Convex Optimization and Modeling

Introduction and a quick repetition of analysis/linear algebra

First lecture, 12.04.2010

Jun.-Prof. Matthias Hein

Organization of the lecture

Advanced course, 2+2 hours, 6 credit points

- **Exercises:**
  - time+location: Friday, 16-18, E2.4, SR 216
  - teaching assistant: Shyam Sundar Rangapuram
  - weekly exercises, theoretical and practical work,
  - practical exercises will be in Matlab (available in the CIP-pools),
  - 50% of the points in the exercises are needed to take part in the exams.
- **Exams:**
  - End-term: 28.7.
  - Re-exam: to be determined
- **Grading:** An exam is passed if you get at least 50% of the points. The grading is based on the best out of end-term and re-exam.

The course is based (too large extent) on the book

Boyd, Vandenberghe: Convex Optimization

The book is freely available: http://www.stanford.edu/ boyd/cvxbook/

Other material:

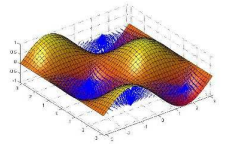
- Bertsekas: Nonlinear Programming
- Hiriart-Urruty, Lemarechal: Fundamentals of Convex Analysis.
- original papers.

For the exercises in Matlab we will use

CVX: Matlab Software for Disciplined Convex Programming

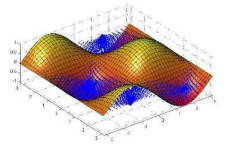
available at: http://www.stanford.edu/ boyd/cvx/ (Version 1.21).

Roadmap of the lecture



- Introduction/Motivation - Review of Linear Algebra
- `Theory:` Convex Analysis: Convex sets and functions
- `Theory:` Convex Optimization, Duality theory
- `Algorithms:` Unconstrained/equality-constrained Optimization
- `Algorithms:` Interior Point Methods, Alternatives
- `Applications:` Image Processing, Machine Learning, Statistics
- `Applications:` Convex relaxations of combinatorial problems

Roadmap for today



- Introduction - Motivation
- Reminder of Analysis
- Reminder of Linear Algebra
- Inner Product and Norms

What is Optimization ?

- we want to find the **best parameters** for a certain problem
  e.g. best investment, best function which fits the data, best tradeoff between fitting the data and having a smooth function (machine learning, image denoising)
- **parameters** underlie restrictions $\Rightarrow$ constraints.
  - total investment limited and positive,
  - images have to be positive, preservation of total intensity

Mathematical Optimization/Programming

$$\min_{x \in D} f(x),$$
$$\text{subject to: } g_i(x) \leq 0,\ i = 1, \ldots, r$$
$$h_j(x) = 0,\ j = 1, \ldots, s$$

- f is the **objective** or **cost** function.
- The domain D of the optimization problem:

$$D = \operatorname{dom} f \bigcap \cap_{i=1}^{r} \operatorname{dom} g_i \bigcap \cap_{j=1}^{s} \operatorname{dom} h_j.$$

- $x \in D$ is **feasible** if the inequality and equality constraints hold at x.
- the **optimal value** p^* of the optimization problem

$$p^* = \inf\{f(x) \,|\, g_i(x) \leq 0,\ i = 1, \ldots, r, \quad h_j(x) = 0,\ j = 1, \ldots, s \quad x \in D\}.$$

Linear Programming

The objective f and the constraints $g_1, \ldots, g_n$, $h_1, \ldots, h_m$ are all **linear**.

Example of Linear Programming: We want to fit a linear function, $\phi(x) = \langle w, x \rangle + b$, to a set of k data points $(x_i, y_i)_{i=1}^k$.

$$\arg\min_{w,b} \sum_{i=1}^{k} \big| \langle w, x_i \rangle + b - y_i \big|$$

This non-linear problem can be formulated as a linear program:

$$\min_{w \in \mathbb{R}^n,\, b,\, \gamma_1, \ldots, \gamma_k \in \mathbb{R}} \sum_{i=1}^{k} \gamma_i,$$

$$\text{subject to: } \langle w, x_i \rangle + b - y_i \leq \gamma_i, \quad i = 1, \ldots, k$$

$$-\big(\langle w, x_i \rangle + b - y_i \big) \leq \gamma_i, \quad i = 1, \ldots, k$$

Note that $\gamma_i \geq \max\big\{ \langle w, x_i \rangle + b - y_i, -\big(\langle w, x_i \rangle + b - y_i \big) \big\} = \big| \langle w, x_i \rangle + b - y_i \big|$.
In particular, at the optimum $\gamma_i = \big| \langle w, x_i \rangle + b - y_i \big|$.

