

# Optimization

Pieter Vanderschueren

Year 2024-2025

# Contents

<b>Mathematical Preliminaries</b>	<b>2</b>
1 Vectors . . . . .	3
2 Independence, Subspaces, Basis and Dimension . . . . .	4
3 Orthogonality and Orthogonal Complements . . . . .	5
4 Matrices . . . . .	6
5 Sequences . . . . .	10
6 Differential Calculus . . . . .	10
<b>Fundamental Concepts</b>	<b>13</b>
7 Fundamental Concepts of Optimization . . . . .	14
8 Types of Optimization Problems . . . . .	15
9 Convex Optimization . . . . .	18
10 The Lagrangian Function and Duality . . . . .	23
<b>Unconstrained Optimization Algorithms</b>	<b>27</b>
11 Optimality Conditions . . . . .	28
12 Estimation and Fitting Problems . . . . .	30
13 Newton Type Optimization . . . . .	31
14 Globalisation Strategies . . . . .	31
15 Calculating Derivatives . . . . .	31
<b>Constrained Optimization Algorithms</b>	<b>32</b>
16 Optimality Conditions for Constrained Optimization . . . . .	33
17 Equality Constrained Optimization Algorithms . . . . .	33
18 Inequality Constrained Optimization Algorithms . . . . .	33
19 Optimal Control Problems . . . . .	33

## Mathematical Preliminaries

# 1 Vectors

## Theorem 1.1: Vector Space $\mathbb{R}^n$

The vector space  $\mathbb{R}^n$  is the set of all  $n$ -dimensional column vectors with real components. The space  $\mathbb{R}^n$  is equipped with the following operations:

- component-wise addition:

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

- scalar multiplication:

$$\alpha x = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

## Theorem 1.2: Dot product

The dot product of two vectors  $x, y \in \mathbb{R}^n$  is defined as the scalar

$$x \cdot y = x^T y = \sum_{i=1}^n x_i y_i.$$

## Theorem 1.3: Norm

A norm  $\|\cdot\|$  on  $\mathbb{R}^n$  is a function  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfying the following:

- non-negativity:  $\|x\| \geq 0$  for any  $x \in \mathbb{R}^n$ ;  $\|x\| = 0$  if and only if  $x = 0$ ,
- positive homogeneity:  $\|\alpha x\| = |\alpha| \|x\|$  for any  $x \in \mathbb{R}^n$  and  $\alpha \in \mathbb{R}$ ,
- triangle inequality:  $\|x + y\| \leq \|x\| + \|y\|$  for any  $x, y \in \mathbb{R}^n$ .

## Theorem 1.4: $\ell_p$ -norms

The class of  $\ell_p$ -norms is defined by

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

which includes the following special cases:

- $\ell_1$ -norm:  $\|x\|_1 = \sum_{i=1}^n |x_i|$ ,
- $\ell_2$ -norm:  $\|x\| = \|x\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}$ ,
- $\ell_\infty$ -norm:  $\|x\|_\infty = \max_{i=1, \dots, n} |x_i|$ .

### Theorem 1.5: Angle between two vectors

The angle  $\angle(x, y)$  between two vectors  $x, y \in \mathbb{R}^n$  is given by

$$\angle(x, y) = \arccos\left(\frac{x \cdot y}{\|x\| \|y\|}\right).$$

We say that two vectors  $x, y \in \mathbb{R}^n$  are

- orthogonal if  $\angle(x, y) = \frac{\pi}{2}$ , i.e.  $x \cdot y = 0$ ,
- aligned if  $\angle(x, y) = 0$ ,
- anti-aligned if  $\angle(x, y) = \pi$ ,
- parallel if  $\angle(x, y) = 0$  or  $\angle(x, y) = \pi$ ,
- at an acute angle if  $\angle(x, y) < \frac{\pi}{2}$ ,
- at an obtuse angle if  $\angle(x, y) > \frac{\pi}{2}$ .

### Theorem 1.6: Cauchy-Schwarz inequality

For any two vectors  $x, y \in \mathbb{R}^n$ , the following inequality holds:

$$|x \cdot y| \leq \|x\| \|y\|.$$

### Theorem 1.7: Hölder inequality

Let  $p \geq 1$ . For any  $x, y \in \mathbb{R}^n$ , the following inequality holds:

$$|x \cdot y| \leq \|x\|_p \|y\|_q,$$

where  $q = \frac{p}{p-1}$  is the Hölder conjugate of  $p$ .

## 2 Independence, Subspaces, Basis and Dimension

### Theorem 2.1: Linear Independence

A set of vectors  $a_1, a_2, \dots, a_n$  in  $\mathbb{R}^n$  is said to be linearly independent if no vector in the collection can be expressed as a combination of the others. In other words

$$\sum_{i=1}^m \lambda_i a_i = 0 \Rightarrow \forall i \in [1, m] : \lambda_i = 0$$

**Theorem 2.2: Subspace**

A nonempty subset  $S$  of  $\mathbb{R}^n$  is called a subspace if for any real numbers  $\lambda_1, \lambda_2$

$$a_1, a_2 \in S \Rightarrow \lambda_1 a_1 + \lambda_2 a_2 \in S$$

A subspace always contains the zero element.

**Theorem 2.3: Span**

Given a set of vectors  $a_1, a_2, \dots, a_n$  in  $\mathbb{R}^n$ , the set of linear combinations of these vectors is called the span, i.e.

$$\text{span}\{a_1, a_2, \dots, a_n\} = \left\{ y \in \mathbb{R}^n \mid y = \sum_{i=1}^n \lambda_i a_i \right\}$$

**Theorem 2.4: Basis**

The subset  $B = \{a_1, a_2, \dots, a_n\}$  of  $\mathbb{R}^n$  is called a basis if it is linearly independent and spans  $\mathbb{R}^n$ , i.e. it is a maximally independent subset of  $\mathbb{R}^n$ .

