cdot P(\omega_j)}{p(x)}$$

# Posterior Probabilities

Bayes rule facilitates estimation/computation of posterior probabilities (otherwise hard to compute), given prior probability, likelihood and evidence.




Posterior probabilities when $P(\omega1) = 2/3$ and $P(\omega2) = 1/3$. Given e.g. that a pattern is measured with feature value x =14, the probability that it belongs to class ω2 is approx. 0.08, whereas the probability that it belongs to ω1 is 0.92. For each x, posterior probabilities sum up to 1.0.

# Selection of the class that has the Highest posterior probability!!!

Choose $\omega_i$ if $P(\omega_i | x) > P(\omega_j | x)$ for all $i = 1,2,\ldots,c$

$P(error) = \min [\ P(\omega_1 | x), P(\omega_2 | x), \ldots, P(\omega_c | x)\ ]$

In case of multiple features, $\mathbf{x}=\{x_1, x_2,\ldots, x_d\}$ then

Choose $\omega_i$ if $P(\omega_i | \boldsymbol{x}) > P(\omega_j | \boldsymbol{x})$ for all $i = 1,2,\ldots,c$

$P(error) = \min [\ P(\omega_1 | \boldsymbol{x}), P(\omega_2 | \boldsymbol{x}), \ldots, P(\omega_c | \boldsymbol{x})\ ]$

# The Loss Function

- Mathematical description of the cost of each choice.
- Are some choices more "expensive" than others?

✔ $\{\omega_1, \omega_2, \ldots, \omega_c\}$: Set of physical states (classes)

✔ $\mathbf{x} = [x_1, \ldots, x_d]^T$: Feature vector

✔ $\{\alpha_1, \alpha_2, \ldots\ \alpha_a\}$: Set of actions. Note that 'a' does not have to be the same as 'c', as we can perform more or less actions than the number of classes. For example, rejection is also a possible action.

✔ $\lambda(\alpha_i | \omega_j\}$: Cost (κόστος) of the action $\alpha_i$ when the real class is $\omega_j$.

✔ $R(\alpha_i | \mathbf{x})$: conditional risk – Expected loss for action $\alpha_i$.

$$\boxed{R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})}$$

Bayes decision chooses the action that minimizes the conditional risk!

# *Bayes decision based on the conditional risk*

1. Calculation of conditional risk $R(\alpha_i | \mathbf{x})$ for each action.
2. Selection of the action with the lowest conditional risk. Suppose that it is action k
3. The overall risk is:

$$R = \int_{\mathbf{x} \in X} R(\alpha_k \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x}$$

4. This is the Bayesian risk, the <u>lowest possible risk</u> that any classifier can have!
5. E.g. classification into one of two classes:

$$\frac{p(x / \omega_1)}{p(x / \omega_2)} \overset{\omega_1}{\geq} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

# *Minimum Error Rate Classification*

- If as action $\alpha i$ we choose the classification in the class $\omega j$, and if all the costs of incorrect classification are equal to one, we have the so-called symmetric or 0-1 selection:

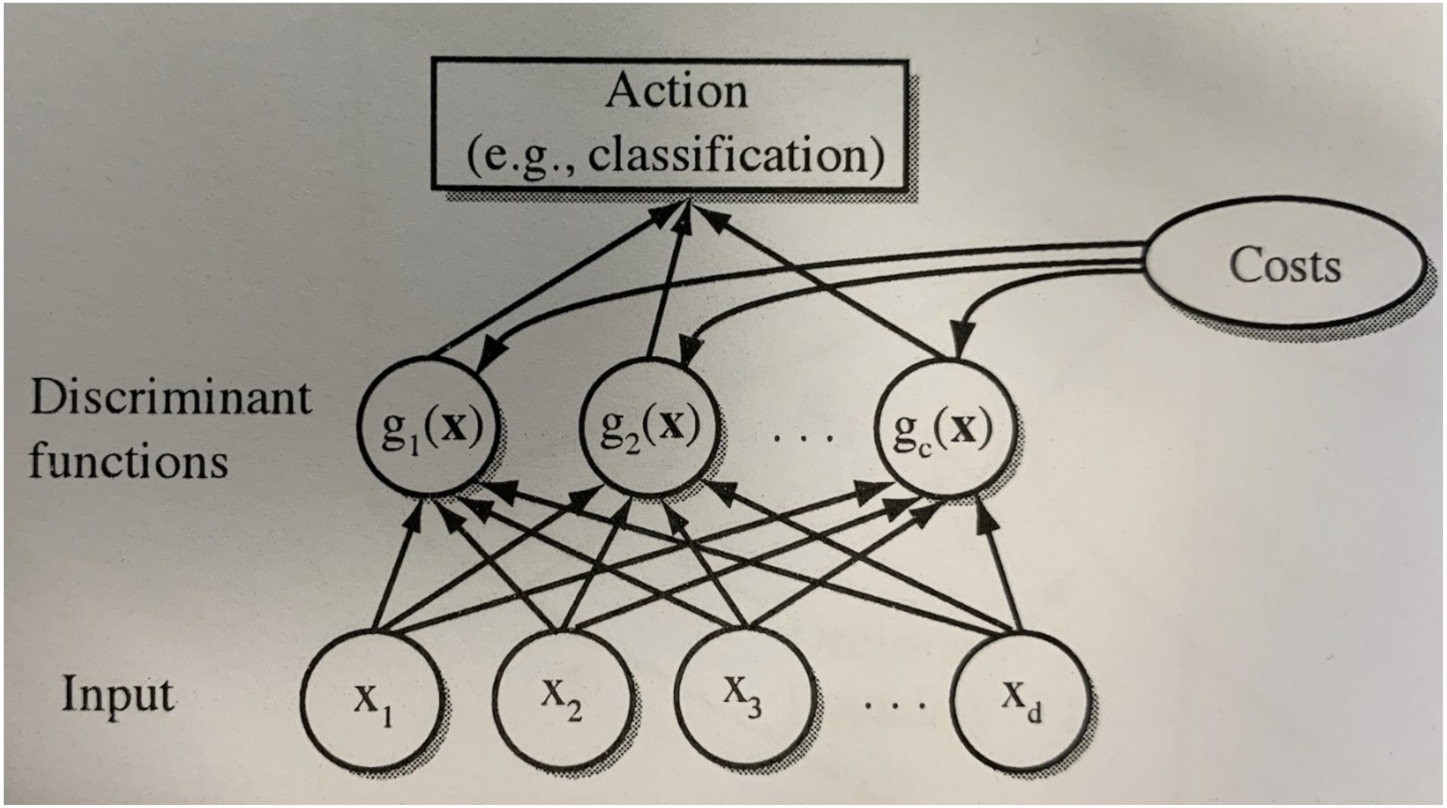
$$\lambda\left(\alpha_i \mid \omega_j\right)=\begin{cases}0, & \text{if } i=j \\ 1, & \text{if } i\neq j\end{cases}$$

- This cost function stipulates zero loss for correct classification, and unit loss for incorrect classification. The corresponding conditional risk corresponding to this cost function is:

$$R(\alpha_i \mid \mathbf{x}) = \sum_{\substack{j\neq i \\ j=1,\ldots,c}} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_j \mid \mathbf{x})$$

which is exactly the probability of error. Obviously, to minimize the risk, we should select the class that maximizes the posterior probability!!!

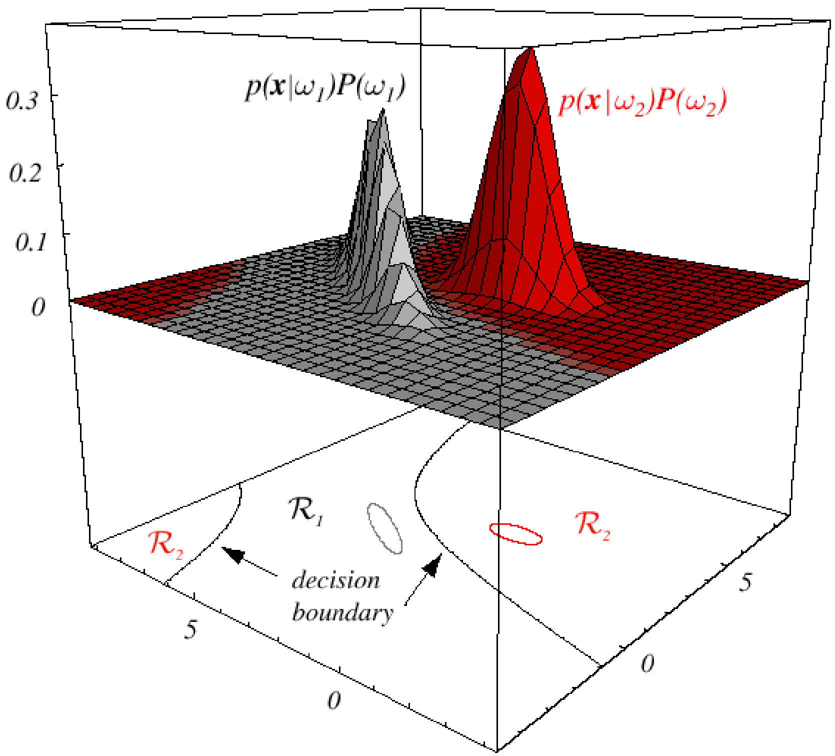
# Discriminant Based Classification



# Discriminant Based Classification

- The discrimination function g(x), separates the classes from each other. This function corresponds the input vector to a class according to the definition: Select class *i* if : $\boxed{g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall i \neq j, \quad i, j = 1,2,\ldots,c}$

- The Bayes rule can be implemented in the form of discrimination functions $$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$$




Any discrimination function creates c decision areas, R1,…, Rc, which are separated by decision surfaces (*επιφάνειες απόφασης*).

Decision areas are not required to be continuous.

