Optimization

Hassan OMRAN

Lecture 1: Introduction and Mathematical Foundations

Télécom Physique Strasbourg
Université de Strasbourg

Information on the course

Lecturer : *Hassan OMRAN*

Email: *hassan.omran@icube.unistra.fr*

Lab sessions : *Hassan OMRAN*
Paul Baksic
Rima Saadaoui
Thibault Poignonec

Teaching hours

- *Lectures:* 12.25*h*
- *Exercises:* 2*h*
- *Lab:* 8*h*

Evaluation

- *Written exam*
- *Reports for lab sessions*

Information on the course

Content

1- Introduction and Mathematical Foundations

Unconstrained Optimization:

2- One-Dimensional Search Methods

3- Multi-Dimensional Search Methods

Linear and Nonlinear Least Squares methods:

4- Least Squares Optimization

Constrained Optimization:

5- Linear Programming

6- Constrained Nonlinear Optimization

Outline of the talk

1. Introduction
2. Optimization problem
3. Mathematical foundations
4. Modeling
5. Optimality conditions
6. Solvers

Why study optimization

*Optimization is the science of finding efficiently the **best (optimal)** solution to a problem while taking into account the imposed **restrictions***

- *An interdisciplinary science : applied mathematics, computer science and engineering*
- *It is a rigorous method for decision making, and essential tool for engineers to deal with technological challenges*
- *Also called Mathematical Programming*
- *Has reached a degree of maturity, and is applied in extremely wide spectrum of applications*
- *An active field of research*

Historical background

- *First optimization problem was probably solved by Elyssa (Didon), a Phoenician princess* [1] *.*
- *She arrived to North Africa with her people running away from a struggle in the city of Tyre in the Levant.*
- *She asked the local Berber king to give them land, and he agreed to give them as much land as could be bounded by a hide of an Ox.*
- *She cut the hide into a very thin and long strip and covered a considerable surface near the sea, and founded the city of Carthage.*
- *She solved the Isoperimetric problem of determining the maximum surface to enclose using a strip of a given length*

Dido Purchases Land for the Foundation of Carthage. Engraving by Matthäus Merian the Elder, in *Historische Chronica*, Frankfurt a.M., 1630. Dido's people cut the hide of an ox into thin strips and try to enclose a maximal domain.

Elyssa and her people marking their land using a hide of an Ox

[1] See *Isoperimetrical Problems* by Sir William Thomson http://math.arizona.edu/ dido/lord-kelvin1894.html

Historical background: Early contributions

- *Calculus: Newton, Leibnitz*
- *Calculus of variations: Bernoulli, Euler, Lagrange and Weierstrass*
- *Lagrange multipliers: Lagrange*
- *Application of steepest descend method: Cauchy*

Early contributions to optimization methods can be traced back to $17^{th} - 19^{th}$ century mathematicians

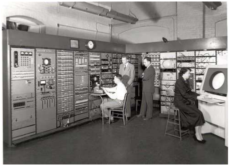
Historical background: WWII

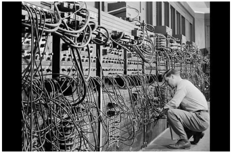
- *The British and the US military faced the problem of allocation of scarce resources*
- *They solicited a team of mathematicians to provide rigorous scientific decision making tools*
- *They gave it the name of Research on Military Operations*
- *Today the field is known as Operations Research and it is applied in many areas: military, economics, health care and transportation systems*

Planning and decision making related problems in WWII, had a great influence on the development of optimization

Historical background: the post-war years till today

- *The invention of digital computer was a paramount in the development of optimization methods*
- *It motivated a massive research efforts, which led to great breakthroughs*
- *Emergence of several domains in optimization*
- *Since then, optimization is a very active research domain*





Electronic Numerical Integrator And Computer (ENIAC)

Historical background: the post-war years till today

Some milestones:

→ *Kantorovich: contribution to formulating linear programming and to economics*[2].

→ *Dantzig: simplex method*

→ *Kuhn and Tucker: contributions to the foundation of nonlinear programming*

→ *Bellman: principle of optimality*

→ *Khachiyan: ellipsoid algorithm*

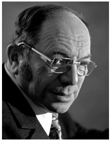
→ *Karmarkar: interior point method*

→ *Rao: fuzzy approaches*

→ *Holland: genetic algorithms*

→ *Dorigo: ant colony methods*

→ *...*



Leonid Kantorovich.

George Dantzig.

[2] He was one of several Nobel Prize laureates in economics who have worked with optimization.

Applications

Optimization is an essential tool in several domains: engineering, economics, health care .. etc

Applications in engineering:

- *Analysis and synthesis of control systems*
- *Trajectory planning of robotic systems*
- *Design and operation of electrical network*
- *Energy efficient transportation systems*
- *Civil engineering structure design*
- *Production and storage planning*
- *Organization of maintenance operation*
- *Operation of chemical processes*
- *Telecommunication network design*
- *Design, planning and control of renewable energy*

Application in health care:

- *Prostheses design and control*
- *Design of rehabilitation systems*
- *Operating room scheduling*
- *Radiation treatment planning*
- *Blood bank management policies*
- *Allocation of donated organs*
- *Vaccine selection algorithm*
- *Human and resource allocation in hospitals*
- *Patients flow optimization in emergency department*

Notation

- $\mathbb{R}$ *denotes the set of real numbers.*
- $\mathbb{R}^+$ *denotes the set of non-negative numbers.*
- $\mathbb{R}^n$ *denotes the Euclidean space of dimension n.*
- $\mathbb{R}^{n \times m}$ *is the set of all $n \times m$ real matrices.*
- *The superscript "T" stands for matrix transpose.*
- *For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the Euclidean inner product is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$. The Euclidean norm is denoted by $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.*
- *For $\mathbf{x} = [x_1 \ x_2 \ldots x_n]^T \in \mathbb{R}^n$ and $\mathbf{y} = [y_1 \ y_2 \ldots y_n]^T \in \mathbb{R}^n$, the inequality $\mathbf{x} \geq \mathbf{y}$ is interpreted in a component-wise manner, i.e., $x_i \geq y_i$ for $i = 1, \ldots, n$.*

Ingredients of an optimization problem

Mathematically speaking, optimization is the problem of finding the minimum (or the maximum [3]) of an objective function subject to constraints.

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ s.t. & \mathbf{x} \in \mathcal{S} \subseteq \mathbb{K} \end{array}$$

- *The elements of the vector* $\mathbf{x}$ *are the* ***decision variables***
- $\mathbb{K}$ *is the universe or* ***domain*** *of the decision variables (example:* $\mathbb{K} = \mathbb{R}^n$*)*
- $f : \mathbb{K} \to \mathbb{R} \cup \{\pm\infty\}$ *is the* ***cost function*** *or* ***objective function***
- *The set* $\mathcal{S} \subseteq \mathbb{K}$ *is the* ***constraints set*** *which describes the* ***feasible solutions***

[3] The maximization of $f(\cdot)$ is equivalent to the minimization of $-f(\cdot)$. Thus, without loss of generality we will consider mainly the minimization problem.

Ingredients of an optimization problem

Mathematically speaking, optimization is the problem of finding the minimum (or the maximum [3]) of an objective function subject to constraints.

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{S} \subseteq \mathbb{K} \end{aligned}$$

- *The elements of the vector $\mathbf{x}$ are the **decision variables***
- *$\mathbb{K}$ is the universe or **domain** of the decision variables (example: $\mathbb{K} = \mathbb{R}^n$)*
- *$f : \mathbb{K} \to \mathbb{R} \cup \{\pm\infty\}$ is the **cost function** or **objective function***
- *The set $\mathcal{S} \subseteq \mathbb{K}$ is the **constraints set** which describes the **feasible solutions***

[3] The maximization of $f(\cdot)$ is equivalent to the minimization of $-f(\cdot)$. Thus, without loss of generality we will consider mainly the minimization problem.

Ingredients of an optimization problem

Mathematically speaking, optimization is the problem of finding the minimum (or the maximum) of an objective function subject to constraints.

Common problem forms:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{g}(\mathbf{x}) \le \mathbf{0} \end{aligned}$$

- *The elements of the vector* $\mathbf{x}$ *are the* ***decision variables***
- $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ *is the* ***cost function*** *or* ***objective function***
- $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$ *is the* ***equality constraints function***
- $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^p$ *is the* ***inequality constraints function***

In this case, the set of feasible solutions is defined by:

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{h}(\mathbf{x}) = \mathbf{0}, \ \ \mathbf{g}(\mathbf{x}) \le \mathbf{0}\}.$$

Ingredients of an optimization problem

Mathematically speaking, optimization is the problem of finding the minimum (or the maximum) of an objective function subject to constraints.

