Mathematical Foundations for Machine Learning | MathFML

Summary

CONTENTS

1. LINEAR ALGEBRA

Definition 1: **Linear combination**

$$\sum_{i=1}^{k} \lambda_i \vec{v}_i$$

1.1. STANDARD BASIS

Basis vectors $\{\vec{e}_1, \vec{e}_2, ..., \vec{e}_n\}$ in $\mathbb{R}^n$ where $\vec{e}_i = \left(0, 0, ..., \underbrace{1}_{i\text{th position}}, ..., 0\right)^\top$

Example 1:

$$\left[\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right] = \mathbb{R}^3$$

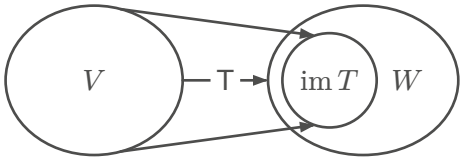
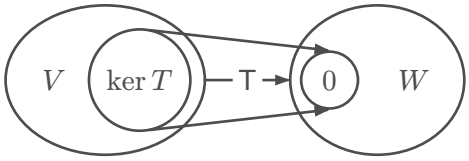
It can be shown that any basis of $\mathbb{R}^n$ consists of exactly n linear independent vectors and that conversely any set of n linear independent vectors is a basis of $\mathbb{R}^n$.

1.2. VECTORS

$$\begin{aligned}
\lambda\vec{0} &= \vec{0} \\
\vec{v} + \vec{0} &= \vec{v} \\
-\vec{v} &= -1 \cdot \vec{v} \\
-\vec{v} + \vec{v} &= \vec{0} \\
(\lambda\mu)\vec{v} &= \lambda(\mu\vec{v}) = \lambda\mu\vec{v} \\
\lambda(\vec{v} + \vec{w}) &= \lambda\vec{v} + \lambda\vec{w} \\
\vec{v} + (\vec{u} + \vec{w}) &= (\vec{v} + \vec{u}) + \vec{w} = \vec{v} + \vec{u} + \vec{w}
\end{aligned}$$

1.3. LINEAR TRANSFORMATIONS

$$T : V \to W$$

Term	Definition	
Image / range	$\operatorname{im} T = \{T(v) \mid v \in V\} = T(V)$	
Kernel / nullspace	$\ker T = \{v \in V \mid T(v) = 0\}$	

***Observations** 1:*

1. $\operatorname{im} T \subset W$
2. $\ker T \subset V$
3. $\vec{0} \in \ker A$

***Example** 2:*

$$A \in \mathbb{R}^{n \times n}, T_A : \begin{cases} \mathbb{R}^n \to \mathbb{R}^n \\ \vec{x} \quad \mapsto A\vec{x} \end{cases}$$

$$\ker A = \left\{ \vec{x} \in \mathbb{R}^n \mid A\vec{x} = \vec{0} \right\} = \left\{ \vec{x} \in \mathbb{R}^n \mid T_A(\vec{x}) = \vec{0} \right\}$$

1.3.1. Affine transformations

All linear transformations share the property, that they map the null vector to null:

$$L_A\left(\vec{0}\right) = \vec{0}$$

For many applications this is undesirable. Therefore linear transformations are often slightly extended to transformations, which are called **affine transformations**. An affine transformation is defined by an $r \times c$ matrix M and an r-dimensional vector b, which, in machine learning and statistics, is usually called **bias-vector** or **intercept**.

$$A_{b,M} : \begin{cases} \mathbb{R}^c \to \mathbb{R}^r \\ x \quad \mapsto b + M \cdot x \end{cases}$$

***Example** 3:* **Straight line in one dimension**

$$A_{b,m} : \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto mx + b \end{cases}$$

1.4. MATRICES

Term	*Definition*
Rank	Maximum number of linearly independent rows or columns $A \in \mathbb{R}^{m \times n} \Rightarrow \operatorname{rank}(A) = \rho(A) \leq n$
Nullity	Number of vectors in the null space $A \in \mathbb{R}^{m \times n} \Rightarrow \operatorname{nullity}(A) = n - \rho(A) = \dim(\ker A)$
Dimension	$\dim(\ker A) = \operatorname{nullity}(A)$ $\vec{x} \in \mathbb{R}^3 \Rightarrow \dim \vec{x} = 3$
Span	Set of all finite linear combinations of the elements of S of a vector space V. $\operatorname{span}(S) = \{\lambda_1\vec{v}_1 + \lambda_2\vec{v}_2 + ... + \lambda_n\vec{v}_n \mid n \in \mathbb{N}, v_1, ..., v_n \in S, \lambda_1, ..., \lambda_n \in K\}$ $\operatorname{span}\{(1,0,0),(0,1,0),(0,0,1)\} = \mathbb{R}^3$

Term	Definition
Column space	Span of the column vectors of a matrix $$\text{colsp}(A) = C(A)$$ $$A \in \mathbb{R}^{m\times n}, \dim(\text{colsp}(A)) = n$$
Row space	Span of the row vectors of a matrix $$\text{rowsp}(A) = C(A^\top)$$ $$A \in \mathbb{R}^{m\times n}, \dim(\text{rowsp}(A)) = m$$
Quadratic Matrix	$M \in \mathbb{R}^{n\times n}$
Symmetric Matrix	$M \in \mathbb{R}^{n\times n} \wedge M^\top = M$

***Observations** 2:*

1. $\text{rank}(T) + \text{nullity}(T) = \dim V$
2. $\dim(\text{im}\, T) + \dim(\ker T) = \dim(\text{domain}\, T)$
3. $\text{rank}(A) = \dim(\text{rowsp}(A)) = \dim(\text{colsp}(A))$
4. $A \in \mathbb{R}^{n\times n} \Rightarrow \dim(\ker A) + \text{rank}\, A = n$
5. $A \in \mathbb{R}^{n\times n}, \ker A = \left\{\vec{0}\right\} \Leftrightarrow \text{rank}\, A = n$

1.4.1. Properties

Term	Rules
Associativity	$(A+B)+C = A+(B+C) = A+B+C$ $(A\cdot B)\cdot C = A\cdot(B\cdot C) = A\cdot B\cdot C$
Distributivity	$C\cdot(A+B) = C\cdot A + C\cdot B$ $(A+B)\cdot C = A\cdot C + B\cdot C$
Commutativity	$A+B = B+A$ $A\cdot B \neq B\cdot A$
Transposing	$(A^\top)^\top = A$ $(A+B)^\top = A^\top + B^\top$ $(\lambda A)^\top = \lambda A^\top$ $(A\cdot B)^\top = B^\top \cdot A^\top$ $A_{ij} = A^\top_{ji}$
Identity matrix	$\mathbb{1}\cdot A = A\cdot\mathbb{1} = A$ for $A \in \mathbb{R}^{n\times n}$
Inverting	$A\cdot A^{-1} = A^{-1}\cdot A = \mathbb{1}$

Term	Rules
Determinate	$\det(\lambda M) = \lambda^n \det(M), M \in \mathbb{R}^{n\times n}$ $\det\begin{pmatrix} A & * \\ 0 & B \end{pmatrix} = \det(A)\cdot\det(B)$ $\det(A\cdot B) = \det(A)\cdot\det(B)$ $\det(A^{-1}) = \frac{1}{\det(A)}$ $\det(A^\top) = \det(A)$
Scalars	$(\lambda+\mu)A = \lambda A + \mu A$ $(\lambda\mu)A = \lambda(\mu A)$ $\lambda(B+C) = \lambda B + \lambda C$ $\lambda(BC) = (\lambda B)C = B(\lambda)C$

1.4.2. Unit matrices

Let M be the set of all 2×2 matrices. How could the basis of M look? $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$

Unit matrices:

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$$M = \{E_{11}, E_{12}, E_{21}, E_{22}\}$$

$$(E_{11})_{22} = 0, (E_{11})_{11} = 1$$

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^3, (\vec{e}_1)_1 = 1$$

1.4.3. Kronecker Delta

$$\delta_{ij} = \begin{cases} 1 \ , \ i = j \\ 0 \ , \ \text{otherwise} \end{cases}$$

***Example** 4:*

$$\delta_{23} = 0, \delta_{11} = 1, (\vec{e}_1)_j = \delta_{1j} = \begin{cases} 1 \ , \ 1 = j \\ 0 \ , \ \text{otherwise} \end{cases}$$

$$\delta_{m1}\delta_{m2} = 0$$

$$\delta_{kl} - \delta_{lk} = 0$$

$$(e_k)_i = \delta_{ki} = \delta_{ik}$$

$$(E_{kl})_{ij} = \delta_{ki}\delta_{lj}$$

$$(E_{rst})_{ijk} = \delta_{ri}\delta_{sj}\delta_{tk}$$

1.4.4. Matrix product

$$A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times r}, A\cdot B \in \mathbb{R}^{n\times r}$$

$$A \cdot B = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ b_{i1} & b_{i2} & \dots & b_{ir} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mr} \end{pmatrix}$$

$$(A \cdot B)_{11} = a_{11}b_{11} + a_{12}b_{12} + \dots + a_{1m}b_{m1} = \sum_{k=1}^{m} a_{1k}b_{k1}$$

$$(A \cdot B)_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

Shorthand:

$$(A \cdot B)_{ij} = \sum_{k} a_{ik}b_{kj}$$

Transposing:

$$\left[(A \cdot B)^\top\right]_{ij} = (A \cdot B)_{ji} = \sum_{k} a_{jk}b_{ik}$$

***Example** 5:* **Let X be a matrix for which X^{-1} exists. Prove that the inverse of $X^\top$ exists and is $\left(X^{-1}\right)^\top$**

$$\begin{aligned} & \mathbb{1}^\top = \mathbb{1} \\ \Rightarrow & \left(X^{-1}X\right)^\top = \mathbb{1} \\ \Leftrightarrow & X^\top\left(X^{-1}\right)^\top = \mathbb{1} \end{aligned}$$

1.5. TENSORS

Vectors = tensors of rank 1, matrices = tensors of rank 2

The standart basis for $\mathbb{R}^{n\times m\times l}$ consists of $n \cdot m \cdot l$ basis vectors E_{ijk} where $1 \leq i \leq n, 1 \leq j \leq m, 1 \leq k \leq l$, such that all components of E_{ijk} are zero except at position i, j, k where the value of the component is 1.

2. RESIDUAL SUM OF SQUARE

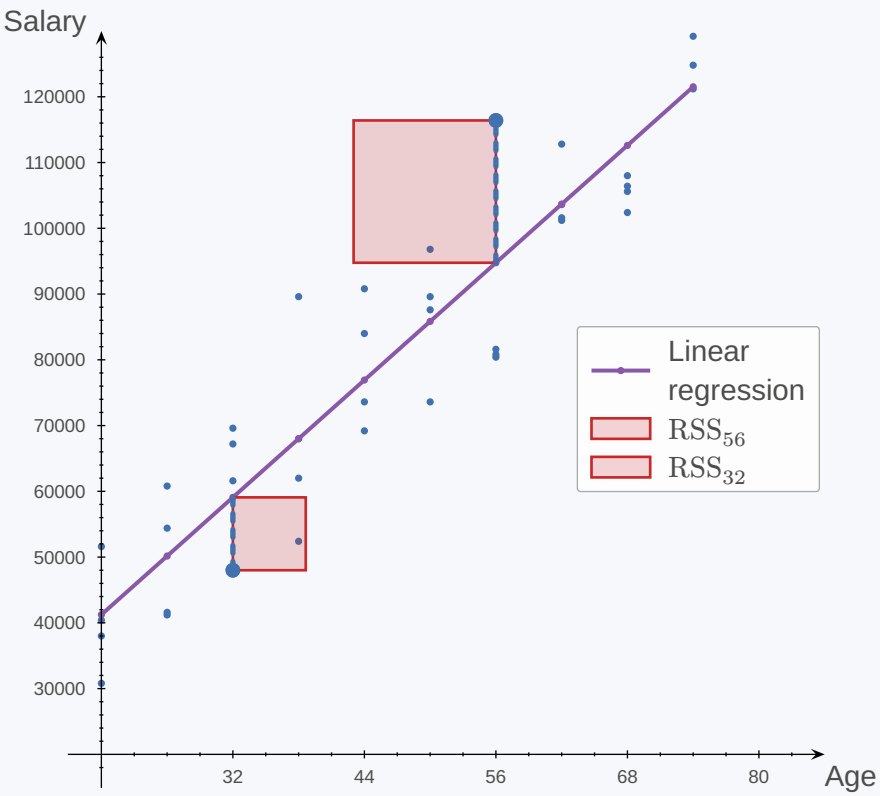
Residual sum of square (RSS) is a statistical method that helps identify the level of discrepancy in a dataset not predicted by a regression model. Thus, it measures the variance in the value of the observed data when compared to its predicted value as per the regression model.

$$RSS = \sum_{i=1}^{N} \underbrace{(y_i - f(x_i))^2}_{\text{Represented as red squares in the example}}$$

$$RSS(m, b) = \sum_{i=1}^{N} (y_i - (mx_i + b))^2 \geq 0, RSS : \mathbb{R}^2 \to \mathbb{R}$$

***Example** 6:* **Linear regression of salaries by age**

$$\mathrm{RSS}(m,b) = \sum_{i=1}^{N} \left((mx_i + b) - y_i\right)^2 \geq 0, \mathrm{RSS} : \mathbb{R}^2 \to \mathbb{R}$$



3. FUNCTIONS OF SEVERAL VARIABLES

Let D and R be two sets. A mapping $f : D \to R$ that associates to each element $x \in D$ exactly one element of $f(x) \in R$ is called function.