### 3 Orthogonality and Orthogonal Complements

**Theorem 3.1: Orthogonality**

A set of nonzero vectors  $a_1, a_2, \dots, a_n$  in  $\mathbb{R}^n$  is said to be orthogonal if the dot product of any two distinct vectors is zero, i.e.

$$\forall i, j \in [1, n] : i \neq j \Rightarrow a_i \cdot a_j = 0$$

Consequently, two subspaces  $S_1$  and  $S_2$  of  $\mathbb{R}^n$  are orthogonal if every vector in  $S_1$  is orthogonal to every vector in  $S_2$ . In that case,  $S_2$  is called the orthogonal complement of  $S_1$  and denoted by  $S_1^\perp$ . The following now holds true for a subspace  $S$  of  $\mathbb{R}^n$ :

$$\mathbb{R}^n = S \oplus S^\perp$$

**Theorem 3.2: Orthonormality**

A set of nonzero vectors  $a_1, a_2, \dots, a_n$  in  $\mathbb{R}^n$  is said to be orthonormal if it is orthogonal and each vector has a unit length, i.e.

$$\forall i, j \in [1, n] : i \neq j \Rightarrow a_i \cdot a_j = 0 \wedge \|a_i\| = 1$$

## 4 Matrices

### Example 4.1: Types of matrices

A matrix  $A \in \mathbb{R}^{m \times n}$  is said to be

- the zero matrix, denoted by  $0$ , if all its entries are zero,
- a square matrix if  $m = n$ ,
- the identity matrix if it is square and all its diagonal entries are one,
- a diagonal matrix if all its off-diagonal entries are zero,
- an upper triangular matrix if all its entries below the diagonal are zero,
- a lower triangular matrix if all its entries above the diagonal are zero,
- a symmetric matrix if it is square and  $A = A^T$ ,
- an orthogonal matrix if it is square and  $AA^T = A^T A = I$ ,
- a non-singular matrix if it is square and there exists another square matrix  $B \in \mathbb{R}^{n \times n}$ , the inverse of  $A$ , such that

$$AB = BA = I$$

- a dyadic matrix if it is of the form  $A = uv^T$  for some vectors  $u, v \in \mathbb{R}^n$ .

### Theorem 4.1: Range

Given a matrix  $A \in \mathbb{R}^{m \times n}$ , the range of  $A$  is the set of  $m$ -dimensional vectors that can be expressed as  $Ax$  for some  $n$ -dimensional vector  $x$ , and we denote it by

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$$

In other words, it is the set of vectors that can be expressed as linear combinations of the columns of  $A$ .

### Theorem 4.2: Kernel

Given a matrix  $A \in \mathbb{R}^{m \times n}$ , the kernel of  $A$  is the set of  $n$ -dimensional vectors that are mapped to the zero vector by  $A$ , and we denote it by

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$$

In other words, it is the set of vectors that are orthogonal to the columns of  $A$ .

#### Theorem 4.3: Rank

Given a matrix  $A \in \mathbb{R}^{m \times n}$ , the rank of  $A$  is the dimension of its range, i.e.

$$\text{rank}(A) = \dim(\mathcal{R}(A))$$

**Note:** The sum of the dimensions of the range of  $A$  and the null space (kernel) of  $A$  is equal to the number of columns  $n$ :

$$\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A)) = n$$

#### Theorem 4.4: Fundamental theorem of Linear Algebra

For any matrix  $A \in \mathbb{R}^{m \times n}$ , it holds that  $\mathcal{N}(A) \perp \mathcal{R}(A^T)$  and  $\mathcal{N}(A^T) \perp \mathcal{R}(A)$ , therefore we have

$$\mathbb{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A^T)$$

$$\mathbb{R}^m = \mathcal{N}(A^T) \oplus \mathcal{R}(A)$$

#### Theorem 4.5: Singular Value Decomposition

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

$$A = U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

where  $U \in \mathbb{R}^{m \times m}$  and  $V \in \mathbb{R}^{n \times n}$  are orthogonal matrices,  $U_1 \in \mathbb{R}^{m \times r}$ ,  $V_1 \in \mathbb{R}^{n \times r}$  and

$$\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$$

is a diagonal matrix, where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$  are the singular values of  $A$ . The columns of  $U$  and  $V$  are called the left and right singular vectors of  $A$ , respectively.

#### Property 4.1: Singular Value Decomposition

The singular value decomposition of a matrix  $A \in \mathbb{R}^{m \times n}$  gives orthogonal bases for the four fundamental subspaces related to  $A$ :

$$\begin{aligned} \mathcal{R}(A) &= \mathcal{R}(U_1), & \mathcal{N}(A^T) &= \mathcal{R}(U_2) \\ \mathcal{R}(A^T) &= \mathcal{R}(V_1), & \mathcal{N}(A) &= \mathcal{R}(V_2) \end{aligned}$$



**Theorem 4.6: Moore Penrose Pseudo Inverse**

Assume  $J \in \mathbb{R}^{m \times n}$  and that the singular value decomposition (SVD) of  $J$  is given by  $J = U\Sigma V^T$ . Then, the Moore-Penrose pseudo-inverse  $J^+$  is given by

$$J^+ = VS^+U^T,$$

where for

$$S = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \\ \hline 0 & \dots & \dots & 0 & \dots & \dots & 0 \end{bmatrix} \quad \text{holds} \quad S^+ = \begin{bmatrix} \sigma_1^{-1} & & & & & & 0 \\ & \sigma_2^{-1} & & & & & \vdots \\ & & \ddots & & & & \vdots \\ & & & \sigma_r^{-1} & & & 0 \\ & & & & 0 & & \vdots \\ & & & & & \ddots & \vdots \\ & & & & & & 0 \end{bmatrix}$$

**Property 4.2: Moore Penrose Pseudo Inverse**

If matrix  $A \in \mathbb{R}^{m \times n}$

- is non-singular then

$$A^+ = A^{-1},$$

- has full column rank, that is  $r = n$ , then

$$A^+A = VV^T = I,$$

i.e.  $A^+$  is a left inverse of  $A$ ,

- has full row rank, that is  $r = m$ , then

$$AA^+ = UU^T = I,$$

i.e.  $A^+$  is a right inverse of  $A$ .

**Theorem 4.7: Orthogonal-triangular decomposition (QR)**

If  $A \in \mathbb{R}^{m \times n}$ , then there exists an orthogonal matrix  $Q \in \mathbb{R}^{m \times m}$  and an upper triangular matrix  $R \in \mathbb{R}^{m \times n}$  such that

$$A = QR.$$

#### Theorem 4.8: Eigenvalue decomposition

Any real symmetric matrix  $A \in \mathbb{R}^{n \times n}$  can be decomposed as

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

where  $Q \in \mathbb{R}^{n \times n}$  is orthogonal, i.e.  $Q^T Q = I$ , and  $\Lambda \in \mathbb{R}^{n \times n}$  is diagonal with the eigenvalues of  $A$  on the diagonal. The columns of  $Q$  form an orthonormal set of eigenvectors.

#### Theorem 4.9: Symmetric positive semi-definite matrices

Let  $Q \in \mathbb{S}^n$  be a symmetric matrix. Then the following statements are equivalent:

1.  $Q$  is positive semi-definite, i.e.  $Q \succeq 0$ ,
2. all eigenvalues of  $Q$  are non-negative, i.e.  $\lambda_i(Q) \geq 0$  for all  $i \in [1, n]$ ,
3. all principal minors of  $Q$ , i.e. the determinant of a submatrix obtained from  $Q$  when the same set of rows and columns are stricken out, are non-negative,
4.  $Q$  can be written as  $Q = AA^T$  for some matrix  $A \in \mathbb{R}^{n \times r}$  and  $r$  is the rank of  $Q$ .