Common problem forms:

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ s.t. & h_i(\mathbf{x}) = 0 \quad i \in \{1, \dots, m\} \\ & g_j(\mathbf{x}) \leq 0 \quad j \in \{1, \dots, p\} \end{array}$$

- *The elements of the vector $\mathbf{x}$ are the **decision variables***
- $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ *is the **cost function** or **objective function***
- $h_i : \mathbb{R}^n \to \mathbb{R},\ i \in \{1, \dots, m\}$ *are the **equality constraints functions***
- $g_j : \mathbb{R}^n \to \mathbb{R},\ j \in \{1, \dots, p\}$ *are the **inequality constraints functions***

In this case, the set of feasible solutions is defined by:

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^{n \times n} : h_i(\mathbf{x}) = 0\ i \in \{1, \dots, m\},\ \ g_j(\mathbf{x}) \leq 0\ j \in \{1, \dots, p\}\}.$$

Fundamental questions

What does it mean to solve the following problem ?

$$f^\star = \min_{\mathbf{x}} \quad f(\mathbf{x}) \qquad \text{s.t.} \quad \mathbf{x} \in \mathcal{S} \subseteq \mathbb{K} \tag{1}$$

Recap: Infimum and minimum of functions

Any function $f : \mathcal{S} \subseteq \mathbb{K} \to \mathbb{R}$ has an ***infimum*** $l = \inf_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) \in \mathbb{R} \cup \{-\infty\}$ *which is uniquely defined by the following properties:*

- $l \leq f(\mathbf{x})$ *for all* $\mathbf{x} \in \mathcal{S}$.
- l *is finite: for all* $\epsilon > 0$ *exists* $\mathbf{x} \in \mathcal{S}$ *with* $f(\mathbf{x}) \leq l + \epsilon$.
- $l = -\infty$: *for all* $\epsilon > 0$ *exists* $\mathbf{x} \in \mathcal{S}$ *such that* $f(\mathbf{x}) \leq -\epsilon$.

If there exists $\mathbf{x}_0 \in \mathcal{S}$ *with* $f(\mathbf{x}_0) = l$, *we say that* $f(\cdot)$ *attains its* ***minimum*** $l = \min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x})$.
If it exists, the minimum is defined through the properties:

- $l \leq f(\mathbf{x})$ *for all* $\mathbf{x} \in \mathcal{S}$.
- *There exists some* $\mathbf{x}_0 \in \mathcal{S}$ *such that* $f(\mathbf{x}_0) = l$.

Fundamental questions

- *What is the least cost? Find the **optimal value***

$$f^\star = \inf_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x}) \geq -\infty.$$

Conventions:

  - *If $f^\star = -\infty$ the problem is **unbounded below**.*
  - *If $\mathcal{S} = \mathbb{R}^n$ then the problem is **unconstrained**.*
  - *If $\mathcal{S} = \emptyset$ then the problem is **infeasible**, and we set $f^\star = +\infty$.*

- *Is there an **optimal solution** ? That is, does there exist*

$$\boldsymbol{x}^\star \in \mathcal{S}, \ \ s.t. \ \ f^\star = f(\boldsymbol{x}^\star)$$

*We call $\boldsymbol{x}^\star$ a **minimizer**. In this case, the **minimum** is attained*

$$f^\star = \min_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x})$$

- *The set of all optimal solutions is*

$$\mathcal{S}^\star = \arg\min_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x}) = \{\boldsymbol{x}^\star \in \mathcal{S} : f(\boldsymbol{x}^\star) = f^\star\}$$

Note that $\mathcal{S}^\star \subseteq \mathcal{S}$ is empty if and only if $f^\star$ is not attained at any point $\boldsymbol{x}^\star \in \mathcal{S}$.

- *Note that the optimal solution might not exist. When it exists it might not be unique.*

Local and global of minimizers

Solving the optimization problem (1) is finding $f^\star$ and then (if $f^\star$ is finite) finding $\mathbf{x}^\star \in \mathcal{S}^*$, or concluding that $\mathcal{S}^* = \emptyset$.

Global minimizer

*A point $\mathbf{x}^\star$ is a **global minimizer** if $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x} \in \mathcal{S} \backslash \{\mathbf{x}^\star\}$.*

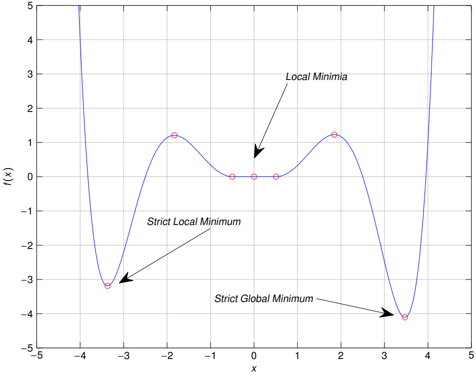
It is often desirable to obtain a global minimizer. However, this might be difficult, that is why several methods look for local minimizers.

Local minimizer

*A point $\mathbf{x}^\star$ is a **local minimizer** if $\exists \epsilon > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for all $\mathbf{x} \in \mathcal{S} \backslash \{\mathbf{x}^\star\}$ such that $\|\mathbf{x} - \mathbf{x}^\star\| < \epsilon$.*

*If we replace $\geq$ by $>$, then we have a **strict global minimizer** and **strict local minimizer**, respectively.*

Local and global of minimizers



Examples of types of minima.

Examples

- *Example of a global minimum:*

$$\begin{aligned} \min_{x} \quad & x^2 \\ s.t. \quad & x \geq -1 \\ & x \leq +1 \end{aligned}$$

- *Example of an infeasible problem:*

$$\begin{aligned} \min_{x} \quad & x^2 \\ s.t. \quad & x \leq -1 \\ & x \geq +1 \end{aligned}$$

Examples

- *Example of an unbounded problem:*

$$\begin{aligned} \min_{x} \quad & x \\ s.t. \quad & x \leq 0 \end{aligned}$$

- *Example of an unattained optimal value*

$$\begin{aligned} \min_{x} \quad & e^{-x} \\ s.t. \quad & x \geq 0 \end{aligned}$$

Here, the infimum is $f^\star = 0$, *but* $\mathcal{S}^\star = \emptyset$, *and the value* 0 *is not attained for a finite* x.

Classifications of optimization problems

Optimization problems can be classified according to several criteria:

Objective function:

- *Function of a single variable*
- *Linear function*
- *Quadratic function*
- *Sum of squares*
- *Smooth nonlinear function*
- *Non-smooth nonlinear function*
- *Convex function*

Constraints:

- *No constraints*
- *Simple bounds*
- *Linear functions*
- *Quadratic function*
- *Smooth nonlinear functions*
- *Non-smooth nonlinear functions*
- *Convex functions*

Classifications of optimization problems

Optimization problems can be classified according to several criteria:

Objective function:

- *Function of a single variable*
- *Linear function*
- *Quadratic function*
- *Sum of squares*
- *Smooth nonlinear function*
- *Non-smooth nonlinear function*
- *Convex function*

Constraints:

- *No constraints*
- *Simple bounds*
- *Linear functions*
- *Quadratic function*
- *Smooth nonlinear functions*
- *Non-smooth nonlinear functions*
- *Convex functions*

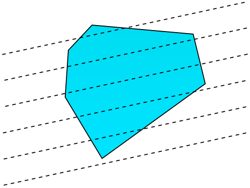
Classifications of optimization problems

Linear Program (LP)

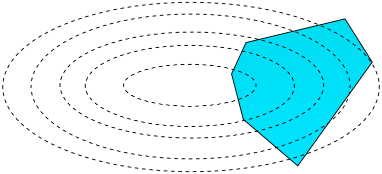
$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \\ & G\mathbf{x} \leq \mathbf{h} \end{aligned}$$

Quadratic Program (QP)

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{q}^T \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \\ & G\mathbf{x} \leq \mathbf{h} \end{aligned}$$



Linear programming on a polytope.



Convex quadratic programming on a polytope.

Classifications of optimization problems

Convex Program

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{x} \in \mathcal{S} \subseteq \mathbb{K} \end{aligned}$$

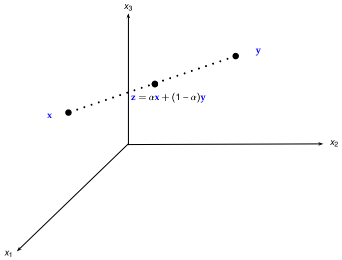
With $f(\cdot)$ a convex function and $\mathcal{S}$ a convex set. The notion of convexity will be presented later.

