# Pattern Recognition - Summary

Friedrich-Alexander-Universität Erlangen-Nürnberg

Winter Term 2025/26

January 9, 2026

## Contents

# 1 Pattern Recognition Basics

## 1.1 Postulates

1. Availability of a representative sample $\omega$ of patterns ${}^i f(x)$ for the given field of problems $\Omega$

$$\omega = \{{}^1 f(x), \ldots, {}^N f(x)\} \subseteq \Omega$$

2. A (simple) pattern has features, which characterize its membership in a certain class $\Omega_\kappa$.
3. Compact domain in the feature space of features of the same class; domains of different classes are (reasonably) separable.
    - small intra-class distance
    - high inter-class distance

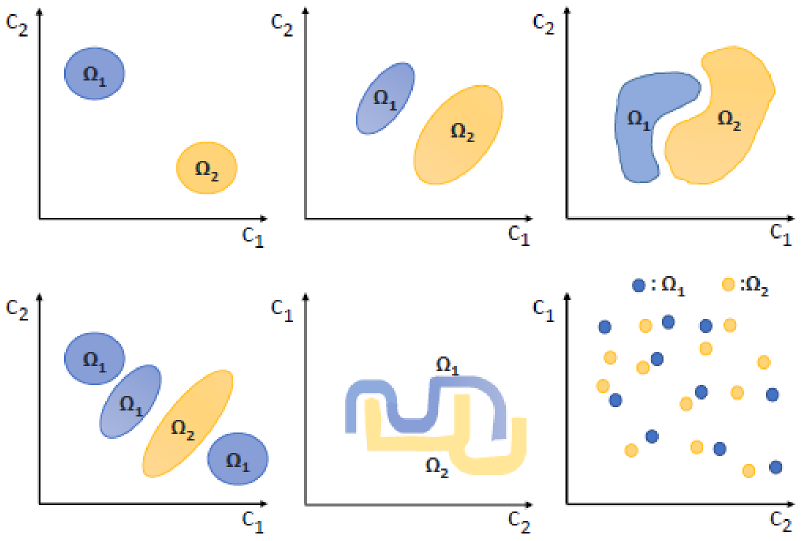


Figure 1: Visualization of compactness in feature space

4. A (complex) pattern consists of simpler constituents, which have certain relations to each other. A pattern may be decomposed into these constituents.
5. A (complex) pattern $\mathbf{f}(\mathbf{x}) \in \Omega$ has a certain structure. Not any arrangement of simple constituents is a valid pattern. Many patterns may be represented with relatively few constituents.
6. Two patterns are similar if their features or simpler constituents differ only slightly.

## 1.2 Definitions

| | |
|---|---|
| $\mathbf{x} \in \mathbb{R}^d$ : | $d$-dimensional feature vector |
| $y$ : | class number (usually $y \in \{0,1\}$ or $y \in \{-1,+1\}$) |
| $p(y)$ : | prior probability of pattern class $y$ |
| $p(\mathbf{x})$ : | evidence (distribution of features in $d$-dimensional feature space) |
| $p(\mathbf{x}, y)$ : | joint probability density function (pdf) |
| $p(\mathbf{x}\|y)$ : | class conditional density |
| $p(y\|\mathbf{x})$ : | posterior probability |

## 1.3 Statistics and Maths

### 1.3.1 Bayes' Theorem

$$\begin{aligned}\underbrace{p(\mathbf{x}, y)}_{\text{joint pdf}} &= \underbrace{p(y)}_{\text{prior}} \cdot \underbrace{p(\mathbf{x}|y)}_{\text{class conditional pdf}} \\ &= \underbrace{p(\mathbf{x})}_{\text{evidence}} \cdot \underbrace{p(y|\mathbf{x})}_{\text{posterior}}\end{aligned}$$

From this we can compute the posterior probability:

$$p(y|\mathbf{x}) = \frac{p(y) \cdot p(\mathbf{x}|y)}{p(\mathbf{x})} \quad \text{with} \quad p(\mathbf{x}) = \sum_{y'} p(y') \cdot p(\mathbf{x}|y') \tag{1.1}$$

**Note on Marginalization**
The evidence $p(\mathbf{x})$ is a marginal of $p(\mathbf{x}, y)$:

- We get $p(\mathbf{x})$ by marginalizing $p(\mathbf{x}, y)$ over $y$:

$$p(\mathbf{x}) = \sum_{y} p(y) \cdot p(\mathbf{x}|y)$$

- Accordingly, we get $p(y)$ by marginalizing $p(\mathbf{x}, y)$ over $\mathbf{x}$, i.e.:

$$p(y) = \int p(\mathbf{x}, y) d\mathbf{x}$$

Basically, marginalization means summing/integrating out the unwanted variable.
Notice: $y$ is a discrete random variable whereas $\mathbf{x}$ is a continuous random vector (summation vs. integration).

### 1.3.2 Covariance Matrix

The covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ of a $d$-dimensional random vector $\mathbf{x}$ is defined as:

$$\Sigma_{ij} = \mathrm{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] \tag{1.2}$$

**Structure:** The covariance matrix is organized as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix} \tag{1.3}$$

The diagonal elements $\sigma_{ii}$ represent the variances of the individual components $x_i$ of $\mathbf{x}$, while the off-diagonal elements $\sigma_{ij}$ represent the covariances between components $x_i$ and $x_j$.

**Practical Computation:**

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^{N} (\vec{x}_i - \vec{\mu}_i)(\vec{x}_i - \vec{\mu}_i)^T \tag{1.4}$$

For a specific class only use the samples belonging to that class.
Alternatively, in matrix form with centered data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \mathbf{X}\mathbf{X}^T \tag{1.5}$$

**Properties:**

- $\mathbf{\Sigma}$ is symmetric: $\mathbf{\Sigma} = \mathbf{\Sigma}^T$
- $\mathbf{\Sigma}$ is positive semi-definite: $\mathbf{z}^T\mathbf{\Sigma}\mathbf{z} \geq 0$ for all $\mathbf{z} \in \mathbb{R}^d$
- Diagonal elements $\sigma_{ii}$ are the variances of the individual components $x_i$ of $\mathbf{x}$
- Off-diagonal elements $\sigma_{ij}$ are the covariances between components $x_i$ and $x_j$

### 1.3.3 The Gaussian

A $d$-dimensional Gaussian (Normal) distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is defined as:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{1.6}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector.
**Note**:

- The probability for an interval $[\mathbf{a}, \mathbf{b}]$ is given by the integral of the pdf over this interval: $\int_{\mathbf{a}}^{\mathbf{b}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x}$
- The Integral over the whole space is 1: $\int_{-\infty}^{\infty} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} = 1$
- The probability at a single point is 0: $\mathcal{N}(\mathbf{x}_0; \boldsymbol{\mu}, \mathbf{\Sigma}) = 0$

### 1.3.4 Maximum Likelihood Estimation (MLE)

- Assume mutually independent and identically distributed (i.i.d.) samples $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N\}$
- The observations $x$ are kept fixed
- We optimize the log-likelihood function w.r.t. the parameters $\boldsymbol{\theta}$ of the distribution $p(\mathbf{x}|\boldsymbol{\theta})$

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \prod_{n=1}^{N} p(\mathbf{x}^n|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log p(\mathbf{x}^n|\boldsymbol{\theta})
\end{aligned} \tag{1.7}$$

### 1.3.5 Maximum A-posteriori (MAP)

- The probability density function of the parameters $p(\boldsymbol{\theta})$ is known (prior knowledge)
- We optimize the posterior probability of the parameters given the data

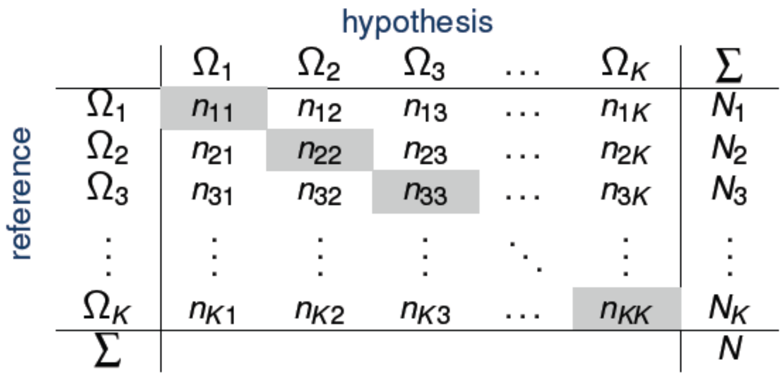
$$\begin{aligned}
\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\ p(\theta|x) \\
&= \underset{\theta}{\operatorname{argmax}}\ \frac{p(\theta)p(x|\theta)}{\sum_{\theta} p(\theta)p(x|\theta)} \\
&= \underset{\theta}{\operatorname{argmax}}\ \log p(\theta) + \log p(x|\theta)
\end{aligned} \tag{1.8}$$

## 1.4 Performance Evaluation

### 1. General Classification ($K$ Classes)

Feature vectors are used as input for the classifier, resulting in a discrete class index.

- **Confusion Matrix:** Visualization of the predictions vs. references.


hypothesis
reference

| | $\Omega_1$ | $\Omega_2$ | $\Omega_3$ | $\dots$ | $\Omega_K$ | $\Sigma$ |
|---|---|---|---|---|---|---|
| $\Omega_1$ | $n_{11}$ | $n_{12}$ | $n_{13}$ | $\dots$ | $n_{1K}$ | $N_1$ |
| $\Omega_2$ | $n_{21}$ | $n_{22}$ | $n_{23}$ | $\dots$ | $n_{2K}$ | $N_2$ |
| $\Omega_3$ | $n_{31}$ | $n_{32}$ | $n_{33}$ | $\dots$ | $n_{3K}$ | $N_3$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $\Omega_K$ | $n_{K1}$ | $n_{K2}$ | $n_{K3}$ | $\dots$ | $n_{KK}$ | $N_K$ |
| $\Sigma$ | | | | | | $N$ |

Figure 2: Confusion matrix with absolute frequencies for a $K$-class problem

**Evaluation Metrics:**

- **Accuracy / Recognition Rate (RR):**

$$\text{RR} := \frac{1}{N} \sum_{\kappa=1}^{K} n_{\kappa\kappa} \cdot 100\%$$

- **Recall and Precision** (for a specific class $\kappa$):

$$\text{recall}_\kappa = \frac{n_{\kappa\kappa}}{\sum_{i=1}^{K} n_{\kappa i}} = \frac{n_{\kappa\kappa}}{N_\kappa}$$

$$\text{precision}_\kappa = \frac{n_{\kappa\kappa}}{\sum_{i=1}^{K} n_{i\kappa}}$$

- **(Unweighted) Average Recall (UAR):**

$$\text{UAR} := \frac{1}{K} \sum_{\kappa=1}^{K} \frac{n_{\kappa\kappa}}{N_\kappa} \cdot 100\%$$

**2. Special Case: Binary Classification (2 Classes)**

For 2 classes (Positive / Negative), the confusion matrix simplifies to:

| | | **Reference** | |
|---|---|---|---|
| | | Positive | Negative |
| **Hyp.** | Positive | True Positive (TP) | False Positive (FP) |
| | Negative | False Negative (FN) | True Negative (TN) |

**Performance Measures:**

- **Accuracy:**

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

- **F-measure** (Harmonic mean of recall and precision):

$$F = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$$

- **Detailed Rates:**
  - True Positive Rate (Hit Rate, Recall, **Sensitivity**):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  - False Positive Rate (False Alarm Rate, Fall-out):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

  - Positive Predictive Value (**Precision**):

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

  - True Negative Rate (**Specificity**):

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} = 1 - \text{FPR}$$

## 3. Receiver Operating Characteristic (ROC)

Classification is often a threshold decision (e.g., based on eye pressure).

- The TPR and FPR can be computed for different thresholds $\theta$.
- Higher TPR usually leads to higher FPR (trade-off).

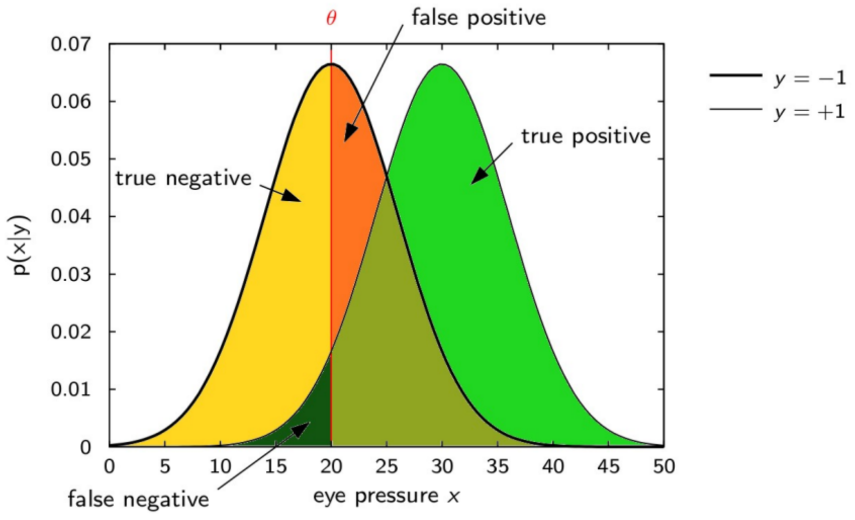


Figure 3: Classification based on threshold $\theta$

**The ROC Curve:** A classifier defines a single 2-d point in the ROC graph. Varying $\theta$ creates the curve.

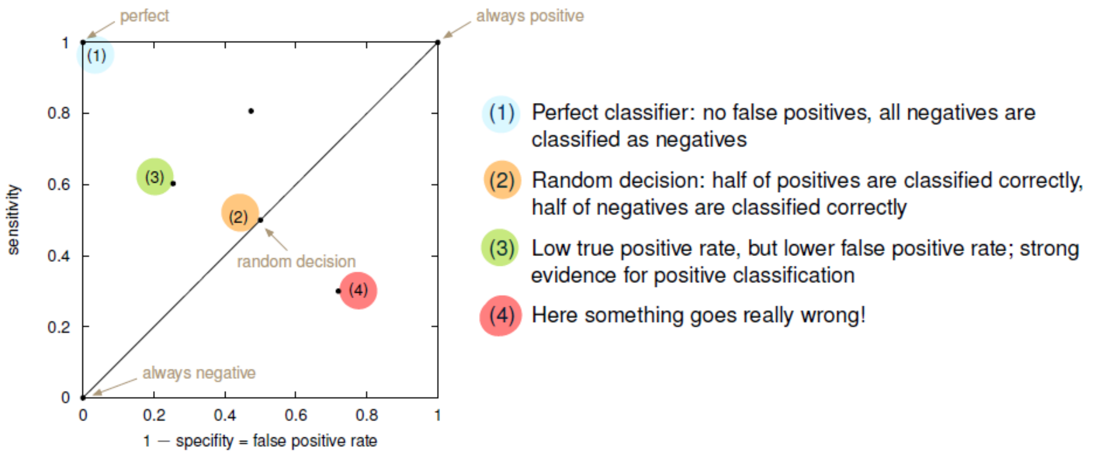


Figure 4: ROC Curve: Sensitivity vs. (1 - Specificity)

**Area Under Curve (AUC):**

- Used to compare classifiers (bigger AUC $\rightarrow$ better classifier).
- TPR and FPR do not depend on the total number of samples.
- Hence, the ROC curve is **independent of the priors** of both classes (as opposed to recall-precision-curves).

## 2 Bayesian Classifier

The Bayesian classifier is based on Bayes' theorem and assumes that the class-conditional densities are known. It computes the posterior probability of each class given the feature vector and assigns the pattern to the class with the highest posterior probability.
**Decision Rule**

$$y^* = \arg\max_y p(y|\mathbf{x}) \tag{2.1}$$

We can rewrite this using Bayes' theorem:

$$p(y|\mathbf{x}) = \frac{p(y)p(\mathbf{x}|y)}{p(\mathbf{x})} \tag{2.2}$$

Since $p(\mathbf{x})$ is the same for all classes (it does not depend on $y$), we can simplify the decision rule to:

$$y^* = \arg\max_y p(y)p(\mathbf{x}|y) \tag{2.3}$$

where $p(y)$ is the prior probability of class $y$ and $p(\mathbf{x}|y)$ is the class-conditional density.
**Optimality of the Bayesian Classifier**
The Bayesian classifier is the optimal classifier if we use a 0-1 loss function. Let $L(y', y)$ be the loss incurred when we decide class $y'$ while the true class is $y$. For a 0-1 loss function:

$$L(y', y) = \begin{cases} 0, & \text{if } y' = y \\ 1, & \text{if } y' \neq y \end{cases} \tag{2.4}$$

**Note:**

- If all class class conditional densities are Gaussian, the Bayesian classifier becomes a Gaussian classifier.
- If all covariance matrices are diagonal, we get a Naive Bayes classifier.
- If all covariance matrices are equal and the Identity matrix, we get a Nearest neighbor classifier.

