EXTENDED READINGS, QUESTIONS, EXERCISES, PROGRAMS

Brno University of Technology
FME, Institute of Mathematics

Optimization I
sop, vo1p, opm

Pavel Popela

February 15, 2017

Chapter 1

Linear and nonlinear models

Content: The following themes will be studied:

- Introduction (Lecture 1–2 below)
- Convex Analysis (3–9)
- Optimality Conditions (10, 17–20)
- Lagrangian Duality (24–25)
- Introduction to Algorithms
  - Unconstrained Problems (11–16)
  - Problems with Constraints (21–23), (26–28)

Detail lecture themes: The order and list will be updated during the flow of lectures (they are numbered below).

1. Organization + Basic concepts
2. Modelling + Solving NLPs (principles)
3. 4. Convex sets
5. 6. Polyhedral sets
7. 8. Linear programming and simplex method
9. 10. Convex functions
11. Unconstrained optimization

Algorithms : general approach and examples

12. Line search methods
13. Multidimensional search without using derivatives
14. 15. 16. Multidimensional search using derivatives (involving advanced methods)
17. 18. 19. Constrained optimization (KKT conditions)
20. Constraints qualification
21. From feasible directions to successive approximations.
22. Projection-based algorithms
23. Reduced gradient algorithms
24. 25. Lagrangian duality
26. Penalty function/based methods
27. Augmented Lagrangian-based methods
28. Barrier functions and their use

Suggested texts: The following books are suggested as extending readings (further interesting books can be found in the university library):

BS93 Bazaraa, M.S., Sherali, H.D., and Shetty, C.M. (1993) *Nonlinear Programming: Theory and Applications*, John Wiley and Sons.

Mi85 Minoux, M. (1985) *Mathematical Programming*, John Wiley and Sons.

BJ90 Bazaraa, M.S., Jarvis, J.J., and Sherali, H.D. (1990) *Linear Programming and Network Flows*, John Wiley and Sons.

Further sources of information: Books: E.g., Fletcher: Practical Methods of Optimization, Wiley.
Journals: Journal of OR, Mathematical Programming.
Interesting WWW pages:

`nlp.fag, http:// www.gams.com`,
`www~math.cudenver.edu/~hgreenbe/glossary/glossary.html`,
`www.maths.mu.oz.au/~worms`,
`mat.gsia.cmu.edu`,
`www.caam.rice.edu/~mathprog`,
`www.informs.org`.

Search also for keywords: nonlinear programming, mathematical programming, optimization ... and share your result with us!

Exercise 1.

What is NEOS?

1.1 Basic concepts

Notation: The following notation is inherited from calculus:

a, b, c	scalars, constants
u, v, x, y, z	variables, unknowns
A, B, X	sets
2^A	set of all subsets of set A
$\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{y}$	column vector, $\mathbf{a}^\top$ row vector
$\mathbf{A}, \mathbf{B}$	matrices (a_{ij}) where $i = 1, ..n, j = 1...n$.
$\emptyset, \{\}$	empty set
$\mathbb{N}, Z, \mathbb{R}$	sets of positive integers, integers, and real numbers
$\in, \subset, \cap, \cup, \bar{S}$	usual set-related symbols
$\wedge, \vee, \Rightarrow, \Leftrightarrow$	usual logical symbols
$\exists x \in X$	there exists x in X
$\forall x \in X$	for any element x of X
$f : \mathbb{R}^n \longrightarrow \mathbb{R}$	a scalar function
$\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$	a vector function
$\nabla f(\mathbf{x})$	the gradient of f at the point $\mathbf{x}$
$\nabla^2 f(\mathbf{x}), \mathbf{H}(\mathbf{x})$	the Hessian of f at $\mathbf{x}$
$E_n, \mathbb{R}^n$	the Euclidean n-dimensional (linear) space
$\|\mathbf{x}\|$	the Euclidean norm
$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$	a distance between $\mathbf{x}$ and $\mathbf{y}$
$\mathcal{N}_\varepsilon(\mathbf{x})$	the epsilon neighbourhood of $\mathbf{x}$
$x_1, ...x_k, ...$ or $\{x_k\}_{k\in\mathbb{N}}$	an infinite sequence
$\lim, \limsup, \liminf$	limit, upper limit, lower limit
$\operatorname{int} S, \operatorname{cl} S, \partial S$	interior, closure, and boundary of set S
$\sup, \inf$	supremum, infimum
$\min, \max$	minimum, maximum
$\leq$	ordering of real numbers
$:=$	define

Definition 2 (Topological concepts)**.**

The following topological concepts are often utilized. Denote $\mathcal{N}_\varepsilon(\mathbf{x}) = \{\mathbf{y}:\|\mathbf{y}-\mathbf{x}\| < \varepsilon\}$ and $S \subset \mathbb{R}^n$:

Closure: $\mathbf{x} \in \operatorname{cl} S \Leftrightarrow \forall\varepsilon{>}0 : S \cap \mathcal{N}_\varepsilon(\mathbf{x}) \neq \emptyset$
Closed set: S is closed $\Leftrightarrow S = \operatorname{cl} S$
Interior: $\mathbf{x} \in \operatorname{int} S \Leftrightarrow \exists\varepsilon{>}0 : \mathcal{N}_\varepsilon(\mathbf{x}) \subset S$
Open set: S is open $\Leftrightarrow S = \operatorname{int} S$
Solid set: S is solid $\Leftrightarrow \operatorname{int} S \neq \emptyset$
Boundary: $\mathbf{x} \in \partial S \Leftrightarrow \forall\varepsilon{>}0 : (\mathcal{N}_\varepsilon(\mathbf{x}) \cap S \neq \emptyset) \wedge (\mathcal{N}_\varepsilon(\mathbf{x}) \cap \bar{S} \neq \emptyset)$
Bounded set: S is bounded $\Leftrightarrow \exists\mathbf{x} \in \mathbb{R}^n, \exists\varepsilon{>}0 : S \subset \mathcal{N}_\varepsilon(\mathbf{x})$
Compact set: S is compact $\Leftrightarrow$ (S is bounded)$\wedge$(S is closed)

Properties: A finite intersection (union) of open sets is an open set. A finite intersection (union) of closed sets is an open set.

$\operatorname{cl} S = S \cup \partial S$ $\quad \mathbb{R}^n = \operatorname{cl} \mathbb{R}^n = \operatorname{int} \mathbb{R}^n$ $\quad S \neq \operatorname{int} S \cup \partial S$ $\quad \emptyset = \operatorname{cl} \emptyset = \operatorname{int} \emptyset$
$\operatorname{int} S = S \setminus \partial S$ $\quad S$ open $\Rightarrow \bar{S}$ closed $\quad \partial S \neq S \setminus \operatorname{int} S$ $\quad S$ closed $\Rightarrow \bar{S}$ open

Exercise 3.

Illustrate concepts and properties by figures. Compare concepts: the closed interval and closed set.

Definition 4 (Continuity-related concepts)**.**

The following concepts that are related to convergence and continuity will be often used:

Convergence: $\mathbf{x}_k \longrightarrow \mathbf{x} \Leftrightarrow \forall \varepsilon{>}0,\ \exists N \in \mathbb{N},\ \forall k \geq N : \|\mathbf{x}_k - \mathbf{x}\| < \varepsilon$.

Cluster point: $\mathbf{x}$ is a cluster point of $\{\mathbf{x}_k\}_{k\in\mathbb{N}} \Leftrightarrow \exists L \subset \mathbb{N} : \mathbf{x}_k \longrightarrow \mathbf{x}, k \in L$. Remark: a convergent sequence has a unique cluster point that is a limit.

Limes superior: A limit superior (upper limit) of $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ is the largest cluster point, i.e. for $\{x_k\}_{k\in\mathbb{N}}$ $y = \limsup \mathbf{x}_k \Leftrightarrow (\forall \varepsilon{>}0, \exists K \in \mathbb{N} : k > K \Rightarrow x_k \leq y + \varepsilon) \wedge (\forall \varepsilon{>}0, \forall K > 0, \exists k > K : x_k \leq y - \varepsilon)$

Limes inferior: A limit inferior (lower limit) of $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ is the smallest cluster point. We may define $\liminf x_k = -\limsup(-x_k)$.

Closedness: The alternative definition of the closed set (or its property): S is closed $\Leftrightarrow (\forall \{\mathbf{x}_k\}_{k\in\mathbb{N}}$ sequence such that $\mathbf{x}_k \in S$, $\mathbf{x}_k \longrightarrow \mathbf{x} \Rightarrow \mathbf{x} \in S)$.

Relative interior: Let $S \subset \mathbb{R}^n$, $\operatorname{aff} S$ denotes the affine hull of S (see Remark 36) and $\operatorname{int} S = \emptyset$ is allowed. Then: $\operatorname{relint} S = \{\mathbf{x} \in S \mid \exists \varepsilon > 0 : \mathcal{N}_\varepsilon(\mathbf{x}) \cap \operatorname{aff} S \subset S\}$

Continuous function: For $f(\mathbf{x})$ continuous at $\mathbf{x}$: $\mathbf{x}_k \longrightarrow \mathbf{x}$ and $f(\mathbf{x}_k) \longrightarrow L$ implies $f(\mathbf{x}) = L$

Lower semicontinuous function: f is lower semicontinuous at $\mathbf{x} \in S \Leftrightarrow \varepsilon > 0\ \exists \delta > 0 : \mathbf{y} \in S$, $\|\mathbf{y} - \mathbf{x}\| < \delta \Rightarrow f(\mathbf{y}) - f(\mathbf{x}) > -\varepsilon$ (or $f(\mathbf{x}) < f(\mathbf{y}) + \varepsilon$).

Exercise 5.

Illustrate concepts by drawing figures in $\mathbb{R}^2$. Compare continuity and lower semicontinuity of f.

General decision problems. Complex decision problems may be described using algebraic concepts. Let $\Gamma = (V, E)$ be an oriented graph describing distribution of decision in time or space. Indices $v \in V$ are vertices (nodes) and $e \in E$ are edges (arcs, arrows). For all nodes, we introduce

$$\mathcal{O}^v = (\mathcal{N}^v, \mathcal{K}_D^v, \{X_K^v\}_{K\in\mathcal{K}_D^v}, \mathcal{C}^v, \mathcal{K}_I^v, \{\preceq_K\}_{K\in\mathcal{K}_I^v}).$$

$\mathcal{O}^v$	a description of decision problem related to $v \in V$
$\mathcal{N}^v$	problem-related decision makers in node v
$K_D^v \subset 2^{\mathcal{N}^v}$	groups of decision makers
$K \in K_D^v$	one group of decision makers
X_K^v	set of feasible decisions in v for K
$C^v \subset \prod_{K \in \mathcal{K}_D^v} X_K^v$	decision situations
	(no interactions among problems related to different nodes)
$\preceq_K^v \subset \mathcal{C}^v \times \mathcal{C}^v$	preference relations
$\mathcal{K}_I^v \subset 2^{\mathcal{N}^v}$	groups of interests

Generally for interacting decision problems: $C^v \subset \mathcal{U} \times \prod_{K \in \mathcal{K}_D^v} X_K^v$ describe decision situations (with interactions among problems) and $\mathcal{T}_{uv} : \mathcal{C}^u \longrightarrow \mathcal{C}^v$ describes the influence of decisions from node u on node v.