Line segments

Line segments

*The **line segment** between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the set of points on the straight line joining the two points:*

$$\{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} = \alpha \mathbf{x} + (1-\alpha)\mathbf{y},\ \alpha \in [0,1]\}$$

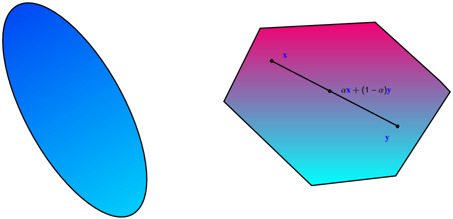


Convex sets

A set $\mathcal{S}$ is convex if

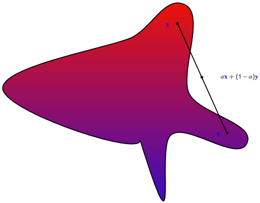
$$\alpha \mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{S}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{S}, \ \forall \alpha \in [0,1]$$

A line segments which starts and ends in $\mathcal{S}$, belongs entirely to $\mathcal{S}$



Examples of convex sets.

Convex sets

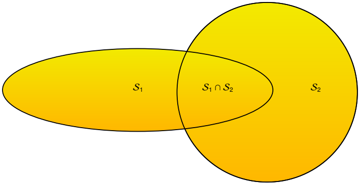


Example of a non convex set.

Convex sets

Theorem 1

The intersection of any collection of convex sets is convex



Intersections of two convex sets.

Note that the union of convex sets is not necessarily convex.

Open, closed, bounded and compact sets

Neighborhood

*A **neighborhood** of point $\mathbf{x} \in \mathbb{R}^n$ is the set*

$$\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| < \epsilon\}, \quad \textit{with } \epsilon > 0.$$

*This is also called a **ball** of radius ϵ.*

Interior point

*A point $\mathbf{x} \in \mathcal{S} \subseteq \mathbb{R}^n$ is said to be an **interior point** of the set $\mathcal{S}$ if the last contains a neighborhood of $\mathbf{x}$. The set of all interior points of $\mathcal{S}$ is called the **interior of** $\mathcal{S}$.*

Boundary point

*A point $\mathbf{x} \in \mathcal{S} \subseteq \mathbb{R}^n$ is said to be a **boundary point** of the set $\mathcal{S}$ if every neighborhood of $\mathbf{x}$ contains a point in $\mathcal{S}$ and a point not in $\mathcal{S}$. The set of all boundary points of $\mathcal{S}$ is called the **boundary of** $\mathcal{S}$.*

Note that a boundary point of $\mathcal{S}$ may or may not be an element of $\mathcal{S}$.

Open, closed, bounded and compact sets

Open sets

*A set $\mathcal{S}$ is said to be **open** if it contains a neighborhood of each of its points.*

Equivalently, each point of $\mathcal{S}$ is an interior point.

Closed sets

*A set $\mathcal{S}$ is said to be **closed** if it contains its boundary.*

Bounded sets

*A set $\mathcal{S}$ is said to be **bounded** if it is contained in a ball of finite radius.*

Compact sets

*A set $\mathcal{S}$ is said to be **compact** if it is closed and bounded.*

Open, closed, bounded and compact sets

Compact sets

*A set $\mathcal{S}$ is said to be **compact** if it is closed and bounded.*

Why are compact sets important ?

Theorem 2 (Theorem of Weierstrass)

*Consider a nonempty **compact** set $\mathcal{S} \subset \mathbb{R}^n$ and let $f : \mathcal{S} \to \mathbb{R}$ be a **continuous** function. Then, $\exists\, \boldsymbol{x}_0 \in \mathcal{S}$ such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ for all $\boldsymbol{x} \in \mathcal{S}$. That is, $f(\cdot)$ achieves its minimum on $\mathcal{S}$.*

Implications

The closeness of $\mathcal{S}$ is crucial. If $\mathcal{S}$ is not closed, a sequence generated in $\mathcal{S}$ may converge to a point outside of $\mathcal{S}$.

For example, consider minimizing $f(x) = \frac{1}{x}$ for $1 > x > 0$. In this case, a numeral algorithm which generates a sequence $\{x_k\}$ such that $1 > x_k > 0$ and $x_k \to 1$ will (for a given accuracy) lead to the false conclusion that 1 is a minimizer.

Hyperplanes

Hyperplanes

Consider the vector $\mathbf{u} = [u_1 \dots u_n]^T \in \mathbb{R}^n$ *with at least one of its elements* $u_i \neq 0$ *and* $v \in \mathbb{R}$*. The set of all points* $\mathbf{x} \in \mathbb{R}^n$ *that satisfy the equation*

$$\mathbf{u}^T\mathbf{x} = u_1x_1 + \cdots + u_nx_n = v$$

is called a ***hyperplane*** *of the space* $\mathbb{R}^n$.

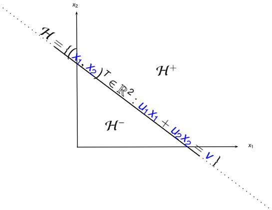
$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{u}^T\mathbf{x} = v\}$$

Consider an arbitrary point $\mathbf{a}$ *from the hyperplane, that is* $\mathbf{u}^T\mathbf{a} = v$*. Then, for any point* $\mathbf{x}$ *in the hyperplane we have:*

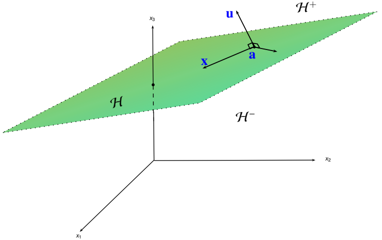
$$\begin{aligned} \mathbf{u}^T\mathbf{x} - v &= 0 \\ \mathbf{u}^T\mathbf{x} - \mathbf{u}^T\mathbf{a} &= 0 \\ \langle \mathbf{u}, \mathbf{x} - \mathbf{a} \rangle &= 0 \end{aligned}$$

Therefore, the hyperplane $\mathcal{H}$ *consists of the points* $\mathbf{x}$ *for which the vectors* $\mathbf{u}$ *and* $\mathbf{x} - \mathbf{a}$ *are orthogonal.*

Hyperplanes



A hyperplane in $\mathbb{R}^2$.



A hyperplane in $\mathbb{R}^3$. $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^3 : u_1 x_1 + u_2 x_2 + u_3 x_3 = v\}$ or equivalently $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^3 : \langle \mathbf{u}^T, \mathbf{x} - \mathbf{a} \rangle = 0\}$.

A hyperplane is not necessarily a subspace of $\mathbb{R}^n$ (unless it passes by the origin). The hyperplane $\mathcal{H}$ devises the space $\mathbb{R}^n$ into two half-spaces:

$$\mathcal{H}^+ = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{u}^T \boldsymbol{x} \geq v\} \qquad \mathcal{H}^- = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{u}^T \boldsymbol{x} \leq v\}$$

Note that half-spaces are convex.

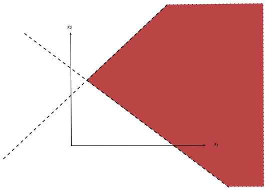
Polytopes and Polyhedras

Polyhedron

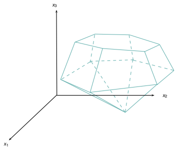
The intersection of a finite number of half-spaces is called a ***polyhedron***

Polytope

A nonempty bounded polyhedron is called a ***polytope****.*



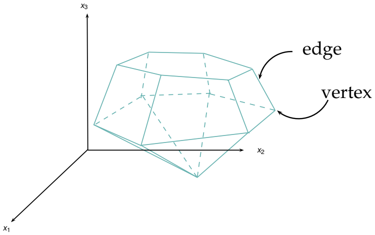
An unbounded polyhedron.



A polytope.

Polytopes and Polyhedras

- *Note that a polyhedron is convex (why ?).*
- *The boundary of any n-dimensional polyhedron consists of a finite number of $(n-1)$-dimensional polyhedra, are called* ***faces****.*
- *Each of these faces has in turn $(n-2)$-dimensional faces, whom are considered to be also faces of the original n-dimensional polyhedron.*
- *Thus, an n-dimensional polyhedron has faces of dimensions $n-1, n-2, \ldots, 1, 0$.*
- *A 1-dimensional face is called an* ***edge****.*
- *A 0-dimensional face is called a* ***vertex****.*



A polytope.