#### Theorem 4.10: Cholesky factorization

If  $A \in \mathbb{R}^{n \times n}$  is symmetric positive definite, then there exists a unique lower triangular matrix  $L \in \mathbb{R}^{n \times n}$  with positive diagonal entries such that

$$A = LL^T.$$

#### Theorem 4.11: Matrix norms

For a matrix  $A \in \mathbb{R}^{m \times n}$  and two vector norms  $\|\cdot\|_p$  and  $\|\cdot\|_q$ , the induced matrix norm is defined as

$$\|A\|_{p,q} = \max_x \{\|Ax\|_q \mid \|x\|_p \leq 1\}.$$

When  $p = q$ , we simply write  $\|A\|_p$ .

#### Example 4.2: Spectral norm

The  $\ell_2$ -induced norm or spectral norm of a matrix  $A \in \mathbb{R}^{m \times n}$  is defined as

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A),$$

where  $\lambda_{\max}(A^T A)$  denotes the largest eigenvalue of  $A^T A$  and  $\sigma_{\max}(A)$  is the largest singular value of  $A$ .

### Example 4.3: Frobenius norm

The Frobenius norm of a matrix  $A \in \mathbb{R}^{m \times n}$  is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} = \sqrt{\text{tr}(A^T A)}.$$

## 5 Sequences

### Lemma 5.1: Convergence

If  $\{x_k\}$  is a non-increasing and bounded below or non-decreasing and bounded above sequence, it converges to a finite real number.

### Theorem 5.1: $O(\cdot)$

For a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we write

$$f(x) = O(g(x))$$

if and only if there exists a constant  $C > 0$  and a neighborhood  $\mathcal{N}$  of 0 such that

$$\forall x \in \mathcal{N} : \|f(x)\| \leq Cg(x),$$

i.e. “ $f$  shrinks as fast as  $g$ ”.

### Theorem 5.2: $o(\cdot)$

For a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we write

$$f(x) = o(g(x))$$

if and only if there exists a neighborhood  $\mathcal{N}$  of 0 and a function  $c : \mathcal{N} \rightarrow \mathbb{R}$  with  $\lim_{x \rightarrow 0} c(x) = 0$  such that

$$\forall x \in \mathcal{N} : \|f(x)\| \leq c(x)g(x),$$

i.e. “ $f$  shrinks faster than  $g$ ”.

## 6 Differential Calculus

### Theorem 6.1: Lipschitz continuity

A mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is Lipschitz continuous with Lipschitz constant  $L$  if

$$\|F(x) - F(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

**Theorem 6.2: Linear mapping**

A mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is linear if

$$\forall x, y \in \mathbb{R}^n, \forall \lambda_1, \lambda_2 \in \mathbb{R} : F(\lambda_1 x + \lambda_2 y) = \lambda_1 F(x) + \lambda_2 F(y).$$

**Theorem 6.3: Affine mapping**

A mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is affine if it is the sum of a linear mapping and a constant vector, i.e.

$$\exists A \in \mathbb{R}^{m \times n}, \exists b \in \mathbb{R}^m : F(x) = Ax + b.$$

**Theorem 6.4: Quadratic function**

A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is quadratic if it can be written as

$$f(x) = \frac{1}{2}x^T Qx + q^T x + c$$

for some matrix  $Q \in \mathbb{R}^{n \times n}$ , vector  $q \in \mathbb{R}^n$ , and scalar  $c \in \mathbb{R}$ .

**Theorem 6.5: First-order Taylor expansion**

Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable at  $x \in \mathbb{R}^n$ . Then for all  $y \in \mathbb{R}^n$ , it holds that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + o(\|y - x\|).$$

**Lemma 6.1: Mean-value theorem**

Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is differentiable. Then for every  $x, y \in \mathbb{R}^n$  there exists a  $\tau \in (0, 1)$  such that

$$f(y) = f(x) + \nabla f(x + \tau(y - x))^T (y - x),$$

Moreover, if  $f$  is continuously differentiable, then

$$f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt$$

**Theorem 6.6: Hessian**

The Hessian of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is the matrix of second partial derivatives, i.e.

$$\nabla^2 f(x) = \left[ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right]_{i,j=1}^n$$

**Theorem 6.7: Second-order Taylor expansion**

Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is twice continuously differentiable at  $x \in \mathbb{R}^n$ . Then for all  $y \in \mathbb{R}^n$ , it holds that

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2).$$

**Theorem 6.8: Jacobian**

The Jacobian of a mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is the matrix of transposed gradients, i.e.

$$J_F(x) = \begin{bmatrix} \nabla f_1(x)^\top \\ \nabla f_2(x)^\top \\ \vdots \\ \nabla f_m(x)^\top \end{bmatrix} \in \mathbb{R}^{m \times n}$$

**Application 6.1: Mean-value theorem for vector-valued functions**

Suppose that  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuously differentiable on  $\mathbb{R}^n$ . Then for every  $x, y \in \mathbb{R}^n$  the following holds:

$$F(y) = F(x) + \int_0^1 J_F(x + t(y - x))(y - x) dt.$$

**Theorem 6.9: Implicit function theorem**

Let  $F : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^n$  be a continuously differentiable mapping of  $x \in \mathbb{R}^n$  and  $p \in \mathbb{R}^m$ . If  $x^* \in \mathbb{R}^n, p^* \in \mathbb{R}^m$  are such that

1.  $F(x^*, p^*) = 0$ ,
2. the partial Jacobian  $J_{F_x}(x^*, p^*)$  is non-singular,

then there exist open sets  $S_{x^*} \subset \mathbb{R}^n, S_{p^*} \subset \mathbb{R}^m$  and a continuously differentiable function  $g : S_{p^*} \rightarrow S_{x^*}$  such that

$$x^* = g(p^*) \quad \text{and} \quad F(g(p), p) = 0 \quad \forall p \in S_{p^*},$$

and

$$J_g(p^*) = -(J_{F_x}(g(x^*), p^*))^{-1} J_{F_p}(g(x^*), p^*).$$

## **Fundamental Concepts**

## 7 Fundamental Concepts of Optimization

### Theorem 7.1: Optimization problem in standard form

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && g(x) = 0, \\ & && h(x) \geq 0. \end{aligned}$$

### Theorem 7.2: Level Set

The set

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

is the level set of  $f$  for the value  $c$ , i.e. the set of all points that map to the same value  $c$ .

### Theorem 7.3: Feasible set

The set

$$\{ x \in \mathbb{R}^n \mid g(x) = 0 \wedge h(x) \geq 0 \}$$

is the feasible set  $\Omega$ , i.e. the set of all points that satisfy the constraints.

