Minimum Error-Rate Discriminant

- In the case of zero-one loss function, the Bayes Discriminant can be further simplified:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \ . \tag{29}$$

Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?

Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?
- **No!**
- Multiply by some positive constant.
- Shift them by some additive constant.

Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?
- **No!**
- Multiply by some positive constant.
- Shift them by some additive constant.
- For monotonically increasing function $f(\cdot)$, we can replace each $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$ without affecting our classification accuracy.
  - These can help for ease of understanding or computability.
  - The following all yield the same exact classification results for minimum-error-rate classification.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_j p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{30}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \tag{31}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \tag{32}$$

Visualizing Discriminants

Decision Regions

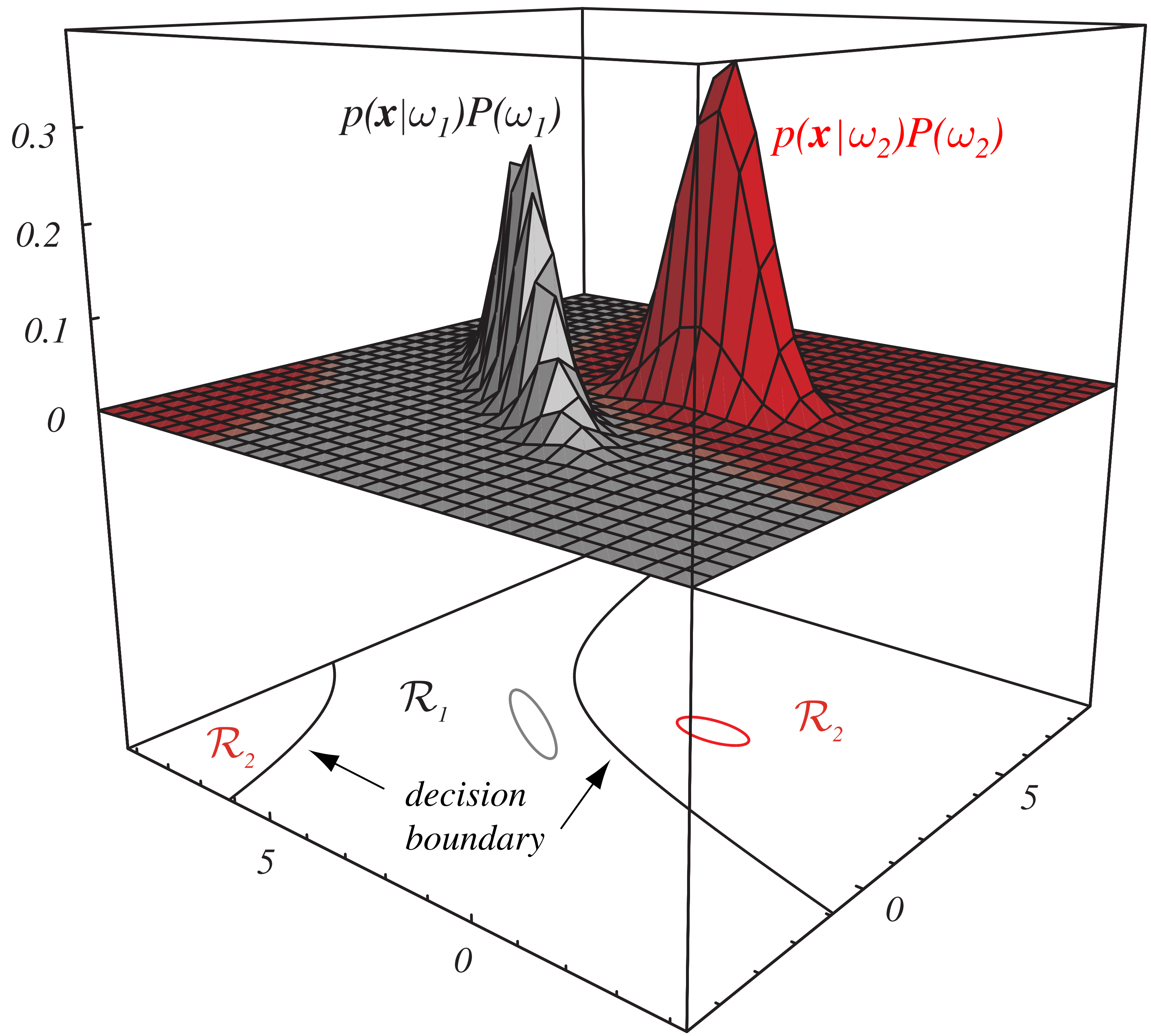
- The effect of any decision rule is to divide the feature space into decision regions.
- Denote a decision region $\mathcal{R}_i$ for ω_i.
- One not necessarily connected region is created for each category and assignments is according to:

$$\text{If } g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \forall j \neq i, \text{ then } \mathbf{x} \text{ is in } \mathcal{R}_i \quad . \tag{33}$$

- **Decision boundaries** separate the regions; they are ties among the discriminant functions.

Visualizing Discriminants

Decision Regions



Two-Category Discriminants

Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad . \tag{34}$$

- What is a suitable decision rule?

Two-Category Discriminants

Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \ . \tag{34}$$

- The following simple rule is then used:

$$\text{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0\text{; otherwise decide } \omega_2. \tag{35}$$

Two-Category Discriminants

Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad . \tag{34}$$

- The following simple rule is then used:

$$\text{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0\text{; otherwise decide } \omega_2. \tag{35}$$

- Various manipulations of the discriminant:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \tag{36}$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \tag{37}$$

Background on the Normal Density

- This next section is a slight digression to introduce the Normal Density (most of you will have had this already).
- The Normal density is very well studied.
- It easy to work with analytically.
- Often in PR, an appropriate model seems to be a single typical value corrupted by continuous-valued, random noise.
- Central Limit Theorem (Second Fundamental Theorem of Probability).
  - The distribution of the sum of n random variables approaches the normal distribution when n is large.
  - E.g., `http://www.stattucino.com/berrie/dsl/Galton.html`

Expectation

- Recall the definition of expected value of any scalar function $f(x)$ in the continuous $p(x)$ and discrete $P(x)$ cases

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx \quad (38)$$

$$\mathcal{E}[f(x)] = \sum_{x} f(x)P(x) \quad (39)$$

where we have a set $\mathcal{D}$ over which the discrete expectation is computed.

Univariate Normal Density

- Continuous univariate normal, or **Gaussian**, density:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad . \tag{40}$$

- The **mean** is the expected value of x is

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x) dx \quad . \tag{41}$$

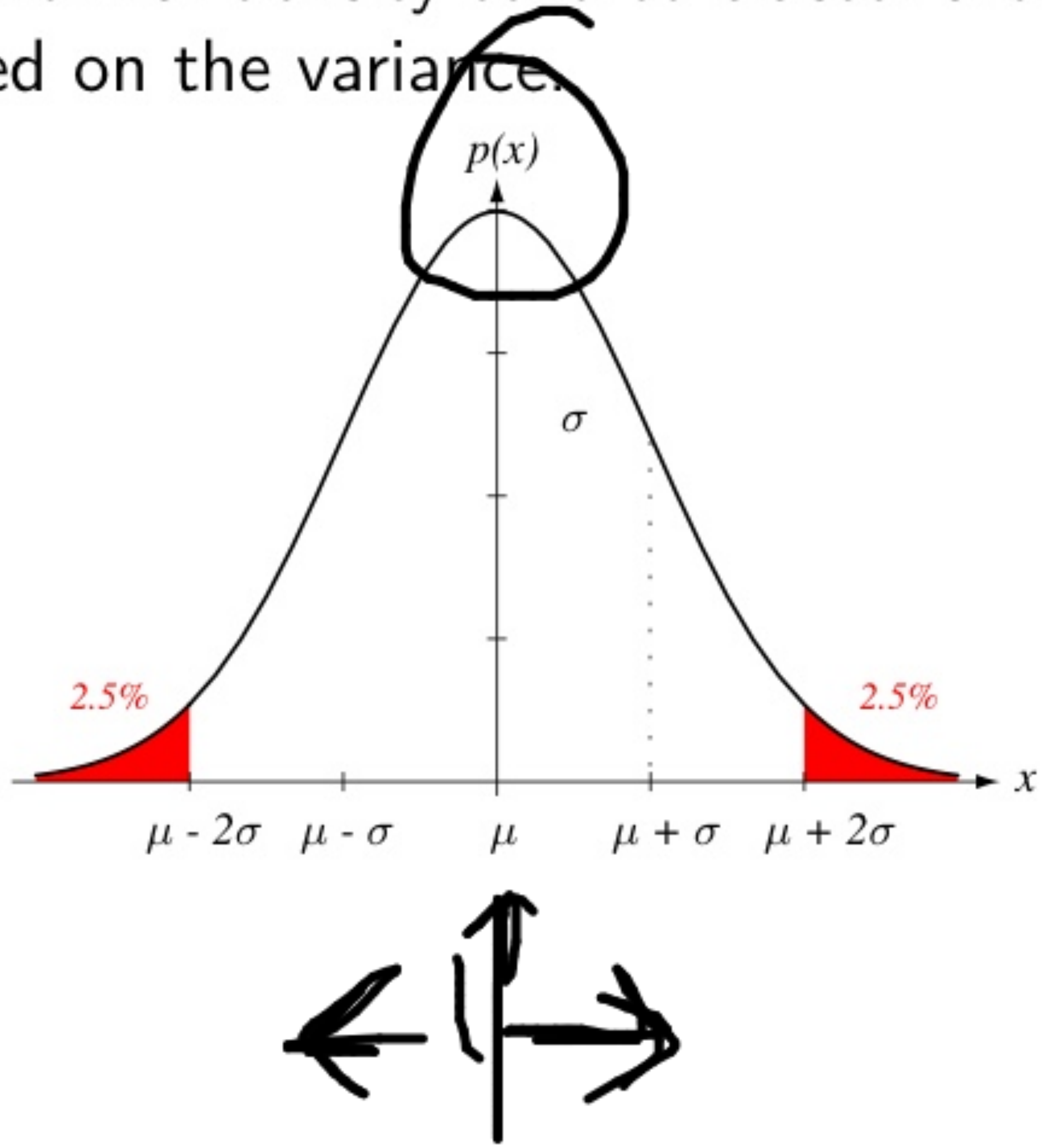
- The **variance** is the expected squared deviation