## 3 Logistic Regression

Logistic Regression is a discriminative model. We model the posterior probability $p(y|\mathbf{x})$ directly instead of modeling the class-conditional densities. We use the logistic (sigmoid) function to map the linear combination of input features to a probability value between 0 and 1.

$$g(x) = \frac{1}{1+e^{-x}} \tag{3.1}$$

$$g'(x) = g(x)(1-g(x)) \tag{3.2}$$

We can express the posterior probability as follows:

$$p(y=0|\mathbf{x}) = \frac{1}{1+e^{-F(\mathbf{x})}} \quad \text{and} \quad p(y=1|\mathbf{x}) = 1 - p(y=0|\mathbf{x}) = \frac{1}{1+e^{F(\mathbf{x})}} \tag{3.3}$$

**Decision Boundary:**
The decision boundary is defined by $F(\mathbf{x}) = 0$. This mean any point $\mathbf{x}$ on the decision boundary satisfies:

$$\log \frac{p(y=0|\mathbf{x})}{p(y=1|\mathbf{x})} = \log \frac{0.5}{0.5} = \log 1 = 0 \tag{3.4}$$

If we plug a Gaussian in for the class conditional densities, we get:

$$\log \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}}{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}} =$$

$$= \frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right)$$

$$= \frac{1}{2}\left(\underbrace{\mathbf{x}^T(\Sigma_1^{-1}-\Sigma_0^{-1})\mathbf{x}}_{\text{Quadratic Term } (\sim x^2)} \underbrace{-2(\boldsymbol{\mu}_1^T\Sigma_1^{-1}-\boldsymbol{\mu}_0^T\Sigma_0^{-1})\mathbf{x}}_{\text{Linear Term } (\sim x)} + \underbrace{(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0)}_{\text{Constant}}\right)$$

This can be formulated as a quadratic equation in $\mathbf{x}$.

$$F(\mathbf{x}) = \underbrace{\mathbf{x}^T A\mathbf{x}}_{\text{Quadratic Term}} + \underbrace{\alpha^T\mathbf{x}}_{\text{Linear Term}} + \alpha_0 = 0 \tag{3.5}$$

Now we can think about the shape of the decision boundary:

- If $\Sigma_0 \neq \Sigma_1$: The decision boundary is quadratic (e.g., hyperbola, parabola, ellipse).
- If $\Sigma_0 = \Sigma_1 = \Sigma$: The quadratic term vanishes and the decision boundary is linear (a hyperplane).

Note: We can lift the dimension of $\mathbf{x}$ to transform a non-linear decision boundary in the original space to a linear decision boundary in the higher-dimensional space.

**Parameter Estimation:**
We can rewrite equation 3.5 as a linear combination by lifting the input vector $\mathbf{x}$ to $\tilde{\mathbf{x}}$:

$$F(\mathbf{x}) = \theta^T\mathbf{x}' = 0 \tag{3.6}$$

We can use the parameter vector $\theta$ to parametrize the logistic function (I'm too lazy to write the tilde everywhere):

$$g(\theta^T\mathbf{x}) = \frac{1}{1+e^{-\theta^T\mathbf{x}}} \tag{3.7}$$

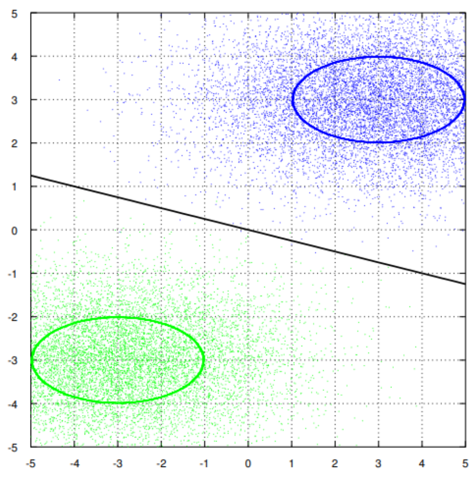


Figure 5: Linear Decision Boundary: Equal Covariance Matrices lead to a linear decision boundary (hyperplane).

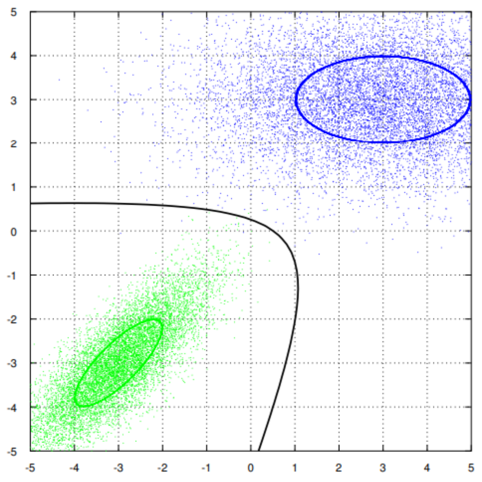


Figure 6: Quadratic Decision Boundary: Unequal Covariance Matrices lead to a quadratic decision boundary (e.g., hyperbola, parabola, ellipse).

The posterior probabilities are now given by:

$$p(y=0|\mathbf{x}) = g(\theta^T\mathbf{x}) \tag{3.8}$$

$$p(y=1|\mathbf{x}) = 1 - g(\theta^T\mathbf{x}) \tag{3.9}$$

We have to eastimate the parameters from a set $S$ of $m$ training samples:

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\} \tag{3.10}$$

Before we start with the estimation, lets reqwrite the psoteriors using a bernoulli distribution:

$$p(y|\mathbf{x}) = (g(\theta^T\mathbf{x}))^{1-y}(1-g(\theta^T\mathbf{x}))^y \tag{3.11}$$

Now we can start. We use maximum likelihood estimation (MLE) and assume the training samples are mutually independent. We compute the log-likelihood function:

$$\begin{aligned}
\mathcal{L}(\theta) &= \log \prod_{i=1}^{m} p(y_i|\mathbf{x}_i) \\
&= \sum_{i=1}^{m} \left( g\left(\theta^T\mathbf{x}_i\right)^{y_i} \left(1-g\left(\theta^T\mathbf{x}_i\right)\right)^{1-y_i} \right) \\
&= \sum_{i=1}^{m} \left( y_i \log g\left(\theta^T\mathbf{x}_i\right) + (1-y_i)\log\left(1-g\left(\theta^T\mathbf{x}_i\right)\right)\right) \\
&= \sum_{i=1}^{m} \left( y_i\theta^T\mathbf{x}_i + \log\left(1-g\left(\theta^T\mathbf{x}_i\right)\right)\right)
\end{aligned}$$

Our goal is to maximize the log-likelihood function or minimize the negative log-likelihood function:

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta) = \arg\min_\theta -\mathcal{L}(\theta) \tag{3.12}$$

We can do this using Newton's method 8.2.4 or gradient descent 8.2.2.

## 4 Naive Bayes

The Naive Bayes (Idiots Bayes) bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is particularly suited for high-dimensional data.
The indepence assumption implies:

$$P(\mathbf{x}|y) = \prod_{i=1}^{d} P(x_i|y) \tag{4.1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is the feature vector and $y$ is the class label.
Using Bayes' theorem, we can formulate the decision rule as:

$$\begin{aligned} y^* &= \arg\max_y p(y|\mathbf{x}) \\ &= \arg\max_y p(y)p(\mathbf{x}|y) \\ &= \arg\max_y p(y) \prod_{i=1}^{d} p(x_i|y) \end{aligned} \tag{4.2}$$

**Decision Boundary:**

- Quadratic, if the classes have different covariance matrices (i.e., $\Sigma_k \neq \Sigma_j$ for some $k \neq j$)
- Linear, if all classes share the same covariance matrix (i.e., $\Sigma_k = \Sigma$ for all $k$)

Note: If all classes share the same covariance matrix and the covariance matrix is the identity matrix $\Sigma = I$, we obtain the **Nearest Neighbor Classifier**.

### 4.1 Statistical Dependency of limited order

If we do not want to assume full independence between features, but want to keep the number of parameters manageable, we can assume statistical dependency of limited order. This means that each feature $x_i$ depends only on a limited number of other features, say $k$ features. This leads to the following formulation:

$$p(\mathbf{x}|y) = \prod_{i=1}^{d} p(x_i|y, x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \tag{4.3}$$

where $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ are the $k$ features that $x_i$ depends on.

**First order dependency (i.e., $k = 1$):**

$$p(\mathbf{x}|y) = p(x_1|y) \cdot p(x_2|y, x_1) \cdot p(x_3|y, x_2) \cdots p(x_d|y, x_{d-1}) \tag{4.4}$$

So each feature depends only on the previous feature in addition to the class label $y$.
In this case the covariance matrix would be a **band matrix** with bandwidth 1:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & \cdots & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & 0 \\ 0 & \sigma_{32} & \sigma_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{dd} \end{pmatrix} \tag{4.5}$$

**Second order dependency (i.e., $k = 2$):**

$$p(\mathbf{x}|y) = p(x_1|y) \cdot p(x_2|y, x_1) \cdot p(x_3|y, x_1, x_2) \cdots p(x_d|y, x_{d-2}, x_{d-1}) \tag{4.6}$$

In this case the covariance matrix would be a **band matrix** with bandwidth 2:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & 0 & \cdots & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \cdots & 0 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \cdots & 0 \\ 0 & \sigma_{42} & \sigma_{43} & \sigma_{44} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{dd} \end{pmatrix} \tag{4.7}$$

Here, each feature depends on the two previous features in addition to the class label $y$.

**Parameter tying:**
Here we assume that the covariance for each feature is the same across all classes, i.e., $\sigma_{i|y=k} = \sigma_i$ for all classes $k$. We can still assume some dependency among the features, e.g., first order dependency. In this case the covariance matrix would be a diagonal matrix with the same entries for each class:

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_{12} & 0 & \cdots & 0 \\ \sigma_{21} & \sigma_2 & \sigma_{23} & \cdots & 0 \\ 0 & \sigma_{32} & \sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_d \end{pmatrix} \tag{4.8}$$

# 5 Discriminant Analysis

## 5.1 Motivation

From chapter 2 we know, that the decison boundary of Gaussian classifiers is quadratic in general. However, if we assume that all classes share the same covariance matrix and $\Sigma = I$ (Identity matrix), the decision boundary not only becomes linear, but we get a Nearest Neighbour classifier. Based on this we seek to find a transformation that maps the data into a space where the class covariances are equal and the Identity matrix.
Another observation we made is, that lifting feature vectors into higher dimensions can result in a linear decision boundary.
And if we already are in higher dimensions, why not try to find a lower-dimensional space that still allows for good class separation?
**Goal:**

1. Transformation that gives $\Sigma = I$ for all classes
2. Lower dimensional space with good class separability
3. Linear decision boundary in the transformed space

## 5.2 From quadratic to linear decision boundaries

By using SVD we can decompose the covariance matrix of the data as follows:

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{U}\mathbf{D}^{1/2})^T = (\mathbf{U}\mathbf{D}^{1/2}) \cdot \mathbf{I} \cdot (\mathbf{U}\mathbf{D}^{1/2})^T \tag{5.1}$$

Where $\mathbf{U}$ is the orthogonal matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues. We want to incorporate this into the Gaussian density function. This requires to compute the inverse and determinant of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T = (\mathbf{U}\mathbf{D}^{-1/2})(\mathbf{U}\mathbf{D}^{-1/2})^T \tag{5.2}$$

$$\det\mathbf{\Sigma} = \prod_i^d d_{i,i} \tag{5.3}$$

Now we can rewrite the Gaussian density function as follows:

$$\begin{aligned}
\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma) &= \frac{1}{\sqrt{\det 2\pi\Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{\sqrt{\det 2\pi\Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \underbrace{(UD^{-\frac{1}{2}}) \cdot I \cdot (UD^{-\frac{1}{2}})^T}_{\Sigma^{-1}} (\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{\sqrt{\det 2\pi\Sigma}} e^{-\frac{1}{2}\left(\underbrace{(D^{-\frac{1}{2}}U^T)\mathbf{x}}_{\text{transformed } \vec{x}} - \underbrace{(D^{-\frac{1}{2}}U^T)\boldsymbol{\mu}}_{\text{transformed } \vec{\mu}}\right)^T \underbrace{I}_{\substack{\text{Covariance matrix} \\ \text{becomes an Identity matrix} \\ \Sigma=1}} \left(\underbrace{(D^{-\frac{1}{2}}U^T)\mathbf{x}}_{\text{transformed } \vec{x}} - \underbrace{(D^{-\frac{1}{2}}U^T)\boldsymbol{\mu}}_{\text{transformed } \vec{\mu}}\right)}
\end{aligned}$$

From this we can read off the transformation that maps the data into a space where the covariance matrix is the Identity matrix:

$$x' = \Phi(x) = \mathbf{D}_y^{-\frac{1}{2}}\mathbf{U}_y^T\mathbf{x} \tag{5.4}$$

The transformation $\Phi(x)$ is called **whitening transformation**.
We can also show that $\mathbf{x}'$ is normally distributed:

$$p(\mathbf{x}'|y) = \mathcal{N}(\mathbf{x}';\boldsymbol{\mu}_{\boldsymbol{y}}', \Sigma_y') = \mathcal{N}(\mathbf{x}';\mathbf{D}_y^{-\frac{1}{2}}\mathbf{U}_y^T\boldsymbol{\mu}_{\boldsymbol{y}}, \mathbf{I}) \tag{5.5}$$

## 5.3 Linear Discriminant Analysis (LDA)

What we have seen above has one big disadvantage: The transformation is class dependent! This means each class gets mapped into its own space. To solve this problem we compute the transformation based on the joint covariance matrix:

1. **ML estimation of the joint covariance matrix:**

$$\widehat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})(\mathbf{x}_i - \boldsymbol{\mu}_{y_i})^T$$

2. **Compute SVD of covariance matrix:**

$$\widehat{\Sigma} = UDU^T$$

3. **Assign transform:**

$$\phi = D^{-\frac{1}{2}}U^T$$

4. **Compute mean vectors for all** $y$

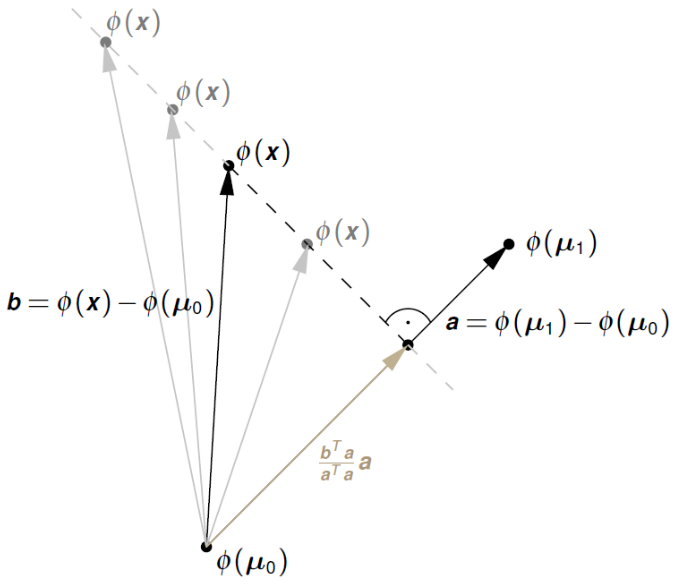
$$\boldsymbol{\mu}'_y = \phi(\boldsymbol{\mu}_y) = D^{-\frac{1}{2}}U^T\boldsymbol{\mu}_y$$

**Output:** feature transform $\phi$, transformed mean vectors $\boldsymbol{\mu}'_y$

**Decision Rule:**

$$\begin{aligned} y^* &= \underset{y}{\operatorname{argmax}}\, p(y \mid \phi(\mathbf{x})) \\ &= \underset{y}{\operatorname{argmax}} \left\{ \log p(y) - \frac{1}{2}(\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}_y))^T(\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}_y)) \right\} \\ &= \underset{y}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\phi(\mathbf{x}) - \phi(\boldsymbol{\mu}_y)\|_2^2 - \log p(y) \right\} \end{aligned}$$

where $\|\cdot\|_2$ denotes the $L_2$ norm. Note that this basically is a biased nearest neighbour classifier in the transformed space.
This gets clearer when looking at the following visualization:



$\phi(\boldsymbol{x})$
$\phi(\boldsymbol{x})$
$\phi(\boldsymbol{x})$
$\phi(\boldsymbol{x})$
$\phi(\boldsymbol{\mu}_1)$
$\boldsymbol{b} = \phi(\boldsymbol{x}) - \phi(\boldsymbol{\mu}_0)$
$\boldsymbol{a} = \phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_0)$
$\frac{b^T a}{a^T a} a$
$\phi(\boldsymbol{\mu}_0)$

Figure 7: Nearest Neighbour classification for two classes

The classification of a new LDA-transformed sample $\Phi(\mathbf{x})$ only depends on the distance of the projected sample on a line connecting the two class means $\boldsymbol{\mu}'_0$ and $\boldsymbol{\mu}'_1$. The direction orthogonal to this line is not relevant for classification.
Based on this observation we can also use the angle between the transformed sample $\Phi(\mathbf{x})$ and $\boldsymbol{\mu}'_1 - \boldsymbol{\mu}'_0$ for classification. Our decision rule can also be written as:

$$y^* = \begin{cases} 0, & \text{if } \phi(\mathbf{x})^T(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_0)) < \frac{1}{2}(\phi(\boldsymbol{\mu}_1)^T\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_0)^T\phi(\boldsymbol{\mu}_0)) \\ 1, & \text{otherwise.} \end{cases}$$

At a first look this might seem overwhelming, but all we do is projecting $\Phi(\mathbf{x})$ onto the line connecting the two class means and makeing a decision based on which side of the midpoint the projection lies.