Derivation: the Jacobian

The Jacobian

Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$, then the matrix $J(\boldsymbol{x})$ defined as

$$f(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix}, \qquad J(\boldsymbol{x}) = Df(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} \tag{1}$$

*is called the **Jacobian matrix**, or the **derivative matrix** of $f(\cdot)$ at $\boldsymbol{x}$.*

Theorem 3

Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$ which is differentiable at $\boldsymbol{x}_0$. The derivative of $f(\cdot)$ is uniquely determined by the $m \times n$ derivative matrix $Df(\boldsymbol{x}_0)$. The best affine approximation of $f(\cdot)$ near $\boldsymbol{x}_0$ is then given by

$$f(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}) + r(\boldsymbol{x}) \quad \text{with} \quad \mathcal{A}(\boldsymbol{x}) = f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0), \quad \text{where} \quad \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{\|r(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0$$

Derivation: the gradient

The gradient

*If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is called **gradient** of $f(\cdot)$ and it is defined by*

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} = Df(\boldsymbol{x})^T \tag{2}$$

Note that for $f : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian (1) *can be written as*

$$J(\boldsymbol{x}) = \begin{bmatrix} Df_1(\boldsymbol{x}) \\ Df_2(\boldsymbol{x}) \\ \vdots \\ Df_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \nabla f_1(\boldsymbol{x})^T \\ \nabla f_2(\boldsymbol{x})^T \\ \vdots \\ \nabla f_m(\boldsymbol{x})^T \end{bmatrix}$$

Derivation: the Hessian

The Hessian

Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. If ∇f is differentiable, then the function $f(\cdot)$ is ***twice differentiable*** *and we call $D^2 f(\boldsymbol{x})$ the* ***Hessian*** *matrix of $f(\cdot)$ at $\boldsymbol{x}$:*

$$D^2 f(\boldsymbol{x}) = D(\nabla f(\boldsymbol{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial {x_2}^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial {x_n}^2}(\boldsymbol{x}) \end{bmatrix} \tag{3}$$

The symmetry of the Hessian

The Hessian matrix is symmetric if the function $f(\cdot)$ is twice continuously differentiable. That is, the components of $f(\cdot)$ have continuous partial derivatives of order 2.

Derivation rules

The chain rule

Let $f : \mathcal{S} \to \mathbb{R}$ be a differentiable function on an open set $\mathcal{S} \subset \mathbb{R}^n$, and $g : (a, b) \to \mathcal{S}$ be a differentiable function on (a, b). The function $h : (a, b) \to \mathbb{R}$ given by $h(t) = f(g(t))$ is differentiable on (a, b), and

$$\dot{h}(t) = \frac{dh}{dt}(t) = Df(g(t))Dg(t) = \nabla f(g(t))^T \begin{bmatrix} \dot{g}_1(t) \\ \vdots \\ \dot{g}_n(t) \end{bmatrix} \tag{4}$$

The product rule

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ defined by $h(\boldsymbol{x}) = f(\boldsymbol{x})^T g(\boldsymbol{x})$. Then, $h(\cdot)$ is also differentiable and

$$Dh(\boldsymbol{x}) = f(\boldsymbol{x})^T Dg(\boldsymbol{x}) + g(\boldsymbol{x})^T Df(\boldsymbol{x}) \tag{5}$$

Derivation: some examples

Example

Consider a given matrix $A \in \mathbb{R}^{m\times n}$ and a given vectors $\boldsymbol{y}_1 \in \mathbb{R}^n$ and $\boldsymbol{y}_2 \in \mathbb{R}^m$

$$D(\boldsymbol{y}_1^T\boldsymbol{x}) = ? \qquad \nabla(\boldsymbol{y}_2^T A\boldsymbol{x}) = ?$$

Example

Consider a given matrix $Q \in \mathbb{R}^{n\times n}$. Find

$$D(\boldsymbol{x}^T Q\boldsymbol{x}) \quad = \quad ?$$

Repeat the question for a symmetric matrix $Q = Q^T$:

Example

Consider a given symmetric matrix $Q = Q^T \in \mathbb{R}^{n\times n}$ and a vector $q \in \mathbb{R}^n$. Find

$$\nabla(\frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x}) = ? \qquad D^2(\frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x}) = ?$$

Derivation: some examples

Example

Consider a given matrix $A \in \mathbb{R}^{m\times n}$ and a given vectors $\mathbf{y}_1 \in \mathbb{R}^n$ and $\mathbf{y}_2 \in \mathbb{R}^m$

$$D(\mathbf{y}_1^T\mathbf{x}) = \mathbf{y}_1^T \qquad \nabla(\mathbf{y}_2^T A\mathbf{x}) = A^T\mathbf{y}_2$$

Example

Consider a given matrix $Q \in \mathbb{R}^{n\times n}$. Find

$$D(\mathbf{x}^TQ\mathbf{x}) = \mathbf{x}^T(Q+Q^T)$$

Repeat the question for a symmetric matrix $Q = Q^T$: $D(\mathbf{x}^TQ\mathbf{x}) = 2\mathbf{x}^TQ$

Example

Consider a given symmetric matrix $Q = Q^T \in \mathbb{R}^{n\times n}$ and a vector $q \in \mathbb{R}^n$. Find

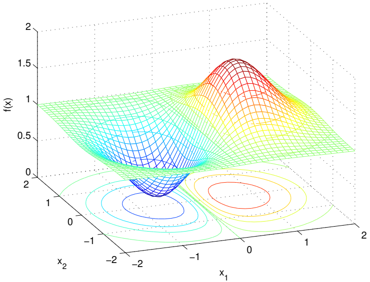
$$\nabla(\frac{1}{2}\mathbf{x}^TQ\mathbf{x} + \mathbf{q}^T\mathbf{x}) = Q\mathbf{x} + \mathbf{q} \qquad D^2(\frac{1}{2}\mathbf{x}^TQ\mathbf{x} + \mathbf{q}^T\mathbf{x}) = Q$$

Level sets

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, and a scalar c. We call the set of points

$$\xi = \{\mathbf{x} : f(\mathbf{x}) = c\}$$

*the **level set** of the function $f(\cdot)$ at level c.*



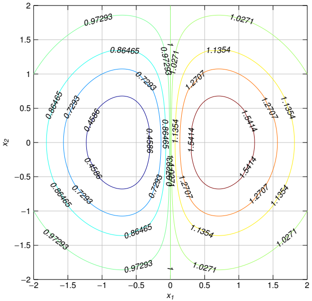
Level sets of $f(\mathbf{x}) = 1 + 2x_1 e^{-x_1{}^2 - x_2{}^2}$.

Level sets

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, and a scalar c. We call the set of points

$$\xi = \{\mathbf{x} : f(\mathbf{x}) = c\}$$

*the **level set** of the function $f(\cdot)$ at level c.*



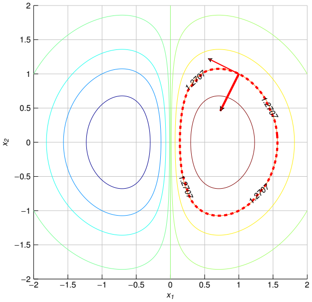
Level sets of $f(\mathbf{x}) = 1 + 2x_1 e^{-x_1^2 - x_2^2}$.

Level sets

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, and a scalar c. We call the set of points

$$\xi = \{\boldsymbol{x} : f(\boldsymbol{x}) = c\}$$

*the **level set** of the function $f(\cdot)$ at level c. Note that the gradient is orthogonal to the level set !*



Level sets of $f(\mathbf{x}) = 1 + 2x_1 e^{-x_1{}^2 - x_2{}^2}$.

Level sets

Consider the curve defined by the level set which passes by $\mathbf{x}_0$*, that is* $\xi = \{\mathbf{x} : f(\mathbf{x}) = f(\mathbf{x}_0)\}$.