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \quad . \tag{42}$$

Univariate Normal Density

Sufficient Statistics

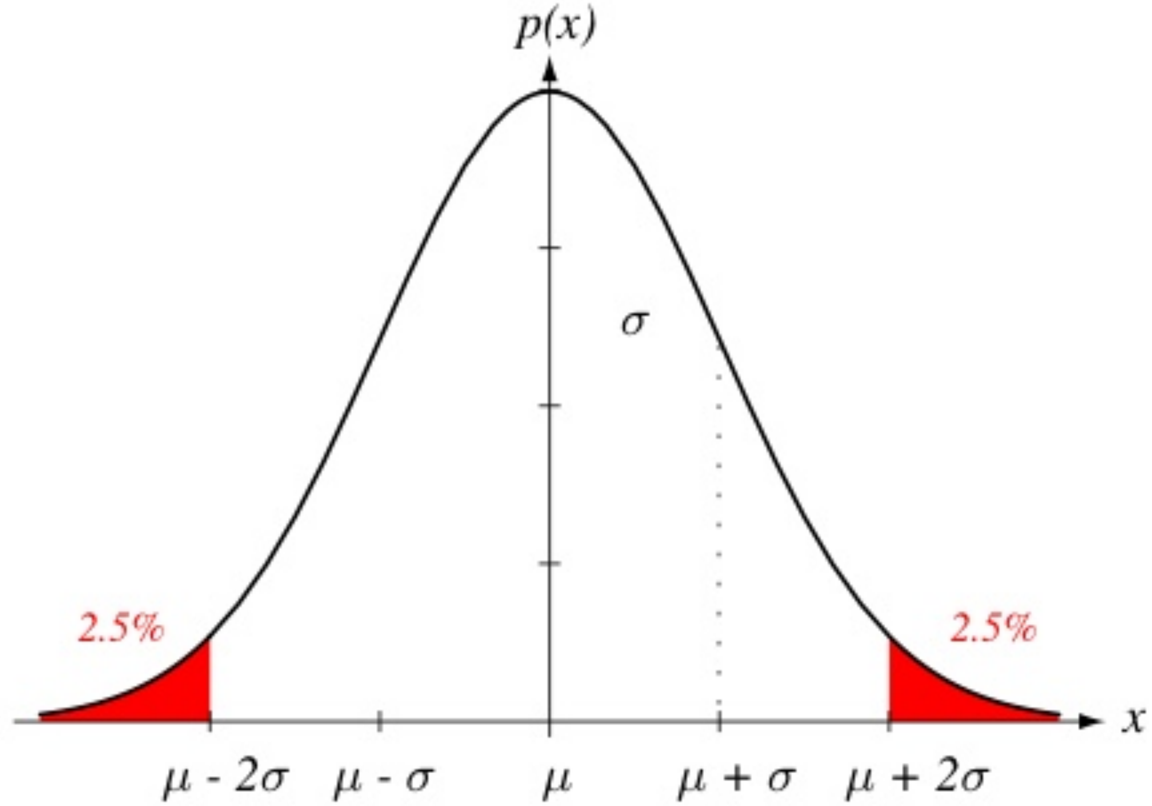
- Samples from the normal density tend to cluster around the mean and be spread-out based on the variance.



Univariate Normal Density

Sufficient Statistics

- Samples from the normal density tend to cluster around the mean and be spread-out based on the variance.



- The normal density is completely specified by the mean and the variance. These two are its **sufficient statistics**.
- We thus abbreviate the equation for the normal density as

$$p(x) \sim N(\mu, \sigma^2) \tag{43}$$

Entropy

- **Entropy** is the uncertainty in the random samples from a distribution.

$$H(p(x)) = -\int p(x) \ln p(x) dx \tag{44}$$

- The normal density has the maximum entropy for all distributions have a given mean and variance.
- What is the entropy of the uniform distribution?

Entropy

- **Entropy** is the uncertainty in the random samples from a distribution.

$$H(p(x)) = -\int p(x) \ln p(x) dx \tag{44}$$

- The normal density has the maximum entropy for all distributions have a given mean and variance.
- What is the entropy of the uniform distribution?
- The uniform distribution has maximum entropy (on a given interval).

Multivariate Normal Density

And a test to see if your Linear Algebra is up to snuff.

- The multivariate Gaussian in d dimensions is written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad . \tag{45}$$

- Again, we abbreviate this as $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$.
- The sufficient statistics in d-dimensions:

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \tag{46}$$

$$\mathbf{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} \tag{47}$$

The Covariance Matrix

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T} p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements σ_{ii} are the variances of the respective coordinate x_i.
- The off-diagonal elements σ_{ij} are the covariances of x_i and x_j.
- What does a $\sigma_{ij} = 0$ imply?

$$\begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ & \end{bmatrix}$$

The Covariance Matrix

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements σ_{ii} are the variances of the respective coordinate x_i.
- The off-diagonal elements σ_{ij} are the covariances of x_i and x_j.
- What does a $\sigma_{ij} = 0$ imply?
- That coordinates x_i and x_j are statistically independent.

The Covariance Matrix

$$\mathbf{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements σ_{ii} are the variances of the respective coordinate x_i.
- The off-diagonal elements σ_{ij} are the covariances of x_i and x_j.
- What does a $\sigma_{ij} = 0$ imply?
- That coordinates x_i and x_j are statistically independent.
- What does $\mathbf{\Sigma}$ reduce to if all off-diagonals are 0?