Decision surfaces satisfy the:
$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$$

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i),$$

# Gaussian PDF

- If the likelihood functions follow the multidimensional Gaussian, then the discrimination function takes the form

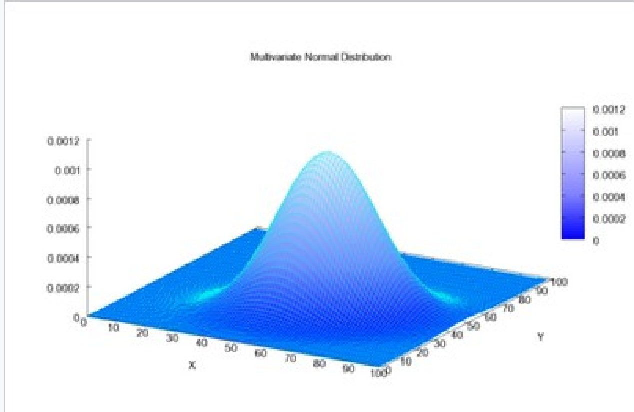
$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}\left[(\mathbf{x}-\mu_i)^T \cdot \Sigma_i^{-1} \cdot (\mathbf{x}-\mu_i)\right] - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i)}$$

- There are 3 cases depending on the covariance matrix

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}):$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

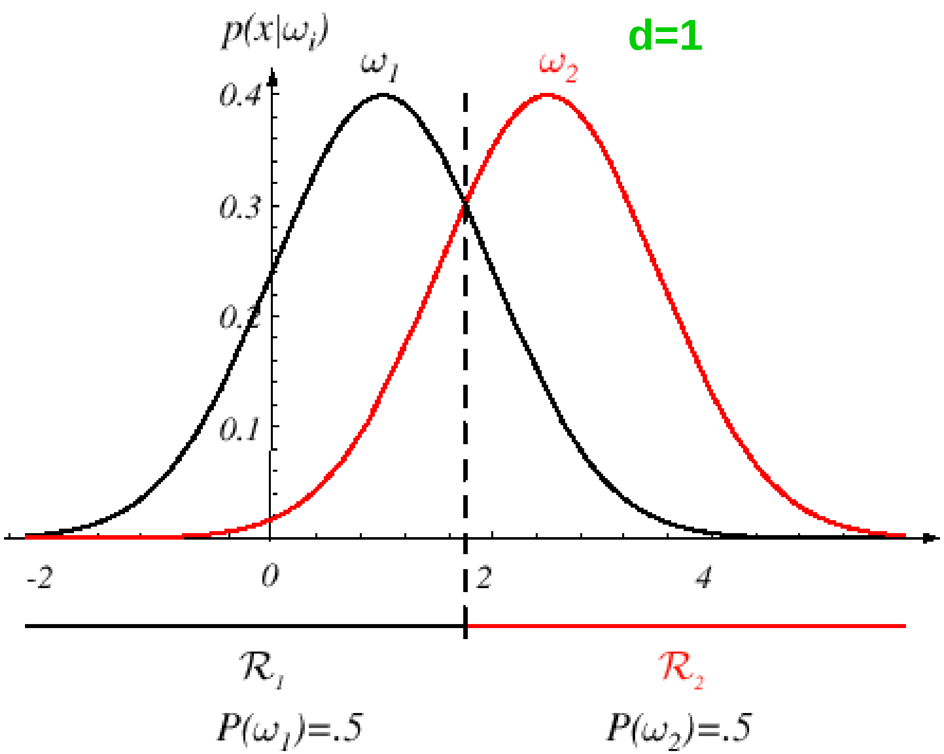
Multivariate

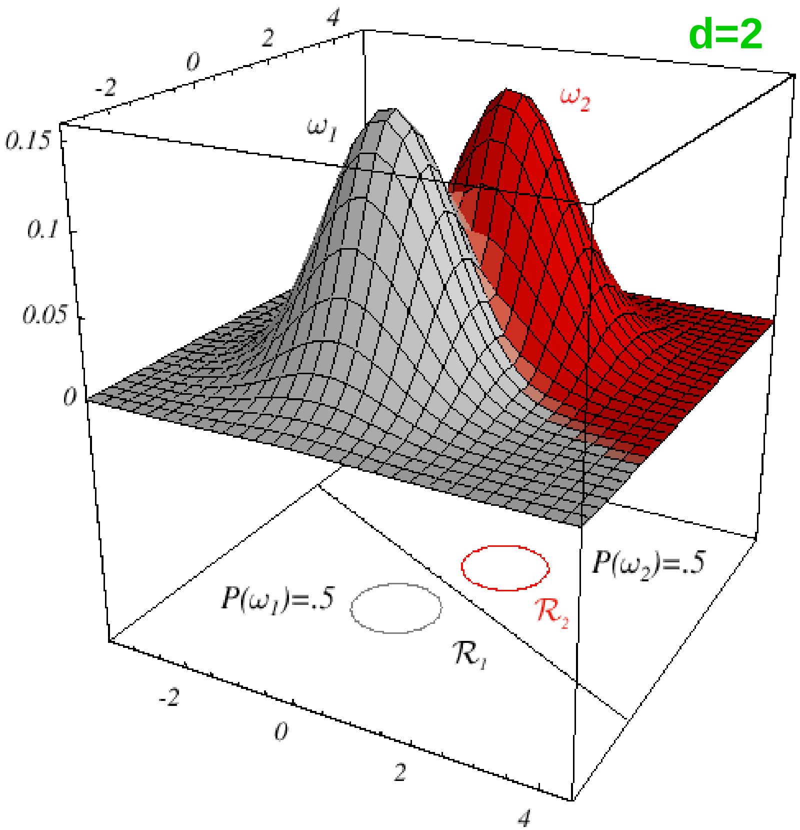



Bivariate normal joint density

# Case 1: $\Sigma_i = \sigma^2 I$

- The features are statistically independent, and all have the same covariance: The samples are in super-spheres of equal size, and the decision surfaces are **super-planes** of dimension d-1.



# Case 1: $\Sigma_i = \sigma^2 I$

**Covariance Matrix**

$$S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{bmatrix}$$

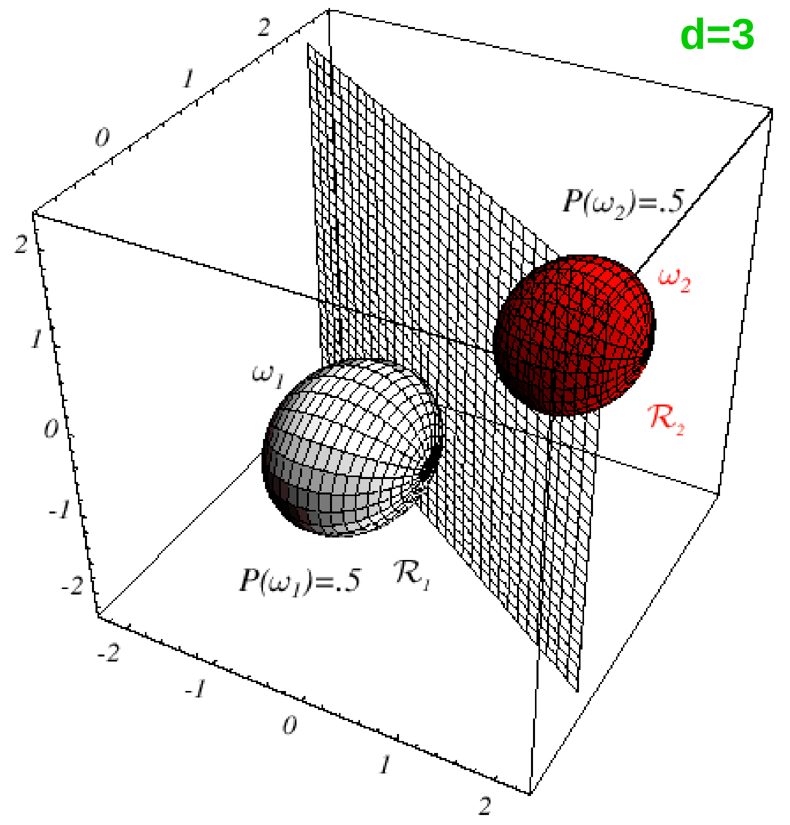
where $s_j^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ : variance of $j^{th}$ variable

$s_{jk} = (1/n)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ : covariance b/w $j^{th}$ & $k^{th}$ variables

$\bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$ : mean of $j^{th}$ variable

# Case 1: $\Sigma_i = \sigma^2 I$

- When d = 3, the samples are in spheres of equal size, and the decision surfaces are **flat**.