Term	*Definition*
Domain of f	D
Range of f	R
Real valued function	$R \subset \mathbb{R}$
Vector valued function	$R \subset \mathbb{R}^m$
A function of n variables	$D \subset \mathbb{R}^n$

3.1. VISUALIZING FUNCTIONS

If $f : D \to R$ is a function from $D \subset \mathbb{R}^n$ to $R \subset \mathbb{R}^m$ its graph is defined as:

$$\mathrm{graph}(f) = \{(\vec{x}, \vec{y}) \in D \times R \mid x \in D \wedge \vec{y} = f(\vec{x})\} \subset D \times R \subset \mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$$

Due to the fact that the visual imagination of humans is limited to at most 3 dimensions, the graph of a function is a useful concept for graphical illustration if and only if $n + m \leq 3$.

***Example** 7:*

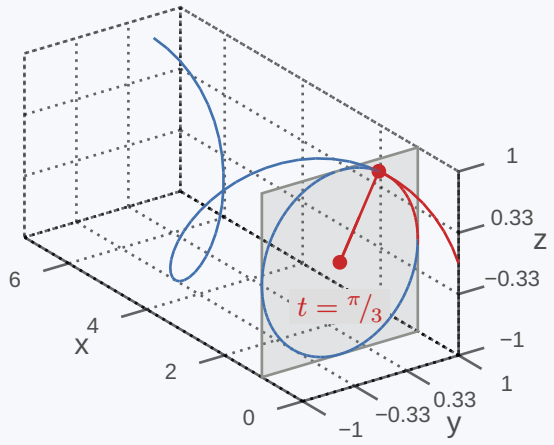
The graph of the function

$$f : \begin{cases} \mathbb{R} \to \mathbb{R}^2 \\ t \mapsto (\cos t, \sin t) \end{cases}$$

is

$$\begin{aligned} \operatorname{graph}(f) &= \{(t, x, y) \in \mathbb{R}^3 \mid (x, y) = f(t) \wedge t \in \mathbb{R}\} \\ &= \{(t, \cos t, \sin t) \in \mathbb{R}^3 \mid t \in \mathbb{R}\} \end{aligned}$$

It is displayed below and illustrates how f maps t to the corresponding values



***Example** 8:*

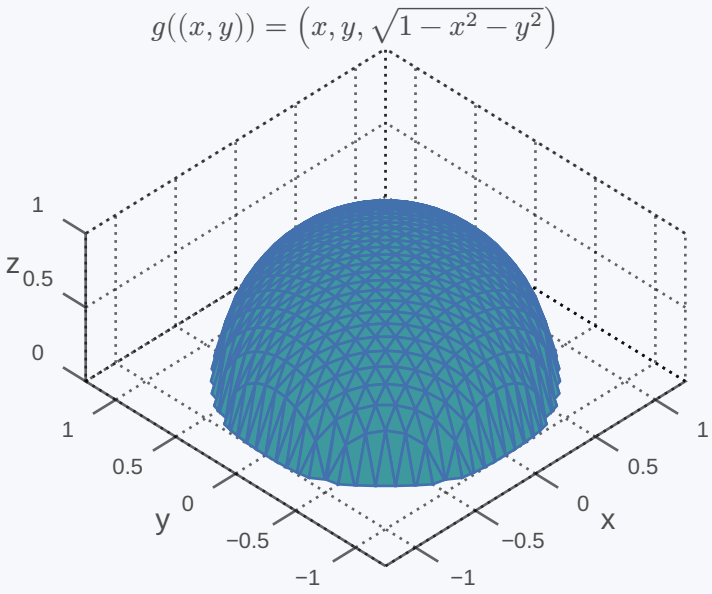
The graph of the function

$$g : \begin{cases} \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\} \to \mathbb{R}^2 \\ (x, y) \qquad\qquad\qquad\qquad \mapsto \sqrt{1 - x^2 - y^2} \end{cases}$$

is

$$\begin{aligned} \operatorname{graph}(g) &= \left\{(x, y, z) \in \mathbb{R}^3 \;\middle|\; z = \sqrt{1 - x^2 - y^2} \wedge x^2 + y^2 \leq 1\right\} \\ &= \left\{\left(x, y, \sqrt{1 - x^2 - y^2}\right) \in \mathbb{R}^3 \;\middle|\; x^2 + y^2 \leq 1\right\} \end{aligned}$$

It is displayed below and illustrates how f maps (x, y) to the corresponding values



3.1.1. Curves and surfaces

Term	*Definition*
Curves	Vector valued functions of one variable, i.e. they map one-dimensional input to multi-dimensional output
n-dimensional curve	A continuous function $f : I \to Z$ that maps an interval $I \subset \mathbb{R}$ into a subset $Z \subset \mathbb{R}^n$
Surfaces	Real valued functions of multiple variables, i.e. they map multi-dimensional input to one dimensional output
n-dimensional hyper-surface	A continuous real valued function $f : D \to Z$ that maps a sufficiently large subset $D \subset \mathbb{R}^n$ into a subset $Z \subset \mathbb{R}$ If $n = 2$ a hypersurface is also simply called surface. If $n = 1$ a hyper-surface is simply a continuous real valued function of one variable.

The graph of both, an n-dimensional curve and an n-dimensional hyper-surface, is a subset of $\mathbb{R}^{n+1}$. A graphical illustration of such a graph therefore requires $n \leq 2$. However, unlike surfaces which are typically illustrated as in Figure 1 , a curve $c : I \to \mathbb{R}^n$ is usually not illustrated through

$$\operatorname{graph} c = \{(t, y) \in R \times R^n \mid y = c(t) \wedge t \in I\}$$

Instead curves are typically visualized by skipping the independent variable t from the graph, i.e. by visualizing the image of the curve

$$c(I) = \{y \in R^n \mid y = c(t) \wedge t \in I\}$$

which means that we are projecting the graph of the curve onto the gray plane that is spanned from the dependent variables. As this kind of plot saves one dimension we can also visualize curves for which the range Z is a subset of $\mathbb{R}^3$.

***Example** 9:*

The graph of the function

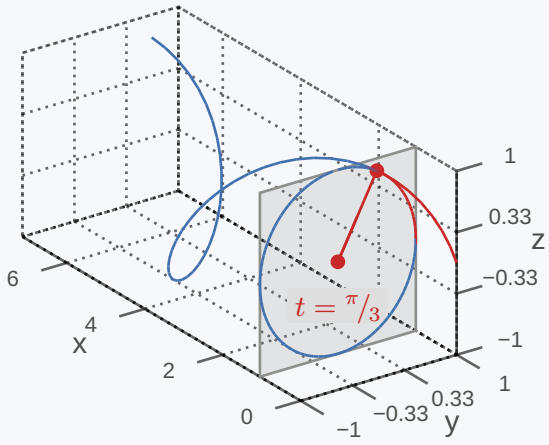
$$f : \begin{cases} \mathbb{R} \to \mathbb{R}^2 \\ t \mapsto (\cos t, \sin t) \end{cases}$$

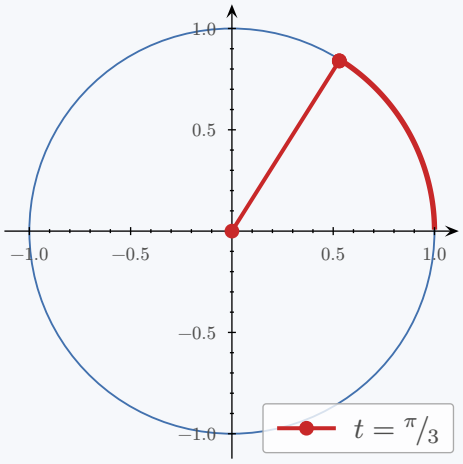
is

$$\begin{aligned} \text{graph}(f) &= \{(t, x, y) \in \mathbb{R}^3 \mid (x, y) = f(t) \wedge t \in \mathbb{R}\} \\ &= \{(t, \cos t, \sin t) \in \mathbb{R}^3 \mid t \in \mathbb{R}\} \end{aligned}$$

and its image is the set

$$\begin{aligned} f(\mathbb{R}) &= \{f(t) \in \mathbb{R}^2 \mid t \in \mathbb{R}\} \\ &= \{(\cos t, \sin t) \in \mathbb{R}^2 \mid t \in \mathbb{R}\} \end{aligned}$$

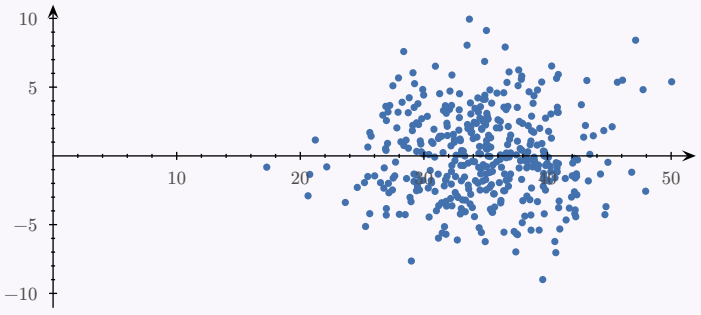




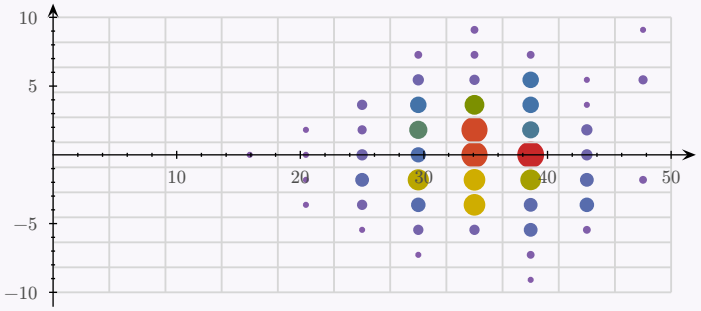
Unlike for image processing, curves are of little importance in machine learning. On the contrary, hypersurfaces belong to the most important functions in machine learning. This is because, to a large extend, machine learning can be thought of being a branch of statistics. In statistics the most important class of functions are probability distributions, which are functions mapping subsets of R^n into R^+.

Definition 2: **Histogram**

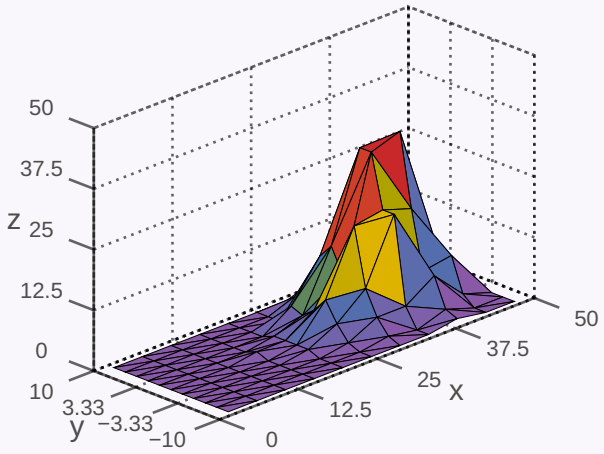
Consider having some amount of 2D datapoints:



We can place a grid over the diagram and count, how many points are in each of the boxes.



We can now plot a diagram that maps the x and y coordinate of the center of each of the boxes to the number of hits of this particular box and thus end up with a diagram that is called a ***histogram***.



We can try to create a histogram with boxes of infinitely small size. The problem with this approach is, that the number of hits per box decreases when the size of the boxes is getting smaller.

In order to compensate that, we therefore need more datapoints. If we do that, the histogram is being transformed into a function, that associates a number to each position on the diagram, that can be interpreted as likelihood for a point to be almost next to this position: the higher this number is, the more likely it is. This function is called the **probability density** of the experiment.