The Covariance Matrix

$$\mathbf{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x}$$

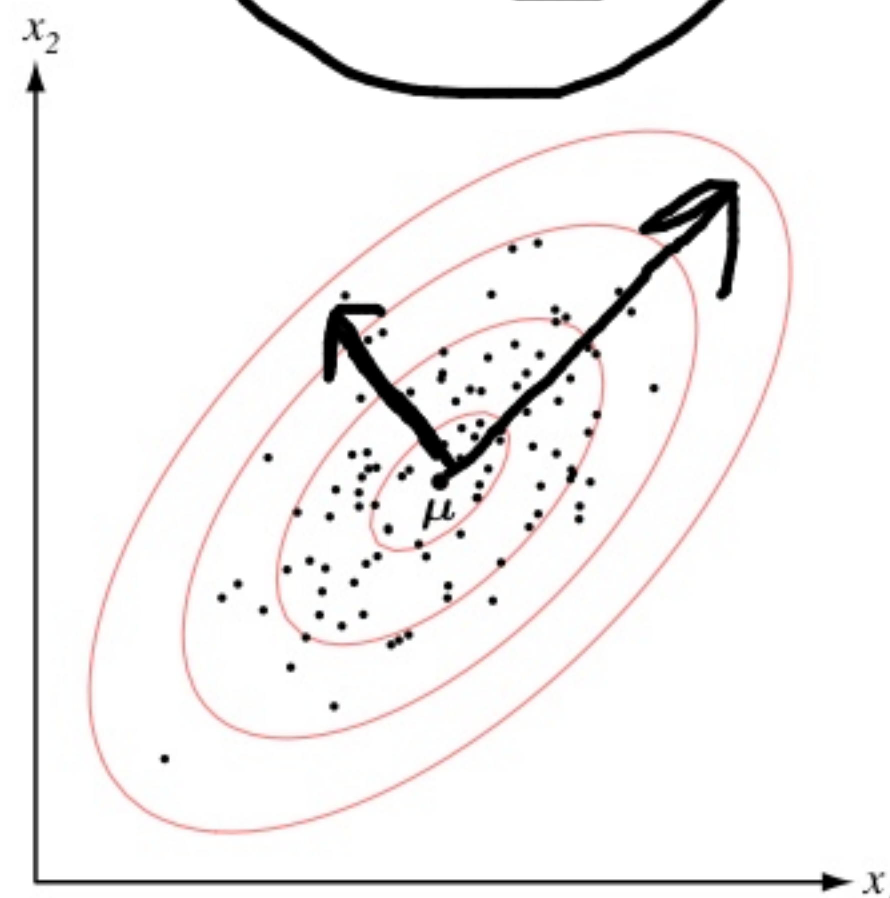
- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements σ_{ii} are the variances of the respective coordinate x_i.
- The off-diagonal elements σ_{ij} are the covariances of x_i and x_j.
- What does a $\sigma_{ij} = 0$ imply?
- That coordinates x_i and x_j are statistically independent.
- What does $\mathbf{\Sigma}$ reduce to if all off-diagonals are 0?
- The product of the d univariate densities.

Mahalanobis Distance

UDV^T eigs std

- The shape of the density is determined by the covariance $\boldsymbol{\Sigma}$.
- Specifically, the eigenvectors of $\boldsymbol{\Sigma}$ give the principal axes of the hyperellipsoids and the eigenvalues determine the lengths of these axes.
- The loci of points of constant density are hyperellipsoids with constant **Mahalonobis distance**:

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \qquad (48)$$

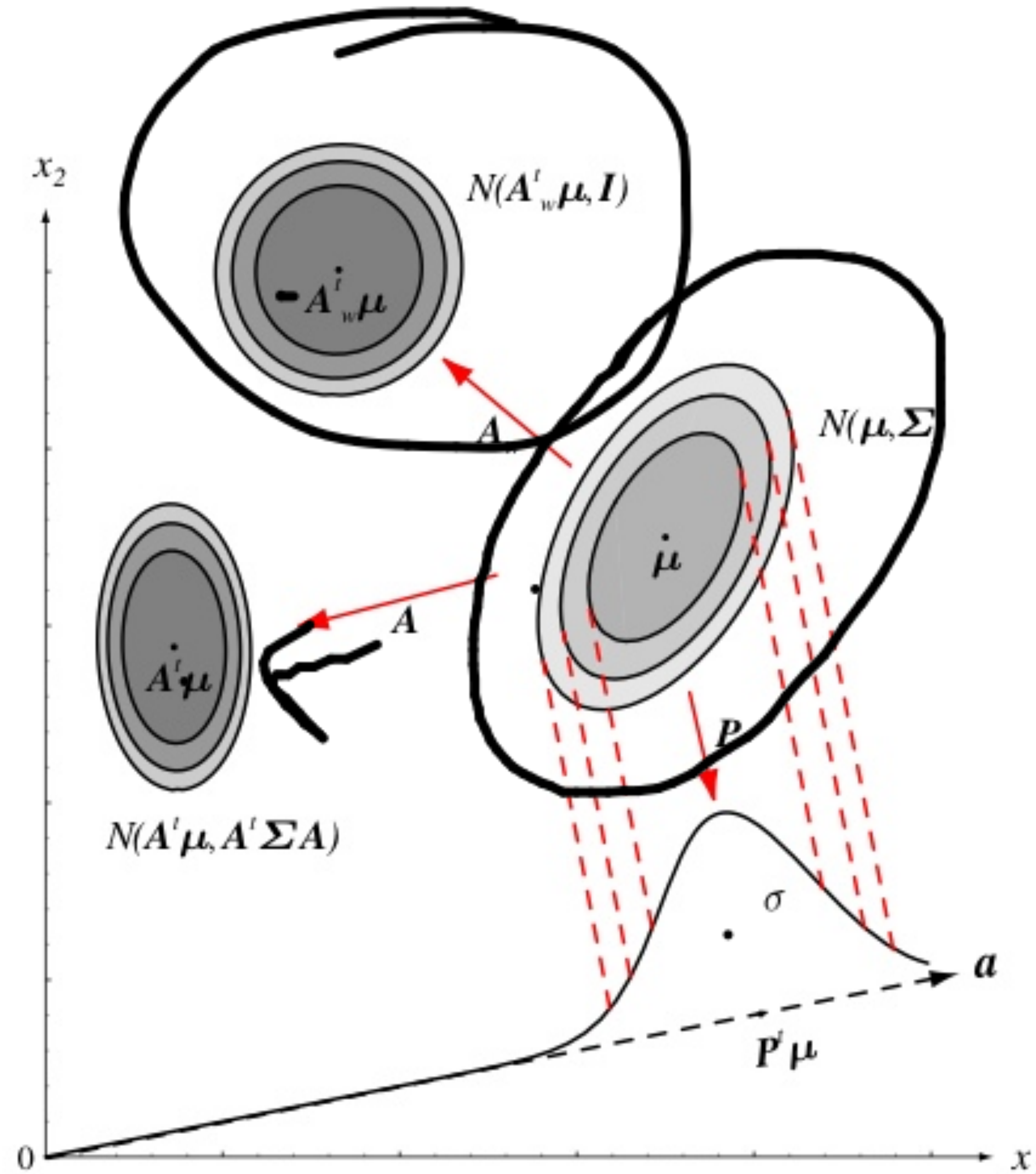


Linear Combinations of Normals

- Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed.
- For $p(\mathbf{x}) \sim N((\mu), \mathbf{\Sigma})$ and $\mathbf{A}$, a d-by-k matrix, define $\mathbf{y} = \mathbf{A}^\mathsf{T}\mathbf{x}$. Then:

$$p(\mathbf{y}) \sim N(\mathbf{A}^\mathsf{T}\boldsymbol{\mu}, \mathbf{A}^\mathsf{T}\mathbf{\Sigma}\mathbf{A}) \quad (49)$$

- With the covariance matrix, we can calculate the dispersion of the data in any direction or in any subspace.



General Discriminant for Normal Densities

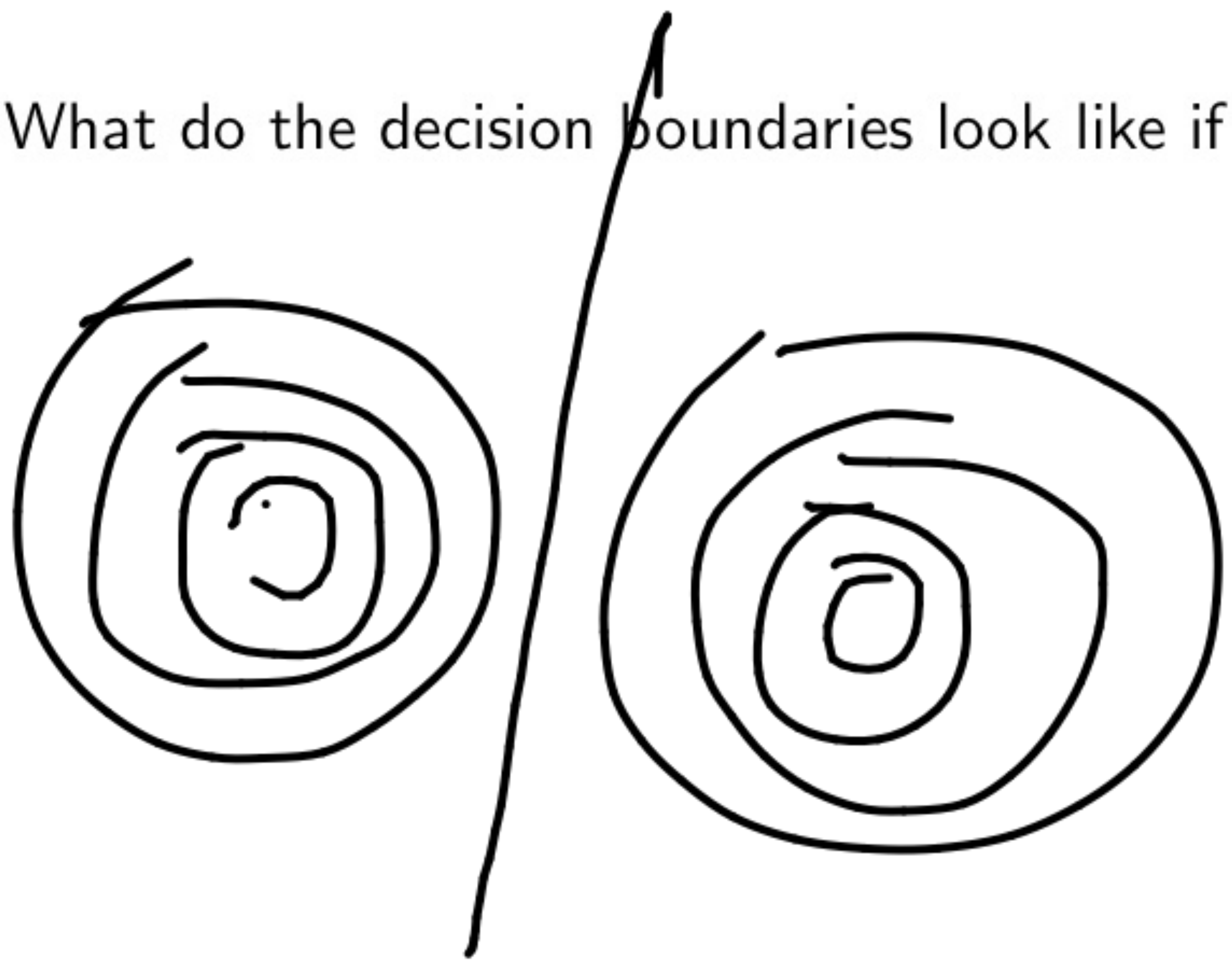
- Recall the minimum error rate discriminant,
  $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$.
- If we assume normal densities, i.e., if $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, then the general discriminant is of the form