Definition 6 (Mathematical program)**.**

Let $S \subset \mathbb{R}^n$ and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Then a mathematical program (MP is further used as the abbreviation) is defined as follows:

$$\min_{\mathbf{x}} \{f(\mathbf{x}) \mid \mathbf{x} \in S\} \tag{1.1}$$

There are various special cases of the general mathematical program ($\mathbf{c}, \mathbf{b}$ are vectors, $\mathbf{A}$ is matrix):

- Linear program (LP): $\min\{\mathbf{c}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$
- Mixed integer linear program (MIP):
  $\exists J \subset \{1, \ldots, n\}$: $\min\{\mathbf{c}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \forall j \in J x_j \in \mathbb{N}\}$
- *Nonlinear program* - see below

Another mathematical programs can be obtained from general cases after some initial modification:

- from optimal control using discretization,
- from multicriteria optimization by scalar reformulation, and
- from stochastic programming by deterministic reformulation.

Definition 7 (Nonlinear program)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, and $X \subset \mathbb{R}^n$. In addition $\circ \in \{\leq, \geq, =\}^m$. Then

$$\min_{\mathbf{x}} \{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \circ \mathbf{0},\ \mathbf{x} \in X\} \tag{1.2}$$

is called a nonlinear program (further used shortcut NLP).

Notice that it can be obtained from mathematical program by assigning: $S = \{\mathbf{x} \in X \mid \mathbf{g}(\mathbf{x}) \circ \mathbf{0}\} = \cap_{i=1}^m S_i = \cap_{i=1}^m \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \circ \mathbf{0}\} = \{\mathbf{x} \in X \mid\ g_i(\mathbf{x}) \leq 0,\ 1 \leq i \leq l,\ g_i(\mathbf{x}) = 0,\ l+1 \leq$

$i \leq m\}$. A standard form of nonlinear program (given by the last form of the feasible set and by minimization) can be obtained by equivalent transformations from any nonlinear program (cf. with linear program standard form, i.e. $\max \longmapsto \min, \leq, \geq \longmapsto =$, etc.

There are also frequently used special cases of nonlinear programs:

- A quadratic program: $\min\{\mathbf{x}^\top \mathbf{H}\mathbf{x} + \mathbf{c}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$.
- A convex program: S is a convex set and f is a convex function.

Descriptive and normative approach. Till now, we utilized a *descriptive approach*. It means that syntax of optimization program is correct, i.e. it says "how mathematical optimization model looks". We turn to a *normative approach* now, saying "what to do" to reach the optimum. We review the ordering concepts.

Ordering of $\mathbb{R}$ and its properties. The couple $(\mathbb{R}, \leq)$ represents a natural ordering of real numbers. There is a binary relation $\leq \subset \mathbb{R} \times \mathbb{R}$ such that it satisfies following axioms:

- $\forall a \in \mathbb{R} : a \leq a$
- $\forall a, b \in \mathbb{R} : (a \leq b) \wedge (b \leq a) \Rightarrow a = b$
- $\forall a, b, c \in \mathbb{R} : (a \leq b) \wedge (b \leq c) \Rightarrow a \leq c$
- In addition $\forall a, b \in \mathbb{R} : (a \leq b) \vee (b \leq a)$.

Definition 8 (max, min, inf, sup)**.**

Let $B \subset \mathbb{R}$, $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, +\infty\}$ where $-\infty$ and ∞ are such elements that satisfy $\forall a \in \mathbb{R} : -\infty \leq a \leq \infty$. Then we define:

$$B_{\min} = \min B \Leftrightarrow (b_{\min} \in B) \wedge (\forall b \in B : b_{\min} \leq b),$$

$$B_{\max} = \max B \Leftrightarrow (b_{\max} \in B) \wedge (\forall b \in B : b_{\max} \geq b),$$

$$\inf B = \max\{c \mid (c \in \mathbb{R}^*) \wedge (\forall b \in B : c \leq b)\},$$

$$\sup B = \min\{c \mid (c \in \mathbb{R}^*) \wedge (\forall b \in B : c \geq b)\}.$$

Denote domain of f as $\mathrm{Dom} f$ and image of f as $\mathrm{Im} f$. Then semantics of MP may be given as $\min_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} = \min \mathrm{Im} f$.

Syntax extensions. Usually, in mathematical programming, one of the following descriptions is used. We describe MP by $\min f(\mathbf{x})$ subject to $\mathbf{x} \in S$ or by $\min_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. To identify the goal of our optimization search, we extend the notation assuming that $\min_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ denotes the set of all minimal objective function values. Similarly, $\mathrm{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ is a set of all minima. So, $? \in \mathrm{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ means the search for at least one minimum (using $=$ instead of $\in$ means to find all minima). Sometimes, we insert the words -glob- or -loc- into argmin to add details about the type of searched minima.

Definition 9 (Scheme for minima)**.**

Let $S \subset \mathbb{R}^n$ and $f : S \longrightarrow \mathbb{R}$. We define that $\mathbf{x}_{\min} \in S$ is a point of

$$\left\{ \begin{array}{c} \text{local} \\ \text{global} \end{array} \right\} \times \left\{ \begin{array}{c} \text{strict} \\ \text{non-strict} \end{array} \right\} \quad \text{minima}$$

$$\Longleftrightarrow$$

$$\left\{ \begin{array}{c} \exists \mathcal{N}_\varepsilon(\mathbf{x}_{\min}) : \forall \mathbf{x} \in S \cap \mathcal{N}_\varepsilon(\mathbf{x}_{\min}) \setminus \{\mathbf{x}_{\min}\} \\ \forall \mathbf{x} \in S \setminus \{\mathbf{x}_{\min}\} \end{array} \right\} \times f(\mathbf{x}_{\min}) \left\{ \begin{array}{c} < \\ \leq \end{array} \right\} f(\mathbf{x})$$

Note that $\mathbf{x}_{\min}$ is an isolated minimum of f on $S \Leftrightarrow \exists \varepsilon > 0 : \exists! \mathbf{x}_{\min} \in \mathcal{N}_\varepsilon(\mathbf{x}_{\min})$.

Example 10 (Use of definition scheme)**.**

Let $f : S \longrightarrow \mathbb{R}$ be a given function with domain S. Then $\mathbf{x}_{\min} \in S$ is a local (non-strict) minimum if and if (iff) $\exists \mathcal{N}_\varepsilon(\mathbf{x}_{\min}) : \forall \mathbf{x} \in S \cap \mathcal{N}_\varepsilon(\mathbf{x}_{\min}) \setminus \{\mathbf{x}_{\min}\} : f(\mathbf{x}_{\min}) \leq f(\mathbf{x})$.

Exercise 11.

Apply the definition scheme 9 to develop other definitions of optimal solutions. Try to reformulate it for maximum.

Theorem 12 (Weierstrass)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, S is a compact set, and $f : S \longrightarrow \mathbb{R}$ continuous function on S. Then mathematical program $\min_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ attains its global minimum (denoted as $\mathbf{x}_{\min}$).

The theorem remains valid if continuity is replaced by 'lower semicontinuity'.

Exercise 13.

Show the importance of the following assumptions of Theorem 12 by counterexamples in $\mathbb{R}^2$: $S \neq \emptyset$, S closed, bounded, and f continuous on S. Does it work for linear programs? Compare formulations: 'must have a global minimum' or 'may have a global minimum' for Weierstrass' Theorem.

Exercise 14.

Why constraint $g(\mathbf{x}) < 0$ is seldom considered in mathematical programming? Hint: Think about Weierstrass' Theorem assumptions and conclusions.

Exercise 15.

If $g(\mathbf{x}) < 0$ is still required then show how to change it to attain $\leq$. Hint: Use 'rounding relaxation' $g(\mathbf{x}) \leq 0$ or ε use $g(\mathbf{x}) \leq -\varepsilon$.

Proof of Theorem 12: Assume S closed and bounded and f continuous on S. Then (from Calculus) f is bounded below on S. Since $S \neq \emptyset \Rightarrow \exists \alpha = \inf\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. Let $0 < \varepsilon < 1 : S_k = \{\mathbf{x} \in S : \alpha \leq f(\mathbf{x}) \leq \alpha + \varepsilon^k\}$. So, $\forall k \in \mathbb{N} : S_k \neq \emptyset$, $\mathbf{x}_k \in S_k$, and $\forall k \in \mathbb{N} : \{\mathbf{x}_k\} \subset S$. Because S is bounded $\Rightarrow \exists L \subset \mathbb{N} : \mathbf{x}_k \longrightarrow \bar{\mathbf{x}}$, $k \in L$. Because S is closed $\Rightarrow \bar{\mathbf{x}} \in S$. Because f is continuous: $\forall k \in L : \alpha \leq f(\mathbf{x}_k) \leq \alpha + \varepsilon^k \Rightarrow \alpha = \lim_{k \in L, k \longrightarrow \infty} f(\mathbf{x}_k) = f(\bar{\mathbf{x}})$. So $f(\bar{\mathbf{x}}) = \inf\{f(\mathbf{x}) \mid \mathbf{x} \in S\} = f(\mathbf{x}_{\min})$ and we have $\mathbf{x}_{\min}$. □

1.2 Modelling

Example 16 (Location of facilities)**.**

Locate centers of activities in optimal way, e.g., imagine Malta towns and markets. Use a map and simplify the problem specification (direct distance considered and no governmental geographic rules given).

Exercise 17.

Formulate a mathematical program. Hint: Begin with a traditional transportation problem.

Modeling recommendations: Answer '10 modelling questions for ever'. The answers may help you to identify model type and properties:

1. How many decision makers participate in decision problem?
2. How many criteria have to be considered?
3. Are decisions spatially or time distributed?
4. Any uncertain parameters have to be taken into account?
5. Which basic types for decision variables must be used (scalars, vectors, functions, etc.)?
6. Which domains are specified for decision variables?
7. Do we have to make a difference between local and global optima?
8. May differentiability or another computational feature of involved functions be effectively employed?
9. Are there linear functions used for problem description?
10. Is it possible to take an advantage from special structures (networks, etc.)?

Modelling in general. Going through successful applications of mathematical programming, we generalize the obtained experience, and we will try to formulate some general rules applicable for model building and solving in mathematical programming. It was observed that every mathematical programming model has its own life-cycle, similar to other models. This life-cycle is composed of several steps. These steps were discussed in detail by Geoffrion. We may use, for instance, the following scheme for the description of modelling life-cycle:

Problem analysis: In this step, the decision problem is identified and analysed:

- First, we must formulate the target of the later-built model.
- Then, the context of the studied problem is analysed to evaluate global consequences of local decisions.
- The extensive discussions among the modelers and experts of the application field follow to define problem structures, parameters, and decision variables.
- Bibliographical references to similar applications may help to classify the considered problem.
- Historical input data and related decisions are collected for model verification. They are classified (e.g., for different technological scenarios) and stored (e.g., in information system databases).
- The results of analysis are formulated using the application specific language as the first step to model building.

Model design: The considered problem is completely classified, the model is built as a mathematical model, and its transformation into the implementation description is realized in the following steps:

- Number of decision makers is defined, and decision variables are identified.
  - The type of decision representation is fixed for each decision variable as scalar, vector, function, etc.
  - Time structure is handled as static or dynamic with discrete or continuous time.
  - Domains of variables are characterized as real, integer, etc.
- Number and the form of decision rules is determined.
  - Artificial variables describing the values of criteria are declared.
  - The optimization criteria are expressed in the form of objective functions.
  - The objective functions are built with general mathematical forms of composed functions combining variables, input parameters, and weighting parameters introduced by a decision maker.
- Sets of decisions, together with the feasible set of possible situations, are specified.
  - The experience of consultants must be used in such a way that historical decisions are also feasible for the built program.
  - Indices and sets of their values assess a model size.
  - Input data parameters are listed, and their certainty is characterized. Primary parameter values are obtained from databases saving analysis results.
  - Derived parameters are calculated from primary data using assignments and mathematical operations to simplify model formulation.
  - A feasible set is described by bounds and constraints.
  - The groups of constraints are ordered by their importance.