### Theorem 7.4: Global minimizer

The point  $x^*$  is a global minimizer if and only if

$$x^* \in \Omega, \forall x \in \Omega : f(x^*) \leq f(x).$$

### Theorem 7.5: Strict global minimizer

The point  $x^*$  is a strict global minimizer if and only if

$$x^* \in \Omega, \forall x \in \Omega \setminus \{x^*\} : f(x^*) < f(x).$$

### Theorem 7.6: Strict local minimizer

The point  $x^*$  is a strict local minimizer if and only if  $x^* \in \Omega$  and there exists a neighborhood  $\mathcal{N}$  of  $x^*$  such that

$$\forall x \in (\Omega \cap \mathcal{N}) \setminus \{x^*\} : f(x^*) < f(x).$$

### Theorem 7.7: Weierstrass

If  $\Omega \in \mathbb{R}^n$  is compact, i.e. limited and closed, and  $f : \Omega \rightarrow \mathbb{R}$  is continuous, then there exists a global minimizer (a solution) of the optimization problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && x \in \Omega. \end{aligned}$$

### Proof 7.1: Weierstrass

Regard the graph of  $f$ ,  $G = \{(x, f(x)) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \Omega\}$ .  $G$  is a compact set, and so is the projection of  $G$  onto its last coordinate, the set  $G = \{s \in \mathbb{R} \mid \exists x \text{ such that } (x, s) \in G\}$ , which is a compact interval  $[f_{\min}, f_{\max}] \subset \mathbb{R}$ . By construction, there must be at least one  $x^*$  such that  $(x^*, f(x^*)) \in G$ . □

## 8 Types of Optimization Problems

### Theorem 8.1: Convex set

A set  $\Omega \subset \mathbb{R}^n$  is convex if

$$\forall x, y \in \Omega, \forall \lambda \in [0, 1] : x + \lambda(y - x) \in \Omega.$$

or if “all connecting lines lie inside the set”.

### Theorem 8.2: Convex function

A function  $f : \Omega \rightarrow \mathbb{R}$  is convex if

$$\forall x, y \in \Omega, \forall \lambda \in [0, 1] : f(x + \lambda(y - x)) \leq f(x) + \lambda(f(y) - f(x)).$$

or if “all secants (i.e. a line segment between two points on the graph) are above graph”. This definition is equivalent to saying that the Epigraph of  $f$ , i.e. the set  $\{(x, s) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \Omega, s \geq f(x)\}$ , is a convex set.

**Note:** a concave function is the same but then with  $\geq$  instead of  $\leq$ .

### Property 8.1: Convex function

If  $f : D \rightarrow \mathbb{R}$  and  $\Omega_f = \{(x, y) \mid x \in D, y \geq f(x)\}$  then the following holds:

$$f \text{ is convex} \Leftrightarrow \Omega_f \text{ is convex.}$$



### Theorem 8.3: Convex optimization problem

An optimization problem with convex feasible set  $\Omega$  and convex objective function  $f : \Omega \rightarrow \mathbb{R}$  is called a convex optimization problem.

### Property 8.2: Globality of local minima of convex optimization problem

For a convex optimization problem, a local minimum is also a global one.

### Proof 8.1: Globality of local minima of convex optimization problem

Regard a local minimum  $x^*$  of the convex optimization problem

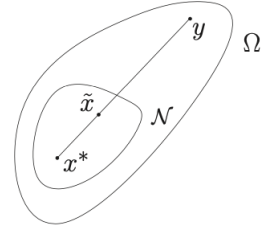
$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && x \in \Omega. \end{aligned}$$

We will show that for any point  $y \in \Omega$  holds  $f(y) \geq f(x^*)$ , i.e.  $x^*$  is a global minimum.

First we choose, using local optimality, a neighborhood  $\mathcal{N}$  of  $x^*$  such that for all  $\tilde{x} \in \Omega \cap \mathcal{N}$  holds  $f(\tilde{x}) \geq f(x^*)$ . Second, we regard the connecting line between  $x^*$  and  $y$ . This line is completely contained in  $\Omega$  due to the convexity of  $\Omega$ . Now we choose a point  $\tilde{x}$  on this line such that it is in the neighborhood, but not equal to  $x^*$ , i.e.  $\tilde{x} = x^* + \lambda(y - x^*)$  for some  $\lambda \in (0, 1)$ , and  $\tilde{x} \in \Omega \cap \mathcal{N}$ . Due to local optimality, we have  $f(\tilde{x}) \geq f(x^*)$ , and due to convexity we have

$$f(\tilde{x}) = f(x^* + \lambda(y - x^*)) \leq f(x^*) + \lambda(f(y) - f(x^*)).$$

It follows that  $\lambda(f(y) - f(x^*)) \geq 0$ , and since  $\lambda \in (0, 1)$ , we have  $f(y) \geq f(x^*)$ , as desired. □



### Theorem 8.4: Unconstrained optimization problem

An optimization problem with no constraints, i.e.  $g(x) = 0$  and  $h(x) \geq 0$  are empty, is called an unconstrained optimization problem.

### Example 8.1: Nonlinear Programming (NLP)

Nonlinear Programming Problems are problems of the following form:

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && g(x) = 0, \\ & && h(x) \geq 0, \end{aligned}$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $g : \mathbb{R}^n \rightarrow \mathbb{R}^p$ , and  $h : \mathbb{R}^n \rightarrow \mathbb{R}^q$ , are assumed to be continuously differentiable at least once.

### Example 8.2: Linear Programming (LP)

When the functions  $f, g, h$  are affine (i.e., they can be expressed in the form  $f(x) = a^T x + b$  for some vector  $a$  and scalar  $b$ ) in the general formulation (see Theorem 8.1), the general NLP gets something easier to solve, namely a Linear Program (LP), which can be written as follows:

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && c^T x \\ & \text{subject to} && Ax - b = 0, \\ & && Cx - d \geq 0, \end{aligned}$$

where  $c \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ ,  $C \in \mathbb{R}^{q \times n}$ , and  $d \in \mathbb{R}^q$ .

### Example 8.3: Quadratic Programming (QP)

When the functions  $g, h$  are affine (as for an LP in Theorem 8.2), and the objective function  $f$  is a linear-quadratic function, the NLP becomes a Quadratic Program (QP), which can be written as follows:

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && c^T x + \frac{1}{2} x^T B x \\ & \text{subject to} && Ax - b = 0, \\ & && Cx - d \geq 0, \end{aligned}$$

where, in addition to the LP parameters,  $B \in \mathbb{R}^{n \times n}$  is the Hessian matrix. Specifically, for the objective function  $f(x) = c^T x + \frac{1}{2} x^T B x$ , the Hessian is given by:

$$\nabla^2 f(x) = B.$$

### Theorem 8.5: Convex QP

If the Hessian matrix  $B$  is positive semi-definite (i.e. if  $\forall z \in \mathbb{R}^n \setminus \{0\} : z^T B z \geq 0$ ) we call the QP a convex QP. Convex QPs are tremendously easier to solve globally than “non-convex QPs” (i.e., where the Hessian  $B$  is not positive semi-definite), which might have different local minima (i.e. have a non-convex solution set).