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (50)$$

$\sigma^2 I$ $\quad \|(x - \mu_i)\|$

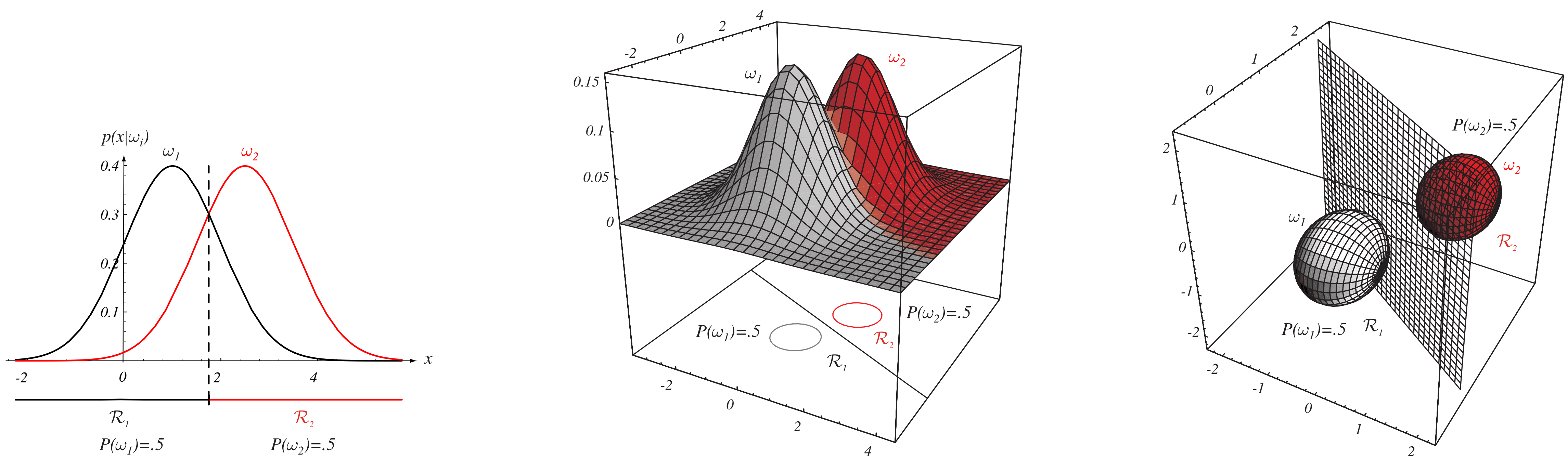
Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$?



Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$?
- They are hyperplanes.



- Let's see why...

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

- The discriminant functions take on a simple form:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \underline{\ln P(\omega_i)} \qquad - \ln \sigma_i^2 \qquad (51)$$

- Think of this discriminant as a combination of two things
  1. The distance of the sample to the mean vector (for each i).
  2. A normalization by the variance and offset by the prior.

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

- But, we don't need to actually compute the distances.
- Expanding the quadratic form $(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu})$ yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^\mathsf{T}\mathbf{x} - 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{x} + \boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \quad . \tag{52}$$

- The quadratic term $\mathbf{x}^\mathsf{T}\mathbf{x}$ is the same for all i and can thus be ignored.
- This yields the equivalent **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^\mathsf{T}\mathbf{x} + w_{i0} \tag{53}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \tag{54}$$

$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{55}$$

- w_{i0} is called the **bias**.

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

Decision Boundary Equation

- The decision surfaces for a linear discriminant classifiers are hyperplanes defined by the linear equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$.
- The equation can be written as

$$\mathbf{w}^\mathsf{T}(\mathbf{x} - \mathbf{x}_0) = 0 \tag{56}$$

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{57}$$

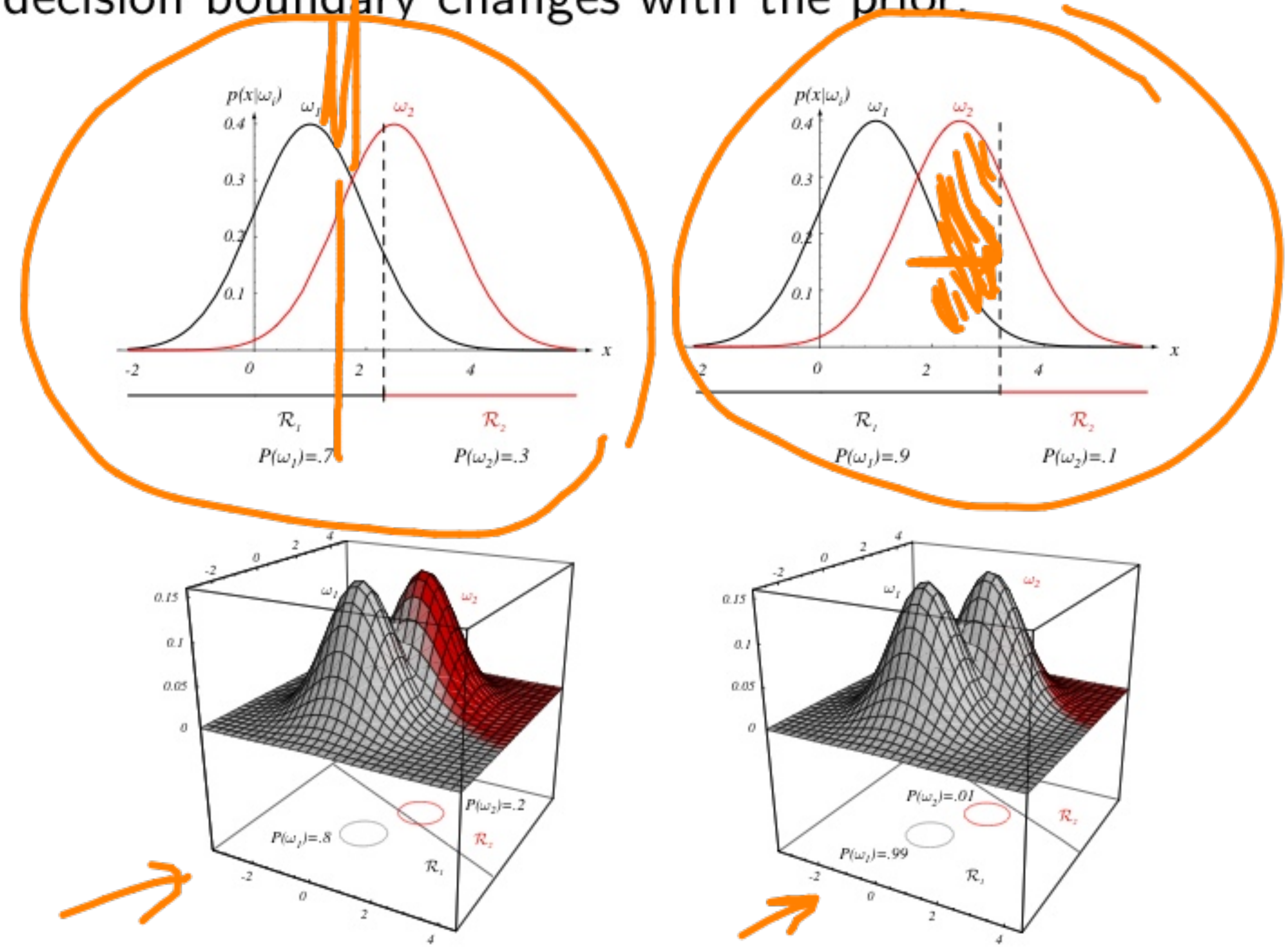
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) \cancel{- \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} \tag{58}$$

- These equations define a hyperplane through point x_0 with a normal vector $\mathbf{w}$.