3.1.1.1. Vector similarity

One approach to decide, whether two images are similar or not, is to first extract features (such as corner-points, edges, etc.) from each of the images, and store these features in a list of numbers, which is called ***feature vector***. It is reasonable to assume, that similar images are mapped to similar feature vectors.

Two vectors v and w are **similar**, if both vectors head into the same direction and are of similar length, so if the difference $v - w$ between both vectors has a small length. In vector geometry, the length of a vector is called a **norm**, and the distance between two vectors is called a **metric**. A norm $\|\cdot\|$ is thus a function that associates with a vector x a positive number $\|x\|$, which is interpreted as length of x.

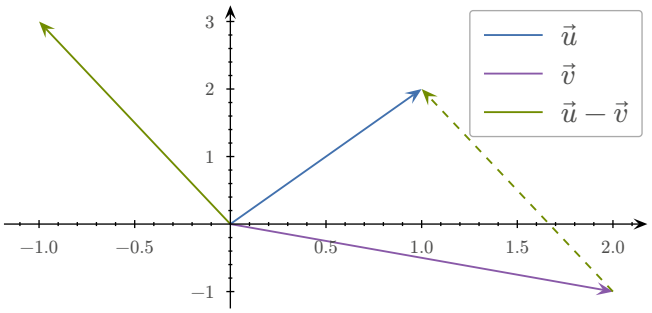
$$\|\cdot\| : \begin{cases} \mathbb{R}^n \to \mathbb{R}^+ \\ x \mapsto \sqrt{\sum_{i=1}^n x_i^2} \end{cases}$$

$= n$-dimensional Euclidean norm or l^2-norm, written also as $\|x\|_2$.

Once a vector space is equipped with a norm $\|.\|$, we can also associate a metric with that vector space, which uses the notion of a length to measure the distance between pairs of vectors:

$$d(\cdot,\cdot): \begin{cases} \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+ \\ (u,v) \quad\quad \mapsto \|u-v\| \end{cases}$$

$=$ metric. The value $d(u,v)$ is identical to the length of the vector that heads from v to u.



3.1.2. Vector valued functions of several variables

In machine learning many functions are vector valued functions of many variables, like linear transformations. Neural networks are typically defined to be a stack of unknown linear transformations, each of which being followed by a rather simple, predefined non-linear operation. Training a neural network is therefore almost equivalent to finding suitable linear transformations that make the network perform the desired operation.

Besides curves and surfaces there are functions that take multi-dimensional input and deliver multi-dimensional output. In order to get an idea how these functions look like, we often display low-dimensional "sections" of the function that can either be displayed as curves or as surfaces. There is however another way, how these functions can be displayed, namely using a **vector field**.

3.2. CALCULUS OF CURVES

***Definition** 3:* **Coordinate functions**

Let $I \subset \mathbb{R}$ be an interval and $c: I \to \mathbb{R}^n$ be an arbitrary curve. Then there are n (contiguous) coordinate functions $c_i: I \to \mathbb{R}$, such that

$$c(t) = \begin{pmatrix} c_1(t) \\ c_2(t) \\ \vdots \\ c_n(t) \end{pmatrix}$$

and vice versa: If n real valued (continuous) functions $c_i: I \to \mathbb{R}$ with a common interval I as domain of definition are given, we can define a curve $c(t)$ whose coordinate functions are $c_1, c_2, ..., c_n$.

The decomposition of curves into coordinate functions allows us to generalize most of the concepts from previous analysis courses to curves:

We consider the curve

$$f: \begin{cases} (0;\infty) \to \mathbb{R}^2 \\ t \quad\quad\quad \mapsto \begin{pmatrix} \cos(t) \\ \frac{\sin(t)}{t} \end{pmatrix} \end{cases}$$

The component functions are

$$f_1(t): \cos(t)$$

$$f_2(t) : \frac{\sin(t)}{t}$$

We can thus argue:

Definition 4: **Continuity**

Since both of the component functions are continuous, f is continuous.

Definition 5: **Limit point**

$$\lim_{t\to 0} f_1(t) = \lim_{t\to 0} \cos(t) = \cos(0) = 1$$
$$\lim_{t\to 0} f_2(t) = \lim_{t\to 0} \frac{\sin(t)}{t} \overset{\text{"0/0"}}{=} \lim_{t\to 0} \frac{\frac{\mathrm{d}}{\mathrm{d}t}\sin(t)}{\frac{\mathrm{d}}{\mathrm{d}t}t} = \lim_{t\to 0} \frac{\cos(t)}{1} = \cos(0) = 1$$
$$\Rightarrow \lim_{t\to 0} f(t) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Definition 6: **Derivative**

$$\frac{\mathrm{d}}{\mathrm{d}t} f_1(t) = \frac{\mathrm{d}}{\mathrm{d}t} \cos(t) = -\sin(t)$$
$$\frac{\mathrm{d}}{\mathrm{d}t} f_2(t) = \frac{\mathrm{d}}{\mathrm{d}t} \frac{\sin(t)}{t} = \frac{\cos(t)\cdot t - \sin(t)\cdot 1}{t^2} = \frac{\cos(t)}{t} - \frac{\sin(t)}{t^2}$$
$$\Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} f(t) = \begin{pmatrix} -\sin(t) \\ \frac{\cos(t)}{t} - \frac{\sin(t)}{t^2} \end{pmatrix}$$

Definition 7: **Linearisation at $t = \pi$**

$$f_1(t) \approx f_1(\pi) + f_1'(\pi)(t-\pi) = \cos(\pi) - \sin(\pi)(t-\pi) = -1$$
$$f_2(t) \approx f_2(\pi) + f_2'(\pi)(t-\pi) = \frac{\sin(\pi)}{\pi} + \left(\frac{\cos(\pi)}{\pi} - \frac{\sin(\pi)}{\pi^2}\right)(t-\pi) = -\frac{1}{\pi}(t-\pi)$$
$$\Rightarrow f(t) \approx f(\pi) + f'(\pi)(x-\pi) = \begin{pmatrix} -1 \\ -\frac{1}{\pi}(t-\pi) \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} - (t-\pi)\begin{pmatrix} 0 \\ \frac{1}{\pi} \end{pmatrix}$$

For any $\vec{a}, \vec{v} \in \mathbb{R}^n$ a curve of the form $g : \begin{cases} \mathbb{R} \to \mathbb{R}^n \\ t \mapsto \vec{a} + t\cdot\vec{v} \end{cases}$ corresponds to a line going through the point a in the direction of v.

Definition 8: **Reparametrization**

Let

$$c : \begin{cases} I \to \mathbb{R}^n \\ t \mapsto c(t) \end{cases}$$

be a curve and $h : J \to I$ an arbitrary bijective function. Then the image of the curve

$$d : \begin{cases} J \to \mathbb{R}^n \\ s \mapsto c(h(s)) \end{cases}$$

is identical to the image of the curve $c(t)$:

$$c(I) = d(J)$$

We call $d(s)$ ***reparametrization*** of $c(t)$ using $t = h(s)$

***Definition** 9:* **Image of a curve**

The image of a curve

$$c : \begin{cases} I \to \mathbb{R}^n \\ t \mapsto c(t) \end{cases}$$

can be envisioned as a road that is passed from a passenger, who reaches position $c(t)$ at time t. Thus:

- The **derivative** $c'(t)$ is a **measure of the speed** of the person at time t.
- The **derivative** $c'(t)$ points into a **direction tangent to the road** at position $c(t)$
- The **norm** $\|c'(t)\|$ of the derivative is **a measure of the speed** with which the person is moving **at time** t.

If we reparametrize the curve c with a bijective function $t = h(s)$ the image of the curve

$$d(s) = c(h(s))$$

is identical to the image of c and thus corresponds to the same road. It is however passed by another person, who is passing point $d(s) = c(h(s))$ with a velocity

$$d'(s) = c'(h(s)) \cdot h'(s)$$

that is $h'(s)$ faster (or slower) than the velocity of the first person at time $t = h(s)$.

3.3. CALCULUS OF REAL VALUED FUNCTIONS IN MANY VARIABLES

3.3.1. Partial derivative and gradient

Let $f : D \to \mathbb{R}$ be an arbitrary real valued function and $x \in D \subset \mathbb{R}^n$ be an arbitrary vector in the domain of f. We consider the domain D of f to be sufficiently large for doing calculus at x, if for every vector $v \in \mathbb{R}^n$ there is a small interval $(-\varepsilon; \varepsilon), \varepsilon > 0$, such that for all $t \in (-\varepsilon; \varepsilon)$ the vector $x + t \cdot v \in D$. This condition guarantees that any curve that hits the domain of f, remains there for at least a short period of time, and is one of the properties that needs to be satisfied in order to denote f as a hyper-surface. The other property that is required for f to be a hyper-surface, is that f is continuous at any point of its domain.

***Definition** 10:* **Partial derivative**

The ***partial derivative*** of f with respect to the i-th coordinate function is defined through

$$\frac{\partial}{\partial x_i} f(x) = \lim_{t \to 0} \frac{f(x + t \cdot e_i) - f(x)}{t}$$

***Definition** 11:* **Gradient**

The ***gradient*** of f is an n-dimensional row vector, whose components correspond to the various **partial derivatives**.

$$\nabla f(x) = \left(\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), ..., \frac{\partial}{\partial x_n} f(x) \right)$$

***Example** 10:* $f : \begin{cases} \mathbb{R} \times \mathbb{R} \to \mathbb{R} \\ (x, y) \mapsto x^2 y^3 \end{cases}$

$$f'_x(x, y) = \frac{\partial f}{\partial x} = 2x \cdot y^3$$

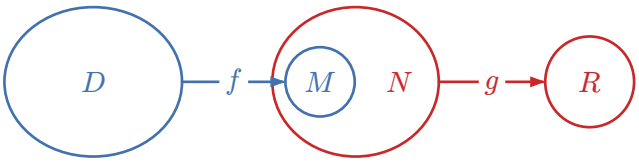
$$f'_y(x, y) = \frac{\partial f}{\partial y} = x^2 \cdot 3y^2$$

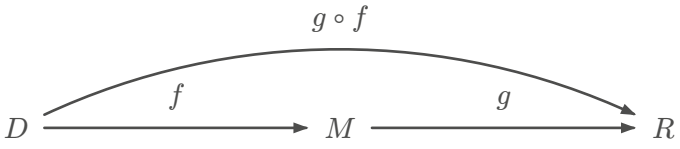
$$\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) = (2x \cdot y^3, x^2 \cdot 3y^2)$$

3.3.2. Compositions of functions

We can use the notion of dependent variables as an alternative representation of functions, this notion allows us to define functions without explicitly stating its arguments.

3.3.3. Commutative diagrams





Let us assume that

$$f : \begin{cases} D \to M \\ x \mapsto f(x) \end{cases}, \quad g : \begin{cases} N \to R \\ y \mapsto g(y) \end{cases}, \quad h : \begin{cases} D \to R \\ x \mapsto g(f(x)) \end{cases} = g \circ f$$

The graph

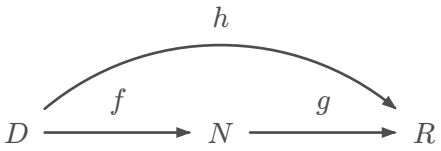
$$D \xrightarrow{f} M \subset N \xrightarrow{g} R$$

illustrates that f is a function from D to M, g is a function from N to R and M is a subset of N. Alternatively, for every set Q, such that $M \subset Q \subset R$ we could also have written

$$D \xrightarrow{f} Q \xrightarrow{g} R$$

indicating that f maps D into a subset of Q, which in turn is a subset of the domain of g and thus can be mapped to R using g.

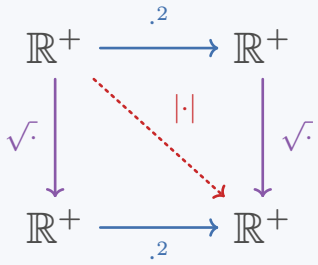
A ***commutative diagram*** is an extension of the previous diagram, in which we add an additional arrow pointing from D to R.



By labeling this additional arrow with h, a commutative diagram can be used to define a function from D to R, that is behaving such, that it maps every $x \in D$ to exactly the same element as the composition of g with f.