### Theorem 8.6: Strictly convex QP

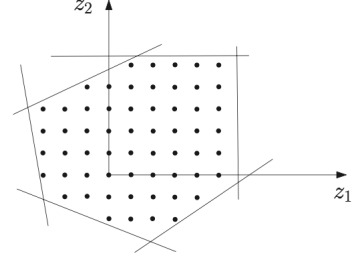
If the Hessian matrix  $B$  is positive definite (i.e. if  $\forall z \in \mathbb{R}^n : z^T B z > 0$ ) we call the QP a strictly convex QP. Strictly convex QPs are a subset of convex QPs (see Theorem 8.5).

### Example 8.4: Mixed-Integer Programming (MIP)

A mixed-integer programming problem or mixed-integer program (MIP) is a problem with both real and integer decision variables. A MIP can be formulated as follows:

$$\begin{aligned} & \underset{\substack{x \in \mathbb{R}^n \\ z \in \mathbb{Z}^m}}{\text{minimize}} && f(x, z) \\ & \text{subject to} && g(x, z) = 0, \\ & && h(x, z) \geq 0, \end{aligned}$$

where  $x \in \mathbb{R}^n$  are the continuous variables, and  $z \in \mathbb{Z}^m$  are the integer variables.



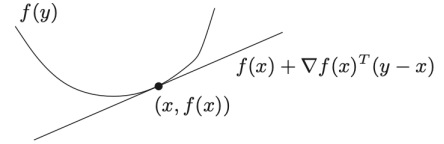
## 9 Convex Optimization

### Theorem 9.1: Convexity for $C^1$ functions

Assume that  $f : \Omega \rightarrow \mathbb{R}$  is continuously differentiable and  $\Omega$  is convex. Then holds that  $f$  is convex if and only if

$$\forall x, y \in \Omega : f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

i.e. tangents lie below the graph.



### Proof 9.1: Convexity for $C^1$ functions

“ $\Rightarrow$ ”: Due to convexity of  $f$  holds for given  $x, y \in \Omega$  and for any  $\lambda \in [0, 1]$  that

$$f(x + \lambda(y - x)) - f(x) \leq \lambda(f(y) - f(x)).$$

and therefore that

$$\nabla f(x)^T (y - x) = \lim_{\lambda \rightarrow 0} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq \frac{f(y) - f(x)}{y - x}.$$

“ $\Leftarrow$ ”: To prove that  $z = x + \lambda(y - x) = (1 - \lambda)x + \lambda y$  holds that  $f(z) \leq (1 - \lambda)f(x) + \lambda f(y)$ , we can use the equation from Theorem 9.1 twice to get

$$f(x) \geq f(z) + \nabla f(z)^T (x - z) \quad \text{and} \quad f(y) \geq f(z) + \nabla f(z)^T (y - z),$$

which yield, when weighted with  $(1 - \lambda)$  and  $\lambda$  respectively, that

$$(1 - \lambda)f(x) + \lambda f(y) \geq f(z) + \underbrace{\nabla f(z)^T [(1 - \lambda)(x - z) + \lambda(y - z)]}_{=0}$$

□

### Theorem 9.2: Generalized Inequality for Symmetric Matrices

We write for a symmetric matrix  $B = B^T$ ,  $B \in \mathbb{R}^{n \times n}$  that “ $B \succeq 0$ ” if and only if  $B$  is positive semi-definite, i.e.,

$$\forall z \in \mathbb{R}^n : z^T B z \geq 0,$$

or, equivalently, if all (real) eigenvalues of  $B$  are non-negative.

### Theorem 9.3: Convexity for $C^2$ functions

Assume that  $f : \Omega \rightarrow \mathbb{R}$  is twice continuously differentiable and  $\Omega$  is convex and open. Then holds that  $f$  is convex if and only if

$$\forall x \in \Omega : \nabla^2 f(x) \succeq 0,$$

i.e. the Hessian of  $f$  is positive semi-definite.

### Proof 9.2: Convexity for $C^2$ functions

To prove that the Theorem 9.1 implies the Theorem 9.3, we can use a second order Taylor expansion of  $y$  at  $x$  in an arbitrary direction  $p$ :

$$f(x + tp) = f(x) + t \nabla f(x)^T p + \frac{t^2}{2} p^T \nabla^2 f(x) p + o(t^2 \|p\|).$$

From this we obtain that

$$p^T \nabla^2 f(x) p = \lim_{t \rightarrow 0} \frac{2}{t^2} \left( \underbrace{f(x + tp) - f(x) - t \nabla f(x)^T p}_{(9.1): \geq 0} \right) \geq 0.$$

Conversely, to prove the other direction, we use the Taylor rest term formula with some arbitrary  $\theta \in (0, 1)$ :

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \underbrace{\frac{1}{2} t^2 (y - x)^T \nabla^2 f(x + \theta(y - x)) (y - x)}_{(9.3): \geq 0}.$$

□

### Property 9.1: Convexity perserving operations on convex functions

The following operations preserve the convexity of a function:

1. Non-negative weighted sum: Suppose that  $\forall i \in [1, m] : f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  are convex functions and  $\forall i \in [1, m] : \lambda_i \geq 0$ . Then the function

$$f(x) = \sum_{i=1}^m \lambda_i f_i(x)$$

is convex.

2. Affine input transformation: If  $f : \Omega \rightarrow \mathbb{R}$  is convex, then also

$$A \in \mathbb{R}^{n \times m} : \tilde{f}(x) = f(Ax + b)$$

is convex on the domain  $\tilde{\Omega} = \{x \mid Ax + b \in \Omega\}$ .

3. Concatenation with monotone convex function: If  $f : \Omega \rightarrow \mathbb{R}$  is convex and  $g : \mathbb{R} \rightarrow \mathbb{R}$  is convex and monotonely increasing, then the composition  $g \circ f$  is convex.
4. Pointwise supremum: The supremum over a set of convex functions  $f_i(x)$ ,  $i \in I$ , where  $I$  can be an infinite set, i.e.,

$$f(x) = \sup_{i \in I} f_i(x)$$

is convex.

5. Composition: Let  $h : \mathbb{R}^m \rightarrow \mathbb{R}$  and  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$  with  $g = (g_1, \dots, g_m)$ . Then  $f(x) = (h \circ g)(x) = h(g(x))$  is convex if any of the followings holds:
- (a)  $h$  is convex and non-decreasing in each argument and  $\forall i \in [1, m] : g_i$  is convex.
  - (b)  $h$  is convex and non-decreasing in each argument and  $\forall i \in [1, m] : g_i$  is concave, i.e.  $-g_i$  is convex.