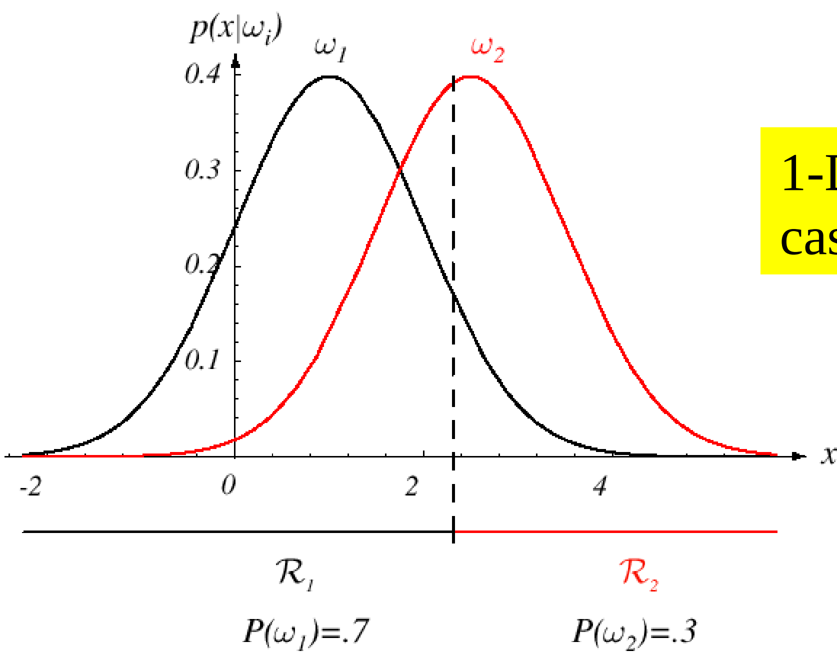


# *Case 1:* $\Sigma_i = \sigma^2 I$

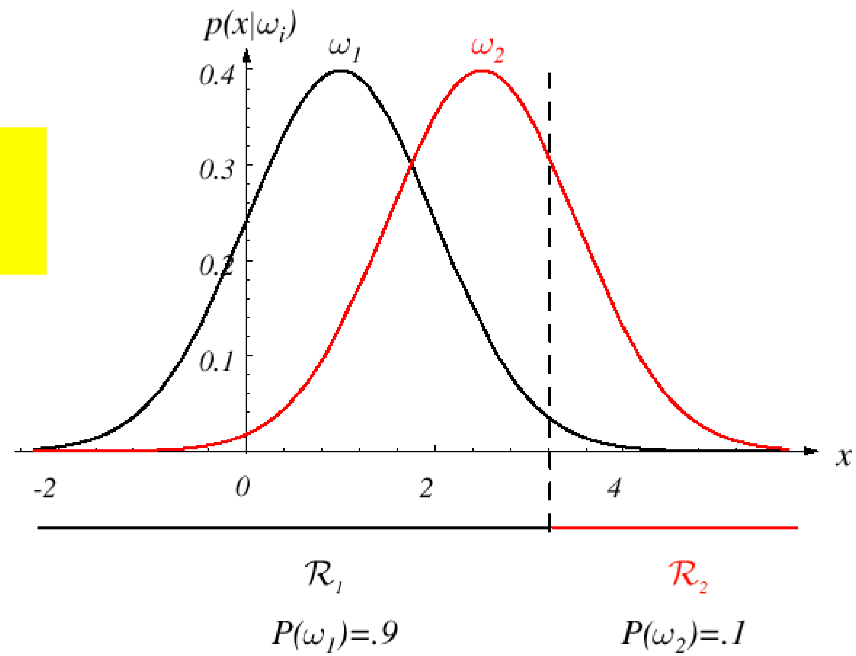
- This case creates linear discriminant functions:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i, \quad w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^T \cdot \boldsymbol{\mu}_i + \ln P(\omega_i)$$




1-D case




**Notice that the prior probabilities move the threshold away from the most probable mean.**

## Case 1: $\Sigma_i = \sigma^2 I$

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[(\mathbf{x}-\mu_i)^T \cdot \Sigma_i^{-1} \cdot (\mathbf{x}-\mu_i)\right] - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(\omega_i),$$

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$$

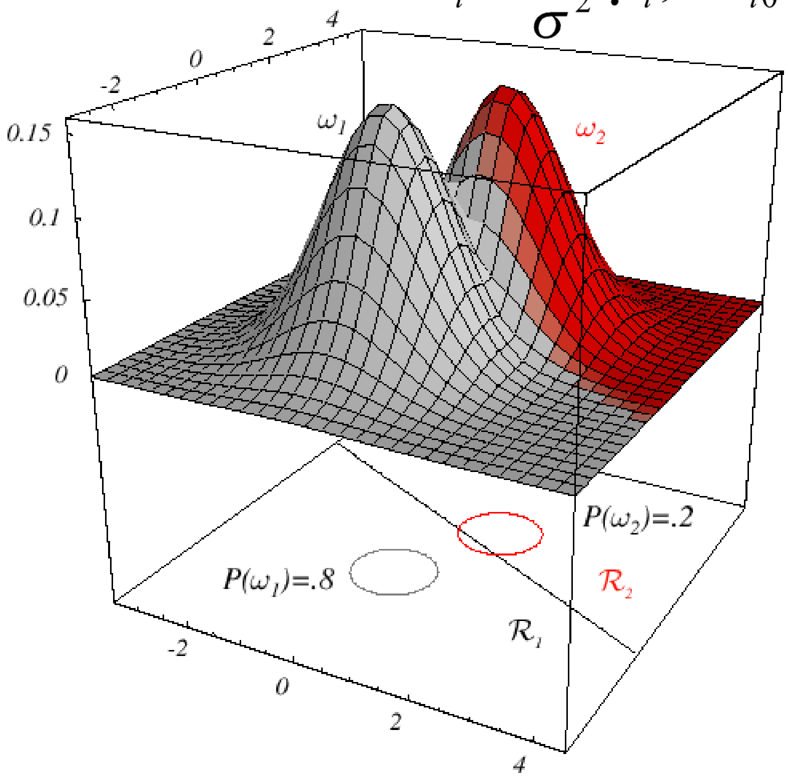
$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i).$$

## Case 1: $\Sigma_i = \sigma^2 I$

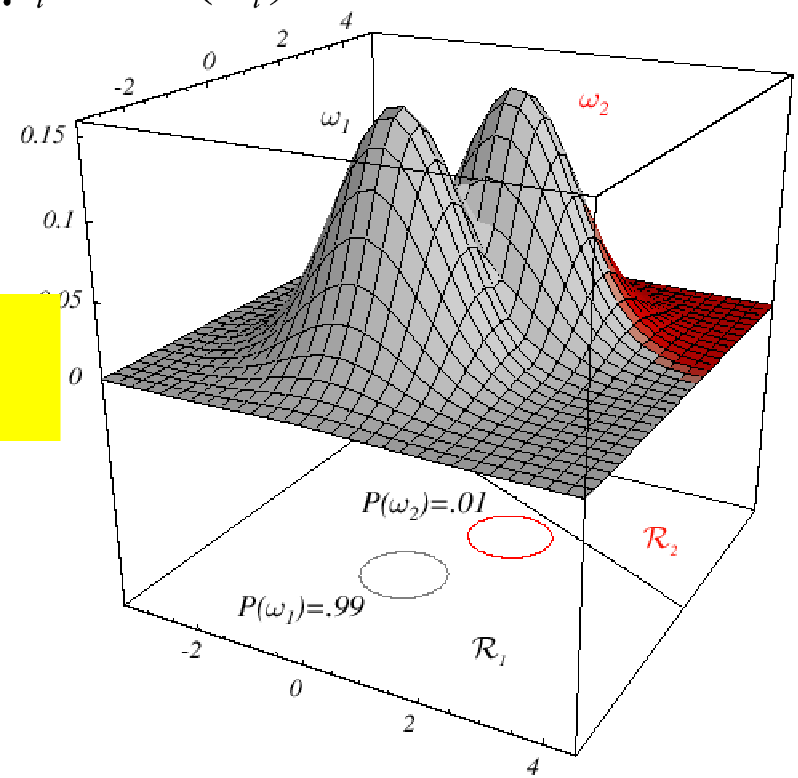
- This case creates linear discrimination functions:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i, \quad w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^T \cdot \boldsymbol{\mu}_i + \ln P(\omega_i)$$




2-D case



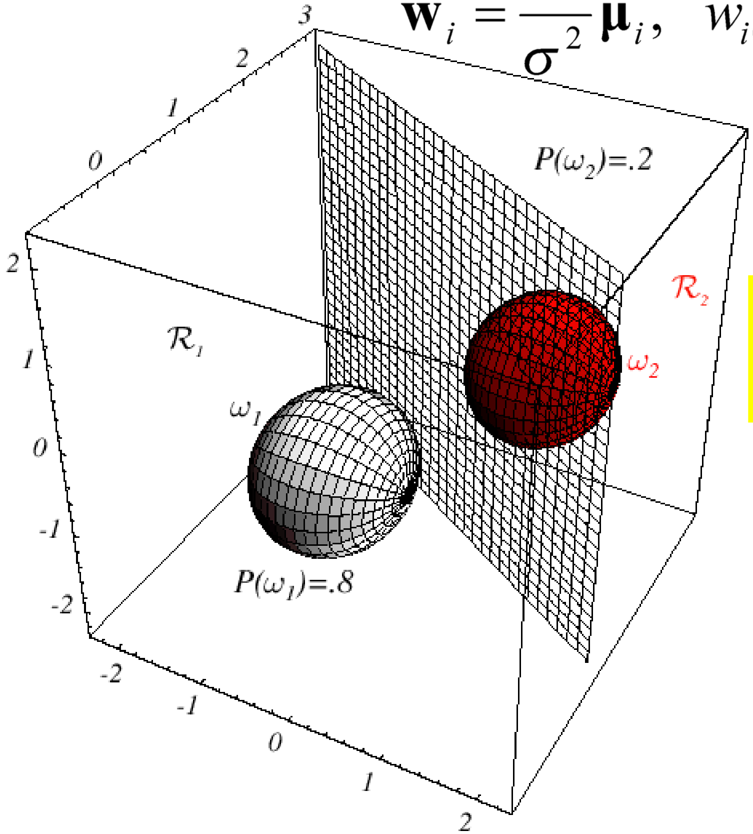

**Notice that the prior probabilities move the decision line away from the most probable mean.**

# Case 1: $\Sigma_i = \sigma^2 I$

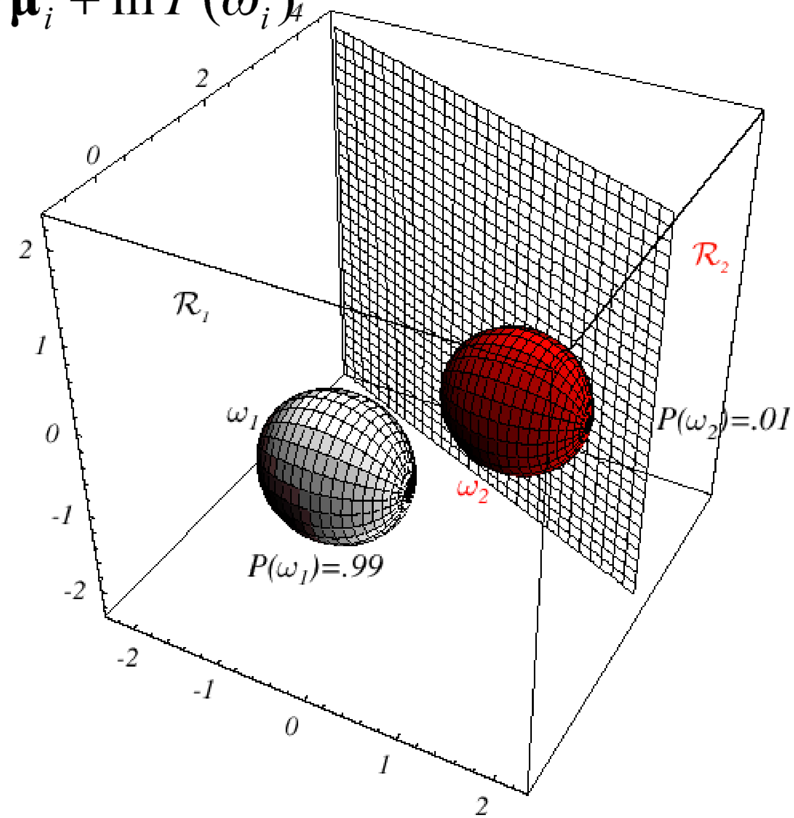
➢ This case creates linear discrimination functions:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i, \quad w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^T \cdot \boldsymbol{\mu}_i + \ln P(\omega_i)$$