***Example** 11:*

$$g(x) = x^2$$
$$f(x) = \sqrt{x}$$
$$f(g(x)) = \sqrt{x^2} \quad = |x|$$
$$g(f(x)) = \left(\sqrt{x}\right)^2 = |x|$$



3.4. CALCULUS OF VECTOR VALUED FUNCTIONS OF MANY VARIABLES

3.4.1. Jacobian matrix and linearisation

In the theory of real valued functions in one variable the linearisation of $f : \mathbb{R} \to \mathbb{R}$ at $x_0 \in \mathbb{R}$ is done via

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0), x \approx x_0$$

and can be interpreted as:

1) First order Taylor polynomial of f, which implies that the linearisation of f at x_0 is that first order polynomial which best approximates f in the vicinity of x_0.
2) Identifying that particular line that fits best to the graph of f in a neighborhood of x_0. This line is known as tangent of f at x_0.
3) Definition of the derivative of f at x_0.

$$f'(x_0) \approx \frac{f(x) - f(x_0)}{x - x_0} \text{ for } x \approx x_0, x \neq x_0$$

tells us that both sides of this equation become equal for $x \to x_0$, and therefore equivalent to the definition of the derivative of f at x_0

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

In the theory of vector valued functions of many variables, the linearisation of a vector valued function in n variables $f : \mathbb{R}^n \to \mathbb{R}^m$, itself has to be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. Further, the right hand side has to be an affine transformation, i.e. a function of the form

$$A_{a,M}(x) = a + M \cdot x$$

in which M is a matrix and a is a vector. Since we expect that $A_{a,M}$ to map $\mathbb{R}^n$ to $\mathbb{R}^m$, we can infer that $a \in \mathbb{R}^m$ and $M \in \mathbb{R}^{m \times n}$.

The **derivative of a vector valued function** $f : \mathbb{R}^n \to \mathbb{R}^m$ can thus be expected to be a matrix $M \in \mathbb{R}^{m \times n}$. This matrix is called ***Jacobian matrix*** and it can be shown that this matrix is comprised of all partial derivatives of all component functions of f.

***Definition** 12:* **Jacobian matrix**

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a some vector valued function and $x_0 \in \mathbb{R}^n$ be a point, where the partial derivatives of all component functions $\partial_{x_k} f_l(x_0)$ exist. The $(m \times n)$-matrix that aggregates all partial derivatives of f at x_0

$$J_f(x_0) = \begin{pmatrix} \partial_{x_1} f_1(x_0) & \partial_{x_2} f_1(x_0) & \dots & \partial_{x_n} f_1(x_0) \\ \partial_{x_1} f_2(x_0) & \partial_{x_2} f_2(x_0) & \dots & \partial_{x_n} f_2(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{x_1} f_m(x_0) & \partial_{x_2} f_m(x_0) & \dots & \partial_{x_n} f_m(x_0) \end{pmatrix}$$

such that the l-th row of $J_f(x_0)$ corresponds to the gradient of the l-th coordinate function f_l, is called ***Jacobian matrix*** of f at x_0. Besides of the symbol $J_f(x_0)$ the following notations are also used for the Jacobin matrix at x_0:

$$f'(x_0) = \frac{\partial f}{\partial x}\bigg|_{x=x_0} = J_f(x_0)$$

***Definition** 13:* **Linearisation**

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a some vector valued function and $x_0 \in \mathbb{R}^n$ be an arbitrary point from the domain of definition of f If $J_f(x_0) = {}^{\partial f}\!/_{\partial x}\big|_{x=x_0}$ is the Jacobin matrix of f at x_0, then

$$f(x) \approx f(x_0) + J_f(x_0) \cdot (x - x_0) \text{ for } x \approx x_0$$

The function on the right-hand side is called ***linearisation*** of f at x_0.

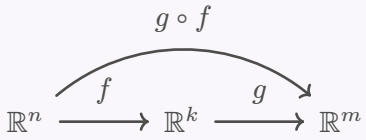
3.5. SUMMARY DERIVATIVES

Function	***Derivative***
$f : \begin{cases} \mathbb{R} \to \mathbb{R}^n \\ x \mapsto y \end{cases}$	$f' : \begin{cases} \mathbb{R} \to \mathbb{R}^n \\ x \mapsto y \end{cases} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x} f_1(x) \\ \frac{\mathrm{d}}{\mathrm{d}x} f_2(x) \\ \vdots \\ \frac{\mathrm{d}}{\mathrm{d}x} f_n(x) \end{pmatrix}$
$f : \begin{cases} \mathbb{R}^n \to \mathbb{R} \\ x \mapsto y \end{cases}$	$f' : \begin{cases} \mathbb{R}^n \to \mathbb{R}^{1 \times n} \\ x \mapsto y \end{cases} = \left(\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), ..., \frac{\partial}{\partial x_n} f(x) \right) = \nabla f$
$f : \begin{cases} \mathbb{R}^n \to \mathbb{R}^m \\ x \mapsto y \end{cases}$	$f' : \begin{cases} \mathbb{R}^n \to \mathbb{R}^{m \times n} \\ x \mapsto y \end{cases} = \begin{pmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(x) & \frac{\partial}{\partial x_2} f_1(x) & \dots & \frac{\partial}{\partial x_n} f_1(x) \\ \frac{\partial}{\partial x_1} f_2(x) & \frac{\partial}{\partial x_2} f_2(x) & \dots & \frac{\partial}{\partial x_n} f_2(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \frac{\partial}{\partial x_2} f_m(x) & \dots & \frac{\partial}{\partial x_n} f_m(x) \end{pmatrix} = J_f$

3.5.1. Chain-rule for vector valued functions in many variables

Definition 14: **Jacobian matrix chain rule**

Let f and g be given such that the following diagram commutes



Then we can calculate the Jacobian matrix of $g \circ f$ using the chain rule

$$\frac{\partial g(f(x))}{\partial x} = \frac{\partial g(f)}{\partial f}\bigg|_{f=f(x)} \cdot \frac{\partial f(x)}{\partial x} = J_g(f(x)) \cdot J_f(x)$$

Definition 15: **Chain rule for hyper-surfaces**

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a hyper-surface and $g : \mathbb{R}^n \to \mathbb{R}^m$ a vector valued function, such that $h = f \circ g$ is defined. Then

$$\nabla h(x) = \nabla f(g(x)) = \nabla f(g)|_{g=g(x)} \cdot \frac{\partial}{\partial x} g(x)$$

Definition 16: **Properties of differentiation of hyper-surfaces**

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be hyper-surfaces. Then

$$\begin{aligned}
\nabla(f+g) &= \nabla f + \nabla g \\
\nabla(f-g) &= \nabla f - \nabla g \\
\nabla(f \cdot g) &= g\nabla f + f\nabla g \\
\nabla \frac{f}{g} &= \frac{g\nabla f - f\nabla g}{g^2}
\end{aligned}$$

4. GEOMETRY OF HYPER-SURFACES

4.1. INTRODUCTION

Term	*Definition*
Training sample	Knowledge of in- and outputs prior to learning
Supervised learning	A process for finding suitable values of parameters that make the system respond as desired for as many training samples as possible
Cost functions	Process of finding the values in supervised learning. Function of n parameters that maps them to a high value if the misclassification is high and low if otherwise: $c : \mathbb{R}^n \to \mathbb{R}$

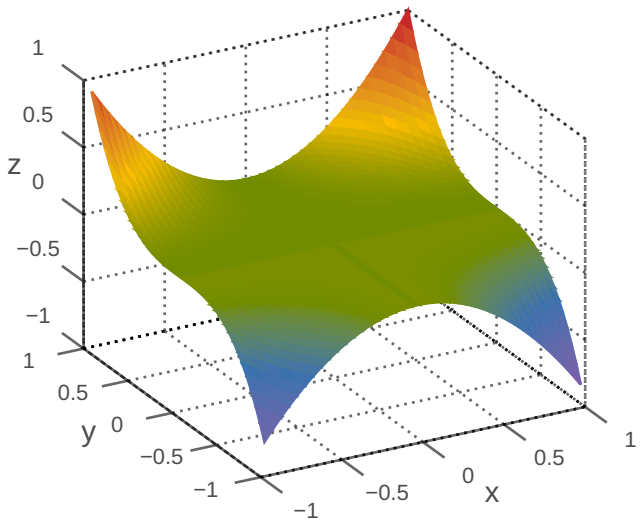
Term	*Definition*
Train	Feed the system with training samples and find the best parameters in the process in searching for the (global) minimum of the cost function

4.1.1. Visualizing surfaces

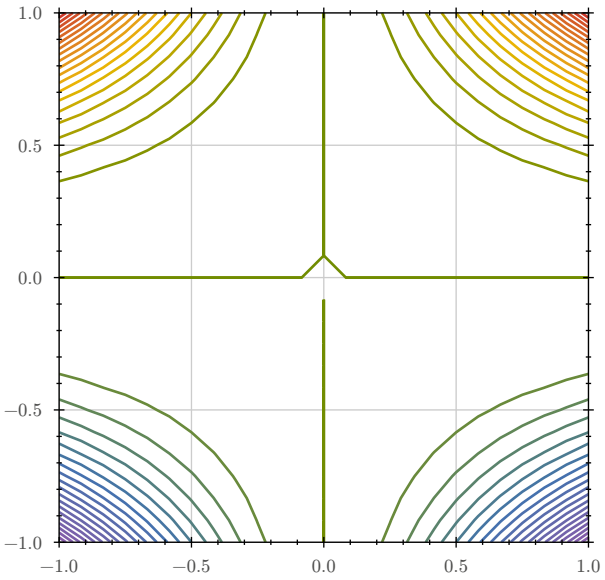
To a large extend, the world we live in can be envisioned as a two dimensional surface that is embedded in three dimensional space. Using this analogy you can take away two important methods for visualizing surfaces: surface-plots (3d) and contour-plots (2d).

$$f(x,y) = x^2y^3$$

Surface-plot



Contour-plot



4.1.2. Tangent space and linearisation of a surface

***Definition** 17:* **Set of all tangent directions**

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a surface and $p = (x_0, y_0, f(x_0, y_0)) \in \operatorname{graph} f$. Then the set of all tangent directions to curves $c(t)$ with $c(t) = p$ that reside entirely inside graph f is given by

$$T = \left\{ \alpha \begin{pmatrix} 1 \\ 0 \\ \partial_x f(x_0, y_0) \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \\ \partial_y f(x_0, y_0) \end{pmatrix} \;\middle|\; \alpha, \beta \in \mathbb{R} \right\}$$

***Definition** 18:* **Tangent space**

The set of all points

$$T = \left\{ \begin{pmatrix} x_0 \\ y_0 \\ f(x_0, y_0) \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ 0 \\ \partial_x f(x_0, y_0) \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \\ \partial_y f(x_0, y_0) \end{pmatrix} \;\middle|\; \alpha, \beta \in \mathbb{R} \right\}$$

that can be reached from p by traveling into a tangent direction, is a 2-dimensional plane, which is called ***tangent space*** of f at p.

***Definition** 19:* **Graph of a linearisation**

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real valued function in n variables and $x_0 \in \mathbb{R}^n$. Then the graph of the linearisation

$$l(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0)$$

is identical to the tangent space T_{p_0} of f at $p_0 = \left(((x_0))_1, ((x_0))_2, ..., ((x_0))_n, f(x_0)\right)^\top$.

4.2. OPTIMIZATION PROBLEMS

4.2.1. Extreme values and stationary points

Let $f : D \to \mathbb{R}$ with $D \subset \mathbb{R}^n$ be a hyper-surface.

Term	***Definition***
Upper bound	Any number $u \in \mathbb{R}$, such that $f(x) \leq u$ for all $x \in D$ is called an ***upper bound*** of f
Lower bound	Any number $l \in \mathbb{R}$, such that $f(x) \geq l$ for all $x \in D$ is called a ***lower bound*** of f
Global maximum	An upper bound u is called ***global maximum value*** or ***absolute maximum value*** of f, if $u \in f(D)$. The value x_0 is called the ***global maximum point*** or ***absolute maximum point***
Global minimum	A lower bound l is called ***global minimum value*** or ***absolute minimum value*** of f, if $l \in f(D)$. The value x_0 is called the ***global minimum point*** or ***absolute minimum point***
Local maximum	Any point $x_0 \in D$ such that, whenever $x \approx x_0 \wedge f(x) \leq f(x_0)$ is called ***local maximum point*** and the value $f(x_0)$ is called ***local maximum value***
Local minimum	Any point $x_0 \in D$ such that, whenever $x \approx x_0 \wedge f(x) \geq f(x_0)$ is called ***local minimum point*** and the value $f(x_0)$ is called ***local minimum value***

***Definition** 20:* **Stationary**

Let $f : D \to \mathbb{R}$ with $D \subset \mathbb{R}^n$ be a hyper-surface that can be differentiated. Any point $x_0 \in D$ with $\nabla f(x_0) = 0$ is called ***stationary***.

***Observations** 3:*

If x_0 is a local extremal point, then f is stationary at x_0, i.e. $\nabla f(x_0) = 0$.

Stationary points can also identify positions, at which a function is becoming flat into one direction without loosing its monotony.