### Proof 9.3: Convexity perserving operations on convex functions

1. TODO (use the definition of convexity).
2. TODO (use the definition of convexity).
3. Recall that  $g$  is a convex and monotonely increasing function, then:

$$\nabla^2(g \circ f) = \underbrace{g''(f(x))}_{\succeq 0} \underbrace{\nabla f(x) \nabla f(x)^T}_{\succeq 0} + \underbrace{g'(f(x))}_{\succeq 0} \underbrace{\nabla^2 f(x)}_{\succeq 0} \succeq 0,$$

i.e.  $g \circ f$  is convex, since the Hessian is positive semi-definite.

4. Epigraph of  $f$  is the intersection of the epigraphs of  $f_i$ , which are convex.
5. Recall that  $g_i$  is convex and that  $h$  is convex and non-decreasing in each argument. Then:

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &= h(g(\lambda x + (1 - \lambda)y)) \\ &\leq h(\lambda g(x) + (1 - \lambda)g(y)) \end{aligned} \tag{1}$$

$$\begin{aligned} &\leq \lambda h(g(x)) + (1 - \lambda)h(g(y)) \\ &= \lambda f(x) + (1 - \lambda)f(y), \end{aligned} \tag{2}$$

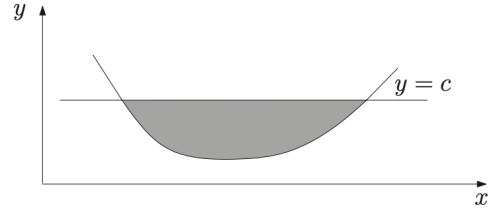
where we used the convexity of  $g_i$  and the fact that  $h$  is non-decreasing to obtain inequality (1), and the convexity of  $h$  to obtain inequality (2). □

### Theorem 9.4: Convexity of Sublevel subsets

If  $f : \Omega \rightarrow \mathbb{R}$  is a convex function, then all its level sets

$$\text{lev}_{\leq \gamma} f = \{ x \in \Omega \mid f(x) \leq \gamma \}$$

are convex.



#### Proof 9.4: Convexity of Sublevel subsets

If  $f(x) \leq c$  and  $f(y) \leq c$ , then for any  $\lambda \in [0, 1]$  holds also

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y) \leq \underbrace{(1 - \lambda)c + \lambda c}_{=c}.$$

□

#### Property 9.2: Convexity perserving operations on convex sets

The following operations preserve the convexity of a set:

1. The intersection of finitely or infinitely many convex sets is convex.
2. Affine image: if  $\Omega$  is convex, then for  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$  also the set  $A\Omega + b = \{y \in \mathbb{R}^m \mid \exists x \in \Omega : y = Ax + b\}$  is convex.
3. Affine pre-image: if  $\Omega$  is convex, then for  $A \in \mathbb{R}^{n \times m}$ ,  $b \in \mathbb{R}^n$  also the set  $\{z \in \mathbb{R}^m \mid Az + b \in \Omega\}$  is convex.

#### Example 9.1: Convex Feasible set

If  $\forall i \in [1, m] : f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  are convex functions, then the set

$$\Omega = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0, i \in [1, m]\}$$

is a convex set, because it is the intersection of sublevel sets  $\Omega_i$  of convex functions  $f_i$ , i.e.

$$\begin{aligned} \Omega &= \bigcap_{i=1}^m \Omega_i \\ &= \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid f_i(x) \leq 0\}. \end{aligned}$$

#### Theorem 9.5: Optimality condition for convex problems

Regard a convex optimization problem with continuously differentiable objective function  $f$ . A point

$x^* \in \Omega$  is a global optimizer if and only if

$$\forall y \in \Omega : \nabla f(x^*)^T (y - x^*) \geq 0.$$

**Proof 9.5: Optimality condition for convex problems**

“ $\Rightarrow$ ”: Assume for the sake of contradiction that  $\exists y \in \Omega : \nabla f(x^*)^T (y - x^*) < 0$ , then we could regard a Taylor expansion

$$f(x^* + \lambda(y - x^*)) = f(x^*) + \lambda \underbrace{\nabla f(x^*)^T (y - x^*)}_{<0} + \underbrace{o(\lambda)}_{\rightarrow 0}.$$

This implies that for sufficiently small positive  $\lambda$ , we have

$$f(x^* + \lambda(y - x^*)) < f(x^*),$$

which contradicts the optimality of  $x^*$ .

“ $\Leftarrow$ ”: Due to the  $C^1$  characterization of convexity of  $f$  in Theorem 9.1, we have for any feasible  $y \in \Omega$ :

$$f(y) \geq f(x^*) + \underbrace{\nabla f(x^*)^T (y - x^*)}_{\geq 0} \geq f(x^*),$$

which implies that  $x^*$  is a global optimizer.

□

### Theorem 9.6: Sufficient Condition for Convex NLP

For a nonlinear optimization problem (NLP) in standard form

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && g_i(x) = 0, i \in [1, m], \\ & && h_i(x) \geq 0, j \in [1, p], \end{aligned}$$

the following conditions are necessary for convexity:

- the objective function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  must be convex,
- the constraint set

$$X = \{ x \in \mathbb{R}^n \mid g(x) = 0, h(x) \geq 0 \}$$

must be convex. Since we know that the intersection of convex sets is convex, we can write  $X$  as the intersection of the sets  $G$  and  $H$ :

$$\begin{aligned} X &= \{x \in \mathbb{R}^n \mid g(x) = 0, h(x) \geq 0\} \\ &= \{x \in \mathbb{R}^n \mid g(x) = 0\} \cap \{x \in \mathbb{R}^n \mid h(x) \geq 0\} \\ &= \left( \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid g_i(x) = 0\} \right) \cap \left( \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid h_i(x) \geq 0\} \right) \\ &= G \cap H \end{aligned}$$

Now we must consider the requirements for the sets  $G$  and  $H$  to be convex:

- Suppose  $-H_i = \{x \in \mathbb{R}^n \mid h_i(x) \leq 0\}$ , the zero sublevel set of the function  $h_i$ . If  $h_i$  is convex, the set  $-H_i$  is convex, as seen in Theorem 9.4. Now, consider the case where  $h_i$  is concave. Since  $h_i$  being concave means that  $-h_i$  is convex, the set  $H_i = \{x \in \mathbb{R}^n \mid h_i(x) \geq 0\}$  is a sublevel set of the convex function  $-h_i$ , and is therefore convex. Thus, the set  $H_i$  is convex when  $h_i$  is concave.
- On the other hand,  $G$  is the level set of  $g_i$ . Therefore it is certainly a convex set whenever  $g_i$  is a affine function

$$\forall i \in [1, m] : g_i(x) = a_i^T x + b_i.$$

TODO: examples

## 10 The Lagrangian Function and Duality

### Theorem 10.1: Primal Optimization Problem

We will denote the globally optimal value of the objective function subject to the constraints as the primal optimal value  $p^*$ , i.e.,

$$p^* = \left( \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad g(x) = 0 \wedge h(x) \geq 0 \right).$$



and we will denote this optimization problem as the primal optimization problem.