Convex Optimization

The objective f and the inequality constraints $g_1, \ldots, g_n$ are **convex**.

The equality constraints $h_1, \ldots, h_m$ are **linear**.

Distance between convex hulls - The hard margin Support Vector Machine (SVM)

We want to separate two classes of points $(x_i, y_i)_{i=1}^k$, where $y_i = 1$ or $y_i = -1$ with a hyperplane such that the hyperplane has maximal distance to the classes.

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \|w\|^2$$

$$\text{subject to: } y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall\, i = 1, \ldots, k$$

This problem has only a feasible solution if the two classes are separable.

Visualization of the SVM solution

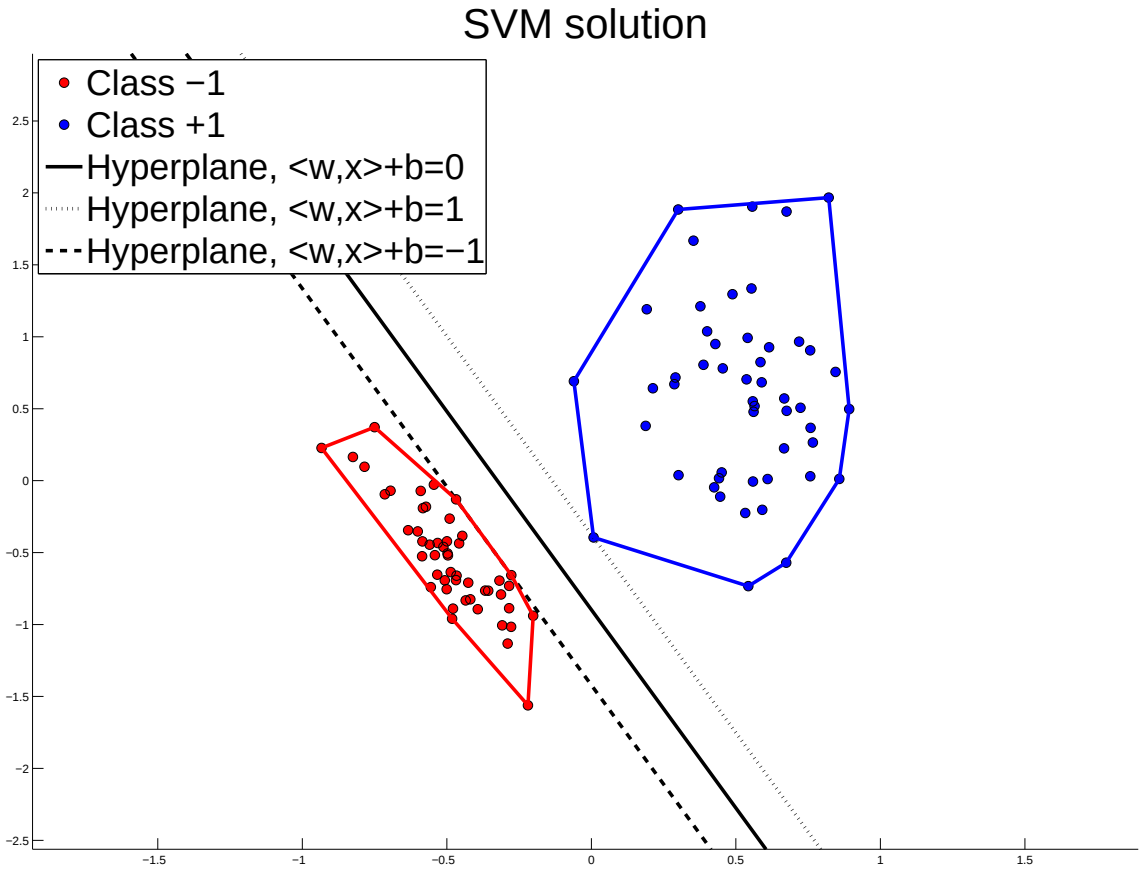
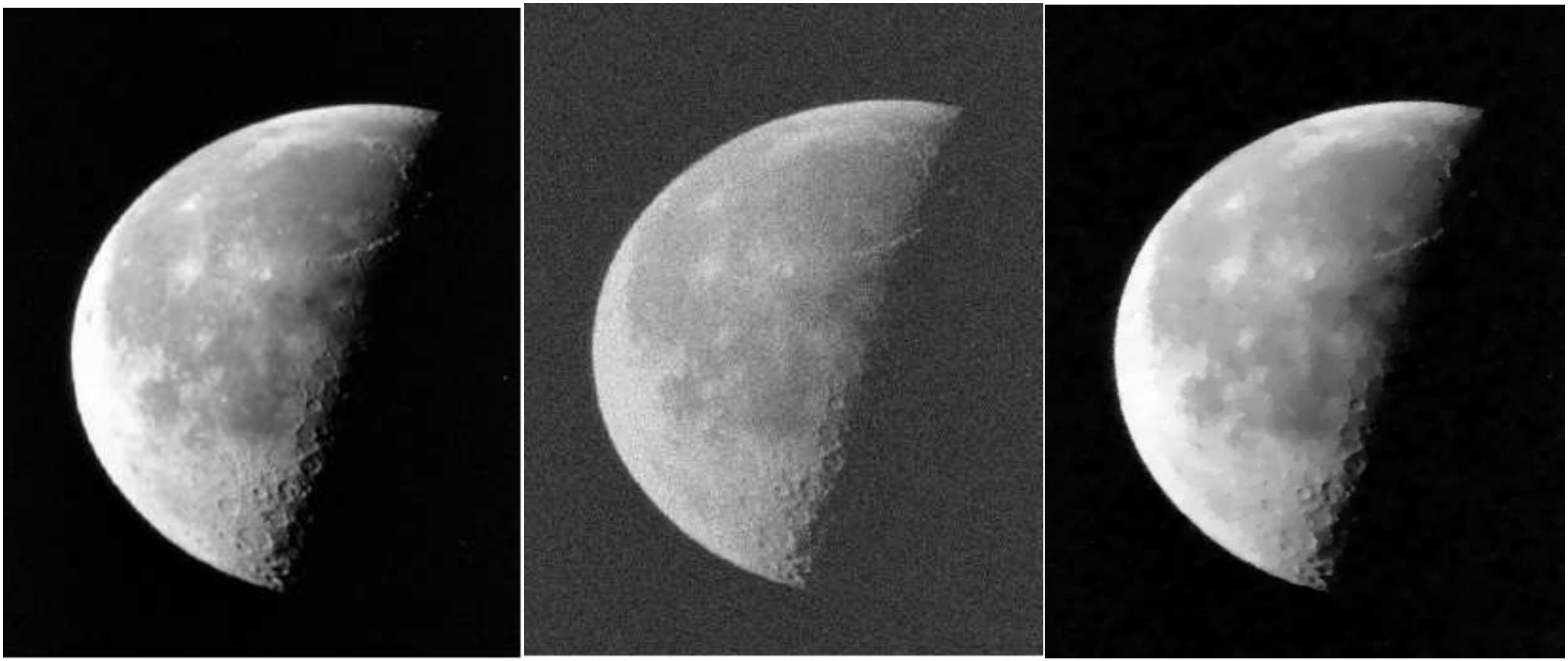