  - Some of them are identified as soft constraints, others as hard constraints.
- The mathematical model properties such as linearity and convexity that help the solution process are studied.
- This model is described using a general modelling language for practical computations.
- The external data sources are assigned to model parameters.
- Model building continues with approximation, relaxation, or transformation that keeps the problem solvable.

Implementation and solution: In this step, the model is completely implemented with the concrete software and hardware platforms. As a result, the solution will be found.

- The mathematical model is implemented for the chosen type of computer and operating system. The appropriate optimization system is selected, a previous general formulation is specified.
- The solver is chosen. Sometimes the expert's opinion is important in this selection, e.g., Schittkowski developed a methodology for the support of a solver selection in nonlinear programming.
- Computing of test problems with simple structure based on historical data set follows.
- Tuning of algorithm parameters for the considered class of problems is realized.
- Permanent connection to other software, as information system data sources, is established.

Presentation and interpretation: In this step, results are collected and analysed.

- The unbounded objective or the infeasible program detect the modelling mistakes, such as missing or contradictory constraints or typing errors in data.
- Results are analysed and evaluated.
- Suggestions for model change are given.

The importance of modelling in the optimization process is also noted by modelers. Practical experience shows that model building in mathematical programming needs approximately 2/3 of the entire time spent with the problem. Model building in mathematical programming and many 'modelling tricks' for different classes of problems were discussed in detail in literature and journals.

Application of problem analysis and model design. The 10 questions introduced earlier may help you to specify a type of the problem. There are two ways to optimization model building:

- Either use existing Input/Output model having the form $\mathbf{y} = \mathbf{h}(\mathbf{x})$ and then: (1) choose $k \in \{1, \ldots, m\}$ and $f(\mathbf{x}) = h_k(\mathbf{x})$ the objective function, (2) specify l_j and u_j lower and upper bounds for $j \in \{1, \ldots, m\} \setminus \{k\}$ and set $g_j = h_j$ for all considered j to get $S = \{\mathbf{x} \mid \forall j : l_j \leq g_j(\mathbf{x}) \leq u_j, \mathbf{x} \in \mathbb{R}^n\}$. Then mathematical program $\min\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ is built.
- Build the model in the bottom-top style. It means to identify (1) the goal $\min z$, (2) the objective $z = f(\mathbf{x})$, (3) the variables $\mathbf{x} \in \mathbb{R}^n$, (4) their bounds $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$, and (5) constraints (I/O (balance) equalities $=$, 'demand'constraints $\geq$, and 'storage' constraints $\leq$.

Example 18 (Continuation – problem analysis)**.**

We choose the second possibility for our example: $\min z$ means to minimize overall transportation costs. $z = f(\mathbf{x}, \mathbf{y}, \mathbf{w})$ costs depend on location of warehouse (x_j, y_j) and on the amount w_j transported from warehouse i to town j. Hint: Compare it with a usual transportation problem and use it partially!

Example 19 (Continuation – model building)**.**

z	cost of total transport is proportional to
	distances and shipments from warehouses
(x_i, y_i)	unknown location of warehouse $i \in \{1, \ldots, m\}$
c_i	known capacity of warehouse $i \in \{1, \ldots, m\}$
(a_j, b_j)	known location of market $j \in \{1, \ldots, n\}$
r_j	known demand at market $j \in \{1, \ldots, n\}$
d_{ij}	unknown distance from warehouse i to market area j
w_{ij}	unknown units shipped from i to j

$$\min z$$

$$z = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} w_{ij}$$

$$\sum_{j=1}^{n} w_{ij} \leq c_i, \ i = 1, \ldots, m$$

$$\sum_{i=1}^{m} w_{ij} \geq r_j, \ j = 1, \ldots, n$$

$$w_{ij} \geq 0$$

Where is the nonlinearity?

$$d_{ij} = \sqrt{(x_i - a_j)^2 + (y_i - b_j)^2} = \|(x_i, y_i) - (a_j, b_j)\|_2$$

Also another norm than$\|\cdot\|_2$may be used

$$l \leq x_i, y_i \leq u, \ i = 1, \ldots, m$$

Exercise 20.

Fill the model with real data for 10 towns, 3 warehouses, e.g. for Malta island! Visualize your data.

Exercise 21.

Use MS Excel for real data computations including optimization.

Useful theoretical properties. At first, we need to know theoretical properties of the model. Weierstrass' Theorem will answer us: 'Does the optimal solution exist?' Other questions remain, e.g. does a unique solution exist?

Exercise 22.

Use Weierstrass' Theorem to recognize whether the optimal solution exists.

Algorithm choice. We must learn more about the algorithms for NLP (this semester), as the simplex method for LP is not applicable. At the beginning, we may restrict ourselves to the use of optimization software. So, how to solve NLP?

Some software tools.

non-specialized	Excel, Matlab, etc.
own programs	hard to compete as you do not know unpublished numerical tricks
procedures	you should know their programming language
solvers	MINOS, CONOPT, etc. Difficult data communication
modeling languages	GAMS, AMPL, AIMMS - the right choice
interfaces	MPL model building, AIMMS report writing

See `www.gams.com` and `www.paragon.nl` for more information. Model more problems as nonlinear programs — see below.

Exercise 23 (Parallelogram)**.**

Formulate a mathematical program (using analytic geometry): Find a point E on the side BC of the triangle ABC such that the parallelogram with vertices D resp. F lying on the sides AB resp. AC has maximal area. Solve it using a computer program (Matlab, Excel, GAMS). Hint: The solution is obviously given by choosing E to be the midpoint of BC. In fact for arbitrary E on BC with $\lambda = BE/BC$ the area of the corresponding parallelogram is the area of (ADEF) $= 2\lambda(1-\lambda)$ multiplying the area of(ABC) and this function is maximal for lambda $= 1/2$.

Exercise 24 (Heron).

Heron (ca. 100 BC) gave a solution to the following problem: On a given line find a point C such that the sum of the distances to the points A and B is minimal. Compare the modeling and solution by mathematical programming with the explanatory and simple geometric approach.
Hint: If A and B lie on opposite sides of the given line, then obviously the intersection of the line with the segment AB is the desired point; otherwise one reflects A in the line getting A' and determines C as the intersection of the line with the segment A'B. Hence: a light ray reflected in the line with angle of incidence and angle of reflection equal takes the shortest possible path from A to B via C. Or: if one wants to go from A to B on the shortest possible route and on the way fetch a pail of water from a (straight) river, then one must solve the same problem.

Exercise 25 (Kepler).

Minimize the cylinder surface S having the constant cylinder volume V.
J. Kepler lived in Prague in 17th century. During his walks, he observed that barrels for wine are sold with only one measurement. Solving this problem, he wrote a paper why two measurements are not necessary for the 'optimal' barrels.

Exercise 26 (Paper box).

Having a sheet of paper, use scissors to create a box with the biggest volume.

Advanced applications of NLP may be found in Bracken, McCormick: Selected applications of nonlinear programming, Wiley, 1968 (a copy available from Dr.Sklenar).

Think about problems that are also implemented in the GAMS system:
WEAPONS.GMS, PROCESS.GMS, CHEM.GMS,
LINEAR.GMS, LEAST.GMS, LIKE.GMS.

Other real-life problems may serve as bases for diploma thesis:

- Melt control problem: Optimize the sequence of charges of input materials to minimize a cost and maximize a quality of produced alloy with respect to constraints involving random losses and goal intervals for the melt composition.
- Irrigation network design: Optimize a reconstructed irrigation pipe network topology, minimize costs, satisfy pressure and flow uncertain demands.
- Heat exchanger parameters optimization: Optimize exchanger design parameters minimizing investment and maintenance costs subject to technical and economical constraints having parameters with values estimated from experimental data.

1.3 Convex sets

1.3.1 Convex sets, combinations, and hulls

Definition 27 (Convex set)**.**

Let $S \subset \mathbb{R}^n$. We say S is a convex set $\Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2, \forall \lambda \in [0;1]: \ \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$.

Definition 27 is very important for understanding and memorizing. "The convex set is a set where you cannot hide yourself." For the convex set, all points of any line segment connecting two points from it again belong to it.

Exercise 28.

Draw figures describing different convex and non-convex sets in $\mathbb{R}^2$. Discuss whether a segment line, triangle, circle, etc. are convex sets. Does exist a concave set? Hint: No!!!

Definition 29 (Convex combination)**.**

Let $\mathbf{x}$, $\mathbf{x}_1$, ..., $\mathbf{x}_k \in \mathbb{R}^n$. Then $\mathbf{x}$ satisfying $\mathbf{x} = \sum_{j=1}^k \lambda_j \mathbf{x}_j$ where $\sum_{j=1}^k \lambda_j = 1$ and $\forall j \in \{1, \ldots, k\} : \lambda_j \geq 0$ is called a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

Specifically: The point $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is a convex combination of $\mathbf{x}_1$ and $\mathbf{x}_2$. Review:

1. The point $\mathbf{x} = \sum_{j=1}^k \lambda_j \mathbf{x}_j$ where $\sum_{j=1}^k \lambda_j = 1$, $\forall j \in \{1, \ldots, k\} : \lambda_j \in \mathbb{R}$ is called an affine combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.
2. The point $\mathbf{x} = \sum_{j=1}^k \lambda_j \mathbf{x}_j$ where $\forall j \in \{1, \ldots, k\} : \lambda_j \in \mathbb{R}$ is called a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

Example 30.

Convexity of the following sets is important in optimization:

1. Let $\mathbf{p} \in \mathbb{R}^n$, $\bar{\mathbf{x}} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, and $\mathbf{p}^\top \bar{\mathbf{x}} = \alpha$ then hyperplane H specified using the normal vector $\mathbf{p}$ and constant α (or fixed point $\bar{\mathbf{x}}$) by $H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{p}^\top \mathbf{x} = \alpha\} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{p}^\top (\mathbf{x} - \bar{\mathbf{x}}) = 0\}$ is a convex set.
2. Let $\mathbf{p} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ then halfspace H^+ defined by $H^+ = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{p}^\top \mathbf{x} \leq \alpha\}$ is a convex set.
3. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ then the set S defined by $S = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is a convex set.
4. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{c} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$ then the feasible set and set of optimal solutions of a linear program $? \in \operatorname{argmin}\{\mathbf{c}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ are convex sets.

Exercise 31.

Using a definition 27 prove the convexity of sets introduced in Example 30. Illustrate all cases by figures in $\mathbb{R}^2$.

Lemma 32 (Properties of convex sets)**.**

Let $S_1, S_2 \subset \mathbb{R}^n$ be convex sets. Then:

1. $S_1 \cap S_2$ is a convex set.
2. $S_1 \oplus S_2 = \{\mathbf{x}_1 + \mathbf{x}_2 \mid \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$ is a convex set.
3. $S_1 \ominus S_2 = \{\mathbf{x}_1 - \mathbf{x}_2 \mid \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$ is a convex set.

Exercise 33.

Prove Lemma 32. The fact that the intersection of convex sets is again a convex set is very useful in optimization. Why? Hint: For example, think about the feasible set of a linear programming problem.

Definition 34 (Convex hull)**.**

Let $S \subset \mathbb{R}^n$, we say that $\mathbf{x} \in \mathbb{R}^n$ belongs to a convex hull of S ($\mathbf{x} \in \operatorname{conv} S$) $\Leftrightarrow \exists k \in \mathbb{N}$, $\exists \lambda_1, \ldots, \lambda_k \geq 0$, $\sum_{j=1}^{k} \lambda_j = 1$, $\exists \mathbf{x}_1, \ldots, \mathbf{x}_k \in S : \mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$.