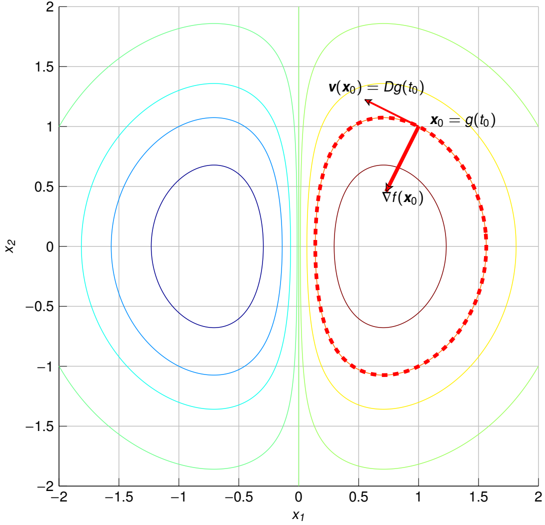
Let $g : \mathbb{R} \rightarrow \mathbb{R}^n$ *be a parameterizing function of the considered level set :* $g(t) = \mathbf{x} \in \xi$ *and* $g(t_0) = \mathbf{x}_0$.
Note that $Dg(t_0) = \mathbf{v}(\mathbf{x}_0)$ *is a tangent vector to* ξ *at* $\mathbf{x}_0$.

Define $h(t) = f(g(t))$. *Note that* $h(t) = f(\mathbf{x}_0) = cte$*, thus* $\frac{dh}{dt}(t_0) = 0$.
By the chain rule at t_0 *we obtain*

$$\begin{aligned} Df(g(t_0))Dg(t_0) &= \frac{dh}{dt}(t_0) \\ Df(\mathbf{x}_0)\mathbf{v}(\mathbf{x}_0) &= 0 \\ \nabla f(\mathbf{x}_0)^T\mathbf{v}(\mathbf{x}_0) &= 0 \\ \langle \nabla f(\mathbf{x}_0), \mathbf{v}(\mathbf{x}_0)\rangle &= 0 \end{aligned}$$

This shows that $\nabla f(\mathbf{x}_0) \perp \mathbf{v}(\mathbf{x}_0)\ \forall \mathbf{x}_0 \in \mathcal{S}$. *We say that* $\nabla f(\mathbf{x}_0)$ *is orthogonal to* ξ.

Level sets

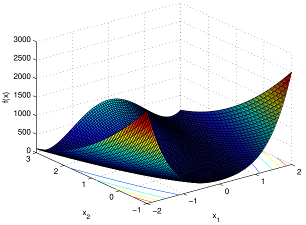


$\boldsymbol{v}(\boldsymbol{x}_0) = Dg(t_0)$
$\boldsymbol{x}_0 = g(t_0)$
$\nabla f(\boldsymbol{x}_0)$
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
x_2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
x_1

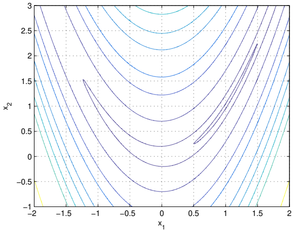
Level sets

The following figures show the graph and the level sets of a famous function called ***Rosenbrock's function***, $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by*

$$f(\boldsymbol{x}) = 100(x_2 - {x_1}^2)^2 + (1 - x_1)^2$$



Graph of Rosenbrock's function.



Level sets of Rosenbrock's function.

Directional derivative

Consider the real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ *and a non-zero vector* $\mathbf{d} = [d_1 \dots d_n]^T \in \mathbb{R}^n$*. The* ***directional derivative*** *of* $f(\cdot)$ *in the direction* $\mathbf{d}$ *is the real-valued function*

$$\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}) = \lim_{\alpha \to 0} \frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha}$$

Theorem 4

For a given $\mathbf{x}$ *and* $\mathbf{d}$*, we have*

$$\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}) = \nabla f(\mathbf{x})^T \mathbf{d} \tag{6}$$

Directional derivative

Proof.

Consider the functions

$$g : \mathbb{R} \to \mathbb{R}^n \quad \textit{defined by} \quad g(\alpha) = \boldsymbol{x} + \alpha \boldsymbol{d}$$

and

$$h : \mathbb{R} \to \mathbb{R} \quad \textit{defined by} \quad h(\alpha) = f(g(\alpha)) = f(\boldsymbol{x} + \alpha \boldsymbol{d})$$

On the one hand

$$\begin{aligned}
\frac{dh}{d\alpha}(0) &= \lim_{\epsilon \to 0} \frac{h(\alpha+\epsilon) - h(\alpha)}{\epsilon}\bigg|_{\alpha=0} \\
&= \lim_{\epsilon \to 0} \frac{h(\epsilon) - h(0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{f(\boldsymbol{x}+\epsilon\boldsymbol{d}) - f(\boldsymbol{x})}{\epsilon} = \frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x})
\end{aligned}$$

On the other hand

$$\begin{aligned}
\frac{dh}{d\alpha}(0) &= Df(g(\alpha))Dg(\alpha)\bigg|_{\alpha=0} = Df(\boldsymbol{x}+\alpha\boldsymbol{d})\boldsymbol{d}\bigg|_{\alpha=0} \\
&= Df(\boldsymbol{x})\boldsymbol{d} = \nabla f(\boldsymbol{x})^T \boldsymbol{d}
\end{aligned}$$

Directional derivative

*Note that for $\|\boldsymbol{d}\| = 1$ the directional derivative represents the **rate of increase** of $f(\cdot)$ in the direction $\boldsymbol{d}$ at point $\boldsymbol{x}$.*

For all $\boldsymbol{d}$ such that $\|\boldsymbol{d}\| = 1$, and by the Cauchy-Schwarz inequality

$$\nabla f(\boldsymbol{x})^T \boldsymbol{d} = \langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle \leq \|\nabla f(\boldsymbol{x})\|$$

and for $\boldsymbol{d} = \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|}$

$$\langle \nabla f(\boldsymbol{x}), \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \rangle = \|\nabla f(\boldsymbol{x})\|$$

The direction of the gradient $\nabla f(\boldsymbol{x})$ indicates the direction of the maximum rate of increase of $f(\cdot)$ at $\boldsymbol{x}$.

Taylor's theorem

Taylor's theorem is an essential tool in optimization. It is the basis for many methods and algorithms.

Theorem 5 (Taylor's theorem)

Consider a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x}_0 \in \mathbb{R}^n$ *and* $\boldsymbol{x} = \boldsymbol{x}_0 + \Delta \in \mathbb{R}^n$. *Then we have that*

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0)\Delta + o(\|\Delta\|) \tag{7}$$

If $f(\cdot)$ *is twice continuously differentiable, then we have*

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0)\Delta + \frac{1}{2}\Delta^T D^2 f(\boldsymbol{x}_0)\Delta + o(\|\Delta\|^2) \tag{8}$$

Taylor's theorem

Reminder: order symbols $\mathcal{O}$ and o

Consider the function $g : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ with Ω a neighborhood of $\mathbf{0}$, which satisfies $g(\mathbf{x}) \neq 0$ if $\mathbf{x} \neq \mathbf{0}$. Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$.

- *$f(\mathbf{x}) = \mathcal{O}(g(\mathbf{x}))$ means that $\frac{\|f(\mathbf{x})\|}{|g(\mathbf{x})|}$ is bounded near $\mathbf{0}$: $\exists K > 0, \sigma > 0$ such that for $\mathbf{x} \in \Omega : \|\mathbf{x}\| < \sigma$ then $\frac{\|f(\mathbf{x})\|}{|g(\mathbf{x})|} \leq K$.*
  *That is, $\mathcal{O}(g(\mathbf{x}))$ represents a function that is **bounded** by a scaled version of g around $\mathbf{0}$.*
- *$f(\mathbf{x}) = o(g(\mathbf{x}))$ means that*
  $$\lim_{\mathbf{x} \to \mathbf{0}, \mathbf{x} \in \Omega} \frac{\|f(\mathbf{x})\|}{|g(\mathbf{x})|} = 0$$
  That is, $o(g(\mathbf{x}))$ represents a function that goes to zero faster than $g(\cdot)$ around $\mathbf{0}$.

Positive definite (semidefinite) matrices

Definition 5

A symmetric matrix $M = M^T \in \mathbb{R}^{n \times n}$, is **positive definite** if the corresponding quadratic form $F_M(\mathbf{d}) = \mathbf{d}^T M\mathbf{d}$ is **positive definite**. That is:

$$\forall \mathbf{d} \neq 0 \in \mathbb{R}^n, \quad \mathbf{d}^T M\mathbf{d} > 0$$

In this case we write $M > 0$.