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

Decision Boundary Equation

- The decision boundary changes with the prior.



General Case: Arbitrary Σ_i

- The discriminant functions are quadratic (the only term we can drop is the $\ln 2\pi$ term):

$$g_i(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{W}_i\mathbf{x} + \boxed{\mathbf{w}_i^{\mathsf{T}}\mathbf{x} + w_{i0}} \tag{59}$$

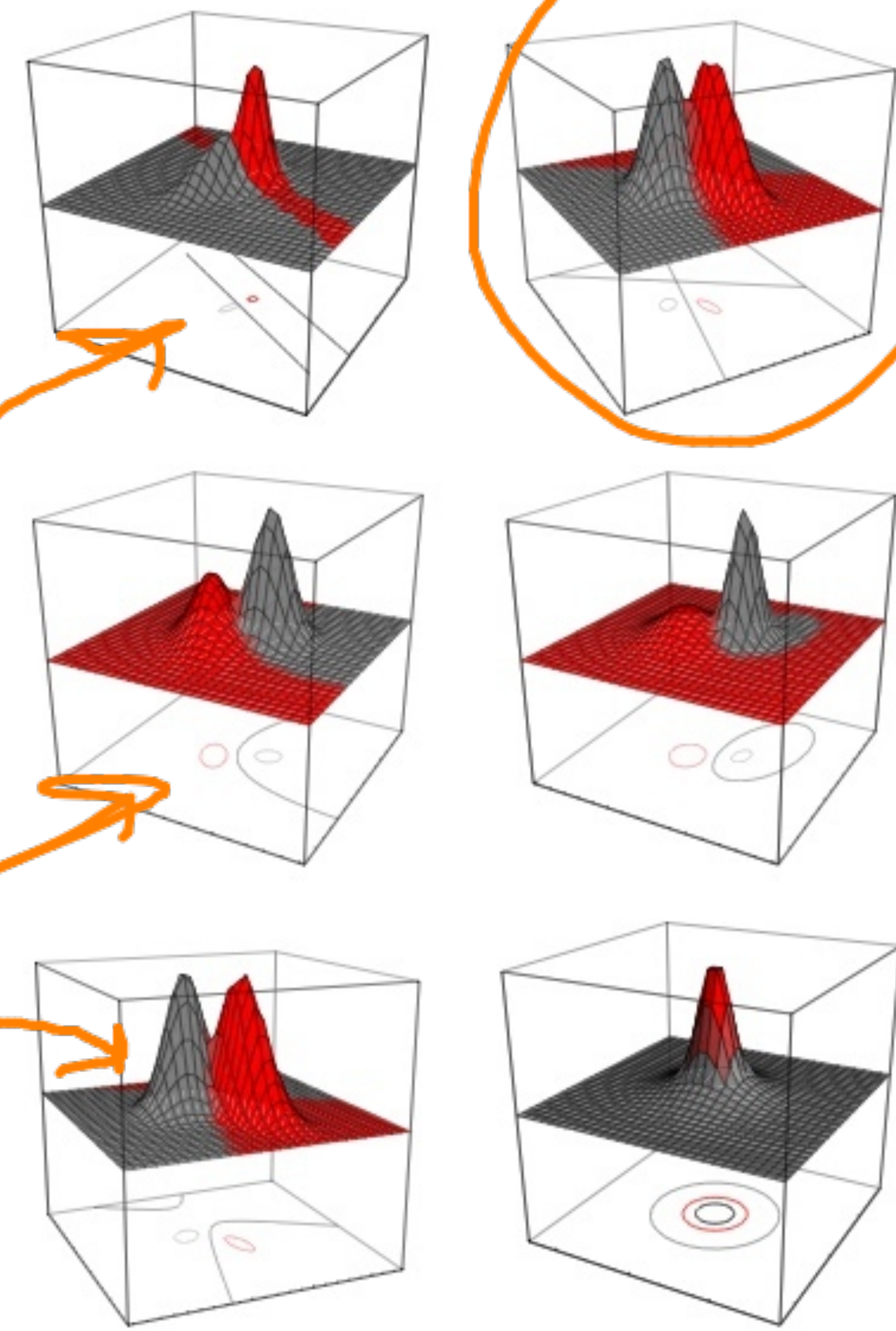
$$\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1} \tag{60}$$

$$\mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i \tag{61}$$

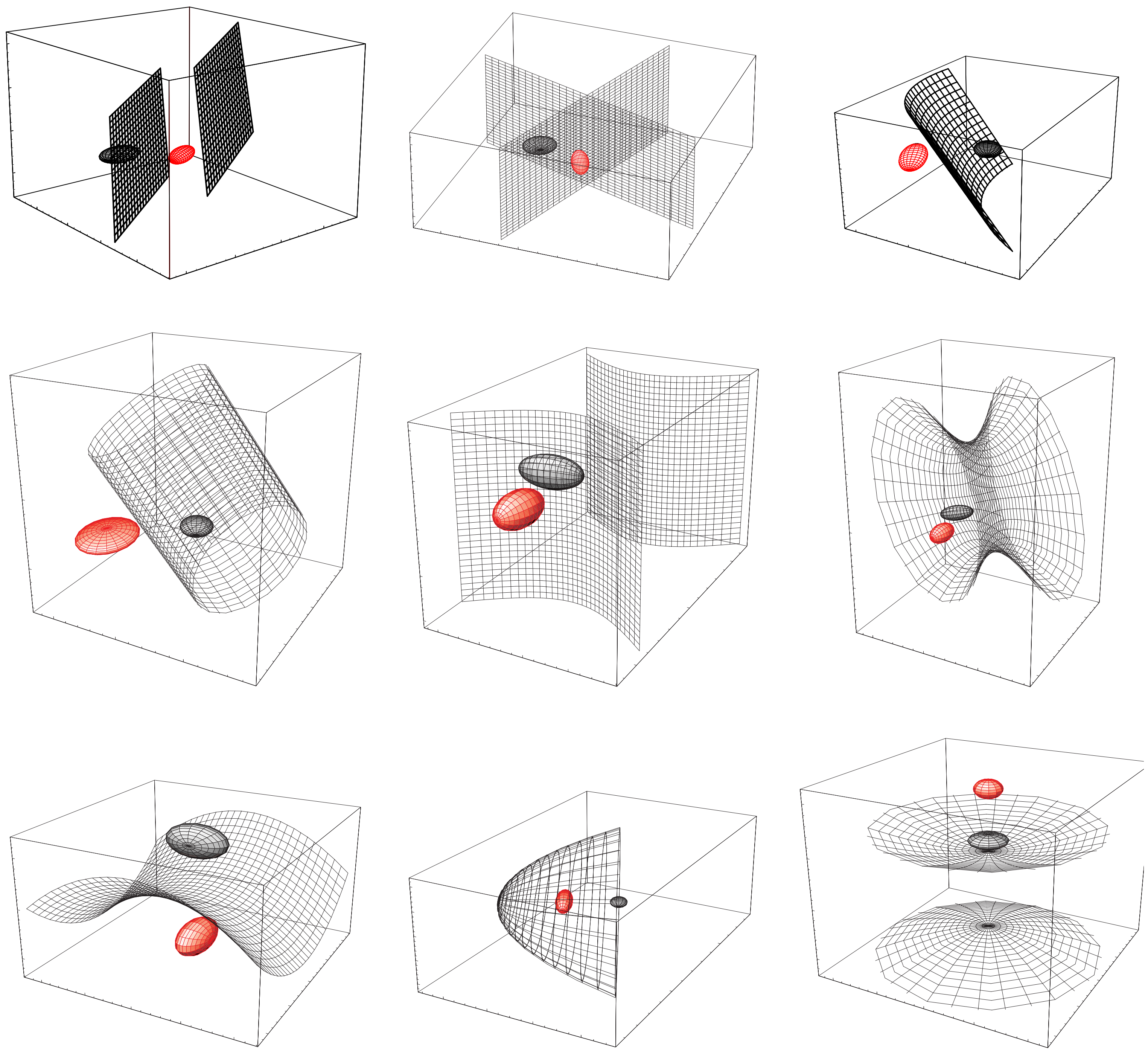
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathsf{T}}\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i) \tag{62}$$

- The decision surface between two categories are **hyperquadrics**.