Exercise 35.

Draw figures describing convex hulls for different sets in $\mathbb{R}^2$.

Remark 36 (Properties of convex hulls)**.**

Remember, understand, and draw figures: Let $S \subset \mathbb{R}^n$ then:

1. $\operatorname{conv} S$ is the smallest convex set containing S (outside specification).
2. $\operatorname{conv} S$ is the intersection of all convex sets containing S.
3. For S convex, $S = \operatorname{conv} S$ is true.
4. $\operatorname{conv} S$ is the set of all convex combinations from S (inside specification).
5. An affine hull $\operatorname{aff} S$ is the set of all affine combinations of points in S. It is the smallest affine subspace containing S.
6. A linear hull $\operatorname{lin} S$ is the set of all linear combinations of points in S. It is the smallest linear subspace.

Exercise 37.

Drawing figures, compare concepts of $\operatorname{conv} S, \operatorname{aff} S$, and $\operatorname{lin} S$ using examples in $\mathbb{R}^2$. Hint: Use two points in $\mathbb{R}^2$.

Definition 38 (Polytope, simplex)**.**

Let $\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1} \in \mathbb{R}^n$ then:

1. $\operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}\}$ is called a polytope.
2. If $\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_1, \ldots, \mathbf{x}_{k+1} - \mathbf{x}_1$ are linearly independent then we say that $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$ are affinely independent.
3. If $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$ are affinely independent then $\operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}\}$ is called a simplex with vertices $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$.
4. There could be no simplex with more than $n+1$ vertices.

Theorem 39 (Caratheodory)**.**

Let $S \subset \mathbb{R}^n$ then $\forall \mathbf{x} \in \operatorname{conv} S: \ \exists\ \lambda_j \geq 0,\ j \in \{1, \ldots, n+1\},\ \sum_{j=1}^{n+1} \lambda_j = 1,\ \exists\ \mathbf{x}_j \in S,$ $j \in \{1, \ldots, n+1\} : \mathbf{x} = \sum_{j=1}^{n+1} \lambda_j \mathbf{x}_j$.

Caratheodory's Theorem says that any point in the convex hull of set S can be represented as a convex combination of <u>at most</u> $n+1$ points in S (remember that n is a dimension of $\mathbb{R}^n$).

Exercise 40.

How many points do you need when $\mathbf{x} \in S$ (not in $\operatorname{conv} S \setminus S$)?

Proof of Theorem 39: Let $S \subset \mathbb{R}^n$ and $\mathbf{x} \in \operatorname{conv} S$. Then (by Definition 34) $\exists k: \ \exists \mathbf{x}_j \in S\ , \exists \lambda_j > 0\ ,$ $j = 1, \ldots, k\ , \sum_{j=1}^{k} \lambda_j = 1$ and $\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$. If $k \leq n+1$ then the theorem is proven. Otherwise, $k > n+1$ and we continue. The main idea is to eliminate one $\lambda_i > 0$. As $k \geq n+2$ then $\mathbf{x}_2 - \mathbf{x}_1, \ldots, \mathbf{x}_k - \mathbf{x}_1$ are linearly dependent (because $k - 1 \geq n + 1$). So $\exists \mu_2, \ldots, \mu_k \in \mathbb{R}$ not all equal zero and satisfying $\sum_{j=2}^{k} \mu_j (\mathbf{x}_j - \mathbf{x}_1) = \mathbf{0}$. We denote $\mu_1 = -\sum_{j=2}^{k} \mu_j$. Then $\sum_{j=1}^{k} \mu_j \mathbf{x}_j = \mathbf{0}$, $\sum_{j=1}^{k} \mu_j = 0$, and $(\mu_1, \ldots, \mu_k)^\top \neq \mathbf{0}$. So, at least one $\mu_j > 0$. Then we may write: $\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \mathbf{0} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j - \alpha \mathbf{0} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j - \alpha \sum_{j=1}^{k} \mu_j \mathbf{x}_j = \sum_{j=1}^{k} (\lambda_j - \alpha \mu_j) \mathbf{x}_j$. Now we choose suitable α to eliminate one nonzero λ_j: $\alpha = \min_{1 \leq j \leq k} \{\frac{\lambda_j}{\mu_j} \mid \mu_j > 0\}$. Assume that $\alpha = \frac{\lambda_i}{\mu_i}$. Because $\lambda_j > 0$ and $\mu_j > 0$ in the minimization term then $\alpha > 0$. Then for other j: (1) $\mu_j \leq 0 \Rightarrow \lambda_j - \alpha \mu_j > 0$, (2) $\mu_j > 0 \Rightarrow \frac{\lambda_j}{\mu_j} \geq \frac{\lambda_i}{\mu_i} = \alpha$, and hence, $\lambda_j - \alpha \mu_j > 0$. In addition $\lambda_i - \alpha \mu_i = 0$. We have $\mathbf{x} = \sum_{j=1 j \neq i}^{k} (\lambda_j - \alpha \mu_j) \mathbf{x}_j$ and $\sum_{j=1 j \neq i}^{k} (\lambda_j - \alpha \mu_j) = 1$. Therefore $\mathbf{x}$ is a convex combination of $k-1$ points in S. This step can be repeated till $k = n+1$ and then the proof is complete. □

1.3.2 Closed and open convex sets

Theorem 41 (About line segment)**.**

Let $S \subset \mathbb{R}^n$ be a convex set, $\operatorname{int} S \neq \emptyset$, $\mathbf{x}_1 \in \operatorname{cl} S$, and $\mathbf{x}_2 \in \operatorname{int} S$. Then $\forall \lambda \in (0;1)$: $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \operatorname{int} S$.

Theorem says: Given a convex set with a nonempty interior, the line segment (without the end points) joining a point in the interior of the set and a point in the closure of the set belongs to the interior of the set.

Proof of Theorem 41: Let $S \subset \mathbb{R}^n$ be convex, $\operatorname{int} S \neq \emptyset$, $\mathbf{x}_1 \in \operatorname{cl} S$, $\mathbf{x}_2 \in \operatorname{int} S$. Because $\mathbf{x}_2 \in \operatorname{int} S \Rightarrow \exists \mathcal{N}_\varepsilon(\mathbf{x}_2) \subset S : \{\mathbf{z} \mid \ \|\mathbf{z} - \mathbf{x}_2\| < \varepsilon\} \subset S$. We denote $\mathbf{y} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for $\lambda \in (0;1)$ chosen arbitrarily. We need to show that if $\mathbf{y} \in \operatorname{int} S$ then $\exists \mathcal{N}_{\varepsilon'} : \mathcal{N}_{\varepsilon'} \subset S$. We choose $\varepsilon' = (1-\lambda)\varepsilon$. So, we want to show $\{\mathbf{z} \mid \|\mathbf{z} - \mathbf{y}\| < (1-\lambda)\varepsilon\} \subset S$. We take $\mathbf{z} : \|\mathbf{z} - \mathbf{y}\| < (1-\lambda)\varepsilon$. We want to show $\mathbf{z} \in S$. It can be done showing that $\mathbf{z}$ is a convex combination of 2 points of S. So, $\exists \mathbf{z}_1$ satisfying

$$\mathbf{x}_1 \in \operatorname{cl} S \Rightarrow \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_1\| < \frac{(1-\lambda)\varepsilon - \|\mathbf{z} - \mathbf{y}\|}{\lambda}\} \cap S \neq \emptyset.$$

Set $\mathbf{z}_2 = \frac{\mathbf{z} - \lambda \mathbf{z}_1}{1-\lambda}$. Then $\mathbf{z} = \lambda \mathbf{z}_1 - (1-\lambda)\mathbf{z}_2$ and because $\mathbf{x}_2 = \frac{\mathbf{y} - \lambda \mathbf{x}_1}{1-\lambda}$:

$$\begin{aligned} \|z_2 - \mathbf{x}_2\| &= \|\frac{\mathbf{z} - \lambda \mathbf{z}_1}{1-\lambda} - \mathbf{x}_2\| = \\ \|\frac{\mathbf{z} - \lambda \mathbf{z}_1 - (\mathbf{y} - \lambda \mathbf{x}_1)}{1-\lambda}\| &= \frac{1}{1-\lambda}\|\frac{\mathbf{z} - \lambda \mathbf{z}_1 - (\mathbf{y} - \lambda \mathbf{x}_1)}{1-\lambda}\| \leq \\ \frac{1}{1-\lambda}(\|\mathbf{z} - \mathbf{y}\| + \lambda\|x_1 - z_1\|) &< \frac{1}{1-\lambda}(\|\mathbf{z} - \mathbf{y}\| + \lambda\frac{(1-\lambda)\varepsilon - \|\mathbf{z} - \mathbf{y}\|}{\lambda}) = \varepsilon. \end{aligned}$$

Then $z_2 \in \operatorname{int} S \Rightarrow \mathbf{z} \in S \Rightarrow \mathbf{y} \in \operatorname{int} S$. □

Exercise 42.

Draw figures and give counterexamples that the theorem does not remain valid when any of its assumptions is omitted.

Corollary 43 (Closure, interior, and convex hull)**.**

Let S be a convex set, so then $\operatorname{int} S$ is convex. Let S be convex and $\operatorname{int} S \neq \emptyset \Rightarrow \operatorname{cl} S$ is convex, $\operatorname{cl}(\operatorname{int} S) = \operatorname{cl} S$, and $\operatorname{int}(\operatorname{cl} S) = \operatorname{int} S$.

Proof: To prove: S is convex and $\operatorname{int} S \neq \emptyset \Rightarrow \operatorname{cl} S$ is convex. So, $\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{cl} S$ arbitrarily. $\exists \mathbf{z} \in \operatorname{int} S$. By theorem 41, we get $\forall \lambda \in (0;1) : \lambda \mathbf{x}_2 + (1-\lambda)\mathbf{z} \in \operatorname{int} S$. We fix $\mu \in (0;1)$: and by theorem $\forall \lambda \in (0;1)$: $\mu \mathbf{x}_1 + (1-\mu)(\lambda \mathbf{x}_2 + (1-\lambda)\mathbf{z}) \in \operatorname{int} S \subset S$. Then $\lambda \longrightarrow 1 \Rightarrow \mu \mathbf{x}_1 + (1-\mu)\mathbf{x}_2 \in \operatorname{cl} S$. □

To prove: S is convex and $\operatorname{int} S \neq \emptyset \Rightarrow \operatorname{cl} \operatorname{int} S = \operatorname{cl} S$. So, we know $\operatorname{cl}(\operatorname{int} S) \subset \operatorname{cl} S$ because $\operatorname{int} S \subset S$. We want to prove $\operatorname{cl}(\operatorname{int} S) \supset \operatorname{cl} S$ Assume $\mathbf{x} \in \operatorname{cl} S$ and $\mathbf{y} \in \operatorname{int} S \neq \emptyset \Rightarrow \forall \lambda \in (0;1)\ \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in \operatorname{int} S$. Then $\lambda \longrightarrow 1^- \Rightarrow \mathbf{x} \in \operatorname{cl} \operatorname{int} S$. □

To prove: S is convex and $\operatorname{int} S \neq \emptyset \Rightarrow \operatorname{int}(\operatorname{cl} S) = \operatorname{int} S$. We know $\operatorname{int} S \subset \operatorname{int}(\operatorname{cl} S)$ because $S \subset \operatorname{cl} S$. We take $\mathbf{x}_1 \in \operatorname{int}(\operatorname{cl} S)$ and want to show that $\mathbf{x}_1 \in \operatorname{int} S$. $\exists \varepsilon > 0 : \|\mathbf{y} - \mathbf{x}_1\| < \varepsilon \Rightarrow \mathbf{y} \in \operatorname{cl} S$. We get $\mathbf{x}_2 \in \operatorname{int} S$ such that $\mathbf{y} = (1+\Delta)\mathbf{x}_1 - \Delta \mathbf{x}_2$ where $\Delta = \frac{\varepsilon}{2\|x_1 - \mathbf{x}_2\|}$ and $(\mathbf{x}_2 \neq \mathbf{x}_1)$. Since $\|y - \mathbf{x}_1\| = \varepsilon/2$ $\Rightarrow \mathbf{y} \in \operatorname{cl} S$. Therefore, we can write $\mathbf{x}_1 = \lambda \mathbf{y} + (1-\lambda)\mathbf{x}_2$ where $\lambda = \frac{1}{1+\Delta} \in (0;1)$. Since $\mathbf{y} \in \operatorname{cl} S$ and $\mathbf{x}_2 \in \operatorname{int} S$ then by Theorem 41, we conclude $\mathbf{x}_1 \in \operatorname{int} S$. □

Exercise 44.