## 5.4 Rank-Reduced Linear Discriminant Analysis

The main target of Rank-Reduced LDA is to find a transformation $\Phi$ that projects features into a lower-dimensional space ($L < K - 1$) where the spread (variance) of the features is maximized. Basically we try to maximize the distance between class means. In terms of an optimization problem this can be written as follows:

$$\mathbf{\Phi}^* = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \left( \frac{1}{K} \sum_{y=1}^{K} (\mathbf{\Phi}\boldsymbol{\mu}'_y - \mathbf{\Phi}\bar{\boldsymbol{\mu}}')^T (\mathbf{\Phi}\boldsymbol{\mu}'_y - \mathbf{\Phi}\bar{\boldsymbol{\mu}}') + \sum_{i=1}^{L} \lambda_i (\|\mathbf{\Phi}_i\|_2^2 - 1) \right)$$

where $\bar{\boldsymbol{\mu}}' = \frac{1}{K}\sum_{y=1}^{K} \boldsymbol{\mu}'_y$ is the global mean of all class means and $\mathbf{\Phi}_i$ is the $i$-th row of the transformation matrix $\mathbf{\Phi}$. The constraints $\|\mathbf{\Phi}_i\|_2^2 = 1$ ensure that the rows of $\mathbf{\Phi}$ are unit vectors and are necessary because $\Phi \to \infty$ would solve.
The solution to this optimization problem is given by an eigenvector-eigenvalue problem.

$$\mathbf{\Sigma}_{\text{inter}}\mathbf{\Phi}^T = \lambda'\mathbf{\Phi}^T \tag{5.6}$$

where $\mathbf{\Sigma}_{\text{inter}}$ is the inter-class scatter matrix, $\mathbf{\Phi}$ is the transformation matrix and $\lambda'$ is the diagonal matrix of eigenvalues.

**Input:** training data: $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_m, y_m)\}$

1. Compute the **covariance matrix of transformed mean vectors**

$$\widehat{\Sigma}_{\text{inter}} = \frac{1}{K} \sum_{y=1}^{K} (\boldsymbol{\mu}'_y - \bar{\boldsymbol{\mu}}')(\boldsymbol{\mu}'_y - \bar{\boldsymbol{\mu}}')^T,$$

   where $\bar{\boldsymbol{\mu}}' = \frac{1}{K} \cdot \sum_{y=1}^{K} \boldsymbol{\mu}'_y$.

2. Compute the $L$ **eigenvectors of the covariance matrix** belonging to the largest eigenvalues.

3. The **eigenvectors are the rows** of the mapping $\mathbf{\Phi}$ from the $(K-1)$- to the $L$-dimensional feature space.

**Output:** matrix $\mathbf{\Phi}$

## 5.5 Fisher Transform

From the postulates or PR we know that it is desireable to have small intra-class distances and large inter-class distances. The Fisher transform tries to find a transformation that optimizes this criterion. We define the following optimization problem:

$$\boldsymbol{r}^* = \underset{\boldsymbol{r}}{\operatorname{argmax}}\, J(\boldsymbol{r}) = \underset{\boldsymbol{r}}{\operatorname{argmax}} \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

With a bit of maths this can be reformulated to:

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}}{\operatorname{argmax}} \frac{\boldsymbol{a}^T \Sigma_{\text{inter}} \boldsymbol{a}}{\boldsymbol{a}^T \Sigma_{\text{intra}} \boldsymbol{a}}$$

Formally this describes the maximization of the Rayleigh ratio for each projection axis $\boldsymbol{a}^*$.

## 5.6 Comments on Dimensionality Reduction

- PCA does not rquire class labels, LDA does
- PCA transformed features are approximately normally distributed (central limit theorem)
- Since LDA is a linear transformation, the transformed features are normally distributed if the original features are normally distributed
- Components of PCA transformed features are uncorrelated and mutually independent
- **John-Lindenstrauss-lemma:** If vectors are projected onto a randomly chosen subspace of sufficient dimension, then the distances between the points are approximately preserved.

# 6 Norms and Norm Dependent Linear Regression

Norms and similarity measures are fundamental concepts in pattern recognition and machine learning. They are essential for defining distances between feature vectors, formulating optimization problems, and regularizing models.

## 6.1 Inner Products and Norms

### 6.1.1 Inner Product

The inner product provides a way to measure the angle and length of vectors.

**Definition: Inner Product**

The inner product of two vectors $\vec{x}, \vec{y} \in \mathbb{R}^d$ is defined as:

$$\langle \vec{x}, \vec{y} \rangle = \vec{x}^T \vec{y} = \sum_{i=1}^{d} x_i y_i \tag{6.1}$$

This concept extends to matrices. For two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, the inner product is defined using the trace operator:

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}(\mathbf{X}^T \mathbf{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} y_{ij} \tag{6.2}$$

### 6.1.2 Norms

A norm is a function $\|\cdot\| : \mathcal{X} \to \mathbb{R}$ that assigns a strictly positive length or size to a vector. To be a valid norm, the function must satisfy specific axioms.

- **Non-negativity:** $\forall \vec{x} : \|\vec{x}\| \geq 0$.
- **Definiteness:** $\|\vec{x}\| = 0$ if and only if $\vec{x} = \vec{0}$.
- **Homogeneity:** $\|a\vec{x}\| = |a| \cdot \|\vec{x}\|$ for any scalar $a \in \mathbb{R}$.
- **Triangle Inequality:** $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$.

**Common Vector Norms:** The general $L_p$-norm ($p \geq 1$) is defined as:

$$\|\vec{x}\|_p = \left( \sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}} \tag{6.3}$$

Specific important instances include:

- **$L_1$-norm (Manhattan):** $\|\vec{x}\|_1 = \sum_{i=1}^{d} |x_i|$. This norm sums the absolute values of the components.
- **$L_2$-norm (Euclidean):** $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$. This corresponds to the standard straight-line distance.
- **$L_\infty$-norm (Maximum):** $\|\vec{x}\|_\infty = \max_i |x_i|$. It takes the value of the largest component magnitude.

*Note on the $L_0$-"norm":* The $L_0$-norm counts the number of non-zero entries in a vector. Although often referred to as a norm, it is **not** a valid norm because it does not satisfy the homogeneity property.

**Quadratic Norms and Mahalanobis Distance:** Given a symmetric positive definite matrix $\mathbf{P}$, the quadratic $L_P$-norm is defined as:

$$\|\vec{x}\|_{\mathbf{P}} = \sqrt{\vec{x}^T \mathbf{P} \vec{x}} = \|\mathbf{P}^{\frac{1}{2}} \vec{x}\|_2 \tag{6.4}$$

If we choose $\mathbf{P} = \mathbf{\Sigma}^{-1}$ (the inverse covariance matrix), the distance induced by this norm corresponds to the **Mahalanobis distance**. This measure accounts for the correlation between variables in a dataset.

### 6.1.3 Unit Balls

The unit ball is the set of all vectors $\vec{x}$ such that $\|\vec{x}\| \leq 1$. The shape of the unit ball visualizes the geometric properties of the norm:

- **$L_1$-norm:** A diamond (rotated square) in 2D. This sharp geometry promotes sparsity in optimization.
- **$L_2$-norm:** A circle in 2D (or sphere in higher dimensions).
- **$L_\infty$-norm:** A square (or hypercube).
- **$L_p$-norm ($p < 1$):** Concave shapes (star-like). Since the set is not convex, these do not form valid norms.

## 6.2 Norm Dependent Linear Regression

Linear regression aims to approximate a target vector $\vec{b}$ using a linear model $\mathbf{A}\vec{x}$. The goal is to minimize the residual vector $\vec{r} = \mathbf{A}\vec{x} - \vec{b}$. The general optimization problem depends on the chosen norm:

$$\hat{\vec{x}} = \text{argmin}_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{b}\| \tag{6.5}$$

Different norms lead to different solutions and robustness properties.

### 6.2.1 Least-Squares Regression ($L_2$-Norm)

Minimizing the squared $L_2$-norm of the residuals is the most common approach:

$$\hat{\vec{x}} = \text{argmin}_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{b}\|_2^2 = \text{argmin}_{\vec{x}} (\mathbf{A}\vec{x} - \vec{b})^T (\mathbf{A}\vec{x} - \vec{b}) \tag{6.6}$$

To solve this, we expand the term and calculate the derivative with respect to $\vec{x}$:

$$f(\vec{x}) = \vec{x}^T \mathbf{A}^T \mathbf{A} \vec{x} - 2\vec{b}^T \mathbf{A} \vec{x} + \vec{b}^T \vec{b} \tag{6.7}$$

$$\nabla_{\vec{x}} f(\vec{x}) = 2\mathbf{A}^T \mathbf{A} \vec{x} - 2\mathbf{A}^T \vec{b} \tag{6.8}$$

Setting the gradient to zero yields the closed-form solution (Normal Equations):

$$\hat{\vec{x}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \vec{b} \tag{6.9}$$

This solution assumes that the columns of $\mathbf{A}$ are linearly independent.

### 6.2.2 Chebyshev and Sum of Absolute Residuals

- **Chebyshev Regression ($L_\infty$):** Minimizes the maximum individual residual ($\max_i |r_i|$). This optimization problem can be reformulated as a Linear Programming (LP) problem.
- **Sum of Absolute Residuals ($L_1$):** Minimizes $\sum |r_i|$. This can also be rewritten as an LP problem. The $L_1$-norm is significantly more **robust to outliers** than the $L_2$-norm, as large errors are not squared.

## 6.3 Regularization

In many cases, the problem is ill-posed or we want to enforce specific properties on the solution vector $\vec{x}$. This is achieved by adding a regularization term to the objective function.

### 6.3.1 Ridge Regression ($L_2$ Regularization)

We minimize the residual error plus the squared Euclidean length of the parameter vector:

$$\min_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{b}\|_2^2 + \lambda\|\vec{x}\|_2^2 \tag{6.10}$$

Geometrically, this constrains the solution to lie within or near an $L_2$ unit ball. It prevents overfitting by keeping weights small.

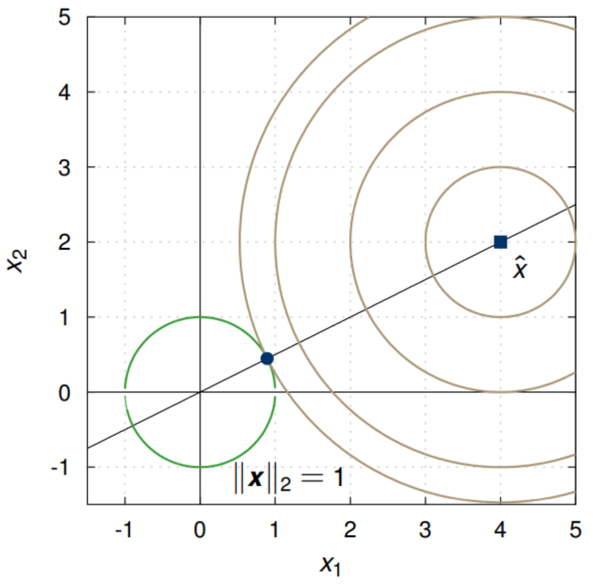


Figure 8: Geometric illustration of the L2-constrained optimization problem

### 6.3.2 Lasso ($L_1$ Regularization)

The Lasso (Least Absolute Shrinkage and Selection Operator) uses the $L_1$-norm for regularization:

$$\min_{\vec{x}} \|\mathbf{A}\vec{x} - \vec{b}\|_2^2 + \lambda\|\vec{x}\|_1 \tag{6.11}$$

Because the $L_1$ unit ball has "corners" (spikes) aligned with the axes, the solution often hits these corners. This forces some coefficients to become exactly zero (solution lies on one of the corrdinate axes), effectively performing **feature selection** and leading to **sparse solutions**. This property is exploited in **Compressed Sensing** to reconstruct signals from few measurements.

## 6.4 Penalty Functions

Instead of strictly using norms, we can define a general penalty function $\phi(r_i)$ for the residuals. The problem becomes:

$$\min \sum_{i=1}^{m} \phi(r_i) \quad \text{subject to} \quad \vec{r} = \mathbf{A}\vec{x} - \vec{b} \tag{6.12}$$

The choice of $\phi$ determines how the algorithm treats errors of different magnitudes.

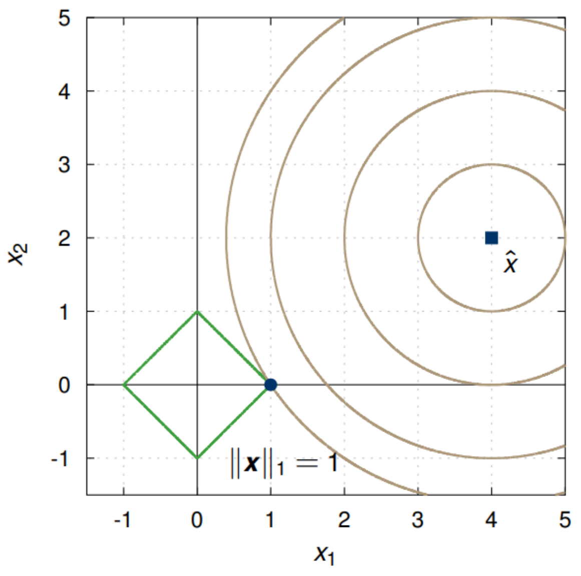


Figure 9: Geometric illustration of the L1-constrained optimization problem

- $L_2$**-Norm:** Parabolic shape. Heavy penalty for large errors (outliers).

$$\phi(r) = r^2 \tag{6.13}$$

- $L_1$**-Norm:** V-shape. Linear penalty for large errors, making it robust.

$$\phi(r) = |r| \tag{6.14}$$

- **Log Barrier:** The cost approaches infinity as residuals approach the limit $a$. This acts as a strict constraint (hard limit).

$$\phi_{Barrier}(r) = \begin{cases} -a^2 \log\left(1 - \left(\frac{r}{a}\right)^2\right) & \text{if } |r| \le a \\ \text{inf} & \text{otherwise} \end{cases} \tag{6.15}$$

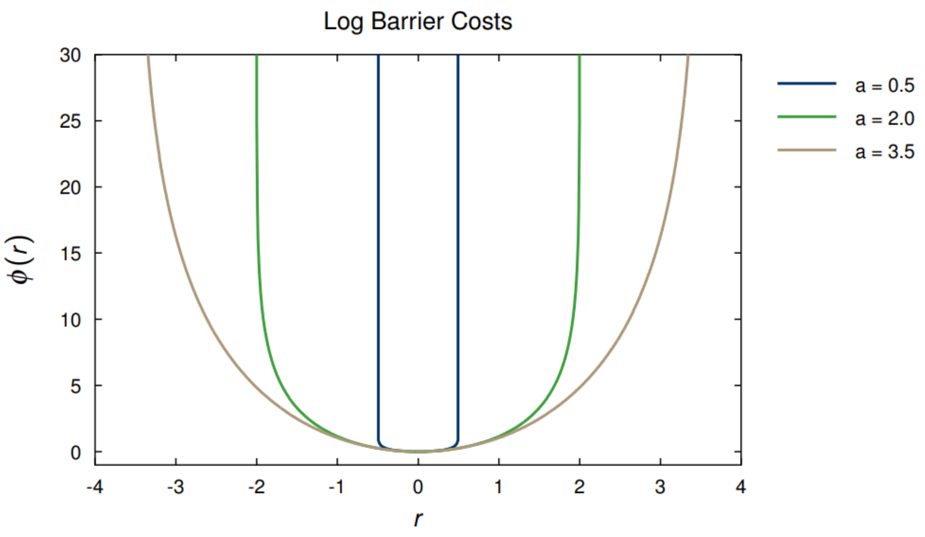


Figure 10: Geometric illustration of the Log-Barrier penalty function

- **Dead Zone:** Costs are zero for small residuals ($|r| \leq a$) and linear afterwards. This ignores small noise ("insensitivity").

$$\phi_{dz}(r) = \begin{cases} 0 & \text{if } |r| \leq a \\ |r| - a & \text{otherwise} \end{cases} \tag{6.16}$$