Figure 1: A linearly separable problem. The hard margin solution of the SVM is shown together with the convex hulls of the positive and negative class. The points on the margin, that is $\langle w, x \rangle + b = \pm 1$, are called **support vectors**.

Unconstrained convex optimization: Total variation denoising

$$\min_f \|Y - f\|^2 + \lambda \|\nabla f\|_1 .$$



Classification of optimization problems:

- **Linear** (good properties, polynomial-time algorithms)
- **Convex** (share a lot of properties of linear problems $\Rightarrow$ good complexity properties)
- **Nonlinear and non-convex** (difficult $\Rightarrow$ global optimality statements are usually not possible)

Instead of

linear versus **nonlinear**

consider

convex versus **non-convex**

problem classes.

Goodies of convex optimization problems:

- many interesting problems can be formulated as convex optimization problems,
- the dual problem of non-convex problems is convex $\Rightarrow$ lower bounds for difficult problems !
- efficient algorithms available - but still active research area.

Goal of this course

- Overview over the theory of convex analysis and convex optimization,
- Modeling aspect in applications: how to recognize and formulate a convex optimization problem,
- Introduction to nonlinear programming, interior point methods and specialized methods.

Properties of sets (in $\mathbb{R}^n$):

Definition 1.

- *A point $x \in C$ lies in the* ***interior*** *of C if $\exists\, \epsilon > 0$ such that $B(x, \varepsilon) \subseteq C$.*
- *A point $x \in C$ lies at the* ***boundary*** *if for every $\varepsilon > 0$ the ball around x contains a point $y \notin C$.*
- *A set C is* ***open*** *if every point x in C is an interior point.*
- *A set C is* ***closed*** *if the complement $\mathbb{R}^n \backslash C$ is open.*
- *A set $C \in \mathbb{R}^n$ is* ***compact*** *if it is closed and bounded*
- *The* ***closure*** *of C is the set C plus the limit elements of all sequences of elements in C.*

$\Rightarrow$ A closed set C contains all limits of sequences of elements in C,

Continuous functions:

Definition 2. *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **continuous** at x if for all $\varepsilon > 0$ there exists a δ such that*

$$\|x - y\| \leq \delta \quad \Longrightarrow \quad \|f(x) - f(y)\| \leq \varepsilon.$$

In particular for a continuous functions we have

$$\lim_{n\to\infty} f(x_n) = f\Big(\lim_{n\to\infty} x_n\Big).$$

Closed functions, Level set

Definition 3. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **closed** if for each α the (sub)level set*

$$L_\alpha = \{x \in \operatorname{dom} f \mid f(x) \leq \alpha\},$$

is closed.

The level set of a discontinuous function need not be closed.

Definition 4. *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *has a* ***local minimum*** *at* x*, if*

$$\exists\, \varepsilon > 0, \ \textit{such that} \quad f(x) \leq f(y), \quad \forall\, y \in B(x, \varepsilon).$$

Properties:

- on a compact set every continuous functions attains its **global maximum/minimum**,
- convex functions are (almost) continuous (except for the boundary).

Discontinuous functions:

- there exist functions which are **everywhere discontinuous**

Dirichlet function: $$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \in \mathbb{R}\backslash\mathbb{Q} \end{cases},$$

only **continuous on the irrational numbers**,

Thomae function: $$f(x) = \begin{cases} 1/q & \text{if } x = \frac{p}{q} \in \mathbb{Q}, \text{ with } \gcd(p,q) = 1, \\ 0 & \text{if } x \in \mathbb{R}\backslash\mathbb{Q} \end{cases}.$$

- discontinuous function: can have **no** global maxima/minima on a compact set ($\text{dom}\, f = [-\frac{\pi}{2}, \frac{\pi}{2}]$, $f(x) = \tan(x)$, $f(\pi/2) = f(-\pi/2) = 0$).
- there exist discontinuous functions which have no local minima/maxima.

Jacobian, Gradient

Definition 5. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $x \in \operatorname{int}\operatorname{dom} f$, the **derivative or Jacobian** of f at x is the matrix $Df(x) \in \mathbb{R}^{m\times n}$ given by*

$$Df(x)_{ij} = \frac{\partial f_i}{\partial x_j}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

The affine function g of z given by

$$g(z) = f(x) + Df(x)(z - x),$$

is the (best) first-order approximation of f at x.

Definition 6. *If $f : \mathbb{R}^n \to \mathbb{R}$ the Jacobian reduces to the **gradient** which we write usually as a (column) vector:*

$$\nabla f(x) = Df(x)^T = \left.\frac{\partial f}{\partial x_i}\right|_x, \quad i = 1, \ldots, n.$$

Second Derivative, Hessian and Taylor's theorem

Definition 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and f twice differentiable and $x \in \operatorname{int}\operatorname{dom} f$, the* ***Hessian matrix*** *of f at x is the matrix $Hf(x) \in \mathbb{R}^{n\times n}$ given by*

$$Hf(x)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n.$$

BV use $\nabla^2 f$ for the Hessian matrix. The quadratic function g of z given by

$$g(z) = f(x) + \nabla f(x)(z - x) + \frac{1}{2}\left\langle z - x, Hf|_x (z - x)\right\rangle,$$

is the (best) second-order approximation of f at x.

Theorem 1 (Taylor second-order expansion)**.** *Let $\Omega \subseteq \mathbb{R}^n$, $f \in C^2(\Omega)$ and $x \in \Omega$, then $\forall\, h \in \mathbb{R}^n$ with $[x, x + h] \subset \Omega$ there $\exists \theta \in [0, 1]$ such that*

$$f(x + h) = f(x) + \langle \nabla f|_x, h\rangle + \frac{1}{2}\langle h, Hf(x + \theta h)h\rangle.$$