Illustrate principle ideas of proofs by drawings in $\mathbb{R}^2$.

1.3.3 Separating hyperplanes

Almost all optimality conditions and duality relationships in nonlinear programming use some sort of separation of convex sets or supporting hyperplanes of convex sets theorems. For them, previous results about $\mathrm{cl}, \mathrm{int}$, and conv are needed together with the following theorem:

Theorem 45 (Minimum distance)**.**

Let $S \subset \mathbb{R}^n, S \neq \emptyset$, $S = \mathrm{cl}\, S$ (i.e. S is closed), $S = \mathrm{conv}\, S$ (i.e. S is convex), and $\mathbf{y} \notin S \Rightarrow \exists! \bar{\mathbf{x}} \in S : \bar{\mathbf{x}} \in \mathrm{argmin}_{\mathbf{x}}\{\|\mathbf{y}-\mathbf{x}\| \mid \mathbf{x} \in S\}$ and $\bar{\mathbf{x}} \in \mathrm{argmin}_{\mathbf{x}}\{\|\mathbf{y}-\mathbf{x}\| \mid \mathbf{x} \in S\}$ $\Rightarrow \forall \mathbf{x} \in S : (\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}}) \leq 0$.

Lemma 46 (Cosine law)**.**

$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n : \|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\mathbf{a}^\top\mathbf{b}$.

Proof: Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and γ is the angle defined by them. Then $\|\mathbf{a}+\mathbf{b}\|^2 = (\|\mathbf{b}\| - \|\mathbf{a}\|\cos\gamma)^2 + (\|\mathbf{a}\|\sin\gamma)^2 = \|\mathbf{a}\|^2(\sin^2\gamma + \cos^2\gamma) + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\cdot\|\mathbf{b}\|\cdot\cos\gamma = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\mathbf{a}^\top\mathbf{b}$ because $\cos\gamma = \frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|\cdot\|\mathbf{b}\|}$ (draw a figure in $\mathbb{R}^2$ and use the Pythagoras' theorem). □

Lemma 47 (Parallelogram law)**.**

$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n : \|\mathbf{a}+\mathbf{b}\|^2 + \|\mathbf{a}-\mathbf{b}\|^2 = 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$.

Proof: By the cosine law: $\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n : \|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\mathbf{a}^\top\mathbf{b}$ and $\|\mathbf{a}-\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a}^\top\mathbf{b}$. By sum of both equalities, we get $2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ and proof is complete. □

Exercise 48.

Assuming that $\|\mathbf{a}\| = \sqrt{\sum_{j=1}^n a_j^2}$ and drawing figures in $\mathbb{R}^2$ explain ideas of both lemmas. Hint: The parallelogram law says that the sum of squared norms of the parallelogram diagonals equals the sum of squared norms of its sides.

Proof of Theorem 45: Existence of $\bar{\mathbf{x}}$: Because $S \neq \emptyset \Rightarrow \exists \hat{\mathbf{x}} \in S$ one fixed point and we define $\bar{S} := \{\mathbf{x} \mid \|\mathbf{y}-\mathbf{x}\| \leq \|\mathbf{y}-\hat{\mathbf{x}}\|\} \cap S$. Because S is closed then $\bar{S}$ is also closed and bounded ($\hat{\mathbf{x}}$ is fixed), and hence, $\bar{S}$ is a compact set. We know that $\|\mathbf{y}-\mathbf{x}\|$ is continuous in $\mathbf{x}$. (Note $\|\mathbf{y}-\mathbf{x}\| = \sqrt{\sum_{j=1}^n (y_j - x_j)^2}$, and for the angle α of $\mathbf{y}-\bar{\mathbf{x}}$ and $\mathbf{x}-\bar{\mathbf{x}}$ then $\cos\alpha = \frac{(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})}{\|\mathbf{y}-\bar{\mathbf{x}}\|\cdot\|\mathbf{x}-\bar{\mathbf{x}}\|}$.) In addition $\inf\{\|\mathbf{y}-\mathbf{x}\| \mid \mathbf{x} \in S\} = \inf\{\|\mathbf{y}-\hat{\mathbf{x}}\| \mid \mathbf{x} \in S\} \cap S$. By Weierstrass' theorem even minimum $\bar{\mathbf{x}}$ exists.

Uniqueness of $\bar{\mathbf{x}}$: We assume $\bar{\mathbf{x}}' \in S : \|\mathbf{y}-\bar{\mathbf{x}}\| = |\mathbf{y}-\bar{\mathbf{x}}'\| = \gamma$ and we show that $\bar{\mathbf{x}} = \bar{\mathbf{x}}'$. By convexity of S $(\bar{\mathbf{x}}+\bar{\mathbf{x}}')/2 \in S$. We check $\mathbf{y} - \frac{\bar{\mathbf{x}}-\bar{\mathbf{x}}'}{2}\| = \frac{1}{2}\|(\mathbf{y}-\bar{\mathbf{x}}) + (\mathbf{y}-\bar{\mathbf{x}}')\| \leq$ (Schwarz inequality)

$\frac{1}{2}\|\mathbf{y}-\bar{\mathbf{x}}\|+\frac{1}{2}\|\mathbf{y}-\bar{\mathbf{x}}'\|=\gamma$. If $<$ then $\frac{\bar{\mathbf{x}}+\bar{\mathbf{x}}'}{2}$ is nearer and we obtain a contradiction with $\bar{\mathbf{x}}\in\operatorname{argmin}\{\ldots\}$. If $=$ then $\exists\lambda:\mathbf{y}-\bar{\mathbf{x}}=\lambda(\mathbf{y}-\bar{\mathbf{x}}')$. As $\|\mathbf{y}-\bar{\mathbf{x}}\|=\|\mathbf{y}-\bar{\mathbf{x}}'\|=\gamma$ then we have $\mid\lambda\mid=1$. If $\lambda=1$ then $\bar{\mathbf{x}}=\bar{\mathbf{x}}'$, so $\bar{\mathbf{x}}$ is unique. If $\lambda=-1$ then from $\mathbf{y}-\bar{\mathbf{x}}=\lambda(\mathbf{y}-\bar{\mathbf{x}}')$ we get $\mathbf{y}=\frac{\bar{\mathbf{x}}+\bar{\mathbf{x}}'}{2}\in S$, however by assumption $\mathbf{y}\notin S$ and contradiction.

To complete the proof $\forall\mathbf{x}\in S:((\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})\le 0)\Rightarrow\bar{\mathbf{x}}\in\operatorname{argmin}\{\|\mathbf{y}-\mathbf{x}\|\mid\mathbf{x}\in S\}$.

Sufficiency $\Rightarrow$: $\mathbf{x}\in S\Rightarrow\|\mathbf{y}-\mathbf{x}\|^2=\|\mathbf{y}-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x}\|^2=$ by parallelogram law $\|\mathbf{y}-\bar{\mathbf{x}}\|^2+\|\bar{\mathbf{x}}-\mathbf{x}\|^2+2(\bar{\mathbf{x}}-\mathbf{x})^\top(\mathbf{y}-\bar{\mathbf{x}})$. Note that $\|\bar{\mathbf{x}}+\mathbf{x}\|\ge 0$ and $2(\bar{\mathbf{x}}-\mathbf{x})^\top(\mathbf{y}-\bar{\mathbf{x}})=-2(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})\ge 0$ by assumption. Hence: $\|\mathbf{y}-\mathbf{x}\|^2\ge\|\mathbf{y}-\bar{\mathbf{x}}\|^2\Rightarrow\bar{\mathbf{x}}\in\operatorname{argmin}\{\|\mathbf{y}-\mathbf{x}\|\mid\mathbf{x}\in S\}$.

Necessity $\Leftarrow$: $\mathbf{x}\in\operatorname{argmin}\{\|\mathbf{y}-\mathbf{x}\|\mid\mathbf{x}\in S\}\Rightarrow\forall\mathbf{x}\in S:\|\mathbf{y}-\mathbf{x}\|\ge\|\mathbf{y}-\bar{\mathbf{x}}\|$, so $\|\mathbf{y}-\mathbf{x}\|^2\ge\|\mathbf{y}-\bar{\mathbf{x}}\|^2$. $\mathbf{x}\in S$ is chosen arbitrarily, S is convex then $\forall\lambda\in[0;1]:\lambda\mathbf{x}+(1-\lambda)\bar{\mathbf{x}}=\bar{\mathbf{x}}+\lambda(\mathbf{x}-\bar{\mathbf{x}})\in S$. So $\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2\ge|\ \mathbf{y}-\bar{\mathbf{x}}\|^2$. By Lemma 47 $\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2=\|\mathbf{y}-\bar{\mathbf{x}}\|^2+\lambda^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2-2\lambda(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})$. Together we have $\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2-\lambda^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2+2\lambda(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})=\|\mathbf{y}-\bar{\mathbf{x}}\|^2\le\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2$. Therefore, $2\lambda(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})\le\lambda^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2$. It is trivially satisfied for $\lambda=0$. Assuming $\lambda>0$ and multiplying the inequality by $1/(2\lambda)$, we get: $(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}})\le\frac{\lambda}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2$. The right-hand-side (RHS) is greater than 0. As $\lambda\longrightarrow 0^+$ (because $\lambda\in[0;1]$) then the RHS approaches 0 and the theorem is proven. □

Definition 49 (Hyperplane)**.**

Let $\alpha\in\mathbb{R}$, $\mathbf{p}\in\mathbb{R}^n$, and $\mathbf{p}\neq\mathbf{0}$. A hyperplane H in $\mathbb{R}^n$ is defined by $H=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}=\alpha\}$.
It also defines two closed halfspaces $H^+=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}\ge\alpha\}$ and $H^-=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}\le\alpha\}$ and two open halfspaces $H_0^+=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}>\alpha\}$ and $H_0^-=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}<\alpha\}$.

Alternatively, for $\bar{\mathbf{x}}\in H:\mathbf{p}^\top\bar{\mathbf{x}}=\alpha$ and $H=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top\mathbf{x}=\mathbf{p}^\top\bar{\mathbf{x}}\}=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{p}^\top(\mathbf{x}-\bar{\mathbf{x}})=0\}$. Similarly, we may define H^+,H^-,H_0^+,H_0^- alternatively.