Finally, there is a third scenario, which can occur: ***saddle points***. Saddle points require functions that depend on at least two variables, because the domain of such functions is sufficiently large to move into at least two different directions without leaving the graph. In such a case it can happen, that a stationary point appears to be as a maximum value in one direction, and as a minimum value in another direction.

4.2.2. Necessary and sufficient conditions for extremal values

Notation for second derivative:

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}(f(x,y))\right) = \frac{\partial^2 f(x,y)}{\partial x^2} = \partial_{xx} f(x,y)$$

Definition 21: **Hessian matrix**

Let $f : D \to \mathbb{R}$ with $D \subset \mathbb{R}^n$ be a hyper-surface. The matrix $H(x)$, whose components are

$$(H(x))_{ij} = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x)$$

is called ***Hessian matrix***. The Hessian matrix is the generalization of the second order derivation to hyper-surfaces.

Let $z = f(c(t))$, c a curve passing our stationary point $c_0 = (x_0, y_0)$ at $t = t_0$ and

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} = \underbrace{(\dot{c}_1(t_0), \dot{c}_2(t_0))}_{v^\top} \cdot \underbrace{\begin{pmatrix} f_{xx}(c_0) & f_{xy}(c_0) \\ f_{yx}(c_0) & f_{yy}(c_0) \end{pmatrix}}_{H} \cdot \underbrace{\begin{pmatrix} \dot{c}_1(t_0) \\ \dot{c}_2(t_0) \end{pmatrix}}_{v}$$

then

- If $\forall v \in \mathbb{R}^2 \setminus \{\vec{0}\}(v^\top H v > 0)$, then ${}^{\mathrm{d}^2 z}/_{\mathrm{d}t^2} > 0$ defines a ***local minimum*** at c_0 and H is called ***positive definite***
- If $\forall v \in \mathbb{R}^2 \setminus \{\vec{0}\}(v^\top H v \geq 0)$, then ${}^{\mathrm{d}^2 z}/_{\mathrm{d}t^2} \geq 0$ defines either a ***local minimum*** or ***non-extremal*** at c_0 and H is called ***positive semi-definite***
- If $\forall v \in \mathbb{R}^2 \setminus \{\vec{0}\}(v^\top H v < 0)$, then ${}^{\mathrm{d}^2 z}/_{\mathrm{d}t^2} < 0$ defines a ***local maximum*** at c_0 and H is called ***negative definite***
- If $\forall v \in \mathbb{R}^2 \setminus \{\vec{0}\}(v^\top H v \leq 0)$, then ${}^{\mathrm{d}^2 z}/_{\mathrm{d}t^2} \leq 0$ defines either a ***local maximum*** or ***non-extremal*** at c_0 and H is called ***negative semi-definite***
- In both of those cases, i.e. $\forall v \in \mathbb{R}^2 \setminus \{\vec{0}\}(v^\top H v < 0 \vee v^\top H v > 0)$, the point c_0 is a ***saddle point*** and H is called ***indefinite***
- In all other cases it defines nothing and further investigation is needed

***Observations** 4:*

1. The Hessian matrix is always symmetric, i.e. that $\partial_i \partial_j f = \partial_j \partial_i f$
2. Eigenvalues of Hessian matrices will thus always be $\in \mathbb{R}$

4.2.3. Analyzing the Hessian Matrix

Let $M \in \mathbb{R}^{n \times n}$ be a quadratic matrix and

$$q_M(v) = v^\top M v$$

its associated quadratic form. Then the quadratic matrix

$$H = \frac{1}{2}(M + M^\top)$$

is symmetric and satisfies

$$\forall v \in \mathbb{R}^n (q_M(v) = q_H(v))$$

***Definition** 22:* **Monomial**

Let $\alpha \in \mathbb{R}$ and $k_1, k_2, ..., k_n \in \mathbb{N}$. The real valued function

$$f : \begin{cases} \mathbb{R}^n & \to \mathbb{R} \\ (x_1, x_2, ..., x_n) & \mapsto \alpha x_1^{k_1} \cdot x_2^{k_2} \cdot ... \cdot x_n^{k_n} \end{cases}$$

is called ***monomial in** n **variables***. The value α is called the ***coefficient of the monomial***. If $\alpha \neq 0$ the number $k = k_1 + k_2 + ... + k_n$ is called the ***degree of the monomial***. Note that this definition implies that a constant function $f(x) = \alpha \neq 0$ is a monomial of degree zero. As $f(x)$ is also constant in the case $\alpha = 0$, we will also refer to $f(x) = 0$ as a monomial of degree zero, although there are books that refer to $f(x) = 0$ as monomial of degree -1.

***Definition** 23:* **Polynomial of several variables**

A real valued function $p : \mathbb{R}^n \to \mathbb{R}$ that can be written as a sum of finitely many monomials $m_1(x), m_2(x), ...m_r(x)$

$$p(x) = \sum_{l=1}^{r} m_l(x)$$

is called a ***polynomial of several variables***. p is said to be a ***polynomial of degree** k*, if at least one of the monomials has degree k, but none of the monomials has a degree higher than k.

Further we say that p is a ***homogeneous polynomial of degree** k*, if all monomials in the definition of p are of identical degree k, and we refer to a polynomial of degree 0 as ***constant function***, a polynomial of degree 1 as ***linear function***, and a polynomial of degree 2 as ***quadratic function*** or ***quadratic polynomial***.

***Observations** 5:*

1. We can see that any quadratic form is a sum of monomials of degree 2, and hence a homogeneous polynomial of degree 2.

$$\begin{aligned} q_H(v) &= v^\top H v \\ &= \sum_{i=1}^{n} \left(v_i \sum_{j=1}^{n} H_{ij} v_j \right) \\ &= \sum_{i,j=1}^{n} H_{ij} v_i v_j \end{aligned}$$

2. Conversely if $q : \mathbb{R}^n \to \mathbb{R}$ is a homogeneous polynomial of degree 2, it is a sum of monomials of degree 2. As any monomial is of the form $\alpha v_i v_j$, we can simply identify H_{ij} with the factor α in front of the monomial $\alpha v_i v_j$.

$$(v_1, v_2, v_3) \cdot \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

$$
\begin{aligned}
&= (v_1, v_2, v_3) \begin{pmatrix} v_1 h_{11} + v_2 h_{12} + v_3 h_{13} \\ v_1 h_{21} + v_2 h_{22} + v_3 h_{23} \\ v_1 h_{31} + v_2 h_{32} + v_3 h_{33} \end{pmatrix} \\
&= v_1^2 h_{11} + v_1 v_2 h_{12} + v_1 v_3 h_{13} \\
&\quad + v_1 v_2 h_{21} + v_2^2 h_{22} + v_3 v_2 h_{23} \\
&\quad + v_1 v_3 h_{31} + v_2 v_3 h_{32} + v_3^2 h_{33} \\
&= v_1^2 h_{11} + v_2^2 h_{22} + v_3^2 h_{33} + 2 v_1 v_2 h_{12} + 2 v_2 v_3 h_{23} + 2 v_1 v_3 h_{13}
\end{aligned}
$$

4.2.3.1. Technique of completing squares

$$
\begin{aligned}
&\underbrace{x^2}_{\sqrt{\cdot}} - \underbrace{4xy}_{\div 2x} + 1 \\
&= (x - 2y)^2 - (-2y)^2 + 1 \\
&= (x - 2y)^2 - 4y^2 + 1
\end{aligned}
$$

Special case: A polynomial that does not contain any square factors at all, i.e. with a polynomial such as

$$n = xy + 4xz + 2yz$$

it is guaranteed to be indefinite.

Definition 24: **LDL decomposition for positive definite and negative definite matrices**

Let $H \in \mathbb{R}^{n \times n}$ be positive definite or negative definite. Then there exists a diagonal matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}$$

and a lower unit triangular matrix

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \dots & * & 1 \end{pmatrix}$$

such that

$$H + L \Lambda L^\top$$

In the formula of L, the asterisk $*$ represents an arbitrary number. Furthermore, H is positive definite if and only if $\lambda_1, \lambda_2, ..., \lambda_n > 0$, and negative definite, if and only if $\lambda_1, \lambda_2, ..., \lambda_n < 0$.

4.3. APPROXIMATE SEARCH FOR EXTREMAL VALUES

4.3.1. Gradient and contour surfaces

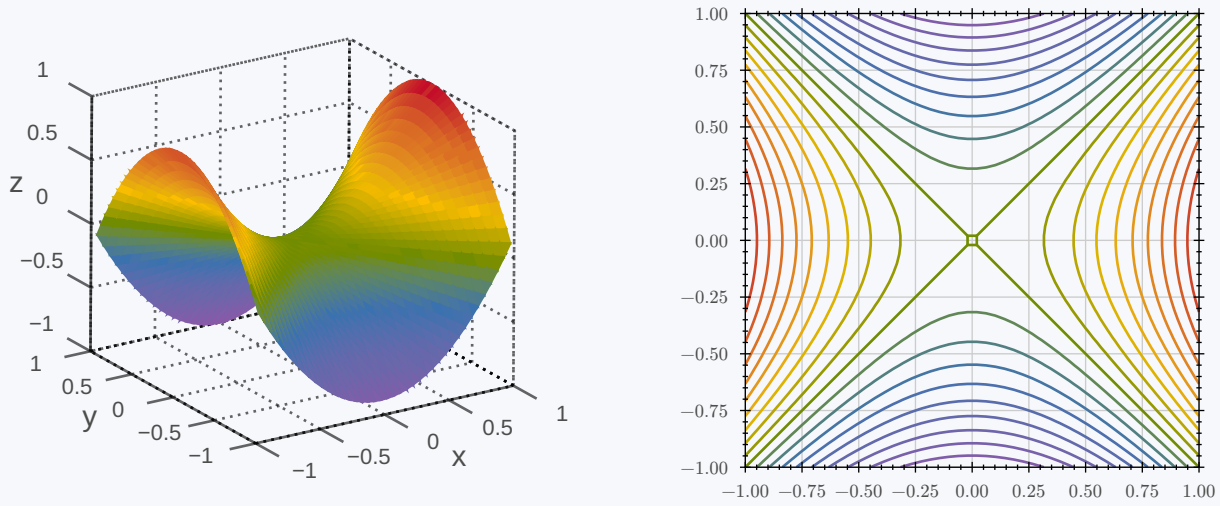
The set of all points of a hyper-surface $f : \mathbb{R}^n \to \mathbb{R}$ for which f results in the same value

$$f^{-1}(c) = \{x \in \mathbb{R}^n \mid f(x) = c\}$$

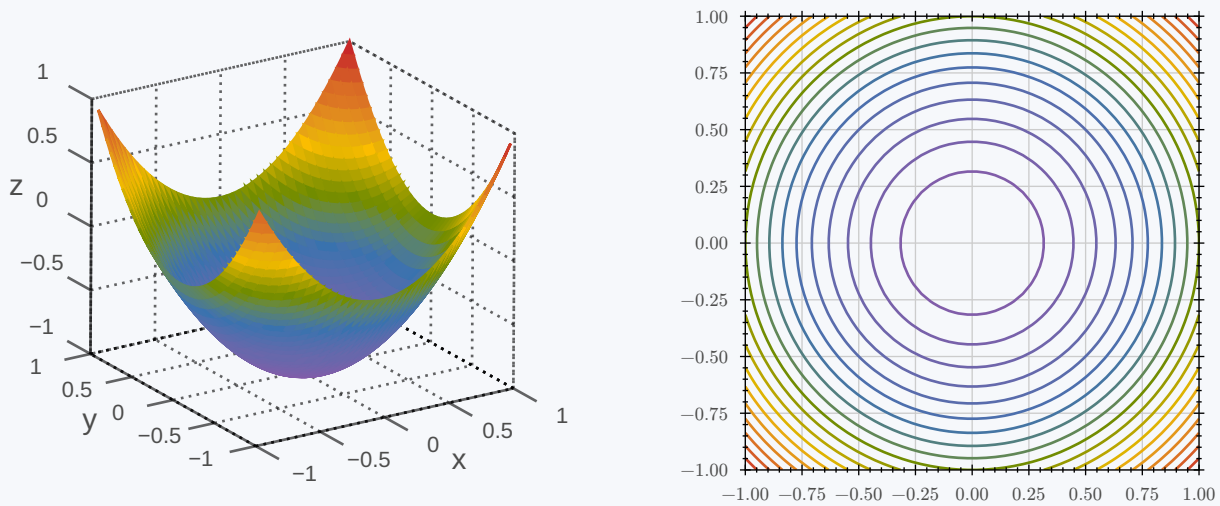
is called ***contour surface***, or a ***contour line***, if $n = 2$.