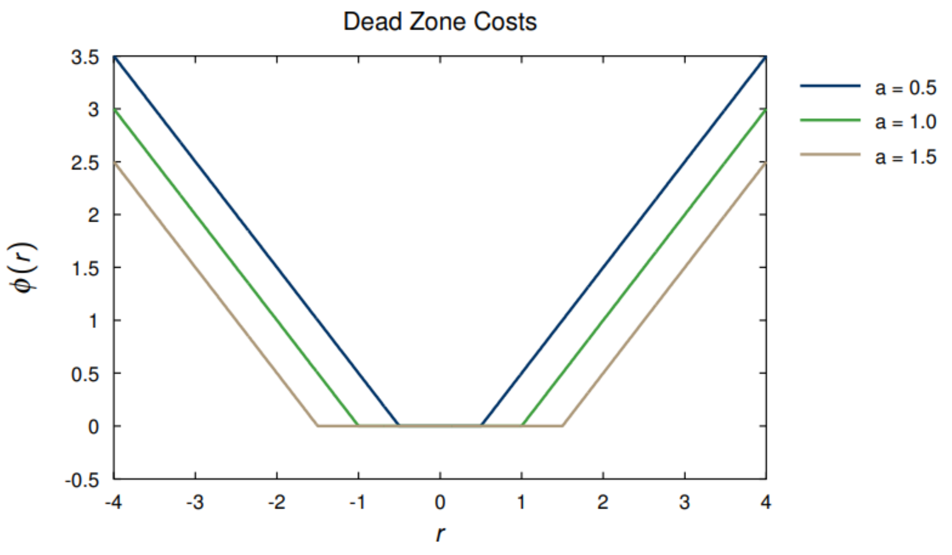


Figure 11: Geometric illustration of the Dead Zones penalty function

- **Large Error:** Quadratic for small errors, but constant for large errors. This completely caps the influence of outliers, but the function is non-convex, which makes optimization difficult.

$$\phi_{LE} = \begin{cases} r^2 & \text{if } |r| \leq a \\ a^2 & \text{otherwise} \end{cases} \tag{6.17}$$

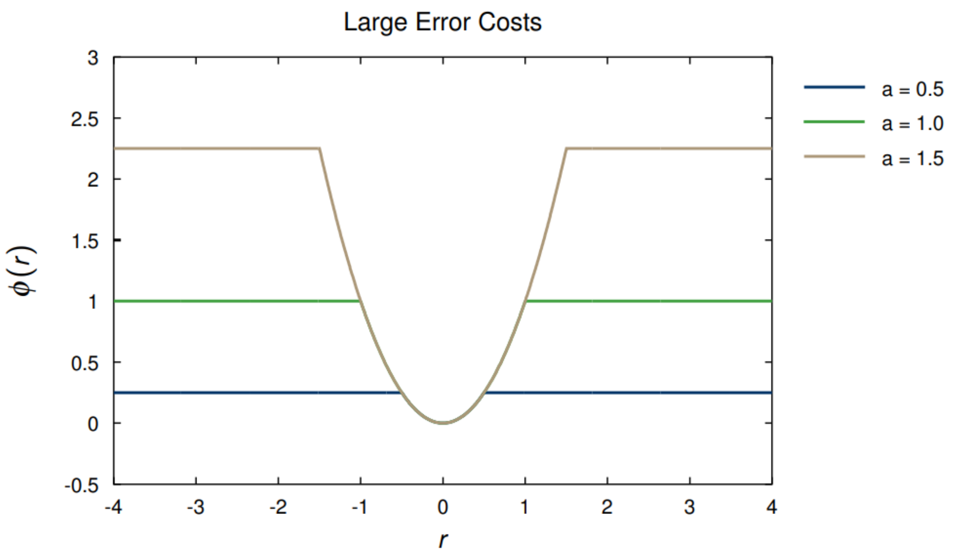


Figure 12: Geometric illustration of the Large Error penalty function

- **Huber Loss:** A hybrid approach. Quadratic for small errors (differentiable at 0) and linear for large errors (robustness).

$$\phi_{\text{Huber}}(r) = \begin{cases} r^2 & \text{if } |r| \le a \\ a(2|r| - a) & \text{otherwise} \end{cases} \tag{6.18}$$

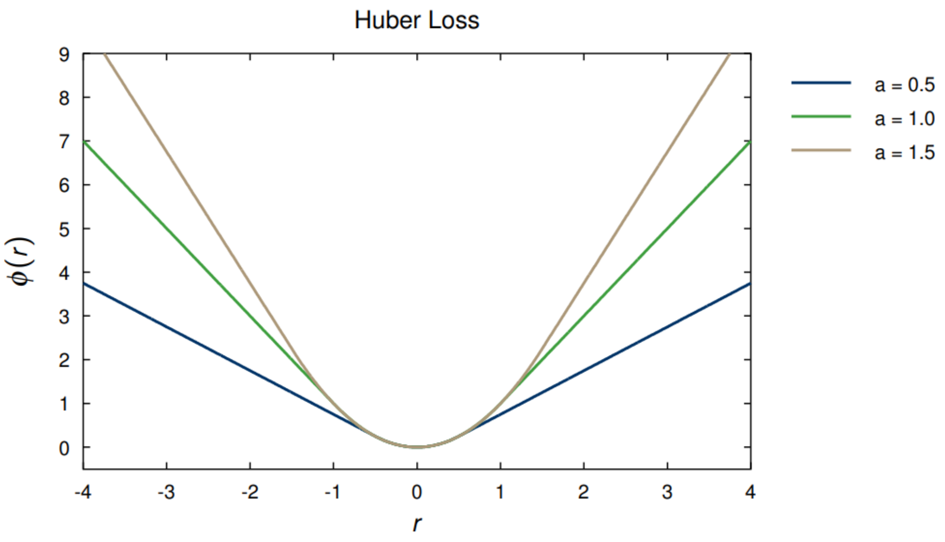


Figure 13: Geometric illustration of the Huber Loss

# 7 Neural Networks: From Perceptron to MLP

## 7.1 Rosenblatt's Perceptron

The Perceptron, introduced by Rosenblatt in 1957, is the fundamental building block of neural networks. It is a linear classifier designed to separate two classes.

### 7.1.1 The Linear Decision Boundary

We assume that the data is linearly separable and the class labels are $y \in \{+1, -1\}$. The decision rule is given by a linear function:

$$y^* = \text{sgn}(\vec{\alpha}^T \vec{x} + \alpha_0) \tag{7.1}$$

where $\vec{\alpha}$ represents the normal vector of the separating hyperplane and $\alpha_0$ is the bias.

### 7.1.2 Objective Function

To find the optimal parameters $\vec{\alpha}$ and $\alpha_0$, we do not simply count the classification errors (which would lead to a discrete and non-differentiable step function). Instead, we minimize the distance of **misclassified** feature vectors to the decision boundary.

Let $\mathcal{M}$ be the set of misclassified feature vectors. The perceptron criterion is defined as:

$$D(\alpha_0, \vec{\alpha}) = -\sum_{\vec{x}_i \in \mathcal{M}} y_i(\vec{\alpha}^T \vec{x}_i + \alpha_0) \tag{7.2}$$

If a point $\vec{x}_i$ is misclassified, the sign of $y_i$ differs from $(\vec{\alpha}^T \vec{x}_i + \alpha_0)$, making the term inside the sum negative. The minus sign ensures that the total sum $D$ is positive. We want to minimize this value.

$$\min_{\alpha_0,\vec{\alpha}} D(\alpha_0, \vec{\alpha}) = -\sum_{\vec{x}_i \in \mathcal{M}} y_i(\vec{\alpha}^T \vec{x}_i + \alpha_0) \tag{7.3}$$

**Optimization Challenges:**

- The set $\mathcal{M}$ changes in every iteration step.
- The cardinality $|\mathcal{M}|$ is a discrete variable, competing with the continuous parameters $\vec{\alpha}$.

### 7.1.3 The Perceptron Learning Algorithm

To minimize the objective function, we use **Stochastic Gradient Descent**. The gradients are:

$$\nabla_{\vec{\alpha}} D = -\sum_{\vec{x}_i \in \mathcal{M}} y_i \vec{x}_i, \quad \frac{\partial D}{\partial \alpha_0} = -\sum_{\vec{x}_i \in \mathcal{M}} y_i \tag{7.4}$$

Since we update after every single misclassification (online learning), the update rule for the $(k+1)$-th step, given a misclassified sample $(\vec{x}_i, y_i)$, is:

$$\begin{pmatrix} \alpha_0^{(k+1)} \\ \vec{\alpha}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \alpha_0^{(k)} \\ \vec{\alpha}^{(k)} \end{pmatrix} + \lambda \begin{pmatrix} y_i \\ y_i \vec{x}_i \end{pmatrix} \tag{7.5}$$

where $\lambda$ is the learning rate (usually set to 1).

**Algorithm 1** Perceptron Learning Algorithm

```
 1: Input: Training data $S = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m)\}$
 2: Initialize: $k = 0$, $\alpha_0^{(0)} = 0$, $\vec{\alpha}^{(0)} = \vec{0}$
 3: repeat
 4:     Select a pair $(\vec{x}_i, y_i)$ from the training set $S$
 5:     if $y_i(\vec{\alpha}^{(k)T} \vec{x}_i + \alpha_0^{(k)}) \leq 0$ then          ▷ Misclassified
 6:         $\vec{\alpha}^{(k+1)} \leftarrow \vec{\alpha}^{(k)} + y_i \vec{x}_i$
 7:         $\alpha_0^{(k+1)} \leftarrow \alpha_0^{(k)} + y_i$
 8:         $k \leftarrow k + 1$
 9:     end if
10: until all samples are correctly classified
11: Output: $\vec{\alpha}^{(k)}$, $\alpha_0^{(k)}$
```

### 7.1.4 Convergence (Novikoff's Theorem)

If the data is linearly separable, the algorithm is guaranteed to converge.

**Theorem: Convergence of Rosenblatt's Perceptron**

Assume there exists an optimal solution $\vec{\alpha}^*$ (with $\|\vec{\alpha}^*\| = 1$) and a margin $\rho > 0$ such that $y_i(\vec{\alpha}^{*T} \vec{x}_i + \alpha_0^*) \geq \rho$ for all samples. Let $M = \max_i \|\vec{x}_i\|_2$. The number of updates $k$ is bounded by:

$$k \leq \frac{(\alpha_0^{*2} + 1)(1 + M^2)}{\rho^2} \tag{7.6}$$

**Proof Sketch:** The proof relies on two observations:

1. The inner product between the current weight vector and the optimal solution grows linearly with each update ($(\vec{\alpha}^{(k)})^T \vec{\alpha}^* \geq k\rho$). This implies alignment.

2. The squared norm of the weight vector grows at most linearly ($\|\vec{\alpha}^{(k)}\|^2 \leq k(1 + M^2)$).
3. Combining these using the Cauchy-Schwarz inequality yields the upper bound.

**Result:** The number of iterations does *not* depend on the dimension of the feature space, but on the geometric margin $\rho$.

## 7.2 Multi-Layer Perceptrons (MLP)

Since the single Perceptron can only solve linearly separable problems, we extend the concept to Multi-Layer Perceptrons (MLPs).

### 7.2.1 Physiological Motivation & Topology

The MLP is inspired by biological neural networks:

- **Dendrites:** Input channels (feature vectors).
- **Synapses:** Weights (strength of connection).
- **Axon Hillock:** Summation and activation threshold.

An MLP consists of an **Input Layer**, one or more **Hidden Layers**, and an **Output Layer**.

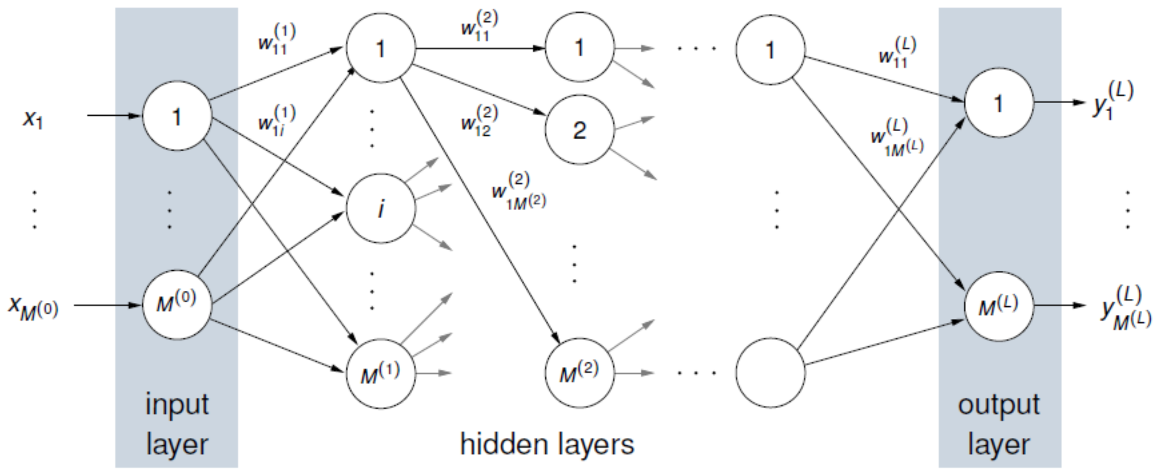


Figure 14: Structure of a Multi-Layer Perceptron (MLP)

- $w_{ij}^{(l)}$: Weight from neuron $i$ in layer $(l-1)$ to neuron $j$ in layer $(l)$.
- $net_j^{(l)}$: The weighted sum input for neuron $j$ in layer $l$.
- $y_j^{(l)}$: The output of neuron $j$ after applying the activation function.

### 7.2.2 Activation Functions

To introduce non-linearity, we apply an activation function $f(\cdot)$ to the net input:

$$y_j^{(l)} = f(net_j^{(l)}) = f\left(\sum_i w_{ij}^{(l)} y_i^{(l-1)} - w_{0j}^{(l)}\right) \tag{7.7}$$

Common functions are:

- **Sigmoid:** $f(\sigma) = \frac{1}{1+e^{-\sigma}}$
- **Tanh:** $f(\sigma) = \tanh(\sigma)$
- **ReLU:** $f(\sigma) = \max(0, \sigma)$

## 7.3 Backpropagation Algorithm

We train the MLP using **Gradient Descent** to minimize a loss function, typically the Mean Squared Error (MSE):

$$\epsilon_{MSE}(w) = \frac{1}{2} \sum_{k=1}^{M^{(L)}} (t_k - y_k^{(L)})^2 \tag{7.8}$$

where $t_k$ is the target value and $y_k^{(L)}$ is the actual output of the last layer $L$. The update rule is:

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial \epsilon}{\partial w_{ij}^{(l)}} \tag{7.9}$$

### 7.3.1 Derivation of the Gradients

We calculate the gradient using the chain rule, starting from the output layer and propagating the error backwards ("Backpropagation").

**1. Output Layer ($l = L$):** We define the **sensitivity** $\delta_k^{(L)}$ as the derivative of the error with respect to the net input. For the output layer, this depends directly on the target values $t_k$.

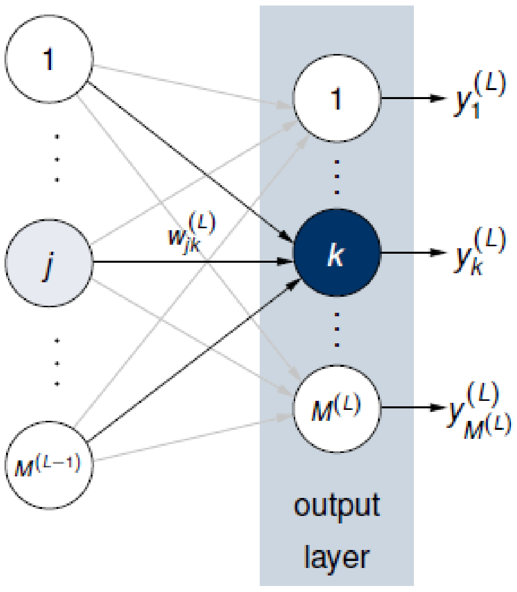


Figure 15: Backpropagation at the Output Layer: The error signal is computed directly from the output $y_k^{(L)}$ and the target $t_k$.