Definition 6

A symmetric matrix $M = M^T \in \mathbb{R}^{n \times n}$, is **positive semidefinite** if the corresponding quadratic form $F_M(\mathbf{d}) = \mathbf{d}^T M\mathbf{d}$ is **positive semidefinite**. That is:

$$\forall \mathbf{d} \in \mathbb{R}^n, \quad \mathbf{d}^T M\mathbf{d} \geq 0$$

In this case we write $M \geq 0$

Positive definite (semidefinite) matrices

Why symmetric matrices ? We can find non symmetric matrices such that $\boldsymbol{d}^T M\boldsymbol{d} > 0$:

Example

$$M = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix}^T M \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = d_1^2 + d_2^2 + (d_1 + d_2)^2 > 0, \forall \boldsymbol{d} = [d_1 d_2]^T \neq 0$$

Note that we did not need the matrix $M \in \mathbb{R}^{n\times n}$ to be symmetric to define a quadratic form. However, since

- $\forall \boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{d}^T M\boldsymbol{d}$ *is scalar, thus* $\boldsymbol{d}^T M\boldsymbol{d} = (\boldsymbol{d}^T M\boldsymbol{d})^T = \boldsymbol{d}^T M^T \boldsymbol{d}$
- $M = \frac{M+M^T}{2} + \frac{M-M^T}{2}$

$\Rightarrow \boldsymbol{d}^T M\boldsymbol{d} = \boldsymbol{d}^T \frac{M+M^T}{2}\boldsymbol{d}$*, and thus there is no loss of generality in considering symmetric matrices to define quadratic forms. Also we will be interested in studying Hessian matrices which, as we have seen, are symmetric under some conditions.*

Example

$$\boldsymbol{d}^T \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} \boldsymbol{d} = \boldsymbol{d}^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \boldsymbol{d}$$

Positive definite (semidefinite) matrices

Nice properties of symmetric matrices:

Theorem 7

All eigenvalues of a symmetric matrix $M = M^T \in \mathbb{R}^{n\times n}$ are real.

Theorem 8

Any real $M = M^T \in \mathbb{R}^{n\times n}$ has a set of n eigenvectors that are mutually orthogonal.

For any $M = M^T \in \mathbb{R}^{n\times n}$, the set of eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ forms an orthogonal basis for $\mathbb{R}^n$. The normalized set $\Rightarrow \{\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_n\}$ defines an orthonormal basis. Then, we define the matrix $P = [\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \cdots, \tilde{\mathbf{v}}_n]$ which satisfies

$$P^T P = I, \quad \text{thus} \quad P^T = P^{-1} \tag{7}$$

Positive definite (semidefinite) matrices

Theorem 9

A symmetric matrix $M = M^T \in \mathbb{R}^{n\times n}$ is positive definite (or positive semidefinite) if and only if all eigenvalues of M are positive (or nonnegative).

Proof

Consider the matrix P defined by the normalized eigenvectors of M such as in (7). $\forall \boldsymbol{d} \in \mathbb{R}^n$ define $\boldsymbol{y} = [y_1 \dots y_n]^T = P^{-1}\boldsymbol{d} = P^T\boldsymbol{d}$. Then,

$$\boldsymbol{d}^T M\boldsymbol{d} = \boldsymbol{y}^T P^T M P\boldsymbol{y}$$

Note that

$$\begin{aligned} P^T M P &= P^{-1}MP \\ &= P^{-1}M[\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2, \dots, \tilde{\boldsymbol{v}}_n] = P^{-1}[M\tilde{\boldsymbol{v}}_1, M\tilde{\boldsymbol{v}}_2, \dots, M\tilde{\boldsymbol{v}}_n] \\ &= P^{-1}[\lambda_1\tilde{\boldsymbol{v}}_1, \lambda_2\tilde{\boldsymbol{v}}_2, \dots, \lambda_n\tilde{\boldsymbol{v}}_n] \end{aligned}$$

with λ_i $i \in \{1, 2, \dots, n\}$ are the eigenvalues of M.

Positive definite (semidefinite) matrices

Proof (Cont.)

$$P^T M P \quad = \quad P^{-1} P \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Thus

$$\mathbf{d}^T M \mathbf{d} = \mathbf{y}^T P^T M P \mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2$$

From the last equation we can see that the scalar $\mathbf{d}^T M \mathbf{d}$ is strictly positive (nonnegative) for all $\mathbf{d} \neq \mathbf{0}$ iff all λ_i are strictly positive (nonnegative), which completes the proof.

□

Positive definite (semidefinite) matrices

*Detecting the positivity of a **symmetric** matrix:*

- *The singe of the eigenvalues (Theorem 9)*
- *Sylvester's Criterion*

Theorem 10

A symmetric matrix

$$M = M^T = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & & \vdots \\ m_{n1} & \cdots & m_{nn} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

*is positive definite if and only if the **leading principal minors** of M are positive. That is*

$$\Delta_1 = m_{11} > 0, \quad \Delta_2 = det \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} > 0, \quad \cdots \quad \Delta_n = det\, M > 0$$

Positive definite (semidefinite) matrices

Theorem 11 (Rayleigh's Inequality)

consider a positive definite matrix $M = M^T \in \mathbb{R}^n$, then $\forall \mathbf{d} \in \mathbb{R}^n$

$$\lambda_{min}\|\mathbf{d}\|^2 \leq \mathbf{d}^T M\mathbf{d} \leq \lambda_{max}\|\mathbf{d}\|^2$$

Proof.

Since M is symmetric matrix then all its eigenvalues are real (Theorem 7)

$$\lambda_{min} \leq \lambda_1 \leq \cdots \leq \lambda_n = \lambda_{max}$$

Moreover, its eigenvectors define an orthonormal basis $\{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \cdots, \tilde{\mathbf{v}}_n\}$ (Theorem 8). By writing $\mathbf{d} = \sum_{i=1}^n \theta_i \tilde{\mathbf{v}}_i$ we get

$$\|\mathbf{d}\| = \sum_{i=1}^n \theta_i^2, \qquad \mathbf{d}^T M\mathbf{d} = \sum_{i=1}^n \theta_i^2 \lambda_i$$

Thus

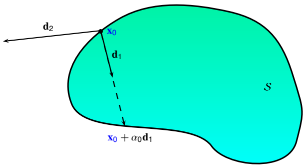
$$\lambda_{min}\sum_{i=1}^n \theta_i^2 \leq \quad \mathbf{d}^T M\mathbf{d} = \sum_{i=1}^n \theta_i^2 \lambda_i \quad \leq \lambda_{max}\sum_{i=1}^n \theta_i^2$$

□

Feasible direction

Feasible direction

*Consider the set $\mathcal{S}$. A nonzero vector $\mathbf{d} \in \mathbb{R}^n$ is called a **feasible direction** at $\mathbf{x}_0 \in \mathcal{S}$ if there exists $\alpha_0 > 0$ such that $\mathbf{x}_0 + \alpha \mathbf{d} \in \mathcal{S}\ \forall \alpha \in [0, \alpha_0]$.*



An illustration of feasible directions: $\mathbf{d}_1$ is a feasible direction and $\mathbf{d}_2$ is not a feasible direction.

Note that if $\mathbf{x}_0$ is an interior point, then all nonzero directions are feasible.

Modeling: a simple example

Finding the mathematical form starting from a real life problem is known as the modeling

A diet problem

Choose the quantities x_1 x_2 of two types of food for a daily need of a person.

- *One unit of food i costs c_i euros.*
- *One unit of food i contains p_i units of protein and v_i units of vitamins.*
- *A healthy diet requires at least p_{min} units of protein and v_{min} units of vitamins.*

How to choose x_1 and x_2 in order to obtain an economic but healthy diet ?

$$\begin{aligned} \min_{x} \quad & c_1 x_1 + c_2 x_2 \\ s.t. \quad & p_1 x_1 + p_2 x_2 \geq p_{min} \\ & v_1 x_1 + v_2 x_2 \geq v_{min} \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

Consider the parameters: $c_1 = 1, c_2 = 1, p_1 = 2, p_2 = 4, v_1 = 5, v_2 = 2, p_{min} = 8, p_{max} = 10$

Modeling: a simple example

Finding the mathematical form starting from a real life problem is known as the modeling

A diet problem

Choose the quantities x_1 x_2 of two types of food for a daily need of a person.

- *One unit of food i costs c_i euros.*
- *One unit of food i contains p_i units of protein and v_i units of vitamins.*
- *A healthy diet requires at least p_{min} units of protein and v_{min} units of vitamins.*

How to choose x_1 and x_2 in order to obtain an economic but healthy diet ?

$$\begin{aligned} \min_{\mathbf{x}} \quad & c_1 x_1 + c_2 x_2 \\ s.t. \quad & p_1 x_1 + p_2 x_2 \geq p_{min} \\ & v_1 x_1 + v_2 x_2 \geq v_{min} \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

Consider the parameters: $c_1 = 1$, $c_2 = 1$, $p_1 = 2$, $p_2 = 4$, $v_1 = 5$, $v_2 = 2$, $p_{min} = 8$, $p_{max} = 10$.