### Theorem 10.2: Lagrangian Function and Lagrange Multipliers

We define the Lagrangian function to be

$$\begin{aligned}\mathcal{L}(x, \lambda, \mu) &= f(x) + \sum_{i=1}^l \lambda_i g_i(x) + \sum_{i=1}^m \mu_i h_i(x) \\ &= f(x) + \lambda^T g(x) + \mu^T h(x).\end{aligned}$$

Here, we have introduced Lagrange multipliers or dual variables  $\lambda \in \mathbb{R}^l$  and strictly positive  $\mu \in \mathbb{R}^m$ .

### Lemma 10.1: Lower Bound Property of Lagrangian

If  $\tilde{x}$  is a feasible point of (10.1) and  $\mu \geq 0$ , then

$$\mathcal{L}(\tilde{x}, \lambda, \mu) \leq f(\tilde{x}).$$

### Proof 10.1: Lower Bound Property of Lagrangian

Since  $\tilde{x}$  is feasible, we have  $g(\tilde{x}) = 0$  and  $h(\tilde{x}) \geq 0$ . Therefore, with  $\mu \geq 0$ , we have:

$$\mathcal{L}(\tilde{x}, \lambda, \mu) = f(\tilde{x}) + \lambda^T \underbrace{g(\tilde{x})}_{=0} + \underbrace{\mu^T}_{\geq 0} \underbrace{h(\tilde{x})}_{\geq 0} \leq f(\tilde{x}).$$

□

### Theorem 10.3: Lagrange Dual Function

The Lagrange dual function is defined as the unconstrained infimum of the Lagrangian function over  $x$ , for fixed multipliers  $\lambda$  and  $\mu$ :

$$q(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu).$$

### Lemma 10.2: Lower Bound property of Lagrange Dual

If  $\mu \geq 0$ , then the dual function  $q(\lambda, \mu)$  is a lower bound on the primal optimal value  $p^*$ , i.e.,

$$q(\lambda, \mu) \leq p^*.$$

### Proof 10.2: Lower Bound property of Lagrange Dual

Since the Lagrange function is bounded from below by Lemma 10.1, we have

$$q(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu) \leq f(\tilde{x}) \quad \text{for any feasible } \tilde{x}.$$

Naturally, this inequality holds in particular for the global minimizer  $x^*$ , which yields:

$$q(\lambda, \mu) \leq f(x^*) = p^*.$$

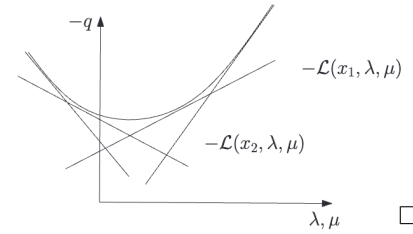
□

### Theorem 10.4: Concavity of Lagrange Dual

The function  $q : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$  is concave, even if the original NLP was not convex.

### Proof 10.3: Concavity of Lagrange Dual

We will show that  $-q$  is convex. The Lagrangian  $\mathcal{L}$  is an affine function in the multipliers  $\lambda$  and  $\mu$ , which in particular implies that  $-\mathcal{L}$  is convex in  $\lambda$  and  $\mu$ . Thus, the function  $-q(\lambda, \mu) = \sup_x -\mathcal{L}(x, \lambda, \mu)$  is the supremum of convex functions in  $\lambda$  and  $\mu$  that are indexed by  $x$ , and therefore convex.



□

### Theorem 10.5: Dual Problem

The dual problem is defined as the convex maximization problem, i.e.,

$$d^* = \left( \max_{\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} q(\lambda, \mu) \quad \text{s.t.} \quad \mu \geq 0 \right).$$

### Theorem 10.6: Weak Duality

Consider a primal-dual pair. Then, the following inequality holds:

$$d^* \leq p^*.$$

### Theorem 10.7: Strong Duality

If the primal optimization problem is convex and the Slater condition (see Theorem 10.8) holds, then strong duality holds, i.e.,

$$d^* = p^*.$$

**Theorem 10.8: Slater Condition**

If there exist one feasible point  $\tilde{x}$  such that all non-linear inequalities are strictly satisfied of a primal convex optimization problem hold, then the Slater condition is satisfied. More explicitly, for a convex problem we must have affine equality constraints,  $g(x) = Ax + b$ , and the inequality constraint functions can be either affine or concave functions, thus without loss of generality assume that the first  $q_1 \leq q$  inequalities are affine and the remaining ones concave. Then the Slater condition holds if and only if there exists an  $\tilde{x}$  such that

$$\begin{aligned} A\tilde{x} + b &= 0, \\ h_i(\tilde{x}) &\geq 0, \quad \text{for } i = 1, \dots, q_1, \\ g_i(\tilde{x}) &> 0, \quad \text{for } i = q_1 + 1, \dots, q. \end{aligned}$$

## Unconstrained Optimization Algorithms

## 11 Optimality Conditions

### Theorem 11.1: Unconstrained optimization Problems

We define the unconstrained optimization problem as

$$\min_{x \in D} f(x),$$

where we regard objective function  $f : D \rightarrow \mathbb{R}$  that are defined on some open domain  $D \subseteq \mathbb{R}^n$ .

### Theorem 11.2: Stationary Point

A point  $\tilde{x}$  is called a stationary point of  $f$  if and only if

$$\nabla f(\tilde{x}) = 0.$$

### Theorem 11.3: Descent Direction

A vector  $p \in \mathbb{R}^n$  is called a descent direction at  $x$  if

$$\nabla f(x)^T p < 0.$$

### Theorem 11.4: First Order Necessary Conditions (FONC)

If  $x^* \in D$  is a local minimizer of  $f : D \rightarrow \mathbb{R}$  and  $f \in C^1$  then

$$\nabla f(x^*) = 0.$$

### Proof 11.1: First Order Necessary Conditions (FONC)

Let us assume for contradiction that  $\nabla f(x^*) \neq 0$ . Then  $p = -\nabla f(x^*)$  would be a descent direction in which the objective could be improved, as follows:

As  $D$  is open and  $f \in C^1$ , we could find a  $t > 0$  that is small enough so that for all  $\tau \in [0, t]$  holds  $x^* + \tau p \in D$  and  $\nabla f(x^* + \tau p)^T p < 0$ . By Taylor's theorem, we would have for some  $\theta \in (0, 1)$  that

$$f(x^* + tp) = f(x^*) + \underbrace{t \nabla f(x^* + \theta tp)^T p}_{< 0} < f(x^*).$$