Taylor expansion

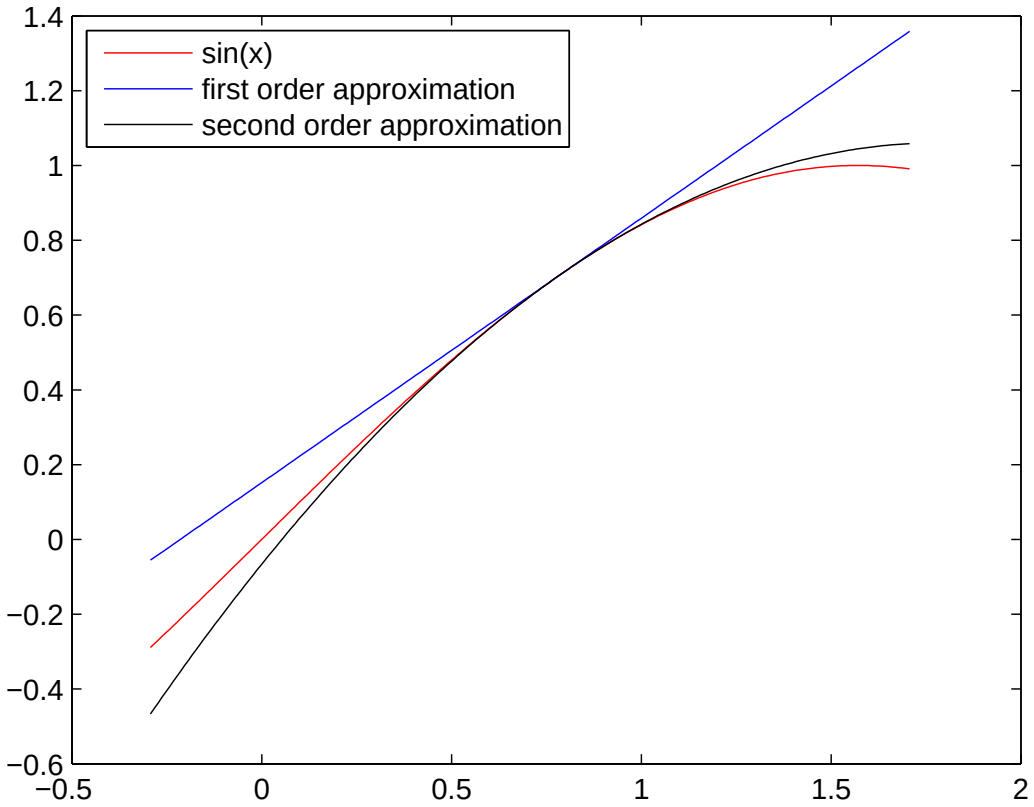


Figure 2: The first and second order Taylor approximation at $x = \frac{\pi}{4}$ of $f(x) = \sin(x)$. $f(\pi/4) = \frac{\sqrt{2}}{2}$, $f'(\pi/4) = \frac{\sqrt{2}}{2}$, $f''(\pi/4) = -\frac{\sqrt{2}}{2}$.

Range and Kernel of linear mappings

Definition 8. *Let $A \in \mathbb{R}^{m\times n}$. The **range** of A is the subspace of $\mathbb{R}^m$ defined as*

$$\operatorname{ran} A = \{x \in \mathbb{R}^m \,|\, x = Ay,\ y \in \mathbb{R}^n\}.$$

*The dimension of $\operatorname{ran} A$ is the **rank** of A. The **null space or kernel** of A is the subspace of $\mathbb{R}^n$ defined as*

$$\ker A = \{y \in \mathbb{R}^n \,|\, Ay = 0\}.$$

Theorem 2. *One has*

$$\dim \ker A + \dim \operatorname{ran} A = n.$$

Moreover, one has the orthogonal decomposition

$$\mathbb{R}^n = \ker A \oplus \operatorname{ran} A^T.$$

Symmetric Matrices

Every real, symmetric matrix $A \in \mathbb{R}^{n\times n}$ can be written as

$$A = Q\Lambda Q^T,$$

where $Q \in \mathbb{R}^{n\times n}$ is orthogonal ($QQ^T = \mathbb{1}$) and Λ is a diagonal matrix having the eigenvalues λ_i on the diagonal. Alternatively, one can write

$$A = \sum_{i=1}^{n} \lambda_i \, q_i \, q_i^T,$$

where q_i is the eigenvector corresponding to the eigenvalue λ_i. Moreover,

$$\det A = \prod_{i=1}^{n} \lambda_i, \quad \operatorname{tr} A = \sum_{i=1}^{n} \lambda_i.$$