The gradient for the weights $w_{jk}^{(L)}$ connecting the last hidden layer to the output layer is given by:

$$\frac{\partial \epsilon_{MSE}}{\partial w_{jk}^{(L)}} = \frac{\partial \epsilon_{MSE}}{\partial net_k^{(L)}} \cdot \frac{\partial net_k^{(L)}}{\partial w_{jk}^{(L)}} = -\delta_k^{(L)} \cdot y_j^{(L-1)} \tag{7.10}$$

with the sensitivity:

$$\delta_k^{(L)} = -\frac{\partial \epsilon_{MSE}}{\partial net_k^{(L)}} = \frac{\partial \epsilon_{MSE}}{\partial y_k^{(L)}} \cdot \frac{\partial y_k^{(L)}}{\partial net_k^{(L)}} = (t_k - y_k^{(L)}) \cdot f'(net_k^{(L)}) \tag{7.11}$$

**2. Hidden Layers ($l < L$):** For a hidden layer $l$, we do not have target values. The error $\epsilon$ depends on the output $y_j^{(l)}$ of neuron $j$ only via the neurons $k$ in the **next layer** $(l+1)$ that it is connected to. We apply the chain rule to derive the sensitivity $\delta_j^{(l)}$:

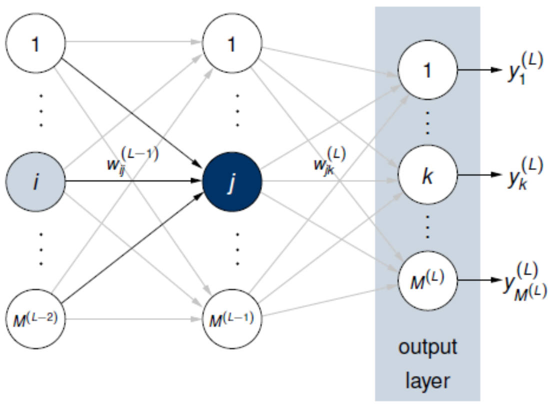


Figure 16: Backpropagation at a Hidden Layer: The error signal $\delta_j$ is derived from the weighted sum of errors $\delta_k$ from the subsequent layer.

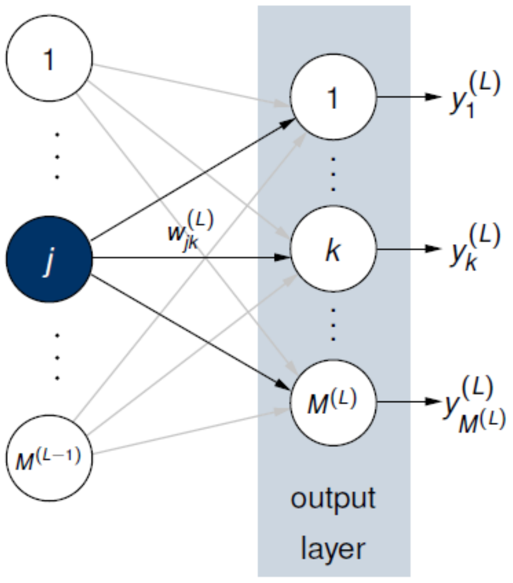


Figure 17: Backpropagation at a Hidden Layer: The sensitivities from the next layer are combined to compute the sensitivity for the current layer.

$$\delta_j^{(l)} = -\frac{\partial \epsilon_{MSE}}{\partial net_j^{(l)}} = -\frac{\partial \epsilon_{MSE}}{\partial y_j^{(l)}} \cdot \frac{\partial y_j^{(l)}}{\partial net_j^{(l)}} \tag{7.12}$$

The second term is simply the derivative of the activation function, $f'(net_j^{(l)})$. The first term, $\frac{\partial \epsilon}{\partial y_j^{(l)}}$, requires summing the contributions from all neurons $k$ in the subsequent layer $(l+1)$ (since $y_j^{(l)}$ feeds into all of them):

$$\frac{\partial \epsilon_{MSE}}{\partial y_j^{(l)}} = \sum_{k=1}^{M^{(l+1)}} \left( \frac{\partial \epsilon_{MSE}}{\partial net_k^{(l+1)}} \cdot \frac{\partial net_k^{(l+1)}}{\partial y_j^{(l)}} \right) \tag{7.13}$$

Here, we identify the terms:

- $\frac{\partial \epsilon_{MSE}}{\partial net_k^{(l+1)}} = -\delta_k^{(l+1)}$ (Sensitivity of the next layer)

- $\frac{\partial net_k^{(l+1)}}{\partial y_j^{(l)}} = w_{jk}^{(l+1)}$ (The weight connecting $j$ to $k$)

Substituting these back yields the recursive formula for the sensitivity:

$$\delta_j^{(l)} = f'(net_j^{(l)}) \cdot \sum_{k=1}^{M^{(l+1)}} \delta_k^{(l+1)} w_{jk}^{(l+1)} \tag{7.14}$$

**Interpretation:** The error $\delta_j^{(l)}$ is the weighted sum of the errors $\delta_k^{(l+1)}$ from the layer above, scaled by the local derivative $f'$. This is the essence of "propagating" the error back.

Finally, the weight update rule remains structurally consistent:

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial \epsilon_{MSE}}{\partial w_{ij}^{(l)}} = \eta \cdot \delta_j^{(l)} \cdot y_i^{(l-1)} \tag{7.15}$$

### 7.3.2 Matrix Notation

The entire forward pass of a fully connected network can be expressed compactly using matrix multiplication. For a 3-layer network:

$$\hat{y} = f_3(W_3 \cdot f_2(W_2 \cdot f_1(W_1 \vec{x}))) \tag{7.16}$$

The gradient computation (Backpropagation) then corresponds to a series of vector-matrix multiplications involving the Jacobian matrices of the layers.

# 8 Optimization

Optimization is crucial for many solutions in pattern recognition (e.g., training classifiers). We distinguish between unconstrained and constrained optimization problems.

## 8.1 Convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ and for all $\theta \in [0, 1]$, the following condition holds:

**Definition: Convexity & Concavity**

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if the domain $\text{dom}(f)$ is a convex set and if $\forall \vec{x}, \vec{y} \in \text{dom}(f)$, and $\theta$ with $0 \leq \theta \leq 1$, we have:

$$f(\theta\vec{x} + (1-\theta)\vec{y}) \leq \theta f(\vec{x}) + (1-\theta) f(\vec{y}) \tag{8.1}$$

A function is **concave** if $-f$ is convex.

**Implication:** For convex functions, any local minimum is also a global minimum. This property is particularly useful because gradient-based methods won't get stuck in suboptimal local minima.

## 8.2 Unconstrained Optimization

Here, we aim to find the minimum of a function $f(\vec{x})$ without any restrictions on $\vec{x}$. Typically, we assume $f$ is twice differentiable and convex.

$$\vec{x}^* = \text{argmin}_{\vec{x}}\, f(\vec{x}) \tag{8.2}$$

A necessary and sufficient condition for the minimum is the zero-crossing of the gradient:

$$\nabla f(\vec{x}^*) = 0 \tag{8.3}$$

Since a closed-form solution is often impossible, we use iterative approaches:

$$\begin{aligned} \text{initialization:} \quad & \vec{x}^{(0)} \\ \text{iteration step:} \quad & \vec{x}^{(k+1)} = \vec{x}^{(k)} + t^{(k)} \Delta\vec{x}^{(k)} \end{aligned}$$

where $\Delta\vec{x}^{(k)}$ is the **search direction** and $t^{(k)}$ is the **step size**.

### 8.2.1 Finding a Suitable Step Size (Line Search)

Choosing the correct step size $t^{(k)}$ is crucial:

- **Too small:** Convergence is extremely slow.
- **Too large:** The algorithm might overshoot the minimum or diverge.

Instead of finding the exact optimal $t$ (which is computationally expensive), we use **inexact line search** methods like **Backtracking Line Search** (Armijo-Goldstein).

The goal is to find a step size $t$ that, that puts us below the red line, ensuring that the function value decreases sufficiently (not just barely) relative to the step size.

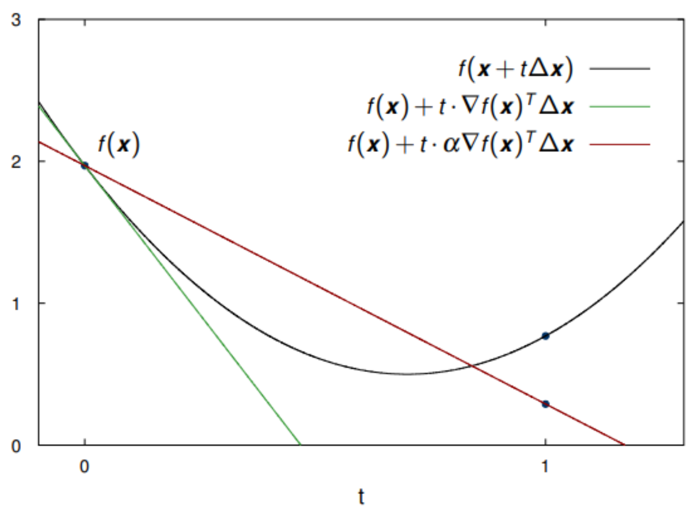


Figure 18: Backtracking line search

**Armijo-Goldstein Condition (Backtracking)**

We start with a large step size ($t = 1$) and iteratively reduce it ($t := \beta t$) until the function value decreases sufficiently. The condition is:

$$f(\vec{x} + t\Delta\vec{x}) \leq f(\vec{x}) + \alpha t \nabla f(\vec{x})^T \Delta\vec{x} \tag{8.4}$$

Where $\alpha \in (0, 0.5)$ defines the required "steepness" of the descent.

### 8.2.2 Gradient Descent

The most natural choice for the search direction is the direction of steepest descent, which is the negative gradient.

$$\Delta\vec{x}^{(k)} = -\nabla f(\vec{x}^{(k)}) \tag{8.5}$$

**Algorithm: Gradient Descent**

1. Set direction: $\Delta\vec{x}^{(k)} = -\nabla f(\vec{x}^{(k)})$
2. Line search (find optimal step size $t$):

   $$t^{(k)} = \text{argmin}_{t\geq 0} f(\vec{x}^{(k)} + t\Delta\vec{x}^{(k)})$$

   (Usually approximated via **Backtracking Line Search** / **Armijo-Goldstein**).
3. Update: $\vec{x}^{(k+1)} = \vec{x}^{(k)} + t^{(k)}\Delta\vec{x}^{(k)}$
4. Repeat until convergence ($\|\vec{x}^{(k)} - \vec{x}^{(k-1)}\| < \epsilon$).

### 8.2.3 Normalized Steepest Descent (General Norms)

Ideally, we want the direction that gives the largest decrease in the linear approximation of $f$. This depends on the chosen norm $\|\cdot\|$.

$$\Delta\vec{x} = \text{argmin}_{\vec{u}}\{\nabla f(\vec{x})^T\vec{u} \mid \|\vec{u}\| = 1\} \tag{8.6}$$

- **$L_2$-Norm:** Direction is $-\nabla f(\vec{x})$ (Standard Gradient Descent). Unit ball is a sphere.

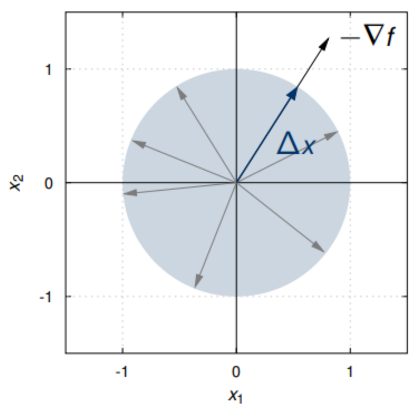


Figure 19: Unit ball in $L_2$ norm

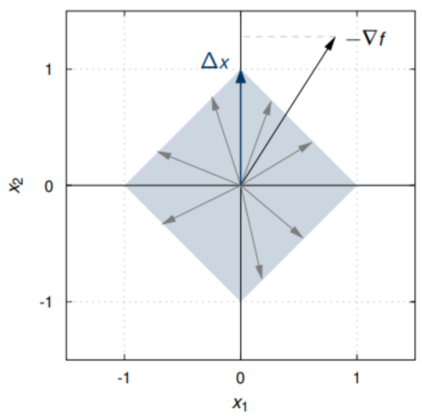


Figure 20: Unit ball in $L_1$ norm

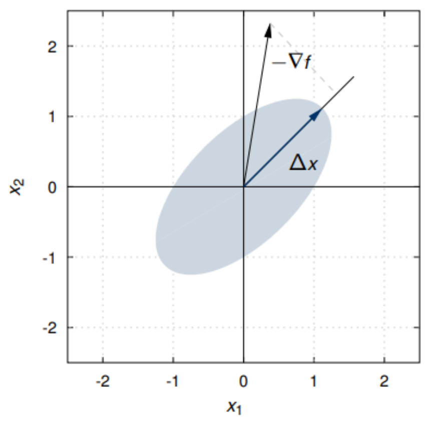


Figure 21: Unit ball in $L_P$ norm

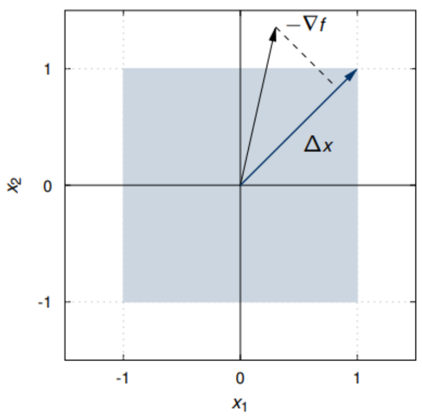


Figure 22: Unit ball in $L_\infty$ norm

- **$L_1$-Norm:** Direction is along the coordinate axis with the largest absolute gradient value (Coordinate Descent). Unit ball is a diamond.
- **$L_\infty$-Norm:** Direction points towards the corners of the unit hypercube (e.g., vector with entries $\pm 1$). Unit ball is a square/cube.
- **$L_P$-Norm (Quadratic Norm):** Defined by a positive definite matrix $\mathbf{P}$ as $\|\vec{u}\|_\mathbf{P} = \sqrt{\vec{u}^T \mathbf{P} \vec{u}}$.

  $$\Delta\vec{x} = -\mathbf{P}^{-1}\nabla f(\vec{x}) \tag{8.7}$$

  This transforms the space to make the contours spherical before taking the gradient step (similar to whitening in LDA). The update direction aligns with the largest principal component in the transformed space.

  **Crucial Connection:** If we set $\mathbf{P} = \nabla^2 f(\vec{x})$ (the Hessian), we effectively perform steepest descent in the local curvature norm, which yields **Newton's Method**.

### 8.2.4 Newton's Method

Newton's method finds the search direction by approximating the function $f(\vec{x})$ locally using a **second-order Taylor polynomial**:

$$f(\vec{x} + \Delta\vec{x}) \approx f(\vec{x}) + \nabla f(\vec{x})^T \Delta\vec{x} + \frac{1}{2}\Delta\vec{x}^T \nabla^2 f(\vec{x}) \Delta\vec{x} \tag{8.8}$$

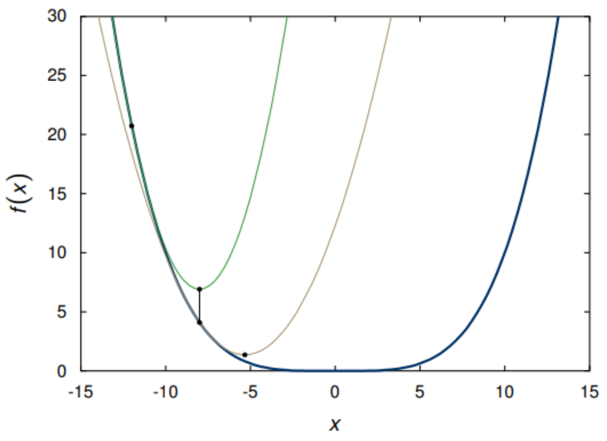


Figure 23: Newtons Method using 2nd order Taylor approximation

We find the optimal step $\Delta\vec{x}$ by setting the derivative of this approximation to zero:

$$\begin{aligned} \nabla_{\Delta\vec{x}}(\dots) &= \nabla f(\vec{x}) + \nabla^2 f(\vec{x})\Delta\vec{x} = 0 \\ \Rightarrow \quad \Delta\vec{x} &= -(\nabla^2 f(\vec{x}))^{-1}\nabla f(\vec{x}) \end{aligned}$$

**Key Difference:** While Gradient Descent assumes the local geometry is a sphere (linear approximation), Newton's Method accounts for the local **curvature** (ellipsoidal geometry) given by the Hessian $\nabla^2 f(\vec{x})$.

## 8.3 Constrained Optimization

We consider the primal optimization problem with constraints:

$$\begin{aligned} \text{minimize} \quad & f_0(\vec{x}) \\ \text{subject to} \quad & f_i(\vec{x}) \le 0, \quad i = 1,\dots,m \\ & h_i(\vec{x}) = 0, \quad i = 1,\dots,p \end{aligned}$$

*Note: $f_0$ is not required to be convex yet.*

### 8.3.1 The Lagrangian & Dual Function

To handle constraints, we define the **Lagrangian** $L$:

$$L(\vec{x},\vec{\lambda},\vec{\nu}) = f_0(\vec{x}) + \sum_{i=1}^{m} \lambda_i f_i(\vec{x}) + \sum_{i=1}^{p} \nu_i h_i(\vec{x}) \tag{8.9}$$

We formulate the **Lagrange dual function** $g(\vec{\lambda},\vec{\nu})$ by minimizing $L$ over $\vec{x}$:

$$g(\vec{\lambda},\vec{\nu}) = \inf_{\vec{x}} L(\vec{x},\vec{\lambda},\vec{\nu}) \tag{8.10}$$

**Properties of the Dual Function**

- $g(\vec{\lambda},\vec{\nu})$ is **always concave**, even if the primal problem is not convex.
- It provides a **lower bound** on the optimal primal value $p^*$: $g(\vec{\lambda},\vec{\nu}) \le p^*$.