3-D case

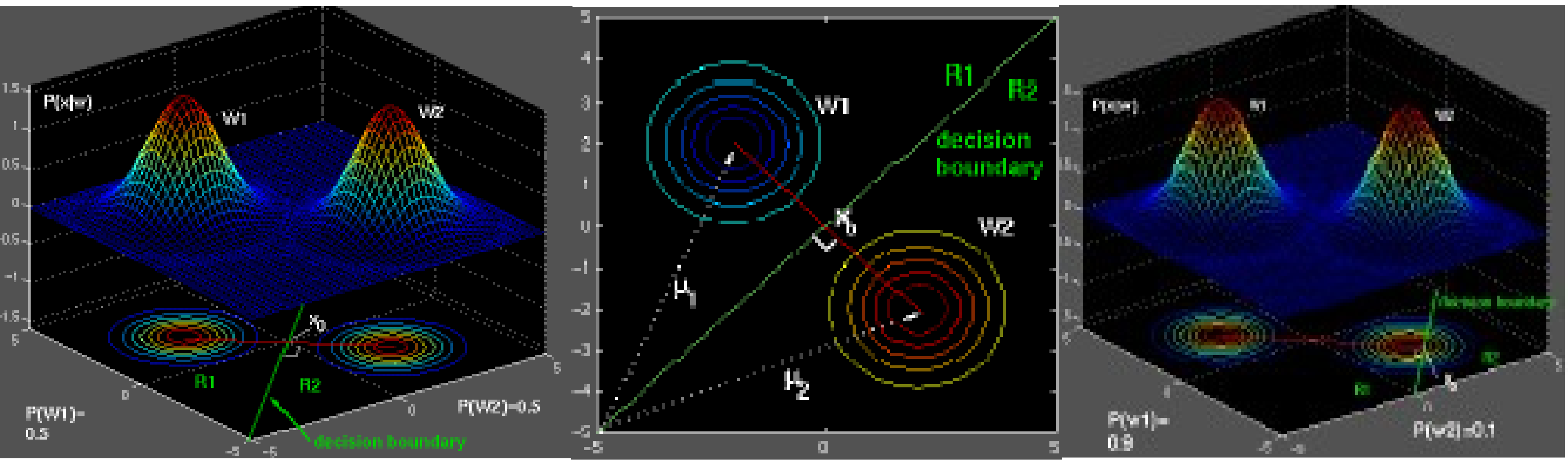



**Notice that the prior probabilities move the decision plane away from the most probable mean.**

# Case 1: $\Sigma_i = \sigma^2 I$

The decision surfaces are hyper-planes defined by the linear equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$, written as $\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)=0$ where:

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j, \quad \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln\left(\frac{P(\omega_i)}{P(\omega_{ji})}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$




Decision Surface: Hyper-plane that passes through the point $\boldsymbol{x}_0$ and is perpendicular to the vector $\boldsymbol{w}$ that connects the mean values $\boldsymbol{\mu}_i$ και $\boldsymbol{\mu}_j$.

# Case 1: $\Sigma_i = \sigma^2 I$

$$g_i(x) = g_j(x)$$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0,$$

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

# Case 2: $\Sigma_i = \Sigma$

- The covariance matrices are arbitrary, but the same for all classes.

The features create hyper-elliptical groups of the same size and shape with centers $\boldsymbol{\mu}_i$.

- Linear decision functions → Hyper-planes as decision surfaces

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}\left[(\mathbf{x}-\mu_i)^T \cdot \Sigma_i^{-1} \cdot (\mathbf{x}-\mu_i)\right] - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i)}$$

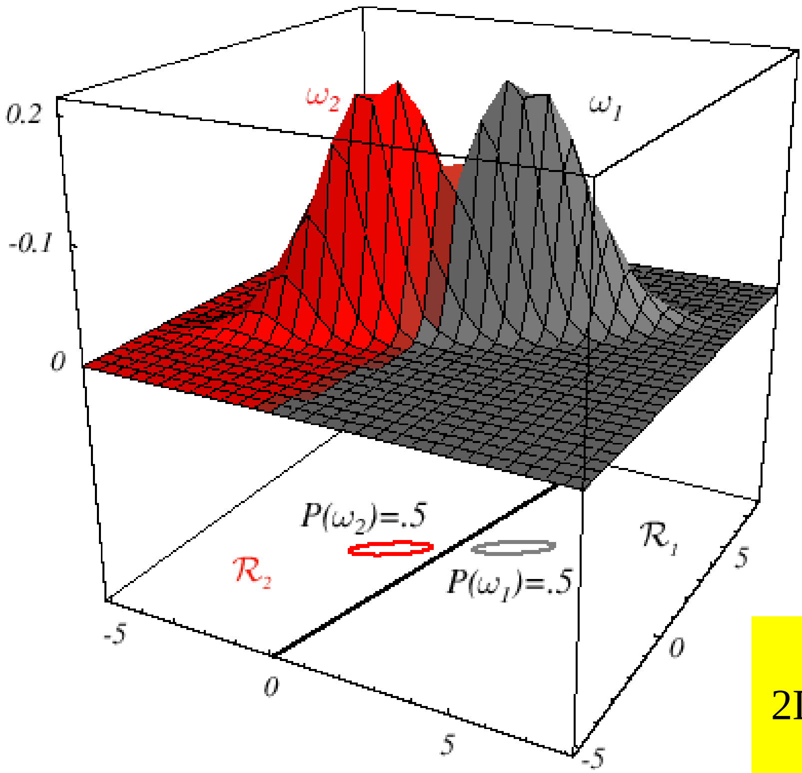
$$\Downarrow$$

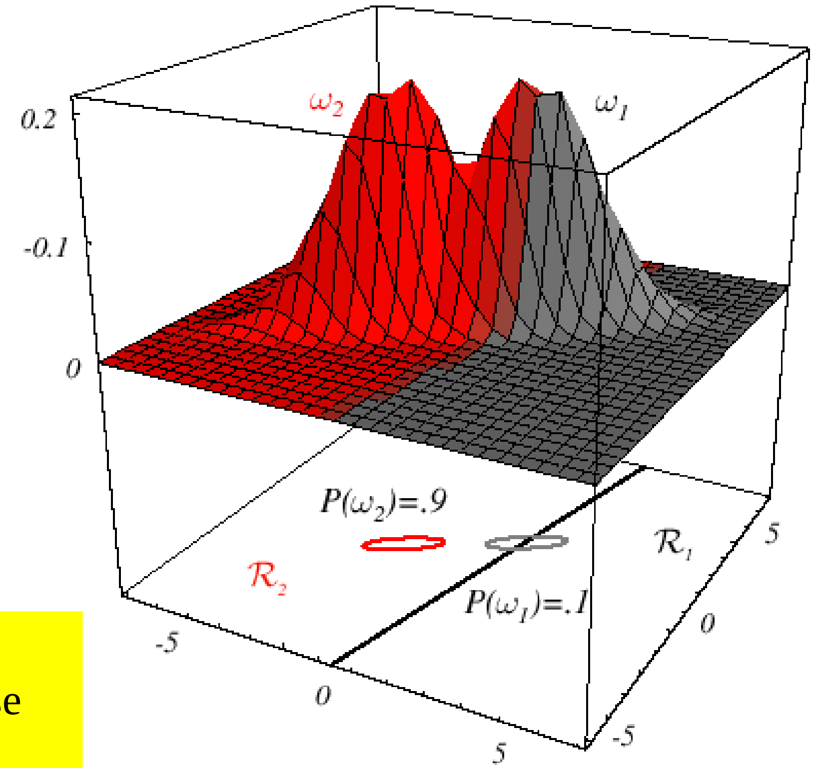
$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}\underbrace{\left[(\mathbf{x}-\mu_i)^T \cdot \Sigma_i^{-1} \cdot (\mathbf{x}-\mu_i)\right]}_{\text{Squared Mahalanobis distance}} + \ln P(\omega_i)}$$

# Case 2: $\Sigma_i = \Sigma$

➢The covariance matrices are arbitrary, but the same for all classes.
The features create hyper-elliptical groups of the same size and shape with centers $\boldsymbol{\mu}_i$.
➢Linear decision functions → Hyper-planes as decision surfaces

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \qquad \mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i, \quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$