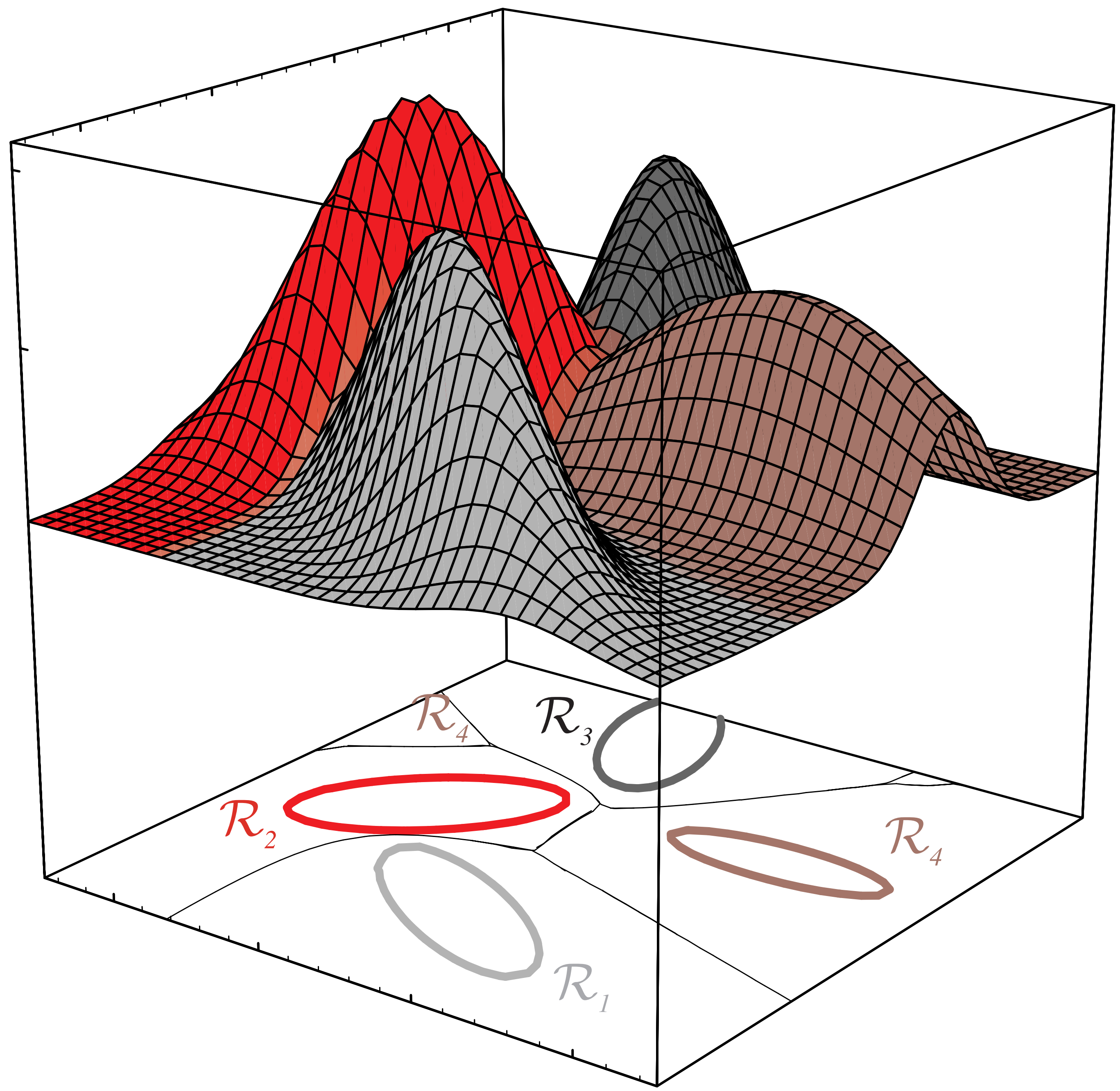
General Case: Arbitrary Σ_i



General Case: Arbitrary Σ_i



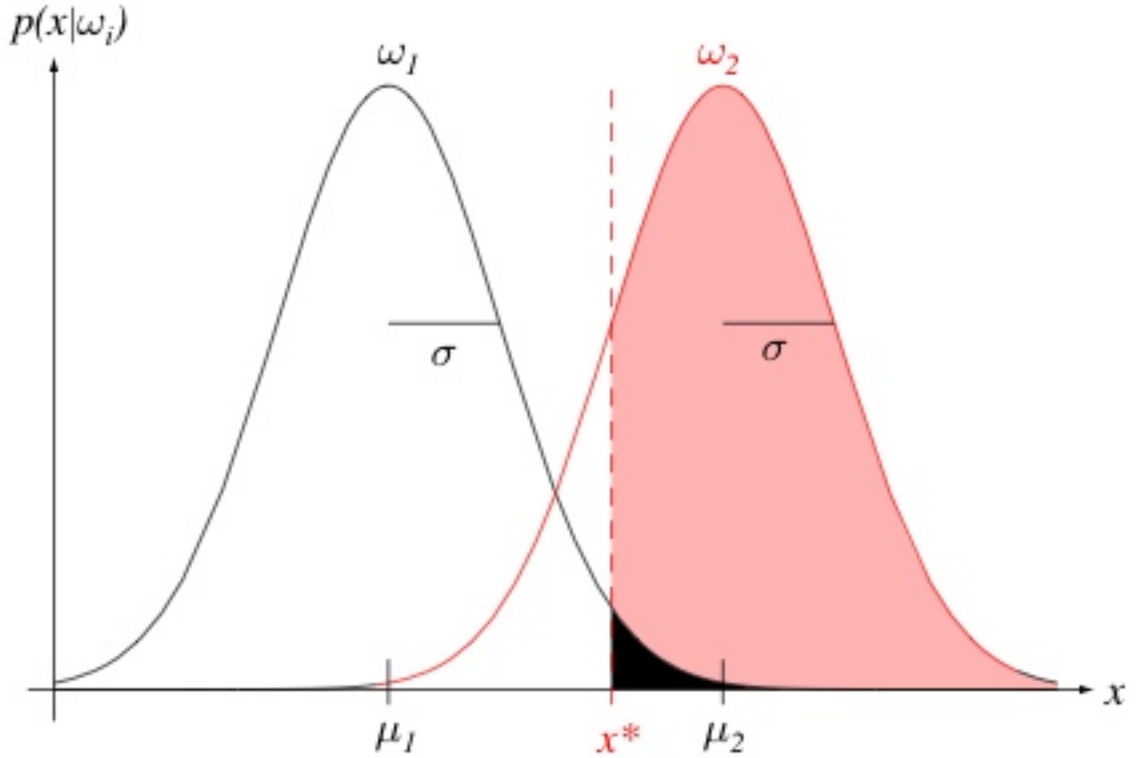
General Case for Multiple Categories



Quite A Complicated Decision Surface!

Signal Detection Theory

- A fundamental way of analyzing a classifier.
- Consider the following experimental setup:



- Suppose we are interested in detecting a single pulse.
- We can read an internal signal x.
- The signal is distributed about mean μ_2 when an external signal is present and around mean μ_1 when no external signal is present.
- Assume the distributions have the same variances, $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$.

Signal Detection Theory

- The detector uses x^* to decide if the external signal is present.
- **Discriminability** characterizes how difficult it will be to decide if the external signal is present without knowing x^*.