This leads to the **Dual Problem**: Find the best lower bound by maximizing $g$.

$$\text{maximize } g(\vec{\lambda},\vec{\nu}) \quad \text{subject to } \vec{\lambda} \succeq 0$$

### 8.3.2 Strong Duality & Slater's Condition

- **Weak Duality:** $d^* \leq p^*$ (Always true). The difference $p^* - d^*$ is the *duality gap*.
- **Strong Duality:** $d^* = p^*$ (Gap is zero). This allows solving the dual problem to find the primal solution.

Strong duality is guaranteed if **Slater's Condition** holds:

**Theorem: Slater's Condition**

For a **convex** optimization problem, strong duality holds if there exists a strictly feasible point $\vec{x}$ such that:

$$f_i(\vec{x}) < 0 \quad \forall i = 1, \dots, m \quad \text{and} \quad A\vec{x} = \vec{b} \tag{8.11}$$

**Refinement:** If constraints are affine (linear), touching the boundary ($f_i(\vec{x}) \leq 0$) is allowed. Only non-linear constraints require strict inequality.

### 8.3.3 Karush-Kuhn-Tucker (KKT) Conditions

If strong duality holds, any optimal pair $(\vec{x}^*, \vec{\lambda}^*, \vec{\nu}^*)$ **must** satisfy the KKT conditions:

1. **Primal Feasibility:**
$$f_i(\vec{x}^*) \leq 0, \quad h_i(\vec{x}^*) = 0$$
2. **Dual Feasibility:**
$$\vec{\lambda}^* \succeq 0$$
3. **Complementary Slackness:** (Crucial for SVMs!)
$$\lambda_i^* \cdot f_i(\vec{x}^*) = 0$$
*Meaning: Either a constraint is active ($f_i(\vec{x}) = 0$) or its multiplier is zero ($\lambda_i = 0$).*
4. **Stationarity:** Gradient of Lagrangian is zero.
$$\nabla f_0(\vec{x}^*) + \sum \lambda_i^* \nabla f_i(\vec{x}^*) + \sum \nu_i^* \nabla h_i(\vec{x}^*) = 0$$

**Conclusion:** For convex problems with strong duality, KKT conditions are necessary and sufficient for optimality. Finding a point that satisfies them means we found the global optimum.

# 9 Support Vector Machines (SVM)

Just like other classifiers (Neural Nets, Nearest Neighbor, etc.), the goal of SVMs is to draw a linear line (decision boundary) to separate classes. But instead of drawing any sufficient line to separate the classes, SVMs aim to find a unique decision boundary that **maximizes the margin (distance)** between each class. The solution to this problem is unique and depends only on the features that are close to the decision boundary.

## 9.1 Hard Margin Problems

The hard margin SVM needs linearly seperable classes. Lets assume there is an affine function

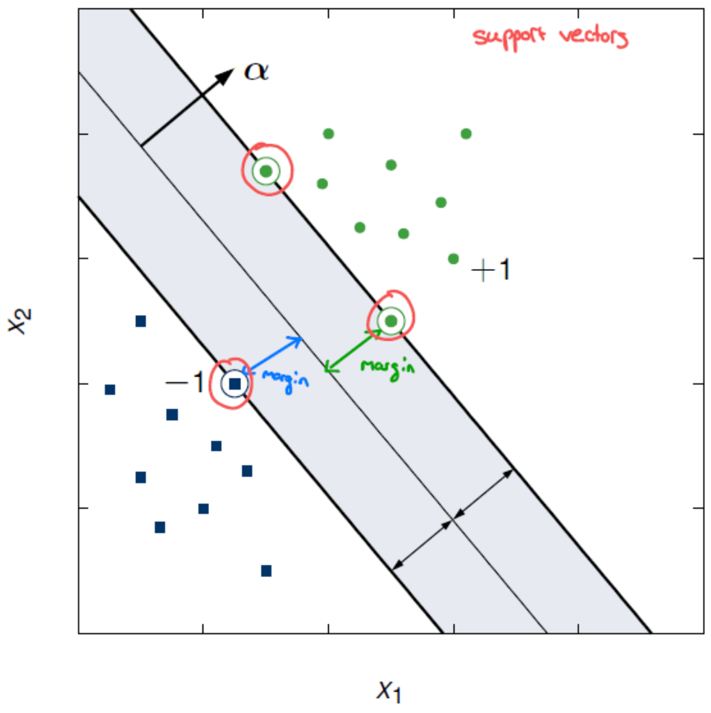


Figure 24: Hard margin SVM

defined as:

$$f(x) = \vec{\alpha}^T x + \alpha_0 \tag{9.1}$$

Where $\vec{\alpha}$ is the normal vector to the decision boundary and $\alpha_0$ is some sort of bias. For any point $x$ on the decision boundary, it holds that $f(x) = 0$.
There are three major points we need to think about to end up with a nice optimization problem:

**1. Introduce margin constraints:**
We need to ensure that all points are classified correctly and therefore lie outside the margin. Therefore, we introduce the following constraints:

- For points of class $+1$: $f(x) \geq 1$
- For points of class $-1$: $f(x) \leq -1$

This means that for a sample point $x_i$ with the label $y_i = +1$, it holds that $f(x_i) \geq 1$. Similarly, for a sample point $x_j$ with the label $y_j = -1$, it holds that $f(x_j) \leq -1$. These two constraints can be combined into one single constraint ($y \in \{+1, -1\}$):

$$y_i f(x_i) - 1 = y_i(\vec{\alpha}^T x_i + \alpha_0) - 1 \geq 0 \quad \forall i \tag{9.2}$$

**2. Define the margin:**
The margin is defined as the distance between the decision boundary and the closest points from either class. To compute the margin width, we take a sample from each class that lies exactly on the margin (i.e., satisfies $y_i f(x_i) - 1 = 0 \quad \forall i$) and subtract them from each other. When we project the resulting vector onto tho normalized normal vector of the hyperplane, we get the margin width:

$$\text{width} = \frac{\vec{\alpha}}{\|\vec{\alpha}\|_2} \cdot (\vec{x}_{y=+1} - \vec{x}_{y=-1}) \tag{9.3}$$

Now we multiply this out:

$$\text{width} = \frac{1}{\|\vec{\alpha}\|_2} (\vec{\alpha}^T \vec{x}_{y=+1} - \vec{\alpha}^T \vec{x}_{y=-1}) \tag{9.4}$$

We know from our margin constraints defined in step 1 that for support vectors (points on the margin), the inequality becomes an equality:

- For the positive support vector $\vec{x}_{y=+1}$:

$$\vec{\alpha}^T \vec{x}_{y=+1} + \alpha_0 = 1 \quad \Rightarrow \quad \vec{\alpha}^T \vec{x}_{y=+1} = 1 - \alpha_0$$

- For the negative support vector $\vec{x}_{y=-1}$:

$$\vec{\alpha}^T \vec{x}_{y=-1} + \alpha_0 = -1 \quad \Rightarrow \quad \vec{\alpha}^T \vec{x}_{y=-1} = -1 - \alpha_0$$

Substituting these expressions back into the width equation:

$$\begin{aligned} \text{width} &= \frac{1}{\|\vec{\alpha}\|_2} ((1 - \alpha_0) - (-1 - \alpha_0)) && (9.5) \\ &= \frac{1}{\|\vec{\alpha}\|_2} (1 - \alpha_0 + 1 + \alpha_0) && (9.6) \\ &= \frac{2}{\|\vec{\alpha}\|_2} && (9.7) \end{aligned}$$

**3. Minimize the norm:**
Since we want to **maximize** the margin width $\frac{2}{\|\vec{\alpha}\|_2}$, this is mathematically equivalent to **minimizing** the length of the normal vector $\|\vec{\alpha}\|_2$. For mathematical convenience (to make derivatives easier later), we minimize the squared norm:

**Primal Optimization Problem (Hard Margin)**

$$\text{minimize} \quad \frac{1}{2}\|\vec{\alpha}\|_2^2 \quad \text{subject to} \quad y_i(\vec{\alpha}^T x_i + \alpha_0) \geq 1 \quad \forall i$$

## 9.2 Soft Margin Problems

In real world applications, data is often not perfectly linearly separable. The Soft Margin SVM relaxes the hard margin constraints by allowing some points to violate the margin or even be misclassified.

Therefore, we introduce slack variables $\xi_i \geq 0$ for each training sample $x_i$, which measure the degree of misclassification (**Hinge Loss**):

- $\xi_i = 0$: point is correctly classified and outside or on the margin
- $0 < \xi_i \leq 1$: point is inside the margin but still correctly classified

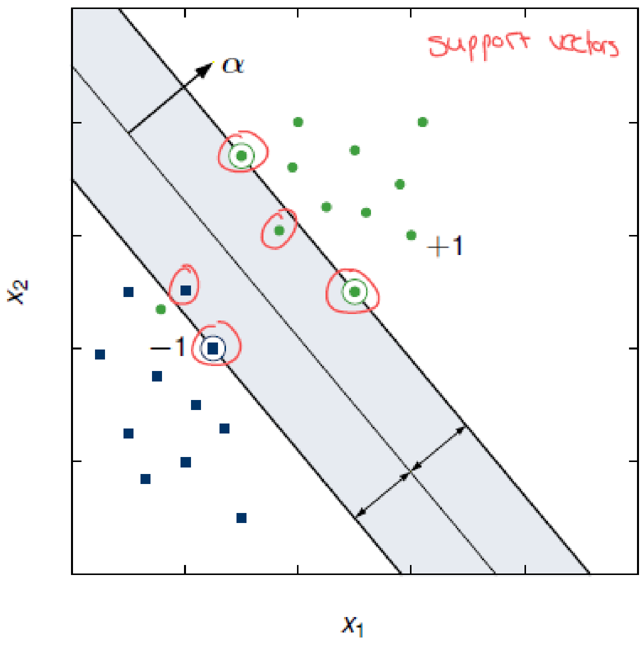


Figure 25: Soft margin SVM

- $\xi_i > 1$: point is misclassified

The margin constraints are now relaxed to:

$$y_i(\vec{\alpha}^T x_i + \alpha_0) \geq 1 - \xi_i \quad \forall i \tag{9.8}$$

The primal optimization problem becomes:

**Primal Optimization Problem (Soft Margin)**

$$\text{minimize} \quad \frac{1}{2}\|\vec{\alpha}\|_2^2 + \mu \sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad -(y_i(\vec{\alpha}^T x_i + \alpha_0) - 1 + \xi_i) \leq 0, \quad -\xi_i \leq 0 \quad \forall i$$

where $\mu > 0$ is a hyperparameter that controls the trade-off between maximizing the margin and minimizing the classification error.

## 9.3 Dual Representation

The primal problem minimizes with respect to $\vec{\alpha}$ and $\alpha_0$. However, solving the **Lagrange Dual Problem** is often more powerful. It reveals that the solution depends *only* on the inner products of the data points, which is the key to the Kernel Trick.

### 9.3.1 Derivation from the Lagrangian

For simplicity, we derive the dual only for the hard margin case here. In the soft margin case, the derivation is similar but includes additional terms for the slack variables. We start with the

Lagrangian of the hard margin problem (using multipliers $\lambda_i \geq 0$):

$$L(\vec{\alpha}, \alpha_0, \vec{\lambda}) = \underbrace{\frac{1}{2}\|\vec{\alpha}\|_2^2}_{\text{primal problem } f_0(x)} \underbrace{- \sum_{i=1}^{m} \lambda_i [y_i(\vec{\alpha}^T \vec{x}_i + \alpha_0) - 1]}_{\text{constraints } f_i(x)} \tag{9.9}$$

To find the dual, we minimize $L$ with respect to the primal variables $\vec{\alpha}$ and $\alpha_0$ (setting gradients to zero):

$$\nabla_{\vec{\alpha}} L = \vec{\alpha} - \sum_{i=1}^{m} \lambda_i y_i \vec{x}_i = 0 \quad \Rightarrow \quad \tilde{\alpha} = \sum_{\mathbf{i=1}}^{\mathbf{m}} \lambda_{\mathbf{i}} \mathbf{y_i} \tilde{\mathbf{x}}_{\mathbf{i}} \tag{9.10}$$

$$\frac{\partial L}{\partial \alpha_0} = - \sum_{i=1}^{m} \lambda_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \lambda_i y_i = 0 \tag{9.11}$$

Substituting (9.10) back into the Lagrangian eliminates $\vec{\alpha}$ and yields the **Dual Objective Function**, which depends only on $\vec{\lambda}$:

**Dual Optimization Problem**

$$\text{maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (\vec{x}_i^T \vec{x}_j)$$

**Subject to:**

$$\lambda_i \geq 0 \quad \forall i, \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i y_i = 0$$

### 9.3.2 From Dual Variables to the Decision Boundary

Once we have solved the Dual Problem and found the optimal Lagrange multipliers $\lambda_i^*$, we need to reconstruct the decision boundary. But which $\lambda_i$ are actually relevant?

This is answered by the **Complementary Slackness** condition (KKT condition 3):

$$\lambda_i \cdot \underbrace{(y_i \cdot (\vec{\alpha}^T \vec{x}_i + \alpha_0) - 1)}_{f_i(\vec{\alpha}, \alpha_0)} = 0 \tag{9.12}$$

Since the product must be zero, we have two cases for every data point $i$:

- **Case 1:** $\lambda_i = 0$. The point is correctly classified and lies somewhere "safe" (not on the margin). These points effectively vanish from our solution.
- **Case 2:** $\lambda_i > 0$. Then the bracket term *must* be zero:

$$y_i(\vec{\alpha}^T \vec{x}_i + \alpha_0) = 1 \tag{9.13}$$

  These points lie exactly on the margin boundaries. We call them **Support Vectors**.

**Conclusion 1 (Sparsity):** The weight vector $\vec{\alpha}$ (see 9.3.1) depends *only* on the support vectors (where $\lambda_i > 0$):

$$\vec{\alpha} = \sum_{i=1}^{m} \lambda_i y_i \vec{x}_i = \sum_{i \in SV} \lambda_i y_i \vec{x}_i \tag{9.14}$$

**Conclusion 2 (The Decision Rule):** We can now write the classification function for a new point $\vec{u}$ purely in terms of training samples and their inner products:

$$f(\vec{u}) = \vec{\alpha}^T\vec{u} + \alpha_0 = \underbrace{\left(\sum_{i=1}^{m} \lambda_i y_i \vec{x}_i\right)^T}_{\vec{\alpha}^T} \vec{u} + \alpha_0 = \sum_{i=1}^{m} \lambda_i y_i (\vec{x}_i^T \vec{u}) + \alpha_0 \tag{9.15}$$

This form is crucial because it allows us to apply the **Kernel Trick** in the next section.

## 9.4 Feature Transform & The Kernel Trick

Both hard and soft margin SVMs can only generate a **linear decision boundary**. This has serious limitations:

- Non-linearly separable data cannot be classified.
- Noisy data can cause problems.
- The formulation deals with vectorial data only.

### 9.4.1 Mapping to a Higher Dimensional Space

To solve this, we map the data into a richer, higher-dimensional feature space using a non-linear transformation $\Phi$:

$$\Phi : \mathbb{R}^d \to \mathbb{R}^D \quad (D \gg d) \tag{9.16}$$

such that the features $\Phi(\vec{x}_i)$ become linearly separable in that new space.

**Example (Quadratic Decision Boundary):** Assume a 2D decision boundary defined by a quadratic function:

$$f(\vec{x}) = \alpha_0 + \alpha_1 x_1^2 + \alpha_2 x_2^2 + \alpha_3 x_1 x_2 + \alpha_4 x_1 + \alpha_5 x_2$$

This is non-linear in the original space $\vec{x} = (x_1, x_2)^T$. However, we can define a transformation $\Phi(\vec{x})$:

$$\Phi(\vec{x}) = (1, x_1^2, x_2^2, x_1 x_2, x_1, x_2)^T \tag{9.17}$$

Now, the decision function becomes linear in the transformed space: $f(\vec{x}) = \vec{\alpha}^T \Phi(\vec{x})$.

### 9.4.2 Applying the Transform to the Dual Problem

Because we reformulated the SVM into its **Dual Representation**, we don't need to explicitly calculate the weight vector $\vec{\alpha}$ in the high-dimensional space. The decision boundary can be rewritten using inner products of the transformed features:

$$f(\vec{x}) = \sum_{i=1}^{m} \lambda_i y_i \langle \Phi(\vec{x}_i), \Phi(\vec{x}) \rangle + \alpha_0 \tag{9.18}$$

The optimization problem changes accordingly:

**Dual Optimization with Feature Transform**

$$\text{maximize} \quad -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \langle \Phi(\vec{x}_i), \Phi(\vec{x}_j) \rangle + \sum_i \lambda_i$$

subject to $\quad \vec{\lambda} \succeq 0, \quad \sum \lambda_i y_i = 0$

Now we have a linear decision boundary in a higher dimensional space!