□

**Theorem 11.5: Second Order Necessary Conditions (SONC)**

If  $x^* \in D$  is a local minimizer of  $f : D \rightarrow \mathbb{R}$  and  $f \in C^2$  then

$$\nabla^2 f(x^*) \succeq 0.$$

**Proof 11.2: Second Order Necessary Conditions (SONC)**

Assume, for the sake of contradiction, that  $\nabla^2 f(x^*) \prec 0$ . This implies the existence of a vector  $p \in \mathbb{R}^n$  such that  $p^T \nabla^2 f(x^*) p < 0$ . In this case, the objective function could be improved in the direction of  $p$ . By choosing a sufficiently small  $t > 0$ , we can ensure that for all  $\tau \in [0, t]$ , the following holds:

$$p^T \nabla^2 f(x^* + \tau p) p < 0.$$

Applying Taylor's theorem, we would have for some  $\theta \in (0, 1)$  that

$$f(x^* + tp) = f(x^*) + \underbrace{t \nabla f(x^*)^T p}_{=0} + \frac{t^2}{2} \underbrace{p^T \nabla^2 f(x^* + \theta p) p}_{<0} < f(x^*),$$

which leads to a contradiction. □

**Theorem 11.6: Convex First Order Sufficient Conditions (cFOSC)**

Assume that  $f : D \rightarrow \mathbb{R}$  is convex and  $f \in C^1$ . If  $x^* \in D$  is a stationary point of  $f$ , then  $x^*$  is a global minimizer of  $f$ .

**Theorem 11.7: Second Order Sufficient Conditions (SOSC)**

Assume that  $f : D \rightarrow \mathbb{R}$  and  $f \in C^2$ . If  $x^* \in D$  is a stationary point of  $f$  and  $\nabla^2 f(x^*) \succ 0$ , then  $x^*$  is a strict local minimizer of  $f$ .

**Proof 11.3: Second Order Sufficient Conditions (SOSC)**

We can choose a sufficiently small closed ball  $B$  around  $x^*$  so that for all  $x \in B$  holds  $\nabla^2 f(x) \succ 0$ . Restricted to this ball, we have a convex problem, so that Theorem 11.6 together with stationarity of  $x^*$  implies that  $x^*$  is a global minimizer of  $f$ . To prove that it is strict, we look for any  $x \in B \setminus x^*$  at the Taylor expansion, which yields with some  $\theta \in (0, 1)$ :

$$f(x) = f(x^*) + \underbrace{\nabla f(x^*)^T (x - x^*)}_{=0} + \frac{1}{2} \underbrace{(x - x^*)^T \nabla^2 f(x^* + \theta(x - x^*)) (x - x^*)}_{>0} > f(x^*).$$

□

### Theorem 11.8: Stability of Parametric Solutions

Assume that  $f : D \times \mathbb{R}^m \rightarrow \mathbb{R}$ , and regard the minimization of  $f(\cdot, \tilde{a})$  for a given fixed value of  $\tilde{a} \in \mathbb{R}^m$ . If  $\tilde{x} \in D$  satisfies the SOSC (see Theorem 11.7), then there is a neighborhood  $\mathcal{N} \subset \mathbb{R}^m$  around  $\tilde{a}$  such that the parametric minimizer function  $x^*(a)$  is well-defined for all  $a \in \mathcal{N}$  (i.e. there is a unique minimizer  $x^*(a)$  for each  $a \in \mathcal{N}$ ), is differentiable in  $\mathcal{N}$ , and  $x^*(\tilde{a}) = \tilde{x}$ . Its derivative at  $\tilde{a}$  is given by

$$\frac{\partial(x^*(\tilde{a}))}{\partial a} = -(\nabla_x^2 f(\tilde{x}, \tilde{a}))^{-1} \frac{\partial(\nabla_x f(\tilde{x}, \tilde{a}))}{\partial a}.$$

Moreover, each such  $x^*(a)$  with  $a \in \mathcal{N}$  satisfies again the SOSC and is thus a strict local minimizer.

### Proof 11.4: Stability of Parametric Solutions

The existence of the differentiable map  $x^* : \mathcal{N} \rightarrow D$  follows from the implicit function theorem applied to the stationarity condition  $\nabla_x f(x^*(a), a) = 0$ . The derivative of  $x^*(a)$  at  $\tilde{a}$  is given by

$$\frac{d(\nabla_x f(x^*(a), a))}{da} = \underbrace{\frac{\partial(\nabla_x f(x^*(a), a))}{\partial x}}_{=\nabla_x^2 f} \frac{\partial x^*(a)}{\partial a} + \frac{\partial(\nabla_x f(x^*(a), a))}{\partial a} = 0$$

The fact that all points  $x^*(a)$  satisfy the SOSC follows from the continuity of the second derivative.  $\square$

## 12 Estimation and Fitting Problems

### Theorem 12.1: Estimation and Fitting Problems

Estimation and fitting problems are optimization problems with a special objective, namely a least squares objective. We define the estimation problem as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|\eta - M(x)\|_2^2,$$

where  $\eta \in \mathbb{R}^m$  is the measurement vector,  $M : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is the model function, and  $x \in \mathbb{R}^n$  is the parameter vector. Many models in estimation and fitting problems are linear functions of  $x$ . If  $M$  is linear,  $M(x) = Jx$ , then  $f(x) = \frac{1}{2} \|\eta - Jx\|_2^2$  which is a convex function, as  $\nabla^2 f(x) = J^T J \succeq 0$ . Therefore local minimizers are found by:

$$\begin{aligned} \nabla f(x) = 0 &\Leftrightarrow J^T Jx^* - J^T \eta = 0 \\ &\Leftrightarrow x^* = \underbrace{(J^T J)^{-1} J^T}_{J^+} \eta \end{aligned}$$

### Theorem 12.2: Pseudo-inverse

$J^+$  is called the pseudo-inverse and is a generalization of the inverse matrix. If  $J^T J \succ 0$ ,  $J^+$  is given

by

$$J^+ = (J^T J)^{-1} J^T.$$

So far,  $(J^T J)^{-1}$  is only defined if  $J^T J \succ 0$ . This holds if and only if  $\text{rank}(J) = n$ , i.e. if the columns of  $J$  are linearly independent.

### 13 Newton Type Optimization

### 14 Globalisation Strategies

### 15 Calculating Derivatives



## Constrained Optimization Algorithms

- 16 Optimality Conditions for Constrained Optimization
- 17 Equality Constrained Optimization Algorithms
- 18 Inequality Constrained Optimization Algorithms
- 19 Optimal Control Problems