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \tag{63}$$

- Even if we do not know μ_1, μ_2, σ, or x^*, we can find d' by using a **receiver operating characteristic** or ROC curve, as long as we know the state of nature for some experiments

Receiver Operating Characteristics

Definitions

- A **Hit** is the probability that the internal signal is above x^* given that the external signal is present

$$P(x > x^* | x \in \omega_2) \tag{64}$$

Receiver Operating Characteristics

Definitions

TP

- A **Hit** is the probability that the internal signal is above x^* given that the external signal is present

$$P(x > x^* | x \in \omega_2) \tag{64}$$

TN

- A **Correct Rejection** is the probability that the internal signal is below x^* given that the external signal is not present.

$$P(x < x^* | x \in \omega_1) \tag{65}$$

Receiver Operating Characteristics

Definitions

- A **Hit** is the probability that the internal signal is above x^* given that the external signal is present

$$P(x > x^*|x \in \omega_2) \tag{64}$$

- A **Correct Rejection** is the probability that the internal signal is below x^* given that the external signal is not present.

$$P(x < x^*|x \in \omega_1) \tag{65}$$

- A **False Alarm** is the probability that the internal signal is above x^* despite there being no external signal present.

$$P(x > x^*|x \in \omega_1) \tag{66}$$

Receiver Operating Characteristics

Definitions

- TP A **Hit** is the probability that the internal signal is above x^* given that the external signal is present

$$P(x > x^*|x \in \omega_2) \tag{64}$$

- TN A **Correct Rejection** is the probability that the internal signal is below x^* given that the external signal is not present.

$$P(x < x^*|x \in \omega_1) \tag{65}$$

- FP A **False Alarm** is the probability that the internal signal is above x^* despite there being no external signal present.

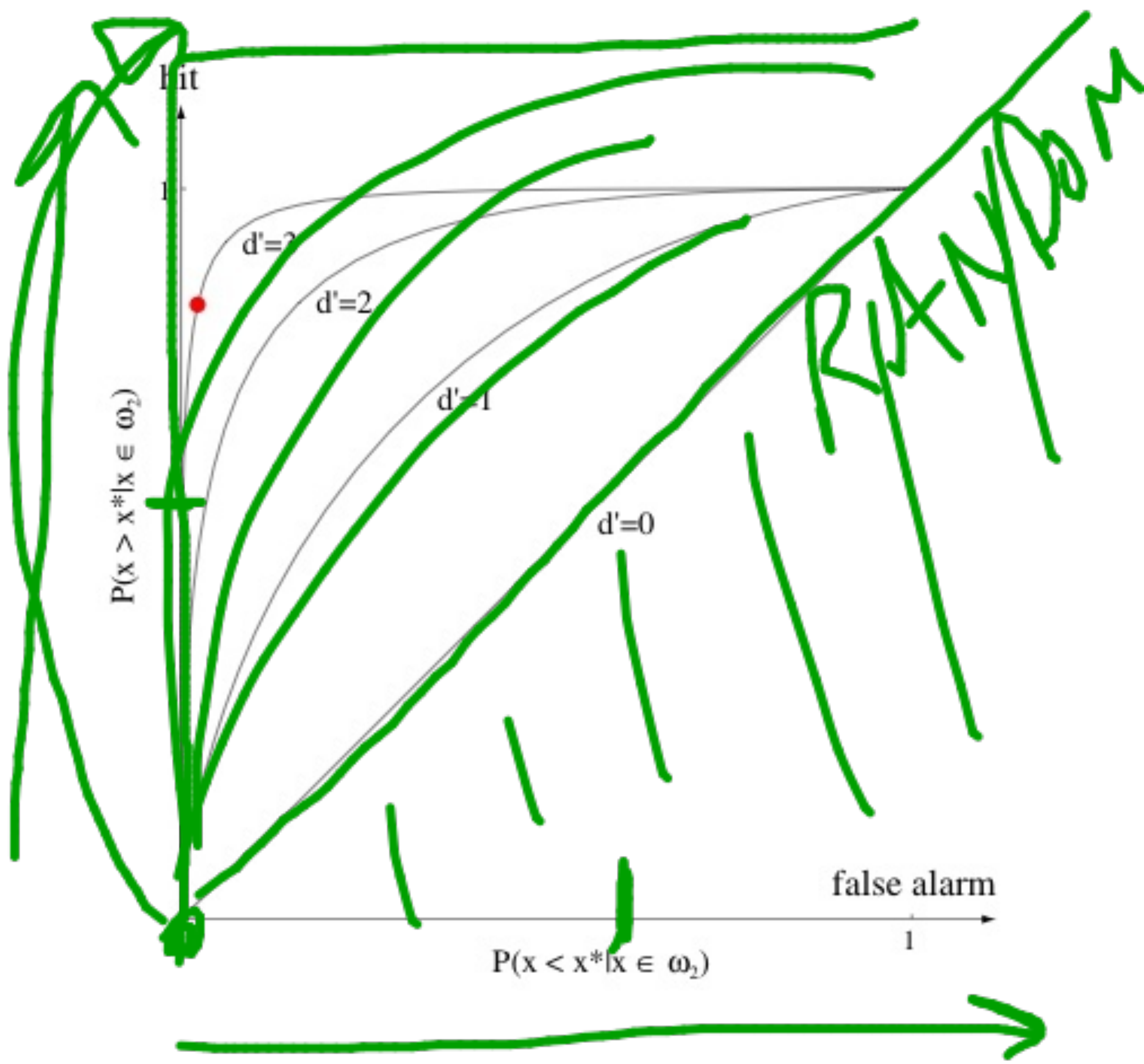
$$P(x > x^*|x \in \omega_1) \tag{66}$$

- FN A **Miss** is the probability that the internal signal is below x^* given that the external signal is present.

$$P(x < x^*|x \in \omega_2) \tag{67}$$

Receiver Operating Characteristics

- We can experimentally determine the rates, in particular the Hit-Rate and the False-Alarm-Rate.
- Basic idea is to assume our densities are fixed (reasonable) but vary our threshold x^*, which will thus change the rates.
- The receiver operating characteristic plots the hit rate against the false alarm rate.
- What shape curve do we want?



Missing Features

- Suppose we have built a classifier on multiple features, for example the lightness and width.
- What do we do if one of the features is not measurable for a particular case? For example the lightness can be measured but the width cannot because of occlusion.

Missing Features

- Suppose we have built a classifier on multiple features, for example the lightness and width.
- What do we do if one of the features is not measurable for a particular case? For example the lightness can be measured but the width cannot because of occlusion.
- **Marginalize!**
- Let $\mathbf{x}$ be our full feature feature and $\mathbf{x}_g$ be the subset that are measurable (or good) and let $\mathbf{x}_b$ be the subset that are missing (or bad/noisy).
- We seek an estimate of the posterior given **just the good features** $\mathbf{x}_g$.

Missing Features

$$P(\omega_i|\mathbf{x}_g) = \frac{p(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} \tag{68}$$

$$= \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}{p(\mathbf{x}_g)} \tag{69}$$

$$= \frac{\int p(\omega_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}_b}{p(\mathbf{x}_g)} \tag{70}$$

$$= \frac{\int g_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}_b}{\int p(\mathbf{x}) d\mathbf{x}_b} \tag{71}$$

- We will cover the Expectation-Maximization algorithm later.
- This is normally quite expensive to evaluate unless the densities are special (like Gaussians).