Let $s : \mathbb{R} \to \mathbb{R}^n$ be a curve, that resides entirely within $f^{-1}(c)$. In this case the real valued function $h(t) = f(s(t))$ is constant, and thus its derivative $h'(t) = 0$ is zero.

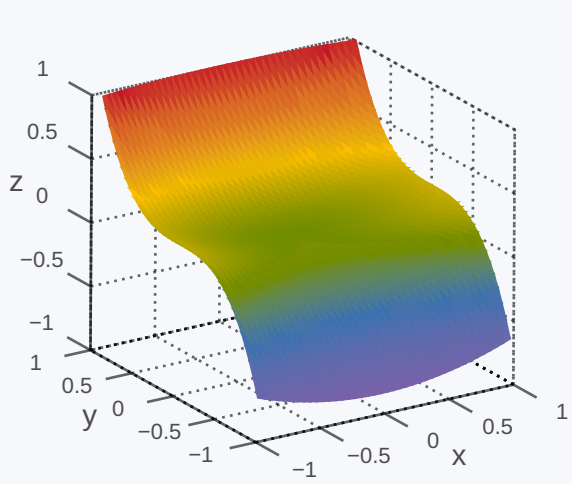
Example *12:* $f(x, y) = x^2 - y^2$

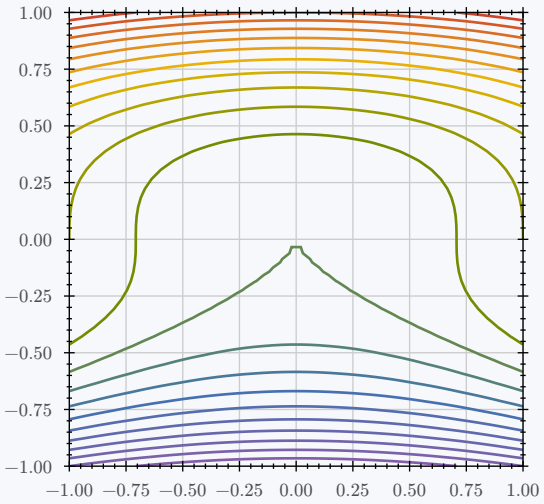


Example *13:* $f(x, y) = x^2 + y^2 - 1$

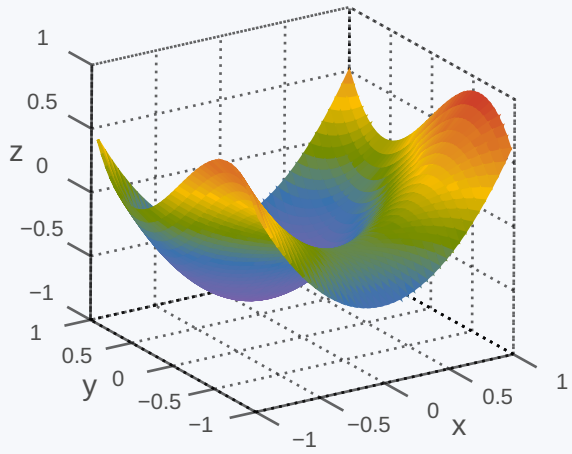


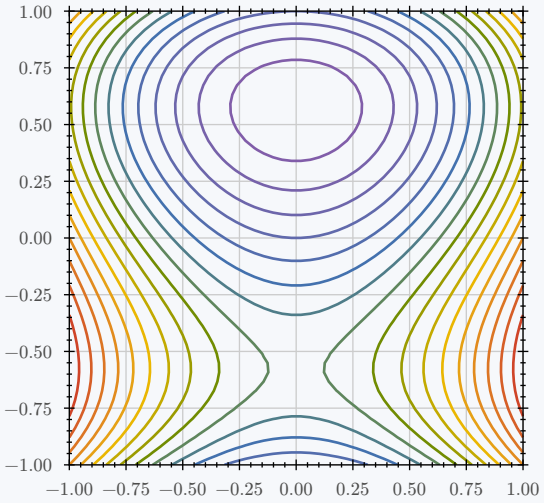
Example 14: $f(x,y) = \frac{1}{5}x^2 + y^3$





Example 15: $f(x,y) = y^3 - y + x^2 - \frac{1}{2}$





$$\gamma : \mathbb{R} \to \mathbb{R}^2, \gamma \subset \mathbb{D}_f = \text{One of the curves on the contour plot}$$

$$\begin{aligned} f(\gamma(t)) &= c \\ &\Rightarrow \nabla f(\gamma(t)) \cdot \gamma'(t) = 0 \\ &= (\nabla f(\gamma(t)))^\top \bullet \gamma'(t) = 0 \\ &\Rightarrow (\nabla f(\gamma(t)))^\top \perp \gamma'(t) \end{aligned}$$

The chain rule for surfaces and curves tells us that

$$h'(t) = \frac{\mathrm{d}}{\mathrm{d}t} f(s(t)) = \nabla f(s(t)) \cdot s'(t)$$

We can thus conclude

$$\nabla f(s(t)) \cdot s'(t) = 0$$

which tells us, that the gradient of f is orthogonal to any curve that completely resides within a contour surface, and thus orthogonal to the contour surface, itself.

$a \bullet b = |a| \cdot |b| \cdot \cos \angle(a, b)$

The formula for the growth of f at x_0

$$\frac{g'(0)}{|v|} = \left|\nabla f(x_0)^\top\right| \cdot \cos \angle\left(\nabla f(x_0)^\top, v\right)$$

tells us:
– In the vicinity of x_0, the values of f grow most rapidly in that direction v, in which $\cos \angle\left(\nabla f(x_0)^\top, v\right)$ is maximal, i.e. in a direction that is parallel to $\nabla f(x_0)^\top$
– In this direction, the growth of f is proportional to ${}^{g'(0)}\!/_{|v|} = \left|\nabla f(x_0)^\top\right|$

4.3.2. Minimizing cost functions using gradient descent

Goal: $\nabla f(x) \approx 0$

Iterative approach:

$$x_{i+1} = x_i - \gamma \cdot \nabla f(x_i)$$

where $\gamma > 0$ is the ***step size*** or ***learning rate*** of gradient descent. Large values of γ allow us to move quickly at the price that we may jump over local minimum points with out notice.

We also need a ***termination criterion*** through which we control the precision $\varepsilon > 0$ of the approximation:

$$\left|x_{i+1} - x_i\right| < \varepsilon$$

or equivalently

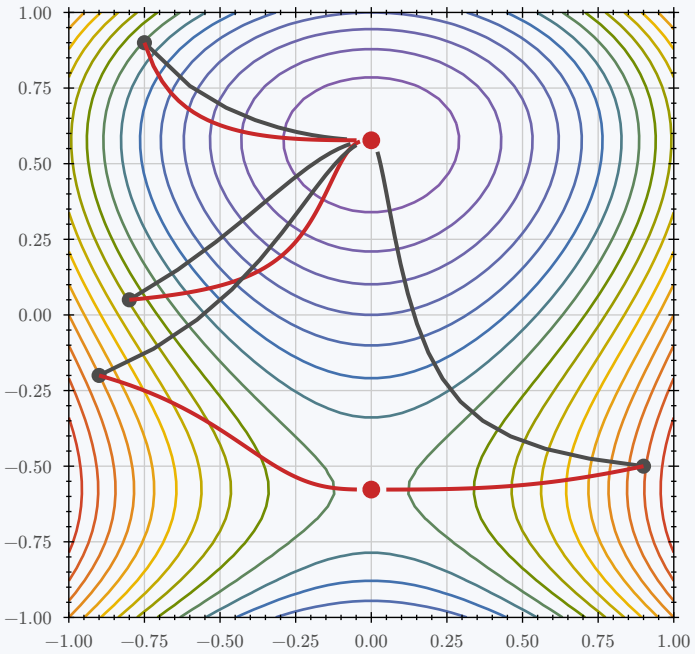
$$|\nabla f(x_i)| < \frac{\varepsilon}{\gamma}$$

leading to gradient descent terminating at a point where $\nabla f(x) \approx 0$. The algorithm is guaranteed to stop at a stationary point, but there is **no guarantee that the stationary point is in fact a local minimum**.

Comparison to Newton's method:

$$x_{i+1} = x_i - \gamma \cdot \underbrace{\left(H_{f(x_i)}\right)^{-1}}_{\substack{\text{Too expensive} \\ \text{in ML!}}} \cdot \nabla f(x_i)$$

***Example** 16:*



$$f(x,y) = y^3 - y + x^2 - \frac{1}{2}$$

Black lines starting from the black points lead to the local minimum, calculated using gradient descent. Red lines are calculated using Newton's method.

The parameters chosen are as follows:

Newton's method: $\gamma = 0.01, \varepsilon = 0.001$
Gradient descent: $\gamma = 0.1, \varepsilon = 0.01$

Procedures for finding good values for ε, γ can include:
– Conjugate gradient method
– Stochastic gradient descent

4.4. PUTTING IT ALL TOGETHER (IMAGE CLASSIFICATION)

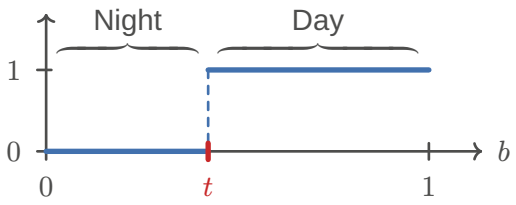
A 128×256 pixel image I with 3 color channels (RGB) can be represented as a $128 \times 256 \times 3$ matrix (tensor of rank 3). The red value of a pixel at row 100, column 12 can be indexed with $I_{100,12,1}$. A function to determine whether the image was taken at day (0) or night (1) would have the following signature:

$$f : \mathbb{R}^{128 \times 256 \times 3} \to \{0, 1\}$$

4.4.1. Feature extraction

Term	*Definition*
Feature	functions encapsulating domain knowledge
Training set	Data samples for which the output category is already known
Supervised learning	Answering how features can be used for predictions by means of a training process usng a training set

There are many different approaches to solving the image classification problem. We could calculate the average brightness b of an image, and then argue, that the image has been taken during the day if b exceeds some threshold t.

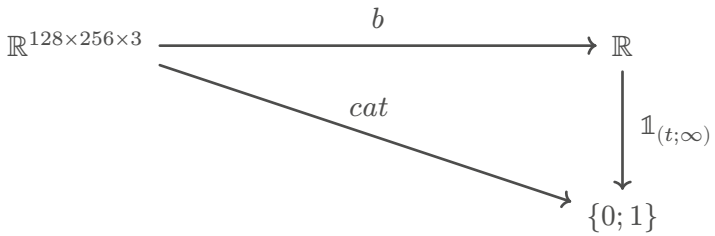


Definition 25: **Indicator function**

Let $A \subset \mathbb{R}^n$ be a set. The ***indicator function*** $\mathbb{1}_A(x)$ is defined as

$$\mathbb{1}_A : \begin{cases} \mathbb{R}^n \to \{0,1\} \\ x \quad \mapsto \begin{cases} 1 \ , \ x \in A \\ 0 \ , \ x \notin A \end{cases} \end{cases}$$

Using indicator functions we can split the categorization function into a ***feature-function*** $b(x)$ which calculates the average brightness and an ***indicator function*** $\mathbb{1}_{(t;\infty)}$ which maps high brightness values to category 1:

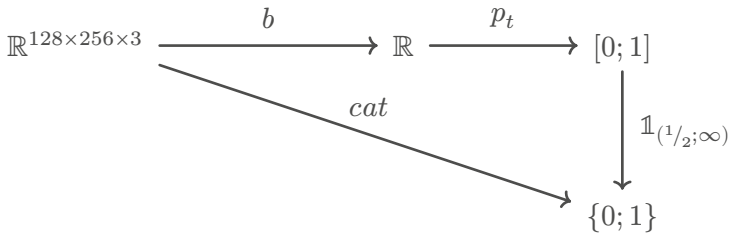


Thus $b(x)$ encapsulates **domain knowledge** and $\mathbb{1}_{(t;\infty)}$ parameterizes **ignorance**, i.e. that we do not yet know a brightness threshold t.

It is the responsibility of the training phase, to find a suitable value of t, such that the system works well for the data provided in the training set. Once we have found such a value and the system also works for other data sets, we say that the system is able to ***generalize***.

4.4.2. Logistic regression

Logistic regression introduces a further layer $p_t(b)$, which calculates the **probability** that an image with a particular feature falls into a specific category. We can still associate the image to the most likely category using the indicator function $\mathbb{1}_{(1/2;\infty)}$. However, in this case, we have the additional information that we are somewhat uncertain about this choice.