Definition 50 (Separating hyperplanes)**.**

Let $S_1,S_2\subset\mathbb{R}^n$ and $S_1\neq\emptyset$, $S_2\neq\emptyset$. Define $H=\{\mathbf{x}\mid\mathbf{p}^\top\mathbf{x}=\alpha\}$. Then we say that:

H separates S_1 and $S_2\Leftrightarrow\exists i\in\{1,2\}:(\forall\mathbf{x}\in S_i:\mathbf{p}^\top\mathbf{x}\ge\alpha)\wedge(\forall\mathbf{x}\in S_{3-i}:\mathbf{p}^\top\mathbf{x}\le\alpha)$.

If in addition $S_1\cup S_2\not\subset H$ then H is said to properly separate S_1 and S_2.

H strictly separates S_1 and $S_2\Leftrightarrow\exists i\in\{1,2\}:(\forall\mathbf{x}\in S_i:\mathbf{p}^\top\mathbf{x}>\alpha)\wedge(\forall\mathbf{x}\in S_{3-i}$ $\mathbf{p}^\top\mathbf{x}<\alpha)$.

H strongly separates S_1 and $S_2\Leftrightarrow\exists\in\{1,2\}\ \exists\varepsilon>0:(\forall\mathbf{x}\in S_i:\mathbf{p}^\top\mathbf{x}\ge\alpha+\varepsilon)\wedge$ $(\forall\mathbf{x}\in S_{3-i}:\mathbf{p}^\top\mathbf{x}\le\alpha)$.

Theorem 51 (Separation of point and convex set)**.**

Let $S\subset\mathbb{R}^n,S\neq\emptyset$, S be closed and convex, and $\mathbf{y}\notin S\Rightarrow\exists\mathbf{p}\in\mathbb{R}^n\neq\mathbf{0}$, $\exists\alpha\in\mathbb{R}:\forall\mathbf{x}\in S:\mathbf{p}^\top\mathbf{x}\le\alpha\wedge\mathbf{p}^\top\mathbf{y}>\alpha$.

Proof: By Theorem 45 $\exists\bar{\mathbf{x}}\in\operatorname{argmin}\{\|\mathbf{y}-\mathbf{x}\|\mid\mathbf{x}\in S\}$ and $\forall\mathbf{x}\in S:(\mathbf{y}-\mathbf{x})^\top(\mathbf{x}-\bar{\mathbf{x}})=(\mathbf{x}-\bar{\mathbf{x}})^\top(\mathbf{y}-\bar{\mathbf{x}})\le 0$. We define $\mathbf{p}=\mathbf{y}-\bar{\mathbf{x}}$, $\alpha=(\mathbf{y}-\bar{\mathbf{x}})^\top\bar{\mathbf{x}}$, and we get $\mathbf{p}^\top\mathbf{x}\le\alpha$. Then $\mathbf{p}^\top\mathbf{y}-\alpha=\mathbf{p}^\top\mathbf{y}-(\mathbf{y}-\bar{\mathbf{x}})^\top\bar{\mathbf{x}}=(\mathbf{y}-\bar{\mathbf{x}})^\top\mathbf{y}-(\mathbf{y}-\bar{\mathbf{x}})^\top\bar{\mathbf{x}}=(\mathbf{y}-\bar{\mathbf{x}})^\top(\mathbf{y}-\bar{\mathbf{x}})=\|\mathbf{y}-\bar{\mathbf{x}}\|^2>0$ because $\mathbf{y}\notin\operatorname{cl}S$, $\bar{\mathbf{x}}\in\operatorname{cl}S$, and $\mathbf{y}\neq\bar{\mathbf{x}}$. □

Exercise 52.

Using figures and examples in $\mathbb{R}^2$, show the importance of assumptions in separating theorems.

Corollary 53 (Intersection of halfspaces)**.**

Let $S \subset \mathbb{R}^n$ be closed and convex $\Rightarrow$ S is the intersection of halfspaces containing S.

Proof: $S \subset \cap H^+$ trivially true. To show $S \supset \cap H^+$, we use a contradiction expecting $\mathbf{y} \notin S, \mathbf{y} \in \cap H^+$. Then by Theorem 51 $\exists H$ hyperplane that strongly separates S ($: S \subset H^+$) and $\mathbf{y}$ ($\mathbf{y} \notin H^+$). So, the contradiction is obtained. □

Corollary 54.

Let $S \subset \mathbb{R}^n$ and $S \neq \emptyset$, $\mathbf{y} \notin \operatorname{cl}\operatorname{conv} S \Rightarrow \exists H$ hyperplane strongly separating S and $\mathbf{y}$.

Proof: Use $\operatorname{cl}\operatorname{conv} S$ instead of S and apply Theorem 51. □

Remark 55 (Equivalent conclusions)**.**

Theorem 51 conclusion is equivalent to:

1. $\exists H$ hyperplane strictly separating S and $\mathbf{y}$.
2. $\exists H$ hyperplane strongly separating S and $\mathbf{y}$.
3. $\exists \mathbf{p} \in \mathbb{R}^n : \mathbf{p}^\top \mathbf{y} > \sup\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S\}$
4. $\exists \mathbf{p} \in \mathbb{R}^n : \mathbf{p}^\top \mathbf{y} < \inf\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S\}$

Exercise 56.

Explain Theorem 51 and the following Corollaries and Remarks drawing figures. Hints: Think about the use of the graphical solution of NLP and LP. Compare sup an $\mathbf{p}^\top \bar{\mathbf{x}}$ values.

Theorem 57 (Farkas)**.**

$\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix, $\mathbf{c} \in \mathbb{R}^n$ is a vector. Denote $S_1 = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0} \wedge \mathbf{c}^\top \mathbf{x} > 0\}$ and $S_2 = \{\mathbf{y} \mid \mathbf{A}^\top \mathbf{y} = \mathbf{c} \wedge \mathbf{y} \geq \mathbf{0}\} \Rightarrow (S_1 \neq \emptyset \wedge S_2 = \emptyset) \vee (S_1 = \emptyset \wedge S_2 \neq \emptyset)$.

Farkas' theorem is extensively used in the derivation of optimality conditions of LP and NLP. It says that either S_1 or S_2 is nonempty (empty).

Proof: Assume that $S_2 \neq \emptyset$. Then $\exists \mathbf{y} \geq \mathbf{0} : \mathbf{A}^\top \mathbf{y} = \mathbf{c}$. We get $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} \leq \mathbf{0}$. (It exists as $\mathbf{x}$ may be chosen $\mathbf{0}$.) Then $\mathbf{c}^\top \mathbf{x} = (\mathbf{A}^\top \mathbf{y})\mathbf{x} = \mathbf{y}^\top \mathbf{A}\mathbf{x}$. Because $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{A}\mathbf{x} \geq \mathbf{0}$ then $\mathbf{y}^\top \mathbf{A}\mathbf{x} \leq \mathbf{0}$, and hence, $\mathbf{c}^\top \mathbf{x} \not> 0$ then $S_1 = \emptyset$.

Assume $S_2 = \emptyset$. It means $S = \{\mathbf{x} \mid \exists \mathbf{y} \in \mathbb{R}^m : \mathbf{x} = \mathbf{A}^\top \mathbf{y}, \mathbf{y} \geq \mathbf{0}\}$ is closed, convex, and $\mathbf{c} \notin S$. By Theorem 51 $\exists \mathbf{p} \in \mathbb{R}^n, \alpha \in \mathbb{R} : \mathbf{p}^\top \mathbf{c} > \alpha$ and $\forall \mathbf{x} \in S : \mathbf{p}^\top \mathbf{x} \leq \alpha$. We know that $\mathbf{0} \in S$ and then $\alpha \geq 0$. As $\alpha \geq \mathbf{p}^\top \mathbf{x} = \mathbf{p}^\top \mathbf{A}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{A}\mathbf{p}$. As $\mathbf{y} \geq \mathbf{0}$ then $\mathbf{A}\mathbf{p}$ must be $\leq \mathbf{0}$ (otherwise we obtain contradiction). So, we have found $\mathbf{p} \in \mathbb{R}^n$ such that $\mathbf{c}^\top \mathbf{p} > \alpha$ and $\mathbf{A}\mathbf{p} \leq \mathbf{0}$. So $\mathbf{p} \in S_1$, and hence, $S_1 \neq \emptyset$. □

Exercise 58.

Set $m = 2$ and $n = 3$. Choose suitable values for $\mathbf{A}$ and $\mathbf{c}$. Denote columns of $\mathbf{A}$ as vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$. Draw two figures. The first one, describing set S_1 in $\mathbb{R}^2$ and the second describing set S_2 indirectly, drawing $\{\mathbf{A}\mathbf{y} \mid \mathbf{y} \geq \mathbf{0}\}$ in $\mathbb{R}^2$.

Farkas' theorem also remains valid for redefined S_1, S_2:

Corollary 59 (Gordon's Theorem)**.**

Let $\mathbf{A}$ be an $m \times n$ matrix. Then either $S_1 = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} < \mathbf{0}\}$ or $S_2 = \{\mathbf{y} \mid \mathbf{A}^\top \mathbf{y} = \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \mathbf{y} \neq \mathbf{0}\}$ is nonempty.

Proof: S_1 can be written as $S_1 = \{\mathbf{x} \in \mathbb{R}^n, s \in \mathbb{R} \mid \mathbf{A}\mathbf{x} + \mathbf{1}s \leq \mathbf{0}\}$ where $\mathbf{1}$ is a vector of m ones. We may rewrite it in the form suitable for S_1 in Theorem 57. We get $\{\mathbf{x}, s \mid (\mathbf{A}, \mathbf{1})(\mathbf{x}^\top, s)^\top \leq \mathbf{0}\}$ and for $\mathbf{e}_{n+1} = (0, \ldots, 0, 1)^\top$ we have $\mathbf{e}_{n+1}^\top (\mathbf{x}^\top, s)^\top > 0$. Then S_2 is specified by $(\mathbf{A}, \mathbf{1})^\top \mathbf{y} = \mathbf{e}_{n+1}$ and $\mathbf{y} \geq \mathbf{0}$. That is $\mathbf{A}^\top \mathbf{y} = \mathbf{0}$, $\mathbf{1}^\top \mathbf{y} = 1$ and $\mathbf{y} \geq \mathbf{0}$. This is equivalent to S_2 description. □

Corollary 60 (Alternatives to Gordon's Theorem)**.**

Let $\mathbf{A}$ be an $m \times n$ matrix ($\mathbf{A} \in \mathbb{R}^{m \times n}$), $\mathbf{B}$ be an $l \times n$ matrix ($\mathbf{B} \in \mathbb{R}^{l \times n}$), and $\mathbf{c}$ be an n vector ($\mathbf{c} \in \mathbb{R}^n$). Then:

1. Either $S_1 = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0}, \mathbf{x} \geq \mathbf{0}, \mathbf{c}^\top \mathbf{x} > 0\}$ or $S_2 = \{\mathbf{y} \mid \mathbf{A}^\top \mathbf{y} \geq \mathbf{c}, \mathbf{y} \geq \mathbf{0}\}$ is nonempty.
2. Either $S_1 = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0}, \mathbf{B}\mathbf{x} = \mathbf{0}, \mathbf{c}^\top \mathbf{x} > 0\}$ or $S_2 = \{(\mathbf{y}^\top, \mathbf{z}^\top) \mid \mathbf{A}^\top \mathbf{y} + \mathbf{B}^\top \mathbf{z} = \mathbf{c}, \mathbf{y} \geq \mathbf{0}\}$ is nonempty.