One can find the eigenvalues via the so-called **Rayleigh-Ritz** principle

$$\lambda_{\min} = \inf_{v\in\mathbb{R}^n} \frac{\langle v, Av\rangle}{\langle v, v\rangle}, \quad \lambda_{\max} = \sup_{v\in\mathbb{R}^n} \frac{\langle v, Av\rangle}{\langle v, v\rangle}.$$

Positive Definite Matrices

Definition 9. *A real, symmetric matrix* $A \in \mathbb{R}^{n \times n}$ *is* ***positive semi-definite*** *if*

$$\langle w, Aw \rangle \geq 0, \qquad \text{for all } w \in \mathbb{R}^n,$$

The real, symmetric matrix $A \in \mathbb{R}^{n \times n}$ *is* ***positive definite*** *if*

$$\langle w, Aw \rangle > 0, \qquad \text{for all } w \in \mathbb{R}^n \text{ with } w \neq 0.$$

Notation:

- S^n: the set of symmetric matrices in $\mathbb{R}^{n \times n}$,
- S^n_+: the set of positive semi-definite matrices,
- S^n_{++}: the set of positive definite matrices.

Singular Value Decomposition

Every real matrix $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = U \Sigma V^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and Σ is a diagonal matrix having the positive **singular values** σ_i on the diagonal.

Facts:

- the singular values σ_i are positive,
- the number of non-zero singular values is equal to the rank of A,
- U contains the left eigenvectors (eigenvectors of AA^T),
- V contains the right eigenvectors (eigenvectors of $A^T A$),
- the singular values are the eigenvalues of AA^T ($A^T A$).

Norms

Definition 10. *Let V be a vector space. A* ***norm*** $\|\cdot\| : V \to \mathbb{R}$ *satisfies,*

- *non-negative:* $\|x\| \geq 0$ *for all* $x \in \mathbb{R}^n$, $\|x\| = 0 \Leftrightarrow x = 0$,
- *homogeneous:* $\|\alpha x\| = |\alpha| \, \|x\|$,
- *triangle inequality:* $\|x + y\| \leq \|x\| + \|y\|$.

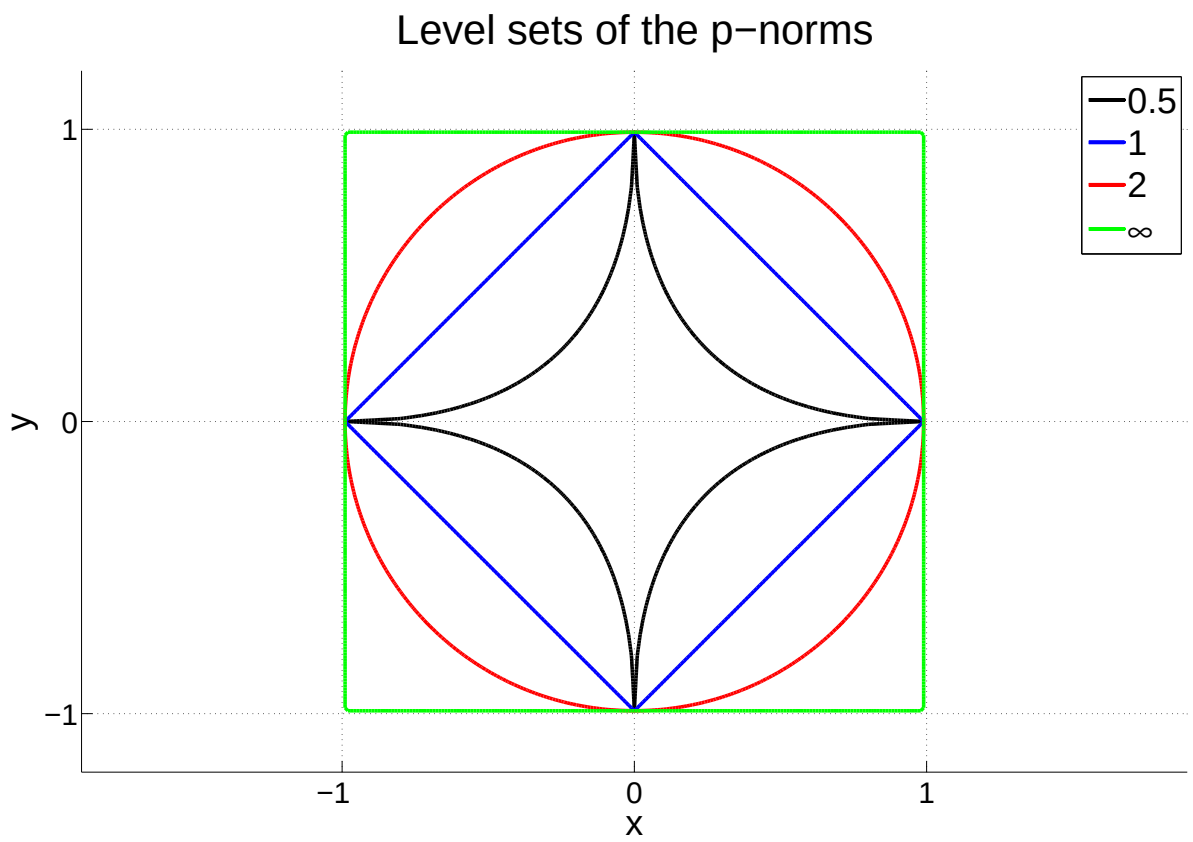
A norm induces a **distance measure(metric)**: $d(x, y) = \|x - y\|$.
In $\mathbb{R}^n$ we have the p-norms ($p \geq 1$)