Modeling: another diet problem

A diet problem

Choose the quantities $\boldsymbol{x} = [x_1 \dots x_5]^T$ *of 5 types of food for a daily need of a person. The quantities should cover the following needs: 2000cal, 55g of Protein and 800g of Calcium, with the least cost possible.*

Food	*Unit*	*Energy(cal)*	*Protein(g)*	*Calcium(mg)*	*Price(Centimes)*	*Quantity*
Cereal	28*g*	110	4	2	30	x_1
Meat	100*g*	200	23	12	100	x_2
Eggs	1	80	6	26	20	x_3
Milk	250*cl*	160	8	285	50	x_4
Vegetables	250*g*	260	14	80	15	x_5

Modeling: another diet problem

A diet problem

Choose the quantities $\boldsymbol{x} = [x_1 \dots x_5]^T$ *of 5 types of food for a daily need of a person. The quantities should cover the following needs: 2000cal, 55g of Protein and 800g of Calcium, with the least cost possible.*

Food	*Unit*	*Energy(cal)*	*Protein(g)*	*Calcium(mg)*	*Price(Centimes)*	*Quantity*
Cereal	28*g*	110	4	2	30	x_1
Meat	100*g*	200	23	12	100	x_2
Eggs	1	80	6	26	20	x_3
Milk	250*cl*	160	8	285	50	x_4
Vegetables	250*g*	260	14	80	15	x_5

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & 30x_1 + 100x_2 + 20x_3 + 50x_4 + 15x_5 \\
s.t. \quad & 110x_1 + 200x_2 + 80x_3 + 160x_4 + 260x_5 \geq 2000 \\
& 4x_1 + 23x_2 + 6x_3 + 8x_4 + 14x_5 \geq 55 \\
& 2x_1 + 12x_2 + 26x_3 + 285x_4 + 80x_5 \geq 800 \\
& x_1 \geq 0, \dots, x_5 \geq 0
\end{aligned}$$

First-Order Necessary Conditions

Theorem 12 (First-Order Necessary Conditions)

Consider $\mathcal{S}$ a subset of $\mathbb{R}^n$ and a continuously differentiable function $f : \mathcal{S} \to \mathbb{R}$. If $\mathbf{x}^\star$ is a local minimizer of $f(\cdot)$ over $\mathcal{S}$, then for any feasible direction $\mathbf{d}$ at $\mathbf{x}^\star$ we have that

$$\mathbf{d}^T \nabla f(\mathbf{x}^\star) \geq 0 \tag{8}$$

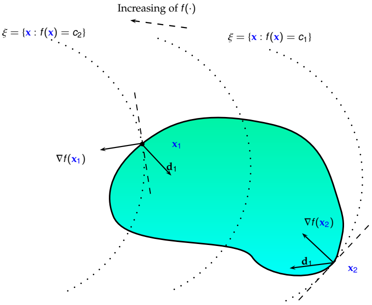


Illustration of first order necessary conditions. The point $\mathbf{x}_1$ does not satisfy the conditions, while $\mathbf{x}_2$ satisfies them.

First-Order Necessary Conditions

Proof.

If $\mathbf{d}^T\nabla f(\mathbf{x}^\star) = 0$ *the proof is trivial. We treat the case* $\mathbf{d}^T\nabla f(\mathbf{x}^\star) \neq 0$. *Consider the local minimizer* $\mathbf{x}^\star$ *and a feasible direction* $\mathbf{d}$ *at* $\mathbf{x}^\star$. *Then* $\exists\alpha_0$ *such that* $\mathbf{x}^\star + \alpha\mathbf{d} \in \mathcal{S}\ \forall\alpha \in [0,\alpha_0]$. *Consider the functions*

$$g : [0,\alpha_0] \to \mathbb{R}^n \quad \textit{defined by} \quad g(\alpha) = \mathbf{x}^\star + \alpha\mathbf{d}$$

and

$$h : [0,\alpha_0] \to \mathbb{R} \quad \textit{defined by} \quad h(\alpha) = f(g(\alpha)) = f(\mathbf{x}^\star + \alpha\mathbf{d})$$

Then, for $\alpha \in (0,\alpha_0)$, *using Taylor's theorem we have*

$$h(\alpha) = h(0) + \dot{h}(0)\alpha + o(\alpha) = h(0) + \underbrace{Df(\mathbf{x}^\star)\mathbf{d}}_{=\mathbf{d}^T\nabla f(\mathbf{x}^\star)}\alpha + o(\alpha)$$

$$f(\mathbf{x}^\star + \alpha\mathbf{d}) = f(\mathbf{x}^\star) + \mathbf{d}^T\nabla f(\mathbf{x}^\star)\alpha + o(\alpha) \tag{9}$$

Thus $\mathbf{d}^T\nabla f(\mathbf{x}^\star)$ *is positive, as if* $\mathbf{d}^T\nabla f(\mathbf{x}^\star) < 0$ *then from* (9) *we can find sufficiently small* α *such that* $f(\mathbf{x}^\star + \alpha\mathbf{d}) < f(\mathbf{x}^\star)$ *which contradicts the fact that* $\mathbf{x}^\star$ *is a local minimizer.*

□

First-Order Necessary Conditions

Proof (Another proof.)

Since $\mathbf{x}^\star$ *is a local minimizer, then for any feasible direction* $\mathbf{d}$*, there exists* $\alpha_0 > 0$ *such that* $\forall \alpha \in (0, \alpha_0)$ *we have*

$$f(\mathbf{x}^\star + \alpha \mathbf{d}) \geq f(\mathbf{x}^\star)$$

which yields

$$\frac{f(\mathbf{x}^\star + \alpha \mathbf{d}) - f(\mathbf{x}^\star)}{\alpha} \geq 0$$

Then by taking the limit as $\alpha \to 0$ *we find*

$$\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}^\star) \geq 0$$

From Theorem 4 we find that $\mathbf{d}^T \nabla f(\mathbf{x}^\star) \geq 0$. □

This has a nice interpretation: if $\mathbf{x}^\star$ *is a local minimizer, then the rate of increase of* $f(\cdot)$ *in any feasible direction is nonnegative.*

First-Order Necessary Conditions

Corollary 13 (The interior point case)

*Consider $\mathcal{S}$ a subset of $\mathbb{R}^n$ and a continuously differentiable function $f : \mathcal{S} \to \mathbb{R}$. Consider an **interior point** $\boldsymbol{x}^\star$. If $\boldsymbol{x}^\star$ is a local minimizer of $f(\cdot)$ over $\mathcal{S}$, then*

$$\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0} \tag{10}$$

Proof.

Exercise □

Second-Order Necessary Condition

Theorem 14 (Second-Order Necessary Conditions)

Consider $\mathcal{S}$ a subset of $\mathbb{R}^n$ and a twice continuously differentiable function $f : \mathcal{S} \to \mathbb{R}$. If $\boldsymbol{x}^\star$ is a local minimizer of $f(\cdot)$ over $\mathcal{S}$, then for any feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^\star$, if $\boldsymbol{d}^T \nabla f(\boldsymbol{x}^\star) = 0$ then

$$\boldsymbol{d}^T D^2 f(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0 \tag{11}$$

Proof

The proof is done by contradiction. Consider the local minimizer $\boldsymbol{x}^\star$ and a feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^\star$. Suppose that $\boldsymbol{d}^T \nabla f(\boldsymbol{x}^\star) = 0$ and $\boldsymbol{d}^T D^2 f(\boldsymbol{x}^\star)\boldsymbol{d} < 0$. Then $\exists \alpha_0$ such that $\boldsymbol{x}^\star + \alpha \boldsymbol{d} \in \mathcal{S}\ \forall \alpha \in [0, \alpha_0]$. Consider the functions

$$g : [0, \alpha_0] \to \mathbb{R}^n \quad \textit{defined by} \quad g(\alpha) = \boldsymbol{x}^\star + \alpha \boldsymbol{d}$$

and

$$h : [0, \alpha_0] \to \mathbb{R} \quad \textit{defined by} \quad h(\alpha) = f(g(\alpha)) = f(\boldsymbol{x}^\star + \alpha \boldsymbol{d})$$

Then, for $\alpha \in (0, \alpha_0)$, using Taylor's theorem we have

$$h(\alpha) = h(0) + \dot{h}(0)\alpha + \frac{1}{2}\ddot{h}(0)\alpha^2 + o(\alpha^2)$$

Second-Order Necessary Condition

Proof (Cont.)

$$h(\alpha) - h(0) = \underbrace{\dot{h}(0)}_{\mathbf{d}^T \nabla f(\mathbf{x}^\star) = 0} \alpha + \frac{1}{2} \underbrace{\ddot{h}(0)}_{\mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} < 0} \alpha^2 + o(\alpha^2)$$

$$= \frac{\alpha^2}{2} \mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} + o(\alpha^2)$$

Note that for sufficiently small $\alpha \in (0, \alpha_0)$ *we have*

$$f(\mathbf{x}^\star + \alpha \mathbf{d}) - f(\mathbf{x}^\star) < 0$$

which contradicts the assumption that $\mathbf{x}^\star$ *is a local minimizer. Thus,* $\mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} \geq 0$.