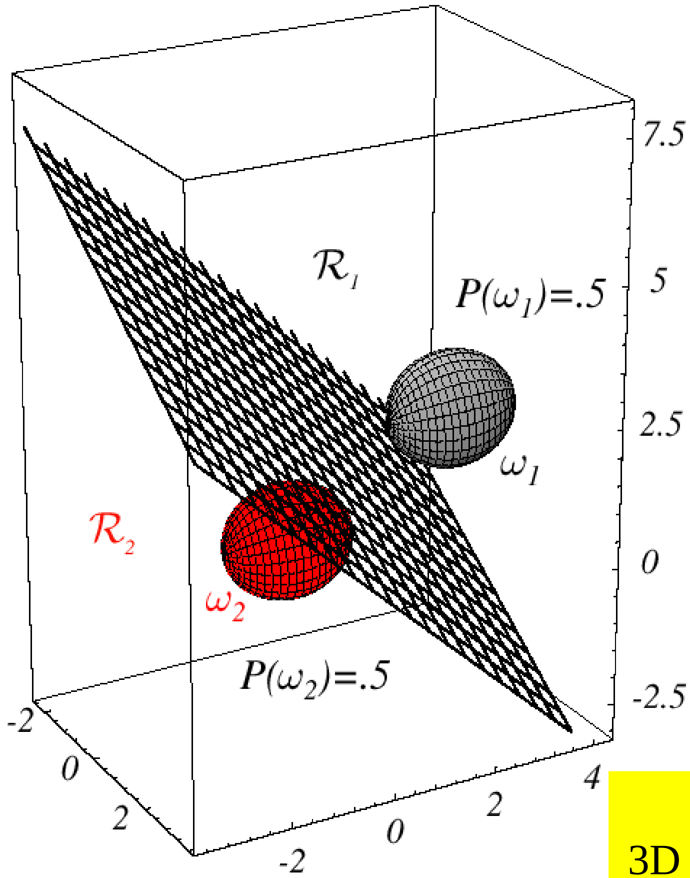



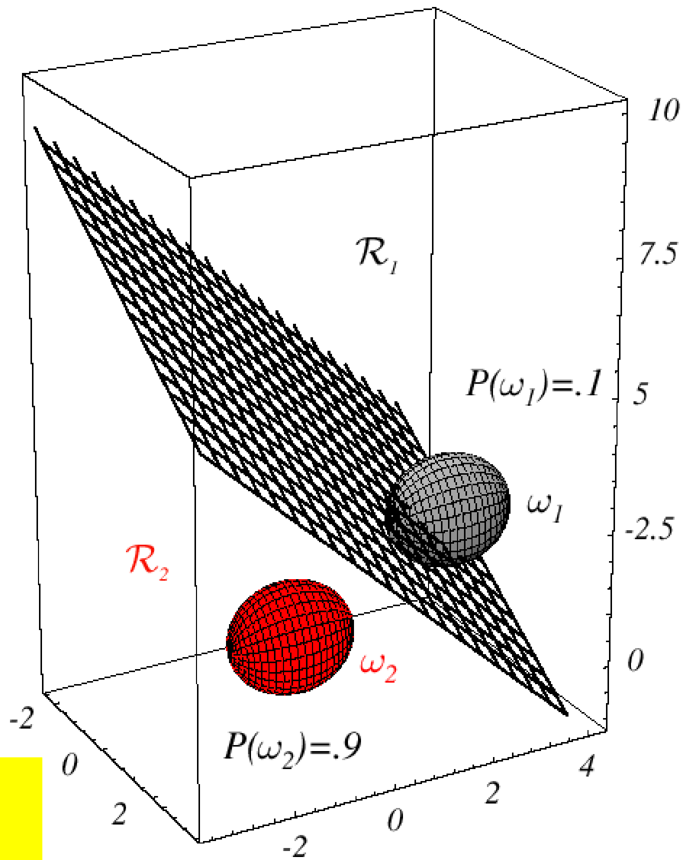

2D Case

# Case 2: $\Sigma_i = \Sigma$

➢The covariance matrices are arbitrary, but the same for all classes.
The features create hyper-elliptical groups of the same size and shape with centers $\boldsymbol{\mu}_i$.

➢Linear decision functions $\rightarrow$ Hyper-planes as decision surfaces



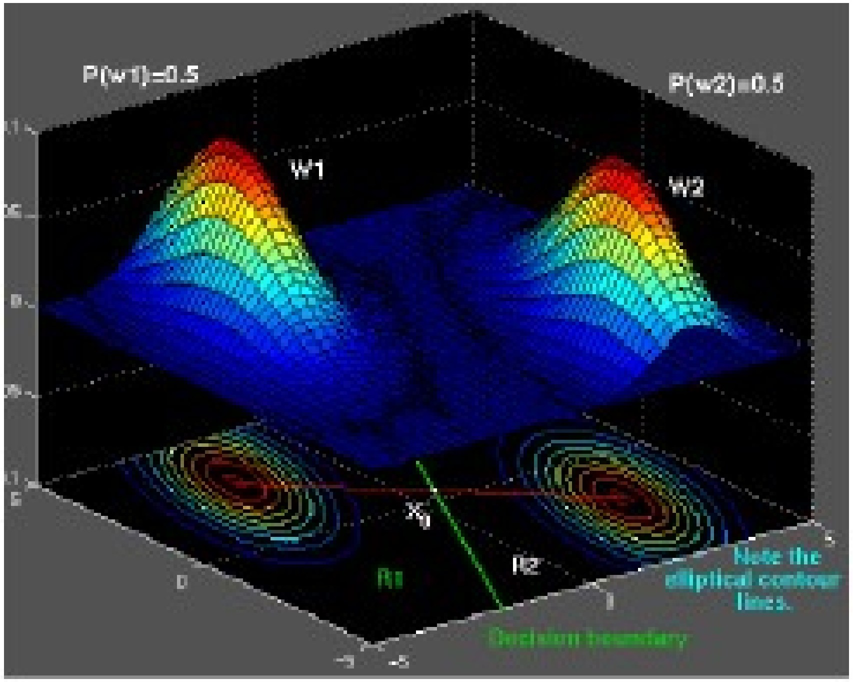



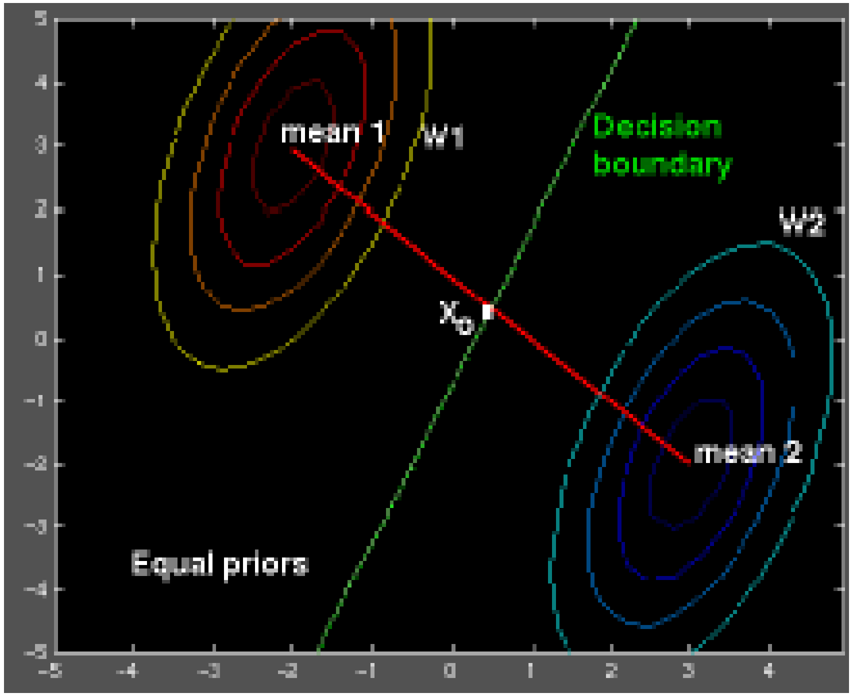

3D Case

# Case 2: $\Sigma_i = \Sigma$

- Decision surfaces $\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)=0$ where:

$$\mathbf{w}=\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j),\quad \mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\left(P(\omega_i)/P(\omega_j)\right)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$




Since $\mathbf{w}=\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$, the decision hyper-plane is **not** perpendicular to the vector **w** that connects the mean values $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$.

# Case 2: $\Sigma_i = \Sigma$

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i).} \tag{57}$$

... are the same for all $c$ classes, then the ln $P(\omega_i)$

$$\boxed{g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0},} \tag{58}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \tag{59}$$

$$\boxed{w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i).} \tag{60}$$

the discriminants are linear, the resulting decision boundaries are again
es (Fig. 2.10). If $\mathcal{R}_i$ and $\mathcal{R}_j$ are contiguous, the boundary between them
quation

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0, \tag{61}$$

$$\boxed{\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \tag{62}$$

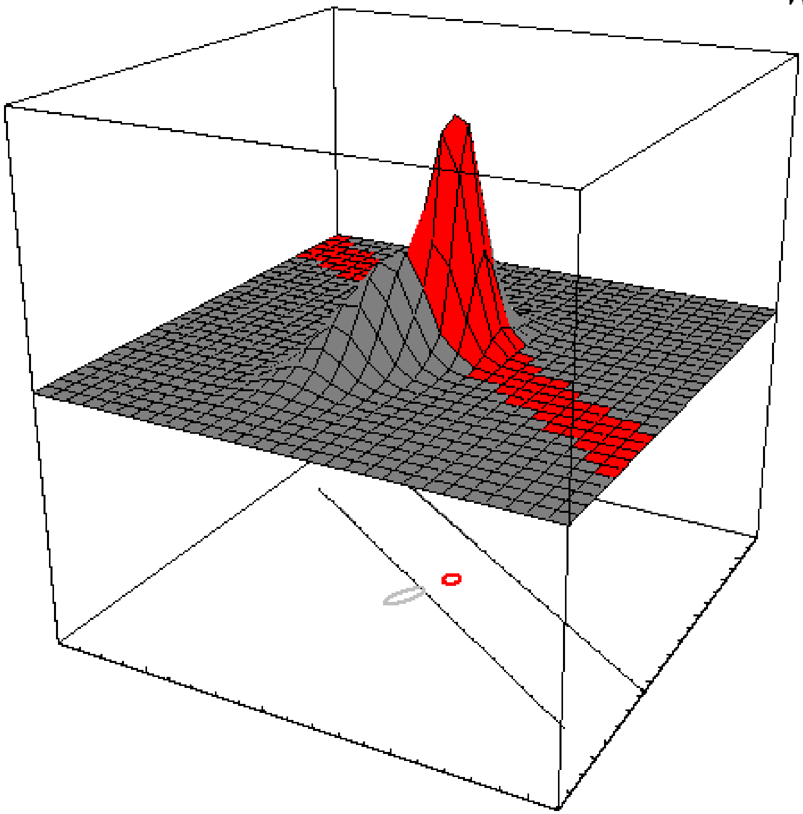
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\,[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{63}$$