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

On matrices $\mathbb{R}^{m \times n}$ this can be defined equivalently:

$$\|X\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}|^p \right)^{\frac{1}{p}}.$$

p-norm



The unit-ball of the p-norms. Note that for $p < 1$, the unit-ball is not convex

$\Rightarrow$ no norm.

Operator/Matrix norm

Definition 11. *Let $\|\cdot\|_\alpha$ be a norm on $\mathbb{R}^m$ and $\|\cdot\|_\beta$ a norm on $\mathbb{R}^n$. The* ***operator-norm*** *of $A : \mathbb{R}^m \to \mathbb{R}^n$ is defined as*

$$\|A\|_{\alpha,\beta} = \sup_{v \in \mathbb{R}^m, \|v\|_\alpha = 1} \|Av\|_\beta .$$

This is equivalent to:

$$\|A\|_{\alpha,\beta} = \sup_{v \in \mathbb{R}^m} \frac{\|Av\|_\beta}{\|v\|_\alpha}.$$

If both norms are Euclidean, then the operator norm is

$$\|A\|_{2,2} = \sigma_{\max}(X) = \sqrt{\lambda_{\max}(A^T A)}.$$

"Proof":

$$\frac{\|Av\|_2}{\|v\|_2} = \sqrt{\frac{\|Av\|_2^2}{\|v\|_2^2}} = \sqrt{\frac{\langle v, A^T Av\rangle}{\langle v, v\rangle}}.$$

Equivalent Norms

Definition 12. *We say that two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space V are* ***equivalent*** *if there exist $a, b > 0$ such that*

$$a\,\|x\|_1 \leq \|x\|_2 \leq b\,\|x\|_1\,, \qquad \forall\, x \in V.$$

Remarks:

- all norms on $\mathbb{R}^n$ are equivalent to each other, e.g.

$$\|x\|_2 = \sqrt{\sum_i x_i^2} \leq \sum_i \sqrt{x_i^2} = \sum_i |x_i| = \|x\|_1 \leq \sqrt{\sum_i |x_i|^2 \sum_i 1} = \sqrt{n}\,\|x\|_2.$$

$$\|x\|_\infty = \max_i |x_i| \leq \sum_i |x_i| = \|x\|_1 \leq n \max_i |x_i| = n\,\|x\|_\infty.$$

- the definition of a continuous function $f : \mathbb{R}^n \to \mathbb{R}^m$ does not depend on the choice of the norm.

Inner Product

Definition 13. *Let V be a vector space. Then an inner product $\langle \cdot, \cdot \rangle$ over $\mathbb{R}$ is a* ***bilinear form*** $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$, *such that*

- *symmetry:* $\langle x, y \rangle = \langle y, x \rangle$,
- *non-negativity:* $\langle x, x \rangle \geq 0$,
- *non-degenerate:* $\langle x, x \rangle = 0 \iff x = 0$,

Remarks:

- An **inner product space** is a vector space with an inner product,
- A complete inner product space is a **Hilbert space**,
- A complete normed space is a **Banach space**.

Inner Product

The standard-inner product on $\mathbb{R}^n$ is given for $x, y \in \mathbb{R}^n$ as

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i,$$

On can extend this to the set of matrices $\mathbb{R}^{n \times m}$, for $X, Y \in \mathbb{R}^{n \times m}$,

$$\langle X, Y \rangle = \operatorname{tr}(X^T Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} Y_{ij}.$$

Clearly, an inner-product induces a norm via: $\|x\| = \sqrt{\langle x, x \rangle}$.