□

Second-Order Necessary Condition

Corollary 15 (The interior point case)

Consider $\mathcal{S}$ a subset of $\mathbb{R}^n$ and a twice continuously differentiable function $f : \mathcal{S} \to \mathbb{R}$. If the ***interior point*** $\boldsymbol{x}^\star$ *is a local minimizer of $f(\cdot)$ over $\mathcal{S}$, then $\nabla f(\boldsymbol{x}^\star) = 0$ and $D^2 f(\boldsymbol{x}^\star)$ is a positive semidefinite matrix.*

Proof.

Note that since $\boldsymbol{x}^\star$ is an interior point then $\forall \boldsymbol{d} \neq 0 \in \mathbb{R}^n$ we obtain a feasible direction. The first fact is a direct result of Corollary 13. The second fact follows from Theorem 14 since

$$\forall \boldsymbol{d} \neq 0 \in \mathbb{R}^n, \qquad \boldsymbol{d}^T D^2 f(\boldsymbol{x}^\star) \boldsymbol{d} \geq 0$$

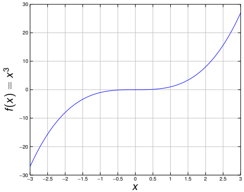
which completes the proof. □

We need sufficient conditions

Let us have a look at examples where the necessary conditions may be satisfied for a point that is NOT a minimizer.

Example

Consider the function $f : \mathbb{R} \to \mathbb{R}$ *defined by* $f(x) = x^3$. *Clearly, the point* 0 *satisfies both the FONC and the SONC, but it is not a minimizer.*



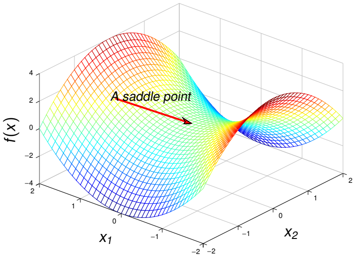
The point 0 satisfies both the FONC and the SONC, but it is not a minimizer.

We need sufficient conditions

Let us have a look at examples where the necessary conditions may be satisfied for a point that is NOT a minimizer.

Example

Consider the function $\mathbb{R}^2 \to \mathbb{R}$ *defined by* $f(\boldsymbol{x}) = {x_1}^2 - {x_2}^2$. *Does the point* 0 *satisfy the FONC ? the SONC ? Is it a minimizer ?*



The point 0 is a saddle point.

Second-Order Sufficient Condition

Theorem 16 (Second-Order Sufficient Condition)

*Consider $\mathcal{S}$ a subset of $\mathbb{R}^n$ and a twice continuously differentiable function $f : \mathcal{S} \to \mathbb{R}$ and an **interior point** $\boldsymbol{x}^\star$. If*

- $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$
- $D^2 f(\boldsymbol{x}^\star) \succ 0$ *(the Hessian matrix at $\boldsymbol{x}^\star$ is positive definite)*

then $\boldsymbol{x}^\star$ is a strict local minimizer of $f(\cdot)$.

Proof

Since $\boldsymbol{x}^\star$ is an interior point, then for any nonzero direction $\boldsymbol{d} \in \mathbb{R}^n$ $\exists \alpha_0$ such that $\boldsymbol{x}^\star + \alpha \boldsymbol{d} \in \mathcal{S}$ $\forall \alpha \in [0, \alpha_0]$. Consider the functions

$$g : [0, \alpha_0] \to \mathbb{R}^n \quad \text{defined by} \quad g(\alpha) = \boldsymbol{x}^\star + \alpha \boldsymbol{d}$$

and

$$h : [0, \alpha_0] \to \mathbb{R} \quad \text{defined by} \quad h(\alpha) = f(g(\alpha)) = f(\boldsymbol{x}^\star + \alpha \boldsymbol{d})$$

Second-Order Sufficient Condition

Proof (Cont.)

Then, for $\alpha \in (0, \alpha_0)$, using Taylor's theorem we have

$$h(\alpha) = h(0) + \dot{h}(0)\alpha + \frac{1}{2}\ddot{h}(0)\alpha^2 + o(\alpha^2)$$

$$\begin{aligned} h(\alpha) - h(0) &= \underbrace{\dot{h}(0)}_{\mathbf{d}^T \nabla f(\mathbf{x}^\star) = 0} \alpha + \frac{1}{2} \underbrace{\ddot{h}(0)}_{\mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} > 0} \alpha^2 + o(\alpha^2) \\ &= \frac{\alpha^2}{2} \mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} + o(\alpha^2) \end{aligned}$$

Since $D^2 f(\mathbf{x}^\star) \succ 0$ then $\mathbf{d}^T D^2 f(\mathbf{x}^\star)\mathbf{d} > 0$ thus for sufficiently small $\alpha \in (0, \alpha_0)$ we have $f(\mathbf{x}^\star + \alpha\mathbf{d}) - f(\mathbf{x}^\star) > 0$, which completes the proof. □

1. Introduction

2. Optimization problem

3. Mathematical foundations

4. Modeling

5. Optimality conditions

6. Solvers

Softwares and tools

- ***Optimization Toolbox*** *: Matlab toolbox for modeling and solving optimization problems*
  https://fr.mathworks.com/help/optim/
- ***YALMIP*** *: Matlab tool for modeling convex and non-convex optimization problems*
  https://yalmip.github.io/
- ***CVX*** *: Matlab tool for modeling convex problems*
  http://cvxr.com/cvx/
- ***AMPL*** *: a powerful modeling tool for wide range of optimization problems, widely used in industry*
- ***GAMS*** *: a high-level modeling system for mathematical programming and optimization*
  https://www.gams.com/
- ***scipy.optimize*** *: a Python package provides several commonly used optimization algorithms*
  https://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html

Softwares and tools

Solvers:

- ***CPLEX*** *: a commercial industry standard solver from IBM*
  https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/
- ***Gurobi*** *: a commercial solver developed by founders of CPLEX*
  http://www.gurobi.com/
- ***MOSEK*** *: a commercial solver, with focus on solving large scale sparse problems*
  http://www.mosek.com/
- ***SeDuMi*** *: a free solver with Matlab interface, widely used for semi-definite programming*
  http://sedumi.ie.lehigh.edu/
- ***SDPT3****: a free solver for semi-definite programming*
  http://www.math.nus.edu.sg/ mattohkc/sdpt3.html

References I

M. D. Mathelin
Cours: Optimisation.
http://icube-avr.unistra.fr/fr/index.php/Optimisation

D. Dimitar
Introduction to optimization.
http://www.aass.oru.se/Research/mro/drdv_dir/seminar_optim_2011.html

C. Scherer
Linear Matrix Inequalities in Control.

E. K. P. Chong and S. H. Zak.
Introduction to Optimization.
Wiley, 2004.

R. Fletcher.
Practical Methods of Optimization.
Wiley, 1987.

S. S. Rao.
Engineering Optimization Theory and Practice.
Wiley, 2009.

References II

J. Nocedal and S. J. Wright
Numerical Optimization.
Springer, 2006.

N. Andreasson, A. Evgrafov and M. Patriksson
An Introduction to Optimization: Foundations and Fundamental Algorithms.
Dover Publications, 2016.

G. B. Dantzig
Linear Programming and Extensions.
Princeton University Press, 1963.

A. Rais and A. Vianaa.
Operations Research in Healthcare: a survey
International Transactions in Operational Research, *2011.*

W. Crown et al
Constrained Optimization Methods in Health Services Research—An Introduction: Report 1 of the ISPOR Optimization Methods Emerging Good Practices Task Force
Value In Health, *2017.*

References III

M. L. Handford and M. Srinivasan.
Robotic lower limb prosthesis design through simultaneous computer optimizations of human and prosthesis costs
Nature, *2016.*