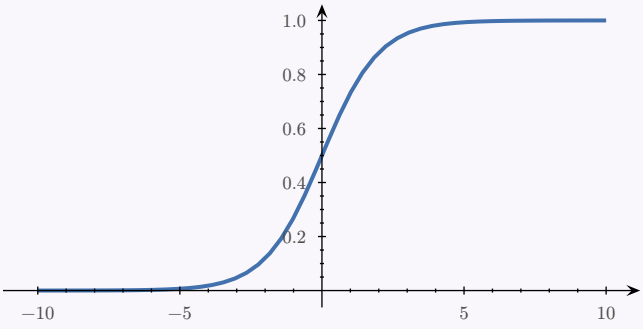
Here, b and $\mathbb{1}_{(1/2;\infty)}$ are deterministic. There is, however, flexibility to later-on adapt the indicator function to a specific use-case, e.g. it might be useful to have a preference for putting objects into category “DAY” by using an indicator function such as $\mathbb{1}_{(1/4;\infty)}$.

In logistic regression the task of identifying a probability distribution $p_t(b)$ that serves our needs is always solved with the help of the sigmoid function:

***Definition** 26:* **Sigmoid function**

The ***sigmoid function*** is defined as follows:

$$\sigma : \begin{cases} \mathbb{R} \to (0;1) \\ t \mapsto \frac{1}{1+e^{-t}} \end{cases}$$



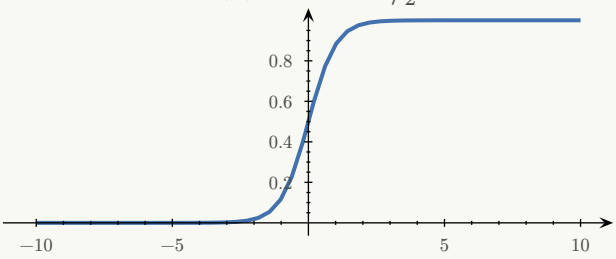
If interpreted as probability distribution, the sigmoid function associates negative values to low probabilities (< 0.5) and positive values to high probabilities (> 0.5), because $\sigma(0) = {}^1/_2$ Therefore the number zero is associated with a probability of 0.5.

***Observations** 6:*

1. The choice of a particular indicator function does not impact the choice of the brightness threshold that is required to calculate p_t
2. Although in logistic regression the sigmoid function is almost always used, other functions exist, as long as they map $\mathbb{R} \to (0;1)$. Another example would be:

$$\tanh : \begin{cases} \mathbb{R} \to (0;1) \\ t \mapsto \frac{e^t - e^{-t}}{e^t + e^{-t}} \end{cases}$$

$f(x) = {}^{\tanh(x)+1}/_2$

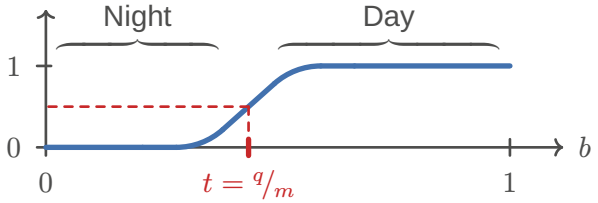


In logistic regression, the sigmoid function is applied to a linear combination of the features, which is eventually adjusted by adding a constant value (known as the ***bias***).

$$p_{m,q}(b) = \sigma(mb - q)$$

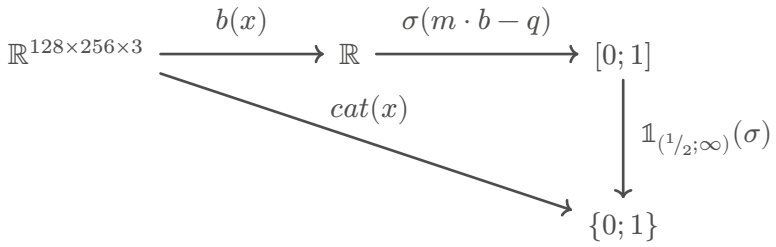
The ***weight*** m controls how strongly b influences the classification; larger $|m|$ makes the sigmoid transition steeper with respect to b and indicates that brightness is a good concept for distinguishing the images. The ***bias correction term*** q shifts the sigmoid left/right along the b axis and thus changes the threshold boundary.

The task of training logistic regression thus boils down to the challenge of finding suitable values for m and b, such that p_m, p is compatible with the training data set.



4.4.3. The maximum likelihood principle

Using logistic regression we now have the following:



where
$b(x) \to$ Average brightness
$\sigma(m \cdot b - q) \to$ Probability to fall into either category
$\mathbb{1}_{(1/2;\infty)}(\sigma) \to$ Map probability to 0 or 1

What still has to be done is to find suitable values for m and q. The likelihood for them being properly chosen is

$$\begin{aligned} &\text{proportional to } p_{m,q}(b(x)) && \text{if } x \text{ falls into category DAY} \\ &\text{proportional to } 1-p_{m,q}(b(x)) && \text{if } x \text{ falls into category NIGHT} \end{aligned}$$

With D and N being the set of images pre-classified as DAY and NIGHT respectively, we can argue that the ***likelihood function***

$$l(m,q) = \underbrace{\left(\prod_{x\in D} p_{m,q}(b(x))\right)}_{\text{Probability of all DAY pictures}} \cdot \underbrace{\left(\prod_{x\in N}(1-p_{m,q}(b(x)))\right)}_{\text{Probability of all NIGHT pictures}}$$

measures the combined likelihood that m and q work well for the entire training set. The ***maximum likelihood principle*** states the best choice of m and q maximizes the likelihood function l.

Instead of calculating J_l, which requires the product rule, we can seek to maximize the ***log-likelihood function*** which poses less of a challenge:

$$\begin{aligned} \ln(l(m,q)) &= \sum_{x\in D}\ln\big(p_{m,q}(b(x))\big) + \sum_{x\in N}\ln\big(1-p_{m,q}(b(x))\big) \\ \frac{\partial}{\partial m}\ln(l(m,q)) &= \sum_{x\in D}\frac{\frac{\partial}{\partial m}\big(p_{m,q}(b(x))\big)}{p_{m,q}(b(x))} - \sum_{x\in N}\frac{\frac{\partial}{\partial m}\big(p_{m,q}(b(x))\big)}{1-p_{m,q}(b(x))} \\ \frac{\partial}{\partial q}\ln(l(m,q)) &= \sum_{x\in D}\frac{\frac{\partial}{\partial q}\big(p_{m,q}(b(x))\big)}{p_{m,q}(b(x))} - \sum_{x\in N}\frac{\frac{\partial}{\partial q}\big(p_{m,q}(b(x))\big)}{1-p_{m,q}(b(x))} \end{aligned}$$

This is possible, because the logarithm is ***strictly monotonic***, and thus preserves the location of maximum and minimum points.

4.4.4. Cost function for logistic regression

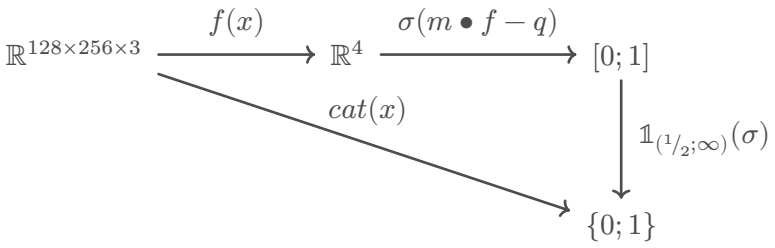
As in machine learning we typically seek to minimize cost, we therefore use the ***negative-log-likelihood*** function as cost function for logistic regression:

$$c(m,q) = -\ln(l(m,q)) = -\left(\sum_{x\in D}\ln\big(p_{m,q}(b(x))\big) + \sum_{x\in N}\ln\big(1-p_{m,q}(b(x))\big)\right)$$

4.4.5. Enhancing the classifier

We can add more ***features***, eg. the average brightness of the individual r, g, b color channels. WARNING: On small amounts of training samples this may result in **overfitting**.

$$f: \begin{cases} \mathbb{R}^{128\times256\times3} \to \mathbb{R}^3 \\ x \mapsto \begin{pmatrix} r(x) \\ g(x) \\ b(x) \end{pmatrix} \end{cases}$$



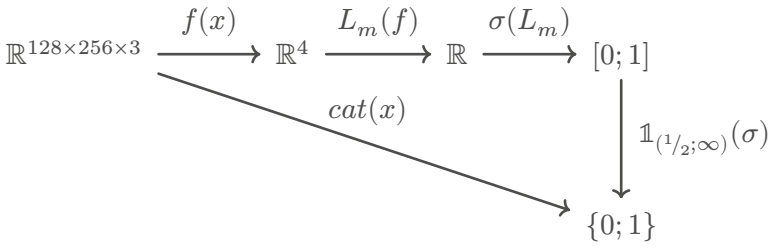
As a consequence, we now have to find suitable values for all components of m and the bias correction term q. By introducing a fourth "constant feature", we can further unify the treatment of bias and features. By using the feature vector

$$f: \begin{cases} \mathbb{R}^{128\times256\times3} \to \mathbb{R}^4 \\ x \mapsto \begin{pmatrix} r(x) \\ g(x) \\ b(x) \\ -1 \end{pmatrix} \end{cases}$$

and using $m = (m_r, m_g, m_b, m_{\text{bias}})$ for the parameters as well as introducing the linear map

$$L_m: \begin{cases} \mathbb{R}^4 \to \mathbb{R} \\ f \mapsto m \cdot f \end{cases}$$

we can further expand the structure of the logistic regression algorithm:



This is the final form of any multi-feature regression algorithm:

1) Extract n features from an input data point
2) Choose one parameter per feature and apply L_m to the feature vector.
3) Apply sigmoid function to obtain probability of the category that input belongs to
4) Use indicator function to associate input with a category

To find the suitable values for m through the feature vectors $f_1, ..., f_N$ that are already mapped to the categories $c_1, ..., c_N$ we can finally minimize

$$\text{cost}(m) = -\left(\sum_{r\in\{i|c_i=1\}} \ln(\sigma(m \cdot f_r))\right) - \left(\sum_{s\in\{i|c_i=1\}} \ln(1-\sigma(m \cdot f_s))\right)$$

5. MULTIVARIATE GAUSSIAN DISTRIBUTION

5.1. PROBABILITY THEORY

Term	***Definition***
Experiment	Procedure thet terminates with a non-deterministic but well defined outcome
Sample space	Set of all possible outcomes $\Omega = \{e_1, e_2, ...\}$
Event	$E \subset \Omega$
Set of all events	$\mathcal{F} = \{E \mid E \subset \Omega\} = P(\Omega)$
Probability density	
Probability measure	A function $$\mathbb{P} : P(\Omega) \to [0; 1]$$ such that the following conditions hold $$\mathbb{P}(\Omega) = 1$$ $$\mathbb{P}(\varnothing) = 0$$ $$\forall A, B \subset \Omega,\ A \cap B = \varnothing\ (\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B))$$ $$\forall A, B \subset \Omega\ (\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B))$$ Its value can be interpreted as probability for the event E to occur when conducting the experiment.

5.1.1. Discrete probability distribution

$$\Omega = \{e_1, e_2, ..., e_n\}, \quad \mathbb{P}(\{e_k\}) = p_k, \quad k \in \{1, 2, ..., n\}$$

Let Ω be an **enumerable** sample space, then $p : \Omega \to [0; 1]$ is called ***discrete probability distribution***, if

$$\sum_{i=1}^{n} p(e_i) = 1$$

Any discrete probability distribution uniquely identifies a probability measure, i.e. the probability measure that maps the event $S \subset \Omega$ to the probability

$$\mathbb{P}(E) = \sum_{e \in E} p(e)$$

5.1.2. Univariate probability distribution

The ***uniform probability distribution*** $\mathrm{unif}(0, 1)$ characterizes an experiment in which one number is chosen from the sample space $\Omega = [0; 1]$ in such a way, that all numbers of Ω have an equal chance to occur.