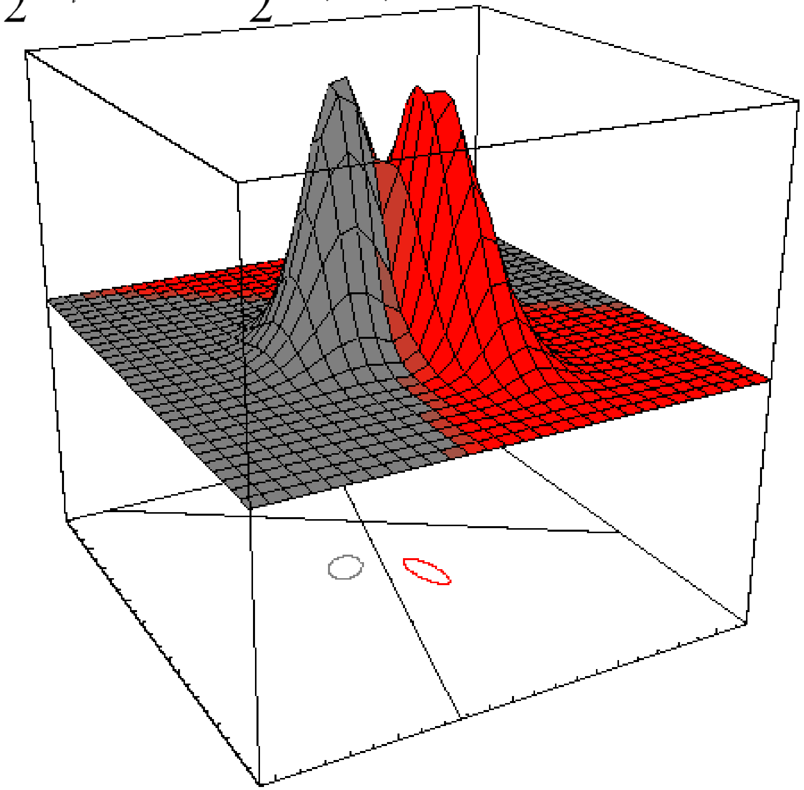
# Case 3: $\Sigma_i$ = any

- Non-linear but <u>squared</u> decision functions.
- Decision surfaces *hyperquadratics (hyper-elliptical, hyper-paraboloid, etc.).*

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \qquad \mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}, \quad \mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

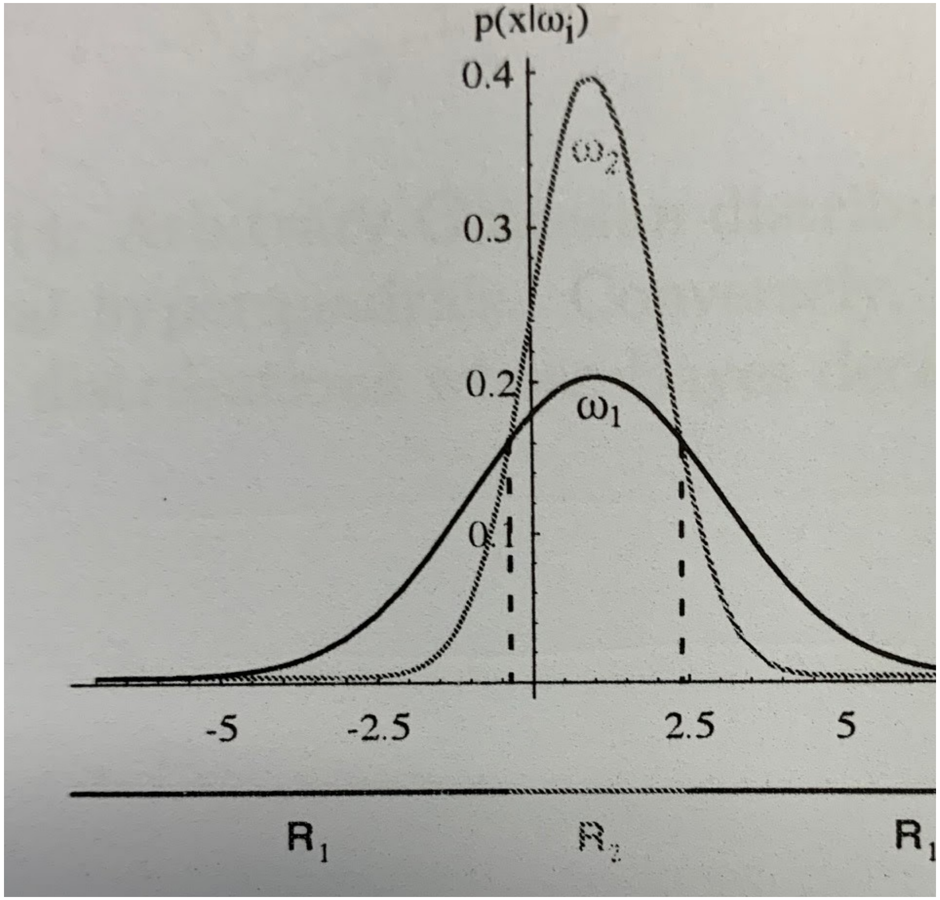
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\mathbf{\Sigma}_i\right| + \ln P(\omega_i)$$





# *Case 3:* $\Sigma_i = \text{any}$

- Non-simply connected decision regions can arise in one dimension for Gaussians having unequal variance



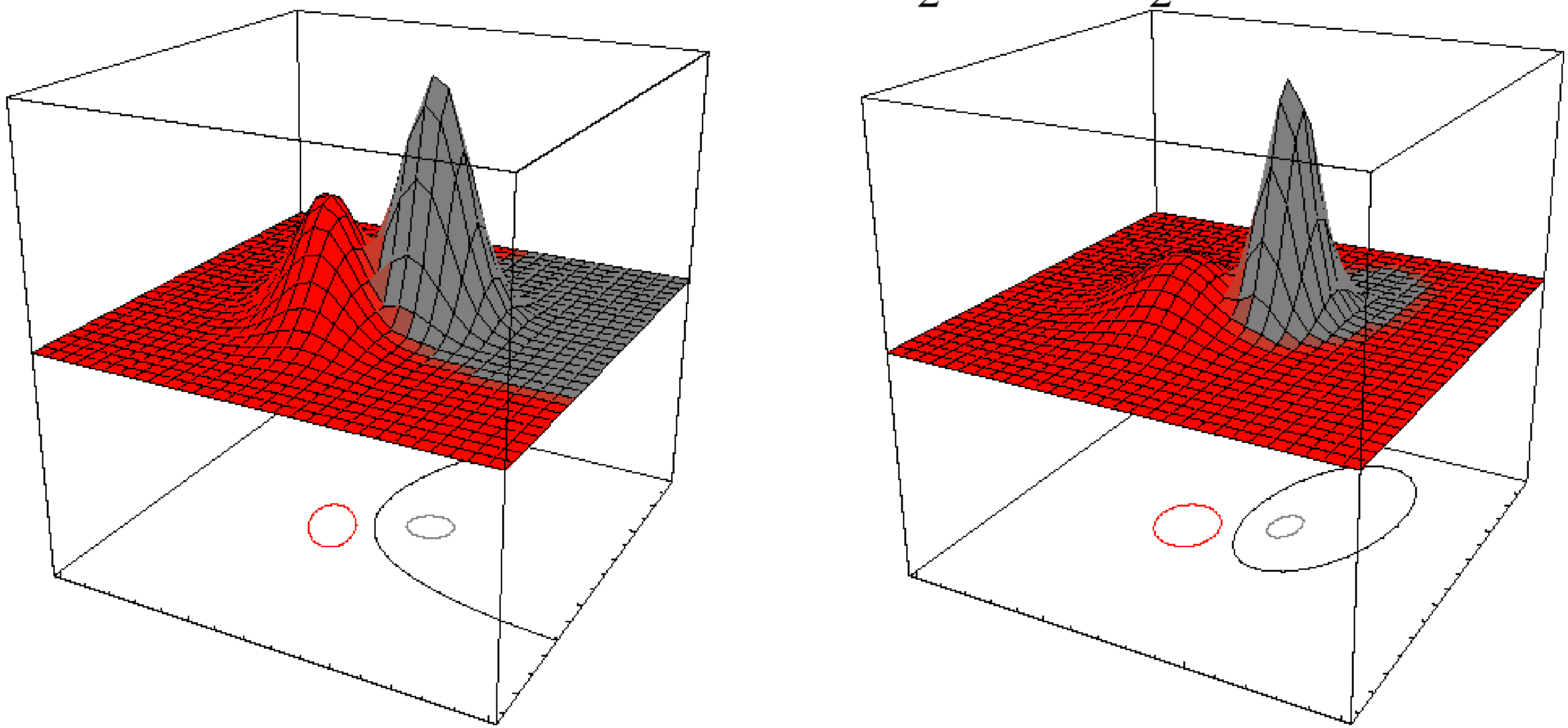
# Case 3: $\Sigma_i$ = any

➢Non-linear but <u>squared</u> decision functions.

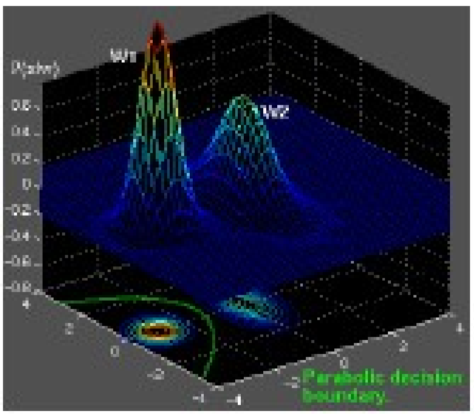
➢Decision surfaces *hyperquadratics (hyper-elliptical, hyper-paraboloid, etc.).*

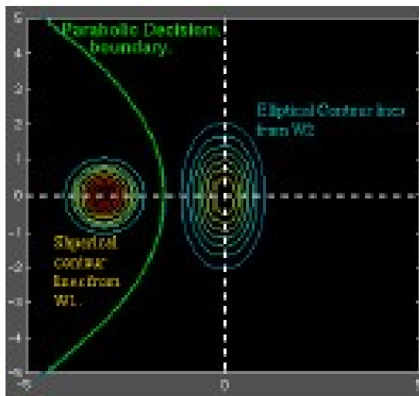
$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \qquad \mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \quad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