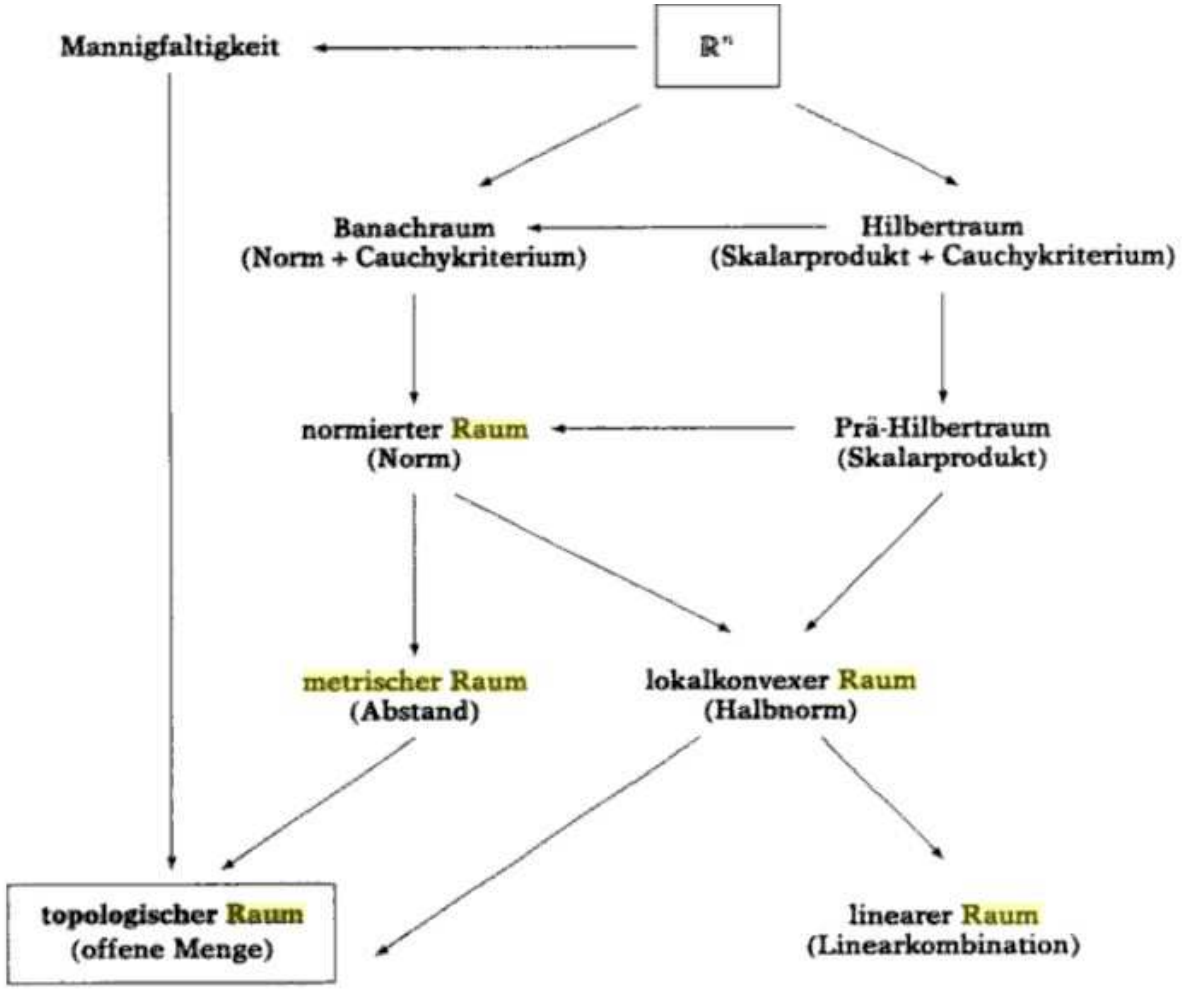
The norm for the inner product on matrices is the **Frobenius norm**

$$\|X\|_F = \sqrt{\operatorname{tr}(X^T X)} = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2 \right)^{\frac{1}{2}}.$$

Every inner product fulfills the **Cauchy-Schwarz inequality**

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|.$$

Hierarchy of mathematical structures



The hierarchy of mathematical structures - an arrow denotes inclusion (e.g. a Banach space is also a metric space or $\mathbb{R}^n$ is also a manifold.) Drawing from "Teubner - Taschenbuch der Mathematik".