Whenever $\Omega \subset \mathbb{R}$, we can use $\mathbb{P}$ to define the ***cumulative distribution function*** (CDF)

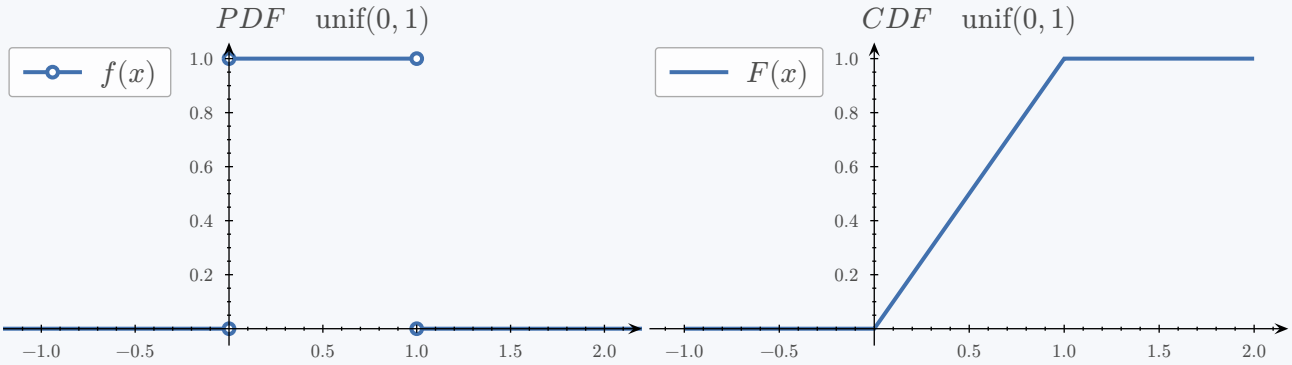
$$F(\alpha) = \mathbb{P}((-\infty; \alpha] \cap \Omega)$$
$$\lim_{\alpha \to -\infty} F(\alpha) = 0$$
$$\lim_{\alpha \to \infty} F(\alpha) = 1$$

Example 17: **Cumulative distribution function F and probability density function f for** $\mathrm{unif}(a,b)$

$$F : \begin{cases} \mathbb{R} \to [0;1] \\ x \mapsto \mathbb{1}_{[a;b]}(x) \cdot \frac{x-a}{b-a} + \mathbb{1}_{(b;\infty)}(x) \end{cases}$$

$$F'(x) = f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{else} \end{cases} = \frac{\mathbb{1}_{(a;b)}(x)}{b-a} = \text{density for } X \sim \mathrm{unif}(a,b)$$

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{\mathbb{1}_{(a;b)}(t)}{b-a}\,\mathrm{d}t \overset{a<x<b}{=} \int_{a}^{x} \frac{1}{a-b}\,\mathrm{d}t = \frac{x-a}{b-a}$$



A function $f : \mathbb{R} \to \mathbb{R}^+$, such that

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1$$

is called ***probability density function*** (PDF). With any PDF we can associate a CDF

$$F(\alpha) = \int_{-\infty}^{\alpha} f(t)\,\mathrm{d}t$$

and a probability measure $\mathbb{P}$ that associates the probability of the event $E \subset \mathbb{R}$ to occur with the area of all points underneath the function $f(t)$ whose t-values reside in E:

$$\mathbb{P}(E) = \int_{t \in E} f(t)\,\mathrm{d}t$$

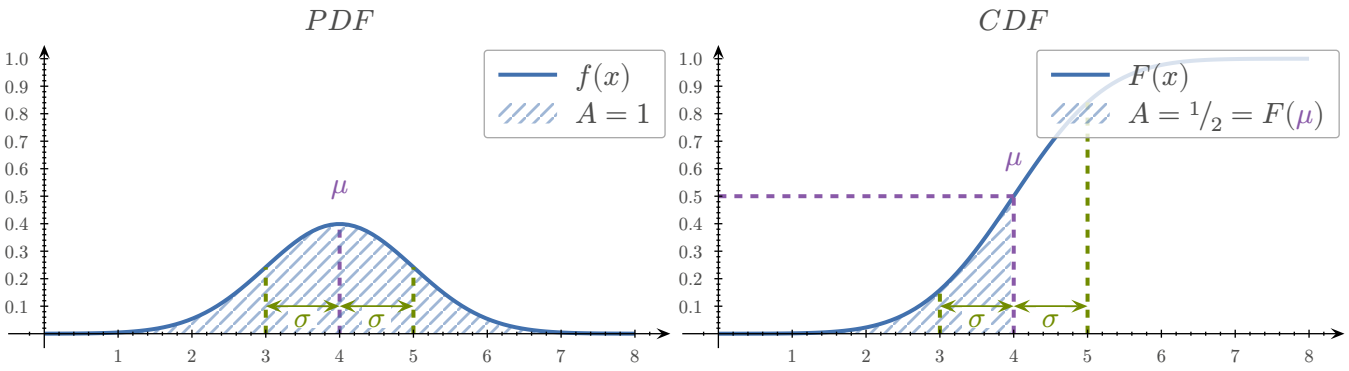
5.1.3. Univariate normal distribution

The probability density of the normal / Gaussian distribution is given by

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

in which μ and σ are parameters, denoted as mean value (μ) and standard deviation (σ).

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,\mathrm{d}t \approx 0.68$$



The value of the CDF evaluated at x represents the area of the PDF from $-\infty$ to x, thus giving the probability of an experiment be in $(-\infty; x]$.

All normal distributions can be defined with the help of the so called ***standard normal distribution***

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

for which $\mu = 0$ and $\sigma = 1$. The general normal distribution can then be expressed in terms of the standard normal distribution using the formula

$$f(x|\mu, \sigma) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

The ***error function*** is defined through

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}t$$

or its Taylor series

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(2k+1)} x^{2k+1}$$

and can be approximated through hyperbolic functions

$$\operatorname{erf}(x) \approx \tanh\left(\frac{2}{\sqrt{\pi}}\left(x + \frac{11}{123}x^3\right)\right)$$

Given this function, one can show, that the CDF of the normal distribution is given by

$$F(x|\mu, \sigma) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right)$$

5.2. ESTIMATION OF PROBABILITY MEASURES

5.2.1. Maximum likelihood principle for normal distributed data

***Example** 18:* **Likelihood of an american person having a certain height**

Using the data from the US National Health and Nutrition Examination Survey (NHANES, 1999-2004) providing us with heights of 3602 adult male persons we can find the likelihood for person k having height x_k is

$$f(x_k \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_k-\mu)^2}$$

and therefore the likelihood for finding 3602 people of height $x_1, x_2, ..., x_{3602}$ is

$$p(\mu, \sigma) = \prod_{k=1}^{3602} f(x_k \mid \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{3602} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{3602}(x_k-\mu)^2}$$

If $x_1, x_2, ..., x_n$ are outcomes of n independent runs of an experiment hat is known to produce normal distributed results, we can estimate the parameters μ and σ^2 of the normal distribution with the help of the maximum likelihood principle:

$$\mu_{\text{MLE}} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2$$

MLE: Maximum Likelihood Estimators

It is possible to slightly improve maximum likelihood estimators through a procedure called ***bias correction***. For instance, the bias-corrected maximum likelihood estimator

$$\sigma^2_{\text{MLE}} = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \mu)^2$$

for the variance of normal distributed data is much more popular than the MLE-estimator, al- though both estimators do not differ much in practice. Much more important than bias correction are the following properties which are demanded by most statistical methods: namely, that subsequent experiments are ***statistically independent*** and ***identically distributed*** (the combination of those terms is often abbreviated with the acronym ***iid***).

5.3. VECTOR VALUED RANDOM VARIABLES

Experiments that have vector valued outcomes instead of having a discrete set as a sample space (eg. $\{\text{DAY}, \text{NIGHT}\}$).

5.3.1. Multiple measurements and statistical independence

In Example 18 , the height x is called a ***random variable***. Typically one uses capital letters to identify random variables and would thus write

$$X \sim \mathcal{N}(\mu = 175.85, \sigma = 7.49)$$

to indicate that the random variable X is normal distributed with mean $\mu = 175.84$ and standard deviation $\sigma = 7.49$. Once the random variable X has been introduced, we use the corresponding lower case letter (here x) as a placeholder for particular values of that random variable.

5.3.1.1. Composite measurements

The experiments of measuring the weight and height of an adult male are not ***statistically independent***: a tall person is usually heavier than a small person. Here the elements of sample space consist of pairs of real numbers, which implies that the associated random variable V is vector-valued: its first component represents the height X of the person and its second component represents his weight Y. Similar to real valued-random variables, we can associate probability densities to vector valued random variables.

However, since the sample space of V is $\mathbb{R}^2$ the probability density of V now needs to be a function of two variables $f_V : \mathbb{R}^2 \to \mathbb{R}$.

5.3.2. Random variables

A random variable is a mapping that maps a (randomly) chosen sample (here: a male adult), to some well defined properties of that sample (here: height or weight value of that person). Formally, random variables are therefore defined to be functions from the set of all possible samples Ω into some other space. The random variables X (=height), Y (=weight) and V (=both values) are therefore functions

$$X : \Omega \to \mathbb{R}$$
$$Y : \Omega \to \mathbb{R}$$
$$V : \Omega \to \mathbb{R}^2$$

We can now use a random variable such as X to define an associated probability measure $\mathbb{P}_X$, whose events are subsets $V \subset \mathbb{R}$ from the range of X, through

$$\mathbb{P}_X(V) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in V\})$$

As we often rely on such probability measures, we also abbreviate it by

$$\mathbb{P}_X(V) = \mathbb{P}(X(\omega) \in V)$$

or

$$\mathbb{P}_X(V) = \mathbb{P}(X^{-1}(V))$$

Definition 27: **Random variables**

Let Ω be a sample space with probability measure $\mathbb{P}$, and S be a set. Then we call any function

$$T : \Omega \to S$$

a ***random variable***. Specifically we call a function

$$X : \Omega \to \mathbb{R} \quad \text{real valued random variable}$$
$$\text{and } V : \Omega \to \mathbb{R}^n \text{ vector valued random variable}$$

For any random variable $T : \Omega \to S$ the probability measure $\mathbb{P}$ on Ω results in an associated probability measure $\mathbb{P}_T$ on S, which is given by

$$\forall Q \subset S \ (\mathbb{P}_T(Q) = \mathbb{P}(T(\omega) \in Q))$$

If $X : \Omega \to \mathbb{R}$ is a real valued random variable, we can also associate a CDF with X, which is given by

$$F_X(x) = \mathbb{P}(X(\omega) \leq x)$$

and in case that F_X is continuous a PDF $f_X(x)$, such that

$$\mathbb{P}(a < X(\omega) < x) = \int_a^b f_X(x)\,\mathrm{d}x$$

Definition 28: **Cumulative distribution function**

If $V : \Omega \to \mathbb{R}^n$ is a vector valued random variable, we define the ***cumulative distribution function*** of V to be

$$F_V(v) = \mathbb{P}(\forall i = 1, 2, ..., n\ (V_i(\omega) \leq v_i))$$

and if F_V is continuous, the probability density function

$$f_V(v) = \frac{\partial}{\partial v_1}\frac{\partial}{\partial v_2}...\frac{\partial}{\partial v_n}F_V(v)$$

***Definition** 29:* **Statistically independent**

Two real valued random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are ***statistically independent***, if the probability density of the random variable

$$V : \begin{cases} \Omega \to \mathbb{R}^2 \\ \omega \mapsto \begin{pmatrix} X(\omega) \\ Y(\omega) \end{pmatrix} \end{cases}$$

satisfies

$$f_V(x, y) = f_X(x) \cdot f_Y(y)$$

5.4. MULTIVARIATE NORMAL DISTRIBUTION

***Definition** 30:* **Multivariate normal distribution**

A vector valued random variable $V : \Omega \to \mathbb{R}^n$ is said to be distributed according to the ***multivariate normal distribution***

$$V \sim \mathcal{N}(\mu, \Sigma)$$

with mean μ and covariance matrix Σ, if the PDF of V is

$$f_V(v) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(v-\mu)^\top \Sigma^{-1}(v-\mu)}$$

***Observations** 7:*

1. $$q_{\Sigma^{-1}}(v - \mu) = -\frac{1}{2}(v - \mu)^\top \Sigma^{-1}(v - \mu)$$
2. Σ is always a positive definite symmetric matrix.

5.5. COORDINATE TRANSFORMATIONS

Cartesian coordinate system: $(O; e_1, e_2)$, where O is the origin and e_1, e_2 are orthogonal basis vectors of same length starting from O.

***Definition** 31:* **Coordinate map**

Let S be an arbitrary set and let U be a sufficiently large subset of $\mathbb{R}^n$. U is sufficiently large, if from any element $x \in U$ it is possible to move a bit into an arbitrary direction without leaving U. A bijective mapping $f : S \to U \subset \mathbb{R}^n$ is called a ***coordinate map*** of S or ***coordinate chart*** of S. The coordinate functions $f_1(s), f_2(s), ..., f_n(s)$ of $f(s)$ are also called the ***coordinates*** of f.

Further, if $U, V \subset \mathbb{R}^n$ are sufficiently large, $f : S \to U$ is a coordinate map and $t : U \to V$ is a bijection, then $g = t \circ f$ is a coordinate map, too, and t is called a ***coordinate transformation***.

As every coordinate map $f : S \to U$ associates elements of S 1-to-1 with elements of U, we can think of a coordinate map as a method of attaching (unique) labels to all elements of S. In the light of this, a coordinate transformation is essentially resulting in a different, but yet unique alternative labeling of the data.