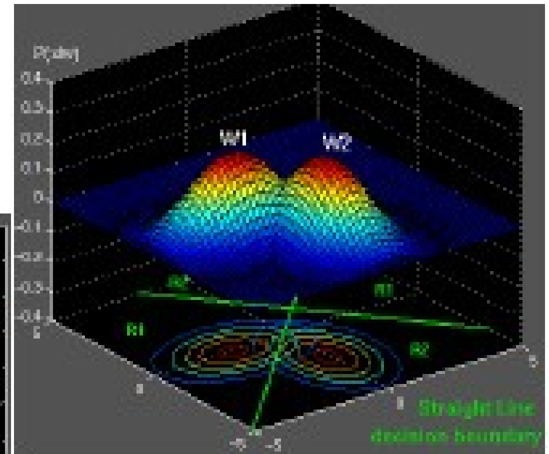
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

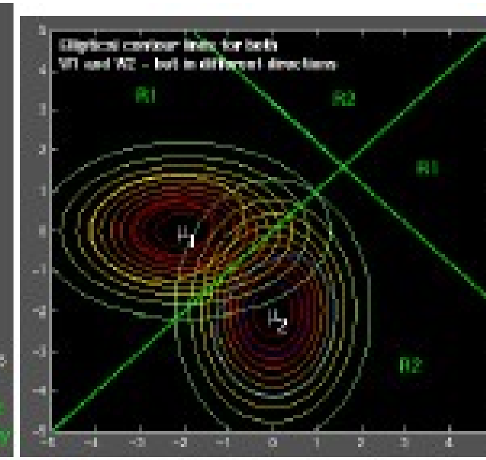


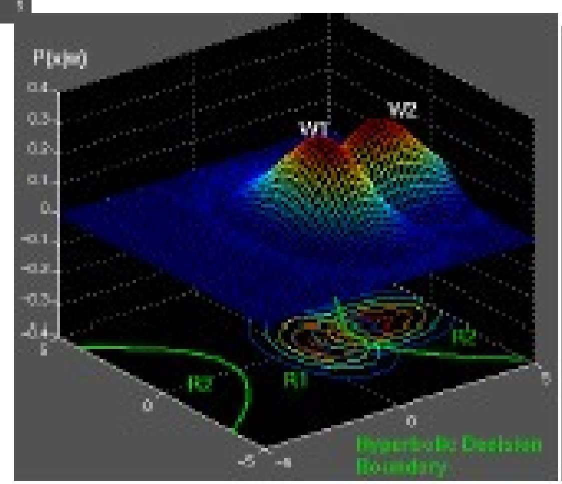
# Case 3: $\Sigma_i = \text{any}$

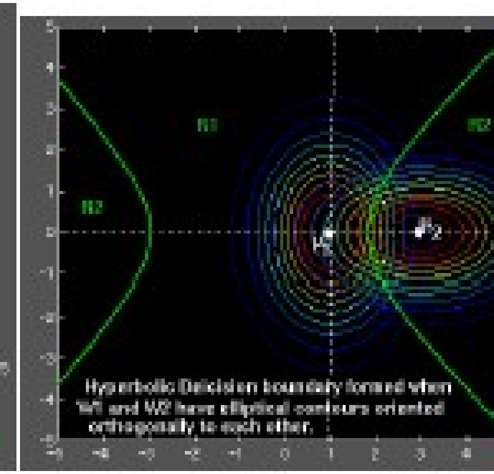


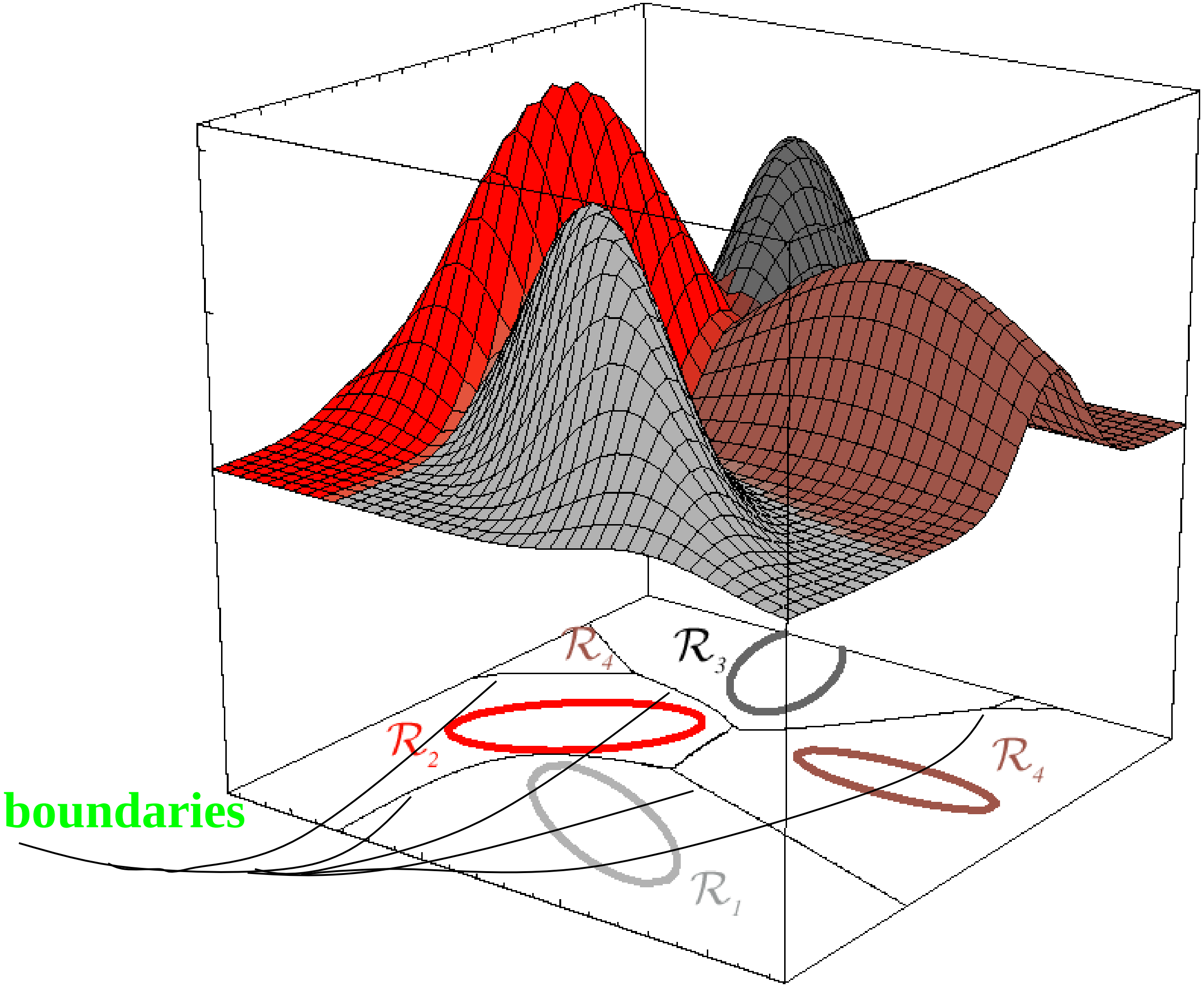



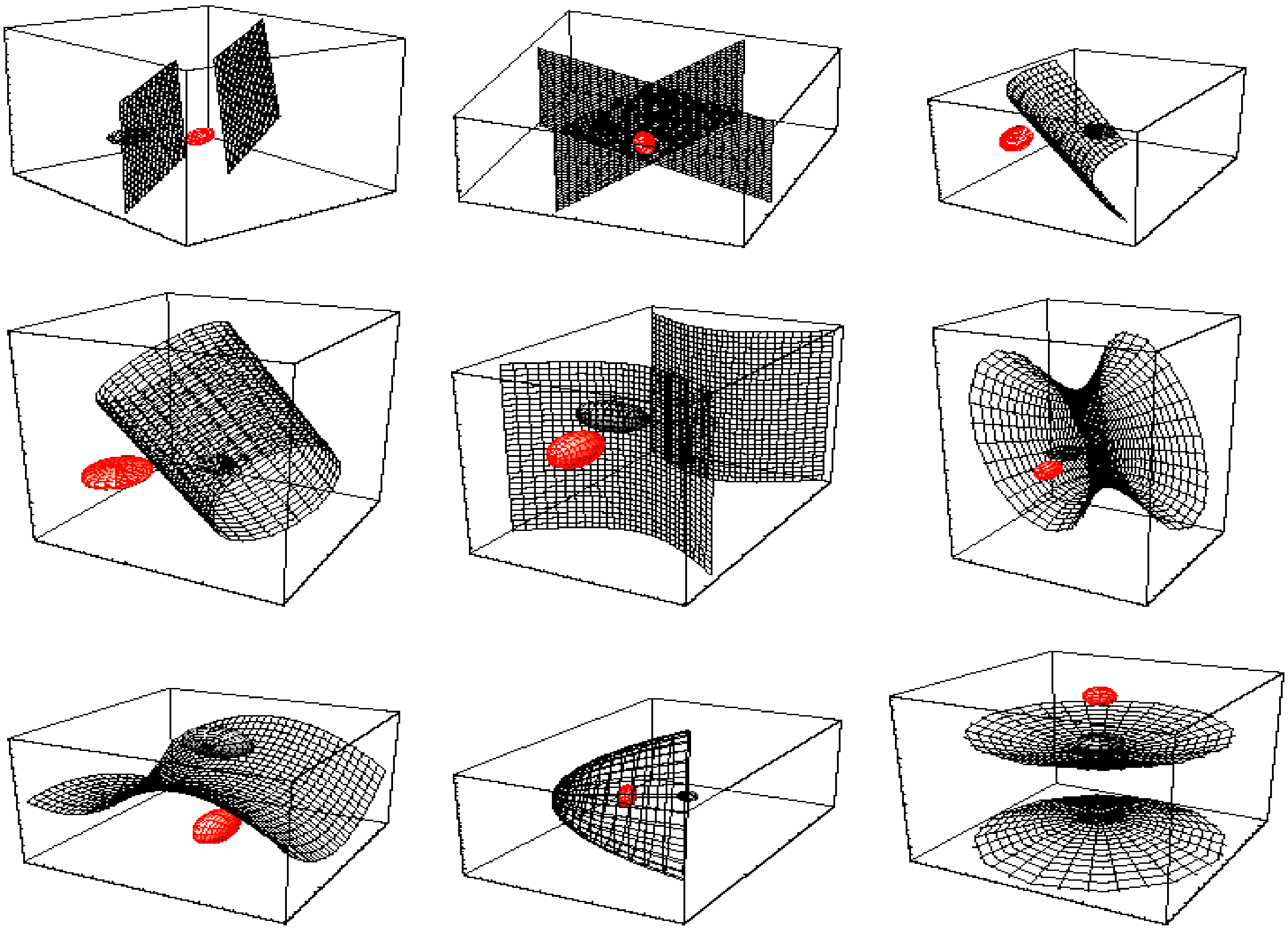
# *Case 3:* $\Sigma_i = \text{any}$

➢ In the case of multiple classes, the boundaries are even more complex:

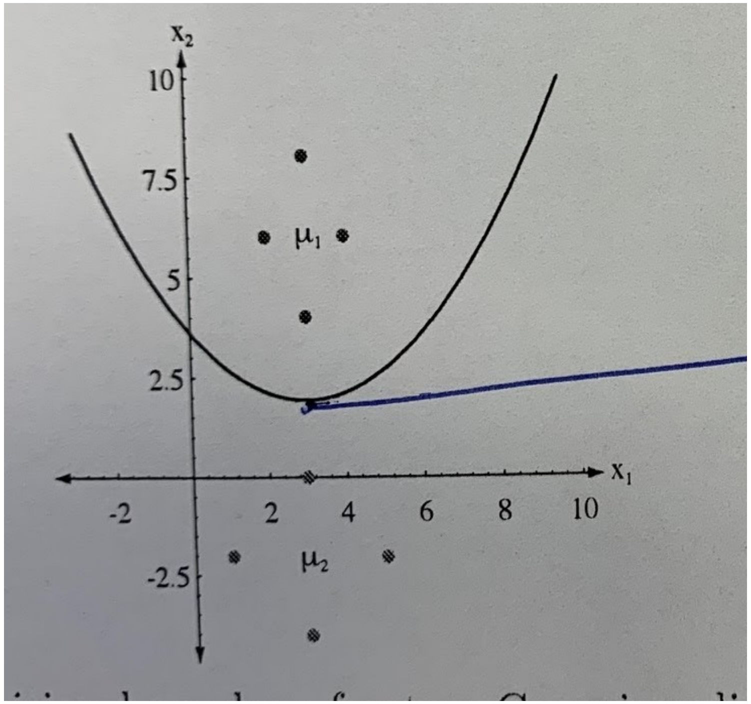


# Case 3: $\Sigma_i = \text{any}$

3-D



# Case 3: $\Sigma_i = \text{any}$

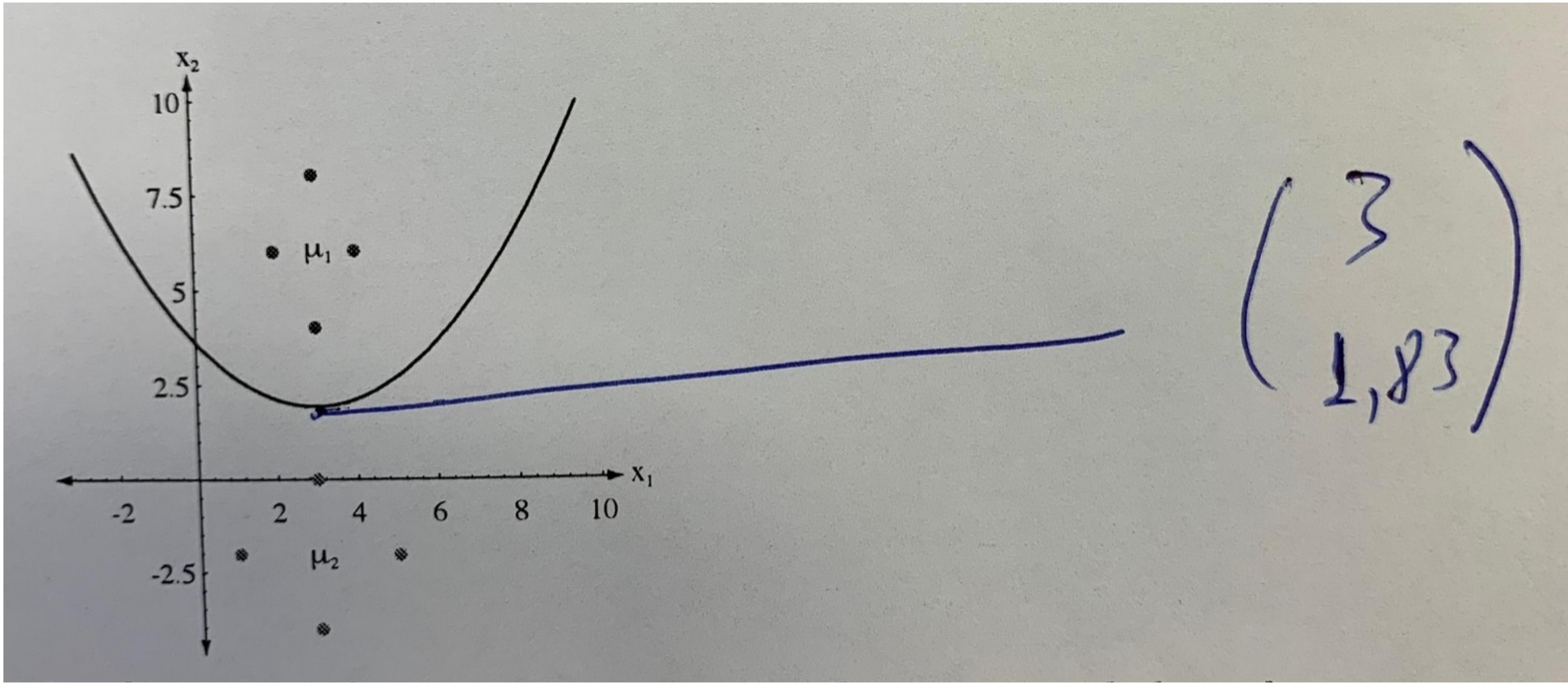



*Example:*
*Decision Regions*
*for 2-D data*

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

# Case 3: $\Sigma_i = \text{any}$



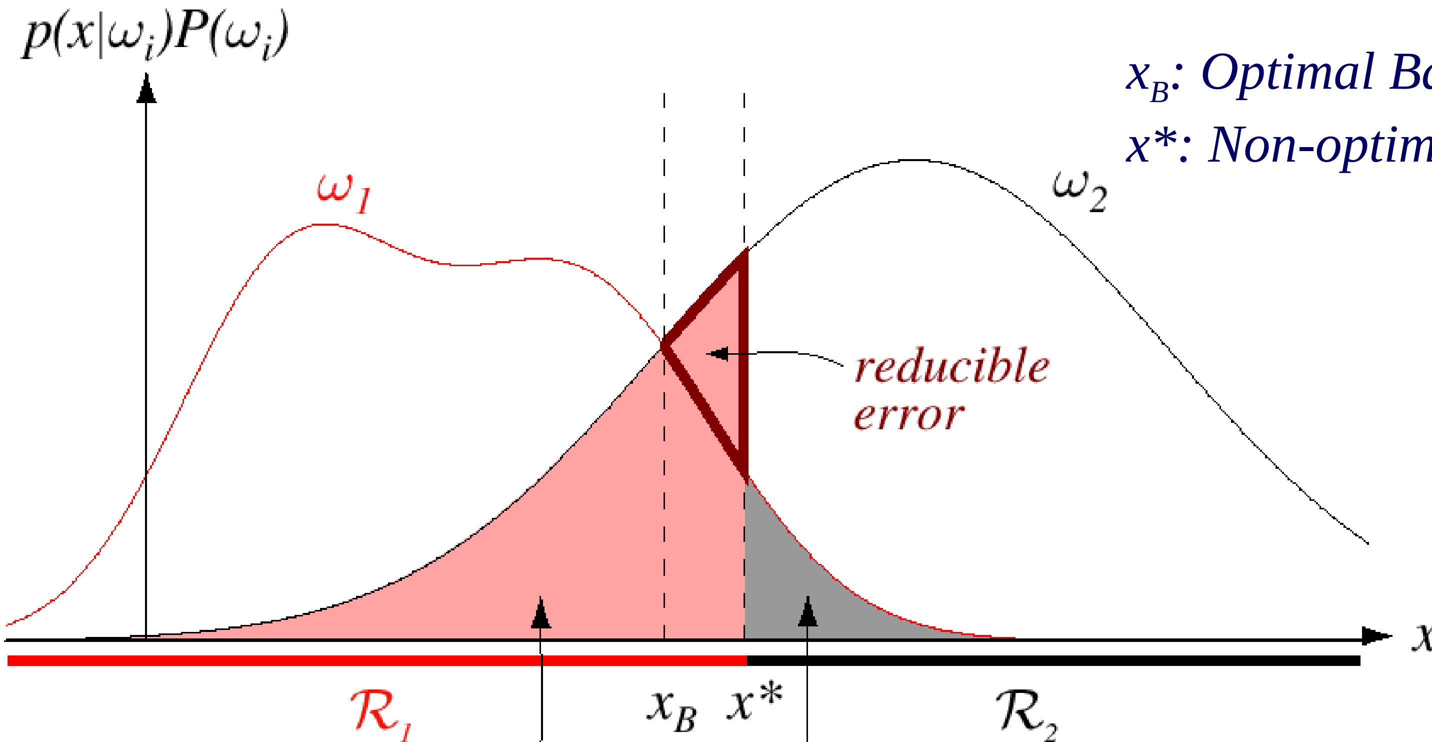

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$$

# Error Probabilities

- In the case of two classes, there are two cases of error:
  - **x** is at $R_1$ while ΚτΦ is $w_2$, and vice versa.

*$x_B$: Optimal Bayes solution*

*$x^*$: Non-optimal solution*




$$\mathbf{P(error)} = \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)\,dx + \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)\,dx$$

$$P(\mathbf{x}\in R_1,\omega_2) = P(\mathbf{x}\in R_1 \mid \omega_2)\cdot P(\omega_2) \qquad P(\mathbf{x}\in R_2,\omega_1) = P(\mathbf{x}\in R_2 \mid \omega_1)\cdot P(\omega_1)$$