### 9.4.3 Kernel Functions

Computing the explicit transformation $\Phi(\vec{x})$ can be computationally very expensive (or even impossible for infinite dimensions). However, since the algorithm **only depends on the inner product**, we can abstract this using a **Kernel Function**:

$$k(\vec{x}, \vec{x}') = \langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle \tag{9.19}$$

Using kernel functions, we avoid modeling the transformation $\Phi(\vec{x})$ explicitly but still get the same result.

**Typical Kernel Functions:**

- **Linear:** $k(\vec{x}, \vec{x}') = \langle \vec{x}, \vec{x}' \rangle$
- **Polynomial:** $k(\vec{x}, \vec{x}') = (\langle \vec{x}, \vec{x}' \rangle + 1)^d$ (The example above corresponds to $d = 2$)
- 
- **Radial Basis Function (RBF) / Gaussian:**

$$k(\vec{x}, \vec{x}') = \exp\left(-\frac{\|\vec{x} - \vec{x}'\|^2}{2\sigma^2}\right) \tag{9.20}$$

- **Sigmoid:** $k(\vec{x}, \vec{x}') = \tanh(a\vec{x}^T\vec{x}' + b)$

# 10 Kernels & Kernel Methods

## 10.1 Feature Transforms

Linear decision boundaries have limitations (cannot classify non-linearly separable data).
**Solution:** Map data into a higher-dimensional feature space using a non-linear transformation $\phi : \mathbb{R}^d \to \mathbb{R}^D$ $(D \geq d)$.

**Example:** Mapping 2D data to 2D (squaring):

$$\phi(\vec{x}) = (x_1^2, x_2^2)^T \tag{10.1}$$

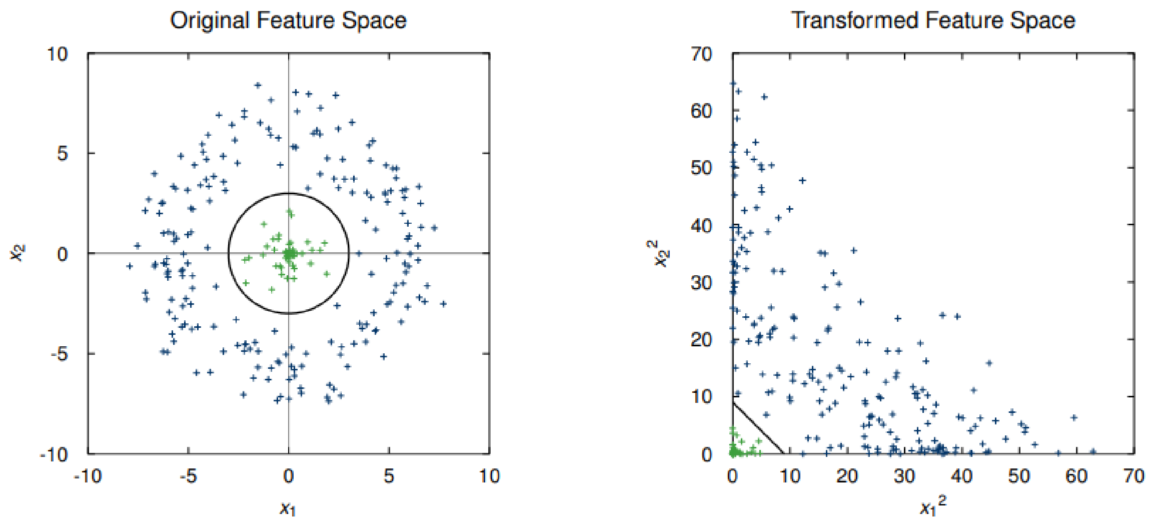


Figure 26: Feature transform: squaring each coordinate

Data that forms a circle in the original space becomes linearly separable in the transformed space.

**Property:** Distances in the transformed space can be computed using only inner products:

$$\|\phi(\vec{x}) - \phi(\vec{x}')\|_2^2 = \langle(\phi(\vec{x}) - \phi(\vec{x}')), (\phi(\vec{x}) - \phi(\vec{x}'))\rangle \tag{10.2}$$

$$= \langle\phi(\vec{x}), \phi(\vec{x})\rangle - 2\langle\phi(\vec{x}), \phi(\vec{x}')\rangle + \langle\phi(\vec{x}'), \phi(\vec{x}')\rangle \tag{10.3}$$

This observation leads us to Kernel Functions.

## 10.2 Kernel Functions & The Kernel Trick

**Definition: Kernel Function**

A kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function that maps a pair of features to a real number. It corresponds to an inner product in some feature space:

$$k(\vec{x}, \vec{x}') = \langle\phi(\vec{x}), \phi(\vec{x}')\rangle \tag{10.4}$$

**The Kernel Trick:** In any algorithm formulated in terms of inner products (like SVM or PCA), we can replace the inner product with a kernel $k(\vec{x}, \vec{x}')$. This avoids the explicit (and expensive) computation of $\phi(\vec{x})$.

**Mercer's Theorem:** For any positve semidefinite kernel function $k$, there exists a feature mapping $\phi$ such that $k(\vec{x}, \vec{x}') = \langle \phi(\vec{x}), \phi(\vec{x}') \rangle$ (kernel function is fully defined by inner products).

**Typical Kernel Functions**

- **Linear:** $k(\vec{x}, \vec{x}') = \langle \vec{x}, \vec{x}' \rangle$
- **Polynomial:** $k(\vec{x}, \vec{x}') = (\langle \vec{x}, \vec{x}' \rangle + 1)^d$
- **RBF (Radial Basis Function):** $k(\vec{x}, \vec{x}') = \exp\left(-\frac{\|\vec{x}-\vec{x}'\|^2}{\sigma^2}\right)$
- **Sigmoid:** $k(\vec{x}, \vec{x}') = \tanh(\alpha \langle \vec{x}, \vec{x}' \rangle + \beta)$

**Kernel Matrix**

**Definition: Kernel Matrix**

For a given set of feature vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$, we define the **Kernel Matrix** $K \in \mathbb{R}^{m \times m}$ as:

$$K_{i,j} = k(\vec{x}_i, \vec{x}_j) = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle \tag{10.5}$$

**Important Properties:**

- The entries of the matrix can be interpreted as **similarity measures** between the feature pairs (because the inner product (german: Skalarprodukt) measures similarity).
- The kernel matrix is always **positive semidefinite** ($K \succeq 0$). This is a necessary condition for a function to be a valid kernel (Mercer's Theorem).

## 10.3 Kernel SVM

In chapter 9, we saw, that standard SVMs can only learn linear decision boundaries. By applying the kernel trick, we can extend SVMs to learn non-linear decision boundaries. Since the decision rule 9.15 and the dual optimization 9.12 problem depend only on inner products, we can replace them with kernel functions. The decision rule becomes:

$$f(\vec{u}) = \sum_{i=1}^{m} \lambda_i y_i k(\vec{x}_i, \vec{u}) + \alpha_0 \tag{10.6}$$

The dual optimization problem becomes:

$$\max_{\vec{\lambda}} \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j k(\vec{x}_i, \vec{x}_j) \tag{10.7}$$

The choice of the kernel function $k$ determines the type of non-linear decision boundary that can be learned. **Result:** By choosing different kernel functions, we can learn a variety of non-linear decision boundaries without explicitly computing the feature transformation $\phi$.

## 10.4 Kernel PCA

Standard PCA finds principal components by diagonalizing the covariance matrix $\Sigma$. In Kernel PCA, we perform PCA in the feature space defined by $\phi$. This allows us to capture non-linear structures in the data. Also the computatuional

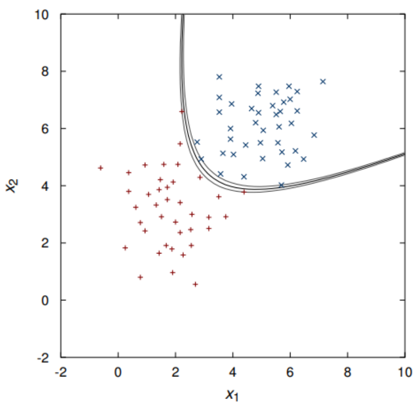


Figure 27: Polynomial Kernel Decision Boundary

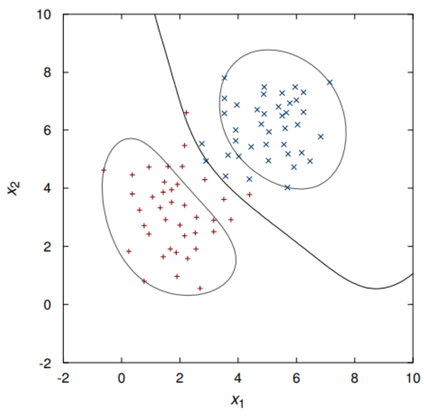


Figure 28: RBF (Gaussian) Kernel Decision Boundary

### 10.4.1 PCA Revisited

The covariance matrix in feature space is (assuming zero mean):

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \vec{x}_i \vec{x}_i^T \tag{10.8}$$

We solve the eigenvalue problem: $\Sigma \vec{e}_i = \lambda_i \vec{e}_i$.

### 10.4.2 Rewriting the Eigenvalue Problem

The eigenvectors $\vec{e}_i$ lie in the span of the feature vectors and can therefore be expressed as a linear combination of feature vectors:

$$\vec{e}_i = \sum_{k=1}^{m} \alpha_{i,k} \vec{x}_k \tag{10.9}$$

Substituting this into the eigenvalue equation:

$$\left( \frac{1}{m} \sum_{j=1}^{m} \vec{x}_j \vec{x}_j^T \right) \cdot \sum_k \alpha_{i,k} \vec{x}_k = \lambda_i \sum_k \alpha_{i,k} \vec{x}_k \tag{10.10}$$

To introduce the kernel, we multiply from the left by $\vec{x}_l^T$ for all $l$:

$$\sum_{j,k} \alpha_{i,k} \underbrace{\vec{x}_l^T \vec{x}_j}_{k(\vec{x}_l, \vec{x}_j)} \underbrace{\vec{x}_j^T \vec{x}_k}_{k(\vec{x}_j, \vec{x}_k)} = m \lambda_i \sum_k \alpha_{i,k} \underbrace{\vec{x}_l^T \vec{x}_k}_{k(\vec{x}_l, \vec{x}_k)} \tag{10.11}$$

Since all feature vectors only appear in the form of inner products, we can replace them with the kernel function $k$:

**Kernel PCA Equation**

$$\sum_{j,k} \alpha_{i,k} k(\vec{x}_l, \vec{x}_j) k(\vec{x}_j, \vec{x}_k) = m\lambda_i \sum_k \alpha_{i,k} k(\vec{x}_l, \vec{x}_k) \tag{10.12}$$

In matrix notation using the Kernel Matrix $K$:

$$K^2 \vec{\alpha}_i = m\lambda_i K \vec{\alpha}_i \tag{10.13}$$

This implies:

$$K \vec{\alpha}_i = m\lambda_i \vec{\alpha}_i \tag{10.14}$$

So now we only need to solve the eigenvalue problem for the kernel matrix $K$!

### 10.4.3 Projection and Centering

- **Computation:** We solve the eigenvalue problem for the $(m \times m)$ kernel matrix $K$.
- **Projection:** The projection $c$ of a transformed feature vector $\phi(\vec{x})$ onto the principal component $\vec{e}_i$ is easily computed by:

$$c = \phi(\vec{x})^T \vec{e}_i = \sum_k \alpha_{i,k} k(\vec{x}, \vec{x}_k) \tag{10.15}$$

- **Centering:** It is assumed that transformed features have zero mean. This is enforced by centering the kernel matrix:

$$\tilde{K}_{i,j} = \tilde{\phi}(\vec{x}_i)^T \tilde{\phi}(\vec{x}_j) = K_{i,j} - \frac{1}{m} \sum_k K_{i,k} - \frac{1}{m} \sum_k K_{k,j} - \frac{1}{m} \sum_k K_{k,l}$$

## 10.5 Kernels for Sequences

Kernels allow us to apply SVMs/PCA to non-vectorial data like text or speech.

- **String Kernels / DTW:** Based on Dynamic Time Warping distance $D$.

$$k(\vec{x}, \vec{y}) = \exp(-D(\vec{x}, \vec{y})) \tag{10.16}$$

- **Fisher Kernels:** Combine generative models (e.g., HMMs) with discriminative classifiers.

$$k(\vec{x}, \vec{x}') = \mathbf{J}_\theta(\vec{x})^T \mathbf{I}^{-1} \mathbf{J}_\theta(\vec{x}') \tag{10.17}$$

where $\mathbf{J}_\theta$ is the Fisher Score (gradient of log-likelihood) and $\mathbf{I}$ is the Fisher Information Matrix.

## 10.6 Laplacian SVM

# 11 Expectation Maximization

Usually we use Methods like Maximum Likelihood Estimation (MLE) 1.3.4 or Maximum A-Posteriori (MAP) 1.3.5 to estimate parameters of probabilistic models. However, these methods tend to fail when we deal with **high dimensional data** or **latent/incomplete variables**.

## 11.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions with unknown parameters. In mathenatical words: we represent our training data with a set of $K$ Gaussian distributions.

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \quad \text{with} \quad \sum_{k=1}^{K} p_k = 1 \tag{11.1}$$

where $p_k$ are the mixture weights, $\boldsymbol{\mu}_k$ are the means, and $\Sigma_k$ are the covariance matrices of the Gaussian components.
The individual Gaussian (probability that the data point $\mathbf{x}_i$ belongs to component $k$) are given by:

$$p_{\text{ik}} = \frac{p_{\text{k}} \mathcal{N}(\mathbf{x}_{\text{i}}|\boldsymbol{\mu}_k, \Sigma_k)}{p(\mathbf{x}_{\text{i}})} \tag{11.2}$$

The ML-Estimates for the parameters are given by:

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^{N} p_{\text{ik}} \tag{11.3}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{N} p_{\text{ik}} \mathbf{x}_{\text{i}}}{\sum_{i=1}^{N} p_{\text{ik}}} \tag{11.4}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{N} p_{\text{ik}} (\mathbf{x}_{\text{i}} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{\text{i}} - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{N} p_{\text{ik}}} \tag{11.5}$$

With all these formulas in hand, we can write out the EM-Algorithm for GMMs:

**Algorithm 2** EM Algorithm for GMM parameter estimation

```
1: **Initialization:** $\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, p_k^{(0)}$
2: $j \leftarrow 0$
3: **repeat**
4:     **Expectation step:** compute new values for $p_{ik}, L$
5:     **Maximization step:** update values for $\boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)}, p_k^{(j)}$
6:     $j \leftarrow j + 1$
7: **until** $L$ is no longer changing
8: **Output:** estimates $\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k, \hat{p}_k$
```

# 12 Expectation Maximization

Usually we use Methods like Maximum Likelihood Estimation (MLE) 1.3.4 or Maximum A-Posteriori (MAP) 1.3.5 to estimate parameters of probabilistic models. However, these methods tend to fail when we deal with **high dimensional data** or **latent/incomplete variables**.

## 12.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions with unknown parameters. In mathematical words: we represent our training data with a set of $K$ Gaussian distributions.

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \quad \text{with} \quad \sum_{k=1}^{K} p_k = 1 \tag{12.1}$$

where $p_k$ are the mixture weights, $\boldsymbol{\mu}_k$ are the means, and $\Sigma_k$ are the covariance matrices of the Gaussian components.
The individual Gaussian (probability that the data point $\mathbf{x}_i$ belongs to component $k$) are given by:

$$p_{\mathrm{ik}} = \frac{p_{\mathrm{k}} \mathcal{N}(\mathbf{x}_{\mathrm{i}}|\boldsymbol{\mu}_k, \Sigma_k)}{p(\mathbf{x}_{\mathrm{i}})} \tag{12.2}$$

The ML-Estimates for the parameters are given by:

$$\hat{p}_k = \frac{1}{N}\sum_{i=1}^{N} p_{\mathrm{ik}} \tag{12.3}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{N} p_{\mathrm{ik}}\mathbf{x}_{\mathrm{i}}}{\sum_{i=1}^{N} p_{\mathrm{ik}}} \tag{12.4}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{N} p_{\mathrm{ik}}(\mathbf{x}_{\mathrm{i}} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_{\mathrm{i}} - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{N} p_{\mathrm{ik}}} \tag{12.5}$$

With all these formulas in hand, we can write out the EM-Algorithm for GMMs:

**Algorithm 3** EM Algorithm for GMM parameter estimation

```
1: Initialization: $\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, p_k^{(0)}$
2: $j \leftarrow 0$
3: repeat
4:     Expectation step: compute new values for $p_{ik}, L$
5:     Maximization step: update values for $\boldsymbol{\mu}_k^{(j)}, \boldsymbol{\Sigma}_k^{(j)}, p_k^{(j)}$
6:     $j \leftarrow j + 1$
7: until $L$ is no longer changing
8: Output: estimates $\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k, \hat{p}_k$
```

## 12.2 Missing Information Principle

EM helps us to find maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.
The joint probability density of the events (observed data $x$ and latent variables $y$) given by the parameters $\theta$ is:

$$p(x, y; \theta) = p(x; \theta)p(y|x; \theta) \leftrightarrow p(x; \theta) = \frac{p(x, y; \theta)}{p(y|x; \theta)} \tag{12.6}$$

We can take the logarithm and get the mathematical expression for the **Missing Information Principle**:

$$\log p(x; \theta) = \log p(x, y; \theta) - \log p(y|x; \theta) \tag{12.7}$$

**Derivation of an update scheme:**
We now take this formulation and derive an iterative update scheme for the parameters $\theta$.