Statistical Independence

- Two variables x_i and x_j are independent if

$$p(x_i, x_j) = p(x_i)p(x_j) \tag{72}$$

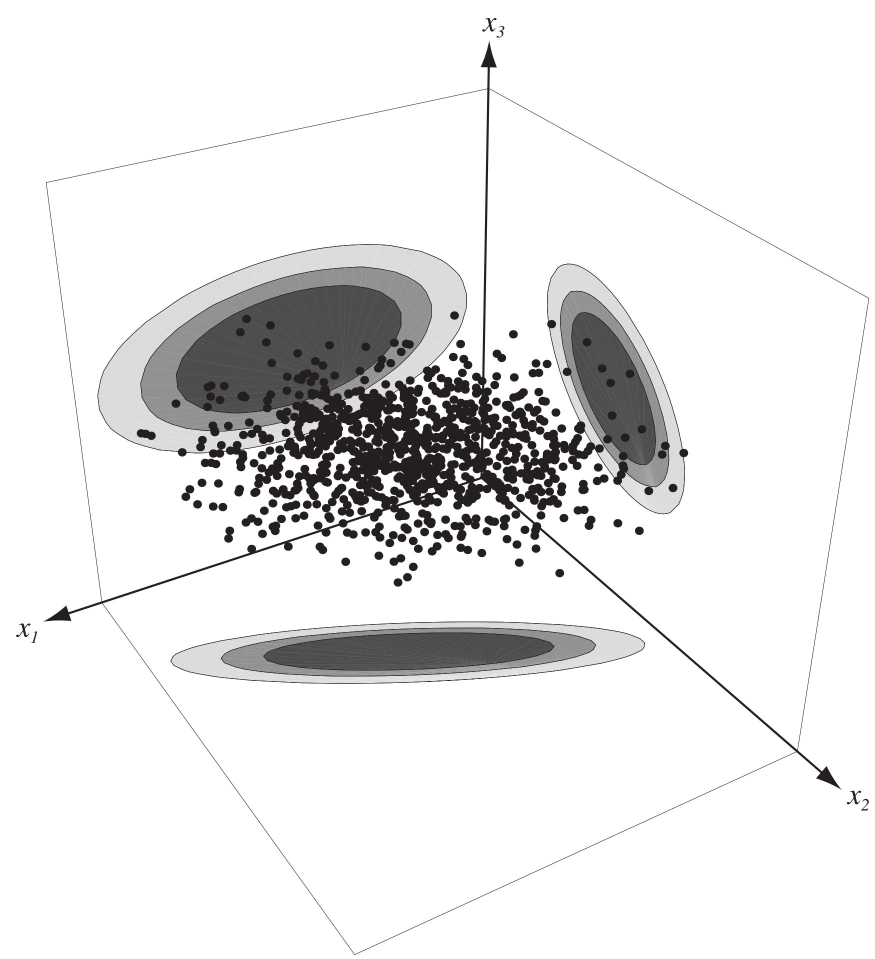
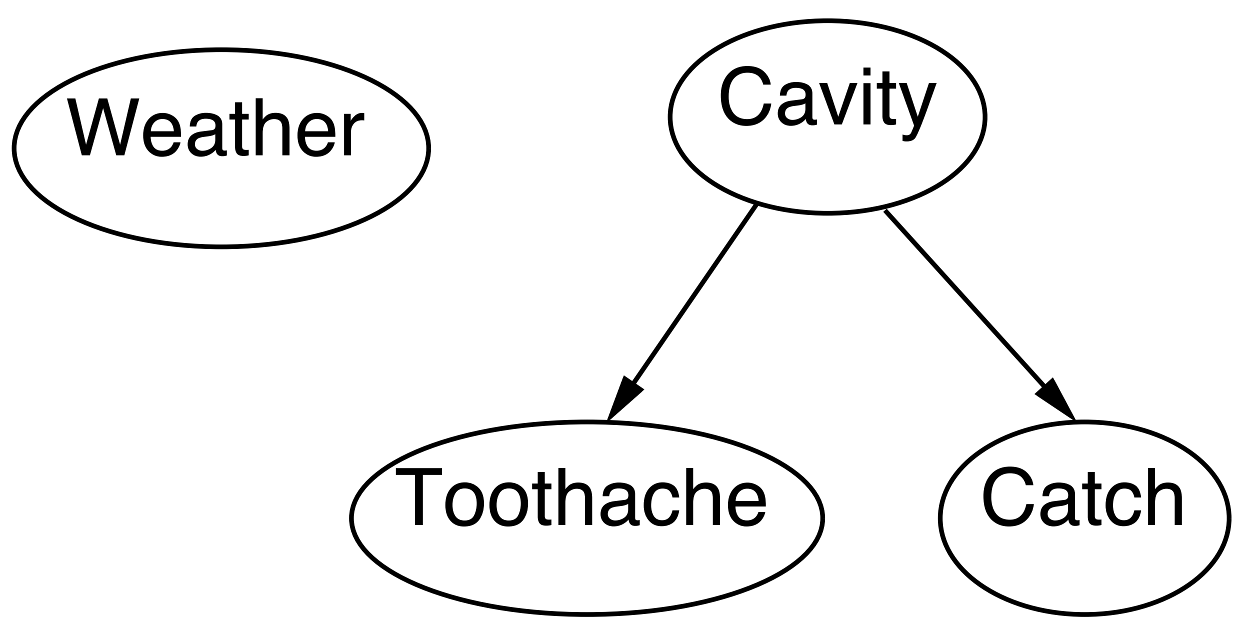


FIGURE 2.23. A three-dimensional distribution which obeys $p(x_1, x_3) = p(x_1)p(x_3)$; thus here x_1 and x_3 are statistically independent but the other feature pairs are not. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Simple Example of Conditional Independence

From Russell and Norvig

- Consider a simple example consisting of four variables: the weather, the presence of a cavity, the presence of a toothache, and the presence of other mouth-related variables such as dry mouth.
- The weather is clearly independent of the other three variables.
- And the toothache and catch are conditionally independent given the cavity (one as no effect on the other given the information about the cavity).

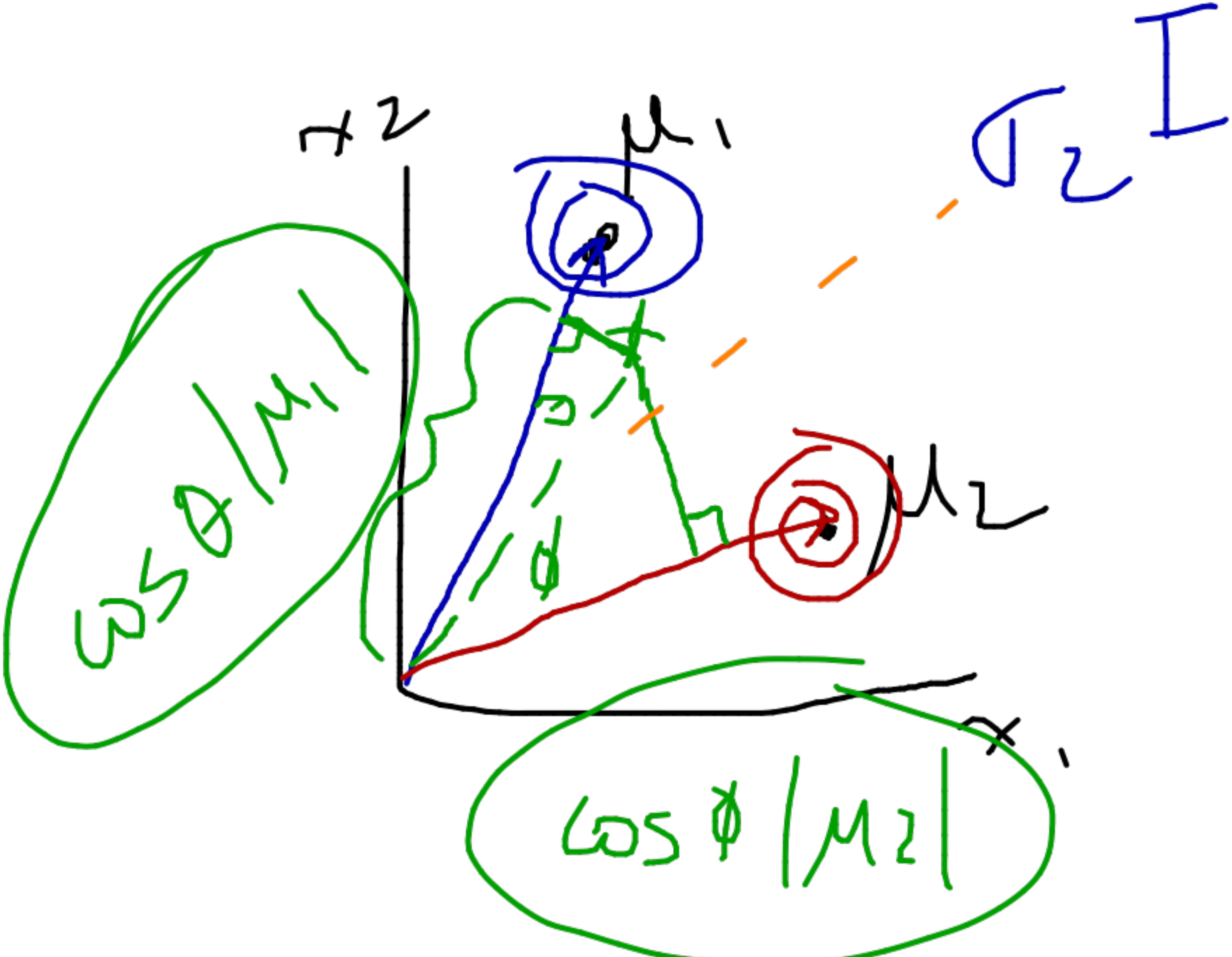


Naïve Bayes Rule

- If we assume that all of our individual features $x_i, i = 1, \dots, d$ are conditionally independent given the class, then we have

$$p(\omega_k|\mathbf{x}) \propto \prod_{i=1}^{d} p(x_i|\omega_k) \tag{73}$$

- Circumvents issues of dimensionality.
- Performs with surprising accuracy even in cases violating the underlying independence assumption.



x2
μ1
σ2 I
μ2
x1
φ
cos φ |μ1|
cos φ |μ2|