Proof: To prove 1.: Introduce slack variables to get equalities for S_2 (replace $\mathbf{A}^\top$ by $(\mathbf{A}^\top - \mathbf{I})$). To prove 2.: Define $\mathbf{z} = \mathbf{z}_1 - \mathbf{z}_2$ where $\mathbf{z}_1 \geq \mathbf{0}$, $\mathbf{z}_2 \geq \mathbf{0}$ in S_2 and replace $\mathbf{A}^\top$ by $(\mathbf{A}^\top, \mathbf{B}^\top, -\mathbf{B}^\top)$. □

1.3.4 Supporting hyperplanes

Definition 61 (Supporting hyperplane)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, $\bar{\mathbf{x}} \in \partial S$. $H = \{\mathbf{x} \mid \mathbf{p}^\top (\mathbf{x} - \bar{\mathbf{x}}) = 0\}$ is called a supporting hyperplane of S at $\bar{\mathbf{x}} \Leftrightarrow S \subset H^+ \vee S \subset H^-$. (Without reference to $\bar{\mathbf{x}}$, we may add $\operatorname{cl} S \cap H \neq \emptyset$.) If $S \not\subset H$ then H is a proper supporting hyperplane of S at $\bar{\mathbf{x}}$.

It is equivalent to $\mathbf{p}^\top \bar{\mathbf{x}} = \inf\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S\}$ or $\mathbf{p}^\top \bar{\mathbf{x}} = \sup\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S\}$.

Theorem 62 (Existence f supporting hyperplane)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, S is convex, $\bar{\mathbf{x}} \in \partial S \Rightarrow \exists H$ supporting S at $\bar{\mathbf{x}}$ $(\exists \mathbf{p} \neq \mathbf{0} : \forall \mathbf{x} \in \operatorname{cl} S : \mathbf{p}^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0)$

Proof: $\bar{\mathbf{x}} \in \partial S \Rightarrow \exists \{\mathbf{y}_k\}$ a sequence such that $\mathbf{y}_k \in \operatorname{cl} S$, $\mathbf{y}_k \longrightarrow \bar{\mathbf{x}}$. $\forall \mathbf{y}_k \ \exists \mathbf{p}_k$, $\|\mathbf{p}_k\| = 1$ such that $\forall \mathbf{x} \in \operatorname{cl} S : \mathbf{p}_k^\top \mathbf{y_k} > \mathbf{p}_k^\top \mathbf{x}$. $\{\mathbf{p}_k\}$ is bounded, so exists $\mathcal{K} \subset \mathbb{N}$ such that $\{p_k\}_{k \in \mathcal{K}}$ has a limit $\mathbf{p}$ and $\|\mathbf{p}\| = 1$. As $\forall k \in \mathcal{K} : \mathbf{p}_k^\top \mathbf{y}_k > \mathbf{p}_k^\top \mathbf{x}$ then $\mathbf{p}_k^\top (\mathbf{x} - \mathbf{y}_k) < 0$. For $k \longrightarrow \infty$ we get $\mathbf{p}^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0$. □

Corollary 63.

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, and:

1. S is convex, $\bar{\mathbf{x}} \notin \operatorname{int} S \Rightarrow \exists \mathbf{p} \in \mathbb{R}^n$, $\mathbf{p} \neq \emptyset : \forall \mathbf{x} \in \operatorname{cl} S : \mathbf{p}^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0$.
2. $\mathbf{y} \notin \operatorname{int} \operatorname{conv} S \Rightarrow \exists H$ separating S and $\mathbf{y}$.
3. $\bar{\mathbf{x}} \in \partial S \cup \partial \operatorname{conv} S \Rightarrow \exists H$ that supports S at $\bar{\mathbf{x}}$.

Exercise 64.

Illustrate Corollary 63 by drawing figures in $\mathbb{R}^2$.

Theorem 65 (Separation of 2 convex sets)**.**

Let $S_1, S_2 \subset \mathbb{R}^n$, $S_1 \neq \emptyset$, $S_1 \neq \emptyset$ be convex sets, $S_1 \cap S_2 = \emptyset \Rightarrow \exists H$ hyperplane separating S_1 and S_2, i.e. $\exists \mathbf{p} \neq \mathbf{0} : \inf\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S_1\} \geq \sup\{\mathbf{p}^\top \mathbf{x} \mid \mathbf{x} \in S_2\}$.

Exercise 66.

Draw an explanatory figure for Theorem 65. Compare with visualization of nonlinear programs.

Proof: $S := S_1 \ominus S_2 = \{\mathbf{x}_1 - \mathbf{x}_2 \mid \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$. As S_1 and S_2 are convex $\Rightarrow S$ convex. Because $S_1 \cap S_2 = \emptyset \Rightarrow \mathbf{0} \notin S$. By Corollary 63.1 $\exists \mathbf{p} \neq \mathbf{0}$, $\forall \mathbf{x} \in S : \mathbf{p}^\top \mathbf{x} \geq 0$, hence $\forall \mathbf{x}_1 \in S_1$, $\forall \mathbf{x}_2 \in S_2 : \mathbf{p}^\top \mathbf{x}_1 \geq \mathbf{p}^\top \mathbf{x}_2$. □

Remark 67.

Gordan's theorem may be derived either from Farkas' theorem or from separating theorems.

Corollary 68.

Let $S_1, S_2 \subset \mathbb{R}^n$, $S_1 \neq \emptyset$, $S_1 \neq \emptyset$, and

1. S_1, S_2 be convex, $\operatorname{int} S_2 \neq \emptyset$, $S_1 \cap \operatorname{int} S_2 = \emptyset \Rightarrow \exists H$ hyperplane separating S_1 and S_2.
2. $\operatorname{int} \operatorname{conv} S_i \neq \emptyset$, $i = 1, 2$ and $\operatorname{int} \operatorname{conv} S_1 \cap \operatorname{int} \operatorname{conv} S_2 = \emptyset \Rightarrow \exists H$ separating S_1 and S_2.

Exercise 69.

Draw illustrating figures for Corollary. Why assumption of 2. are important? Hint: Thing about two straight lines and their intersection!

Proof: To prove 1. replace S_2 with $\operatorname{int} S_2$ in Theorem 65. Similarly prove 2. □

Theorem 70 (Strong separation)**.**

Let S_1, S_2 be closed convex and S_1 is bounded. If $S_1 \cap S_2 = \emptyset \Rightarrow \exists H$ hyperplane strongly separating S_1 and S_2.

Proof: Main idea: Put $S = S_1 \ominus S_2$, show that S is closed, take $\mathbf{0} \notin S$ and strongly separate $\mathbf{0}$ and S. □

Corollary 71.

Let $S_1, S_2 \subset \mathbb{R}^n$, $S_1 \neq \emptyset$, $S_1 \neq \emptyset$, S_1 bounded, $\operatorname{cl} \operatorname{conv} S_1 \cap \operatorname{cl} \operatorname{conv} S_2 = \emptyset \Rightarrow \exists H$ hyperplane strongly separating S_1 and S_2.

Exercise 72.

Why S_1 is assumed to be bounded in Theorem 70? Remove this assumption and create a figure in $\mathbb{R}^2$ showing why the conclusion of theorem does not remain valid.

1.3.5 Convex cones and polarity

Definition 73 (Cone)**.**

$C \subset \mathbb{R}^n$ is called a cone with vertex $\mathbf{0}$ if $\mathbf{x} \in C \Rightarrow \forall \lambda \geq 0 : \lambda \mathbf{x} \in C$.

Definition 74 (Polar cone)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ then the polar cone of S ($\operatorname{pol} S$ or S^*) is given by $\operatorname{pol} S = \{\mathbf{p} \mid \forall \mathbf{x} \in S : \mathbf{p}^\top \mathbf{x} \leq 0\}$.

If $S = \emptyset$ then S^* will be interpreted as $\mathbb{R}^n$.

Exercise 75.

Draw figures in $\mathbb{R}^2$ with convex, nonconvex, and polar cones.

Lemma 76.

Let $S_1, S_2, S_3 \subset \mathbb{R}^n$ be nonempty sets:

1. S^* is a closed convex cone.
2. $S \subset (S^*)^*$ (we also write $(S^*)^* = S^{**} = \text{polpol}S$).
3. $S_1 \subset S_2 \Rightarrow S_2^* \subset S_1^*$.

Theorem 77 (Polar cone property).

Let $C \neq \emptyset$ be a closed convex cone $\Rightarrow C = C^{**}$.

Proof: From Lemma 76: $C \subset C^{**}$. Let $\mathbf{x} \in C^{**}$ and suppose by contradiction that $\mathbf{x} \notin C$. By previous Theorem 51 (separation of set and point) $\exists \mathbf{p} \neq \mathbf{0}$, $\exists \alpha : \forall \mathbf{y} \in C$ (also $\mathbf{0} \in C$, $0 \leq \alpha$): $\mathbf{p}^\top \mathbf{y} \leq \alpha$ and $\mathbf{p}^\top \mathbf{x} > \alpha \geq 0$. Assume $\mathbf{p} \notin C^* : \exists \bar{\mathbf{y}} \in C : \mathbf{p}^\top \bar{\mathbf{y}} > 0$ and also $\forall \lambda \geq 0 : \mathbf{p}^\top (\lambda \bar{\mathbf{y}}) \geq 0$ and $\mathbf{p} \in C^*$. As $\lambda \longrightarrow \infty \Rightarrow \mathbf{p}^\top (\lambda \bar{\mathbf{y}}) \geq 0 \longrightarrow \infty$. Because we have shown $\forall \mathbf{y} : \mathbf{p}^\top \mathbf{y} \leq \alpha$, we see that also $\forall \lambda \geq 0 : \mathbf{p}^\top (\lambda \bar{\mathbf{y}}) \leq \alpha$. It gives the contradiction. Since $\mathbf{x} \in C^{**} := \{\mathbf{u} \mid \forall \mathbf{v} \in C^* : \mathbf{u}^\top \mathbf{v} \leq 0\}$ then $\mathbf{p}^\top \mathbf{x} \leq 0$ but it is a contradiction with $\mathbf{p}^\top \mathbf{x} \geq 0$ and so $\mathbf{x} \in C$. $\square$

Remark 78.

Farkas' theorem can be proven using polar cones. Define $C := \{\mathbf{A}^\top \mathbf{y} \mid \mathbf{y} \geq \mathbf{0}\}$, which is a closed convex cone (explain why) and see that $C^* = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$. Therefore $\mathbf{c} \in C \Leftrightarrow \mathbf{c} \in C^{**}$ (by theorem) where $(\mathbf{c} = \mathbf{A}\mathbf{y})$ and $\mathbf{x} \in C^* \Rightarrow \mathbf{c}^\top \mathbf{x} \leq 0$.

1.3.6 Polyhedral sets, extreme points, extreme directions

Definition 79 (Polyhedral set).

$S \subset \mathbb{R}^n$ is called a polyhedral set if it is the intersection of a finite number of closed halfspaces, i.e. $\exists k \in \mathbb{N}$, $\exists \alpha_i$, $\exists \mathbf{p}_i \in \mathbb{R}^n$, $\mathbf{p}_i \neq \mathbf{0}$, $i = 1, \ldots, k : S = \cap_{i=1}^k \{\mathbf{x} \mid \mathbf{p}_i^\top \mathbf{x} \leq \alpha_i\}$

Remark 80.

(1) Every polytope is a polyhedral set. (2) Bounded polyhedral set is a polytope. (3) The feasible set of LP is a polyhedral set. (4) The convex feasible set of NLP may be contained in a polyhedral set.

Definition 81 (Extreme point)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be convex. $\mathbf{x} \in S$ is called an extreme point (EP) of $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S, \forall \lambda \in (0,1) : \mathbf{x} = \lambda \mathbf{x} - 1 + (1-\lambda)\mathbf{x}_2 \Rightarrow \mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$.