Consider 12.7 at iteration $i+1$. We multiply both sides with $p\left(y|x;\hat{\theta}^{(i)}\right)$ and integrate over the latent variable $y$:

$$\begin{aligned}\int p(y|x;\hat{\theta}^{(i)})\log p(x;\hat{\theta}^{(i+1)})\,dy \quad &= \quad \int p(y|x;\hat{\theta}^{(i)})\log p(x,y;\hat{\theta}^{(i+1)})\,dy - \\ &\quad\;\; \int p(y|x;\hat{\theta}^{(i)})\log p(y|x;\hat{\theta}^{(i+1)})\,dy\end{aligned}$$

**Left Side Simplification:** Since $\log p(x;\hat{\theta}^{(i+1)})$ does not depend on $y$, we can pull it out of the integral:

$$\begin{aligned}\int p(y|x;\hat{\theta}^{(i)})\log p(x;\hat{\theta}^{(i+1)})\,dy &= \log p(x;\hat{\theta}^{(i+1)})\underbrace{\int p(y|x;\hat{\theta}^{(i)})\,dy}_{=1} \\ &= \log p(x;\hat{\theta}^{(i+1)})\end{aligned}$$

**Right Side Definitions:** We define the two terms on the right side as functions of an arbitrary parameter $\theta$:

1. The **Q-function** (Expectation of the complete-data log-likelihood):

$$Q(\theta;\hat{\theta}^{(i)}) = \int p(y|x;\hat{\theta}^{(i)})\log p(x,y;\theta)\,dy \tag{12.8}$$

2. The **H-term** (Expectation of the posterior log-likelihood):

$$H(\theta;\hat{\theta}^{(i)}) = \int p(y|x;\hat{\theta}^{(i)})\log p(y|x;\theta)\,dy \tag{12.9}$$

Putting it all together, the decomposition of the log-likelihood for *any* parameter $\theta$ is:

$$\log p(x;\theta) = Q(\theta;\hat{\theta}^{(i)}) - H(\theta;\hat{\theta}^{(i)}) \tag{12.10}$$

Specifically, for the parameters of the next iteration $\hat{\theta}^{(i+1)}$, this holds as:

$$\log p(x;\hat{\theta}^{(i+1)}) = Q(\hat{\theta}^{(i+1)};\hat{\theta}^{(i)}) - H(\hat{\theta}^{(i+1)};\hat{\theta}^{(i)}) \tag{12.11}$$

**Note on Notation:** It is crucial to distinguish the two roles of the parameters in $Q(\theta;\hat{\theta}^{(i)})$:

- **Second argument $\hat{\theta}^{(i)}$ (Fixed):** These are the parameters from the *previous* iteration. They are treated as constant and are used solely to compute the posterior probabilities $p(y|x;\hat{\theta}^{(i)})$ (the "weights" in the expectation).
- **First argument $\theta$ (Variable):** This is the variable we want to optimize. It appears inside the logarithm $\log p(x,y;\theta)$. In the M-Step, we search for the value of $\theta$ that maximizes this function.

## Why is maximizing Q sufficient?

To justify that we only need to maximize $Q$, we examine the change in log-likelihood between the new parameter $\hat{\theta}^{(i+1)}$ and the old parameter $\hat{\theta}^{(i)}$:

$$\log p(x;\hat{\theta}^{(i+1)}) - \log p(x;\hat{\theta}^{(i)}) = \underbrace{\left(Q(\hat{\theta}^{(i+1)};\hat{\theta}^{(i)}) - Q(\hat{\theta}^{(i)};\hat{\theta}^{(i)})\right)}_{\Delta Q} - \underbrace{\left(H(\hat{\theta}^{(i+1)};\hat{\theta}^{(i)}) - H(\hat{\theta}^{(i)};\hat{\theta}^{(i)})\right)}_{\Delta H}$$

The term $-\Delta H$ corresponds to the difference of the H-terms. We can rewrite this difference:

$$\begin{aligned}
-\Delta H &= H(\hat{\theta}^{(i)}; \hat{\theta}^{(i)}) - H(\hat{\theta}^{(i+1)}; \hat{\theta}^{(i)}) \\
&= \int p(y|x; \hat{\theta}^{(i)}) \log p(y|x; \hat{\theta}^{(i)})\, dy - \int p(y|x; \hat{\theta}^{(i)}) \log p(y|x; \hat{\theta}^{(i+1)})\, dy \\
&= \int p(y|x; \hat{\theta}^{(i)}) \log \frac{p(y|x; \hat{\theta}^{(i)})}{p(y|x; \hat{\theta}^{(i+1)})}\, dy
\end{aligned}$$

This expression is exactly the **Kullback-Leibler Divergence**:

$$-\Delta H = D_{KL}(p(y|x; \hat{\theta}^{(i)}) || p(y|x; \hat{\theta}^{(i+1)}))$$

Since the KL-divergence is always non-negative ($D_{KL} \geq 0$), the term $-\Delta H$ is always $\geq 0$. This means the "entropy part" never decreases the total likelihood gain. Therefore, maximizing $Q$ (making $\Delta Q$ positive) guarantees that the total log-likelihood increases.

**Conclusion: The EM-Algorithm Update Rule**

**Algorithm 4** Expectation Maximization (EM) Algorithm

```
1: **Initialization:** Choose starting parameters $\hat{\theta}^{(0)}$
2: $i \leftarrow -1$
3: **repeat**
4:     $i \leftarrow i + 1$
5:     **Expectation step:**
6:         Calculate the $Q$-function using current parameters $\hat{\theta}^{(i)}$:
               $Q(\theta; \hat{\theta}^{(i)}) := \int p(y|x; \hat{\theta}^{(i)}) \log p(x, y; \theta)\, dy$
7:     **Maximization step:**
8:         Find the new parameters $\hat{\theta}^{(i+1)}$ that maximize $Q$:
               $\hat{\theta}^{(i+1)} \leftarrow \underset{\theta}{\operatorname{argmax}}\, Q(\theta; \hat{\theta}^{(i)})$
9: **until** $\hat{\theta}^{(i+1)} \approx \hat{\theta}^{(i)}$                ▷ Check for convergence
10: **Output:** estimate $\hat{\theta} \leftarrow \hat{\theta}^{(i)}$
```

**Pros and Cons of EM:**

- Pros:
  - Often closed form iteration schemes can be found
  - Numerically robust iteration scheme
  - Constant memory requirements if closed form solutions exist
- Cons:
  - Can converge to local maxima, depending on initialization.
  - May require many iterations to converge, especially for complex models (slow slow slow).

Many optimization problems with EM follow the constraint, that $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$. This can be solved with the method of Lagrange multipliers 8.3.1.

# 13 Independent Component Analysis (ICA)

While Principal Component Analysis (PCA) and Whitening rely on second-order statistics (covariance) to decorrelate data, Independent Component Analysis (ICA) goes a step further. It aims to recover latent variables that are not only uncorrelated but statistically **independent**.

## 13.1 The Cocktail-Party Problem

The classic motivating example for ICA is the "Cocktail-Party Problem". Imagine a room with two speakers $(s_1(t), s_2(t))$ and two microphones at different locations. The microphones record time signals $x_1(t)$ and $x_2(t)$. Since sound travels linearly, the recorded signals are weighted sums of the original sources:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) \end{aligned} \tag{13.1}$$

Here, the coefficients $a_{ij}$ depend on the distances between microphones and speakers. The goal of ICA is to recover the original sources $s(t)$ using only the observations $x(t)$, without knowing the mixing parameters $a_{ij}$ or the source distributions.

## 13.2 Latent Variable Model

We formalize this using a statistical latent variables model. We observe a random vector $\vec{x} \in \mathbb{R}^n$ which is a linear mixture of $n$ independent latent components $\vec{s} \in \mathbb{R}^n$:

$$\vec{x} = A\vec{s} \tag{13.2}$$

where $A \in \mathbb{R}^{n\times n}$ is the constant, unknown **mixing matrix**. The optimization task is to estimate both the mixing matrix $A$ and the realizations of the latent variables $\vec{s}$.

**Assumptions of ICA**

To ensure the model is solvable, we make the following assumptions:

1. The components $s_i$ are statistically independent.
2. The components $s_i$ must have **non-Gaussian** distributions (except for at most one component).
3. The mixing matrix $A$ is square and invertible (for the basic ICA model).

### 13.2.1 Ambiguities

Since both $A$ and $\vec{s}$ are unknown, the model $\vec{x} = A\vec{s}$ has two inherent ambiguities that cannot be resolved without prior knowledge:

1. **Scaling and Sign:** We cannot determine the variance of the independent components. A scalar factor $\alpha$ can be canceled out by dividing the corresponding column of $A$ by $\alpha$. Therefore, we usually restrict $\vec{s}$ to have unit variance ($E\{\vec{s}\vec{s}^T\} = I$). The sign is also undetermined.

2. **Ordering:** There is no natural order of the components. A permutation matrix $P$ can swap components in $\vec{s}$ while permuting columns in $A$ accordingly.

## 13.3 Why Decorrelation is Not Enough

In previous chapters, we used the Whitening Transform (based on PCA) to decorrelate data. For zero-mean data, the whitening transform is given by $\tilde{\vec{x}} = D^{-1/2}U^T\vec{x}$, where $E\{\vec{x}\vec{x}^T\} = UDU^T$.

While whitening ensures that the variables are uncorrelated ($E\{\tilde{\vec{x}}\tilde{\vec{x}}^T\} = I$), it does not guarantee independence.

**The Rotation Problem:** Whitening is only defined up to an orthogonal rotation. If $\tilde{\vec{x}}$ is white, then for any orthogonal matrix $R$, the vector $\hat{\vec{x}} = R\tilde{\vec{x}}$ is also white:

$$E\{\hat{\vec{x}}\hat{\vec{x}}^T\} = RE\{\tilde{\vec{x}}\tilde{\vec{x}}^T\}R^T = RIR^T = I \tag{13.3}$$

Since the Gaussian distribution is rotationally symmetric (determined fully by second-order moments), PCA cannot distinguish between the original axes and rotated axes for Gaussian data. This explains why ICA requires **non-Gaussian** data to identify the unique mixing matrix.

## 13.4 The Principle of Non-Gaussianity

To solve ICA, we use the Central Limit Theorem (CLT). The CLT states that the sum of independent random variables tends toward a Gaussian distribution.

Consider a linear combination of the observed mixture variables $y = \vec{w}^T\vec{x}$. Substituting $\vec{x} = A\vec{s}$ and letting $\vec{z} = A^T\vec{w}$:

$$y = \vec{w}^T A\vec{s} = \vec{z}^T\vec{s} = \sum_i z_i s_i \tag{13.4}$$

This $y$ is a sum of independent components. According to the CLT, this sum is usually "more Gaussian" than the individual source components $s_i$. Conversely, $y$ is **least Gaussian** when it corresponds to exactly one of the independent components (i.e., when only one $z_i$ is non-zero).

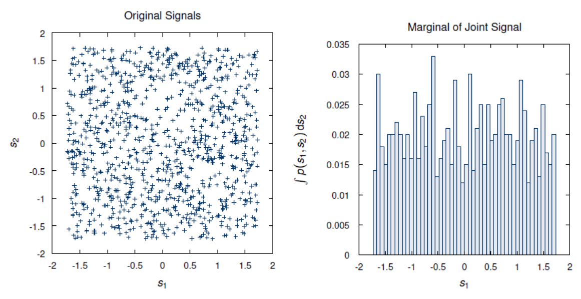


Figure 29: Original Signals: The scatter plot shows a square (independence). The histogram shows a uniform distribution, which is clearly non-Gaussian.

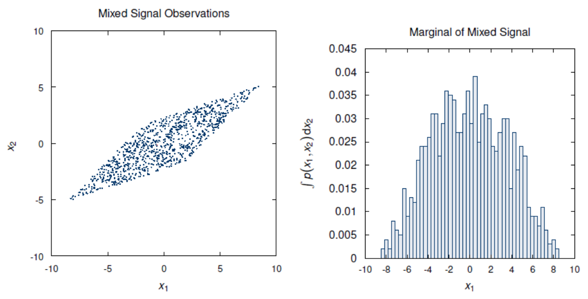


Figure 30: Mixed Signals: The scatter plot becomes a parallelogram (correlation). Due to the CLT, the histogram approaches a Gaussian bell curve.

**Conclusion:** To find the independent components, we must find the weight vector $\vec{w}$ that **maximizes the non-Gaussianity** of $\vec{w}^T\vec{x}$.

## 13.5 Measures of Non-Gaussianity

To perform the maximization described above, we need quantitative measures of how non-Gaussian a distribution is. We assume the variable $y$ is centered and has unit variance.

### 13.5.1 Differential Entropy

The fundamental concept behind many measures of non-Gaussianity is **Entropy**. In information theory, entropy measures the uncertainty or randomness of a random variable. For a continuous random variable $Y$ with probability density function $p(y)$, the **Differential Entropy** $H(Y)$ is defined as:

$$H(Y) = -\int p(y) \log p(y)\, dy \tag{13.5}$$

A crucial theorem for ICA states that for a fixed variance, the **Gaussian distribution maximizes the entropy**. This means the Gaussian distribution is the "most random" or least structured distribution. Distributions that are "spiky" or "flat" (super- or sub-Gaussian) have lower entropy.
**Interpretation:** Entropy can also be interpreted as a measure of code length. $H(Y)$ relates to the average code length required to encode the variable.

### 13.5.2 Kurtosis

Kurtosis is the fourth-order cumulant. For a random variable $y$ with unit variance:

$$\text{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2 = E\{y^4\} - 3 \tag{13.6}$$

- For a Gaussian distribution, $\text{kurt}(y) = 0$.
- **Super-Gaussian** (spiky, e.g., Laplace): $\text{kurt}(y) > 0$.
- **Sub-Gaussian** (flat, e.g., Uniform): $\text{kurt}(y) < 0$.

We optimize the absolute kurtosis $|\,\text{kurt}(\vec{w}^T\vec{x})|$. While theoretically sound, kurtosis is sensitive to outliers because of the fourth power ($y^4$).

### 13.5.3 Negentropy

Entropy $H(y)$ measures randomness. A Gaussian variable has the largest entropy among all distributions with the same variance. Negentropy $J(y)$ is defined to measure the distance to normality:

$$J(y) = H(y_{\text{Gauss}}) - H(y) \tag{13.7}$$

where $y_{\text{Gauss}}$ is a Gaussian variable with the same covariance as $y$. $J(y)$ is always non-negative and zero only if $y$ is Gaussian. Unlike Kurtosis, Negentropy is statistically robust, though computing the entropy requires estimating the pdf, which is computationally difficult. In practice, approximations involving non-quadratic functions $G$ are used:

$$J(y) \approx [E\{G(y)\} - E\{G(y_{\text{Gauss}})\}]^2 \tag{13.8}$$

### 13.5.4 Mutual Information

Another approach is minimizing the Mutual Information (MI) between the estimated components. MI measures the statistical dependence between variables.

$$\text{MI}(\vec{y}) = \sum_{i=1}^{n} H(y_i) - H(\vec{y}) \tag{13.9}$$

For uncorrelated variables, minimizing Mutual Information is equivalent to maximizing the sum of Negentropies of the components.

## 13.6 ICA Estimation Algorithm

A standard algorithm (conceptually similar to FastICA) to estimate one independent component is:

**Algorithm 5** Basic ICA Estimation Algorithm

```
1: Apply centering transform
2: Apply whitening transform
3: $i \leftarrow 1$
4: repeat
5:     Take a random vector $\vec{w}_i$
6:     Maximize non-Gaussianity of $\vec{w}_i^T \vec{x}$ subject to:
7:         $\|\vec{w}_i\| = 1$
8:         $\vec{w}_j^T \vec{w}_i = 0, \quad \forall j < i$
9:     $i \leftarrow i + 1$
10: until $i > n$                    ▷ $n$: number of independent components
11: Use weight matrix: $W = (\vec{w}_1^T, \vec{w}_2^T, \ldots, \vec{w}_n^T)$ to compute $\vec{s}$
12: Output: independent components $\vec{s}$
```

To do this, you have to choose a measure of non-Gaussianity (as described in 13.5).