Exercise 82.

Draw several figures illustrating the concept of an extreme point (EP). Compare it with BFS – Basic Feasible Solution in LP.

Notice that for a compact convex set, any its point can be represented as a convex combination of its extreme points.

Theorem 83 (Characterization of EPs)**.**

Let $S = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ (LP in the standard form), $r(\mathbf{A}) = m$ (rank of $\mathbf{A}$, otherwise we may throw away any redundant rows). Then, $\mathbf{x}$ is an extreme point of S $\Leftrightarrow$ $\mathbf{A}$ can be decomposed by rearranging columns of $\mathbf{A}$ (and rows of $\mathbf{x}$ respectively) in such a way that $\mathbf{A} = (\mathbf{B}, \mathbf{N})$ ($\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}_B + \mathbf{N}\mathbf{x}_N = \mathbf{b}$) satisfying:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix}$$

and $\mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$.

Proof: We have $\mathbf{x}$ defined by theorem formula. $\Rightarrow$ We want to show $\mathbf{x} \in S$ and $\mathbf{x}$ is an EP. It means to show that $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. Let $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\lambda \in (0,1)$. So we have:

$$\mathbf{x}_1 = \begin{pmatrix} \mathbf{x}_{1B} \\ \mathbf{x}_{1N} \end{pmatrix} \mathbf{x}_2 = \begin{pmatrix} \mathbf{x}_{2B} \\ \mathbf{x}_{2N} \end{pmatrix} \quad \text{then} \quad \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{x}_{1B} \\ \mathbf{x}_{1N} \end{pmatrix} + (1-\lambda) \begin{pmatrix} \mathbf{x}_{2B} \\ \mathbf{x}_{2N} \end{pmatrix}.$$

So $\mathbf{B}^{-1}\mathbf{b} = \lambda \mathbf{x}_{1B} + (1-\lambda)\mathbf{x}_{2B}$ and $\mathbf{x}_{1N} = \mathbf{x}_{2N} = \mathbf{0}$. We have $\mathbf{x}_1 \in S \Rightarrow \mathbf{A}\mathbf{x}_1 = \mathbf{b} \Rightarrow \mathbf{B}\mathbf{x}_{1B} = \mathbf{b} \Rightarrow \mathbf{x}_{1B} = \mathbf{B}^{-1}\mathbf{b}$ and $\mathbf{x}_2 \in S \Rightarrow \mathbf{A}\mathbf{x}_2 = \mathbf{b} \Rightarrow \mathbf{B}\mathbf{x}_{2B} = \mathbf{b} \Rightarrow \mathbf{x}_{2B} = \mathbf{B}^{-1}\mathbf{b}$. So, $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$, and hence, $\mathbf{x}$ is an EP.

Conversely, $\mathbf{x}$ is an EP. We suppose (without losing generality) $\mathbf{x} = (x_1, \ldots, x_k, 0, \ldots, 0)^\top$ where $x_1 > 0, \ldots, x_k > 0$. We discuss whether $\mathbf{a}_1, \ldots, \mathbf{a}_k$ columns of $\mathbf{A}$ are linearly independent. If they are dependent then $\exists \lambda_1, \ldots, \lambda_k$ ($\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)^\top \neq \mathbf{0}$) : $\sum_{j=1}^k \lambda_j \mathbf{a}_j = \mathbf{0}$. We construct $\mathbf{x}_1 = \mathbf{x} + \alpha\boldsymbol{\lambda}$ and $\mathbf{x}_2 = \mathbf{x} - \alpha\boldsymbol{\lambda}$ such that $\mathbf{x}_1, \mathbf{x}_2 \geq \mathbf{0}$ (choosing $\alpha > 0$ without loss of generality). So:

$$\begin{aligned} \mathbf{A}\mathbf{x}_1 &= \mathbf{A}\mathbf{x} + \alpha\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}\mathbf{x} + \textstyle\sum_{j=1}^k \lambda_j \mathbf{a}_j = \mathbf{b} \quad \Rightarrow \mathbf{x}_1 \in S \\ \mathbf{A}\mathbf{x}_2 &= \mathbf{A}\mathbf{x} - \alpha\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}\mathbf{x} - \textstyle\sum_{j=1}^k \lambda_j \mathbf{a}_j = \mathbf{b} \quad \Rightarrow \mathbf{x}_2 \in S. \end{aligned}$$

Because $\alpha > 0, \boldsymbol{\lambda} \neq \mathbf{0} : \mathbf{x}_1 \neq \mathbf{x}_2$ and $\mathbf{x} = \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2) \Rightarrow$ then $\mathbf{x}$ is not an EP $\Rightarrow$ contradiction.
So, $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly independent. Assume that $r(\mathbf{A}) = m$. Then $k \leq m$. If $k = m$ then done, otherwise ($k < m$) $\Rightarrow \exists m - k$ columns of $\mathbf{A}$ (of last $n - k$ columns) that with k columns form a linearly independent set of m vectors. After reordering of columns, we get $\mathbf{A} = (\mathbf{B}, \mathbf{N})$ and $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. □

Exercise 84.

For own LP with 4 variables and 2 equality constraints, compute several extreme points. Compare with the graphical solution (for the original 2 variables and 2 inequalities).

Exercise 85.

Discuss types of nonlinear programs whose feasible sets have extreme points. Hint: Draw figures.

Corollary 86 (Finite number of EPs)**.**

The number of extreme points of $S = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is finite and less or equal to $\begin{pmatrix} n \\ m \end{pmatrix}$ (remember that $\mathbf{A} \in \mathbb{R}^{m \times n}$).

Exercise 87.

Prove Corollary, giving the upper bound of number of extreme points.

Theorem 88 (Existence of EP)**.**

Let $S = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \neq \emptyset$, $r(\mathbf{A}) = m \Rightarrow S$ has at least one EP.

Proof: Let $\mathbf{x} \in S$, $\mathbf{x} = (x_1, \ldots, x_k, 0, \ldots, 0)^\top$. If $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly independent then $k \leq m$ and $\mathbf{x}$ is an extreme point. Otherwise $\exists \lambda_1, \ldots, \lambda_k$ $(\boldsymbol{\lambda} = (\lambda 1, \ldots, \lambda_k, 0, \ldots, 0)^\top \neq \mathbf{0} \Rightarrow \exists \lambda_i > 0)$ $\sum_{j=1}^{k} \lambda_j \mathbf{a}_j = \mathbf{0}$. We define: $\alpha = \min_{1 \leq j \leq k} \{\frac{x_j}{\lambda_j} : \lambda_j > 0\} = \frac{x_i}{\lambda_i} > 0$. Consider$\mathbf{x}'$: $x'_j = x_j - \alpha \lambda_j$ for $j = 1, \ldots, k$ and $x'_j = 0$ for $j = k+1, \ldots, n$. Then $x'_j \geq 0$, $j \in \{1, \ldots, k\} \setminus \{i\}$ and $x'_j = 0$ elsewhere. Also: $\sum_{j=1}^{n} \mathbf{a}_j x'_j = \sum_{j=1}^{k} (x_j - \alpha \lambda_j) = \sum_{j=1}^{k} \mathbf{a}_j x_j - \alpha \sum_{j=1}^{k} \mathbf{a}_j \lambda_j = \mathbf{b}$ and $\mathbf{x}'$ has a reduced number of positive components (k-1). Repeat procedure till reaching the linear independence. Then EP will be found. □

Exercise 89.

Build an example showing that for $S = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}\}$ (nonnegativity constraint $\mathbf{x} \geq \mathbf{0}$ omitted) EP does not necessarily exists.

Definition 90 (Direction)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, closed, convex. Then $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{d} \neq \mathbf{0}$ is called a (recession) direction of $S \Leftrightarrow \forall \mathbf{x} \in S, \forall \lambda \geq 0 : \mathbf{x} + \lambda \mathbf{d} \in S$.

Definition 91 (Distinct directions)**.**

Let $\mathbf{d}_1$ and $\mathbf{d}_2$ be two directions of S. They are called distinct directions of $S \Leftrightarrow \forall \alpha > 0 : \mathbf{d}_1 \neq \alpha \mathbf{d}_2$

Definition 92 (Extreme direction)**.**

A direction $\mathbf{d}$ of S is called an extreme direction (ED) of $S \Leftrightarrow (\forall \mathbf{d}_1, \mathbf{d}_2$ directions of S, $\forall \lambda_1, \lambda_2 > 0$: $\mathbf{d} = \lambda_1 \mathbf{d}_1 + \lambda_2 \mathbf{d}_2 \Rightarrow \exists \alpha > 0 : \mathbf{d}_1 = \alpha \mathbf{d}_2)$.

Exercise 93.

Illustrate directions (of S, distinct, and extreme) by figures in $\mathbb{R}^2$.

Theorem 94 (Characterization of EDs)**.**

Let $S = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \neq \emptyset$, $r(\mathbf{A}) = m$. Then, $\bar{\mathbf{d}}$ is an extreme direction of $S \Leftrightarrow \mathbf{A}$ can be decomposed by rearranging columns of $\mathbf{A}$ (and rows of $\mathbf{x}$ respectively) in such a way that $\mathbf{A} = (\mathbf{B}, \mathbf{N})$ $(\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}_B + \mathbf{N}\mathbf{x}_N = \mathbf{b})$ and $\exists \mathbf{a}_j$ column of $\mathbf{N}$, $\exists \alpha > 0$ satisfying:

$$\bar{\mathbf{d}} = \alpha \begin{pmatrix} \mathbf{d}_B \\ \mathbf{d}_N \end{pmatrix} = \alpha \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix}$$

and $\mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0}$ where $\mathbf{e}_j = (e_{ij})_{i=1,\ldots,n-m}$ such that $e_{ij} = 1$ for $i = j$ and $e_{ij} = 0$ for $i = 1, \ldots, n-m$, $i \neq j$.

Proof: We derive it for $\mathbf{d}$ instead fo $\bar{\mathbf{d}}$. With $\mathbf{x} \in S$, $\mathbf{d}$ given in theorem, $\lambda \geq 0$:

$$\mathbf{A}(\mathbf{x} + \lambda \mathbf{d}) = \mathbf{A}\mathbf{x} + \lambda \mathbf{A}\mathbf{d} = \mathbf{b} + \lambda (\mathbf{B}, \mathbf{N}) \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix} = \mathbf{b} + \lambda(-\mathbf{a}_j + \mathbf{a}_j) = \mathbf{b}.$$

Notice that $\mathbf{A}\mathbf{d} = \mathbf{0}$ (so $\mathbf{A}(\mathbf{x} + \lambda \mathbf{d}) = \mathbf{A}\mathbf{x} = \mathbf{b}$) and $\mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0} \Rightarrow \mathbf{d} \geq \mathbf{0} \Rightarrow \mathbf{x} + \lambda \mathbf{d} \geq \mathbf{0}$. So $\mathbf{d}$ is a direction of S.
Now, we show that $\mathbf{d}$ is an extreme direction of S. Let $\mathbf{d}_1$ and $\mathbf{d}_2$ be directions and assume $\mathbf{d} = \lambda_1 \mathbf{d}_1 + = \lambda_2 \mathbf{d}_2$ for $\lambda_1 > 0$ and $\lambda_2 > 0$. We know that at least $n - (m+1)$ components of $\mathbf{d} \geq \mathbf{0}$ are necessarily zeroes. Therefore, the same is valid for $\mathbf{d}_1$ and $\mathbf{d}_2$ corresponding components. Then:

$$\mathbf{d}_i = \alpha_i \begin{pmatrix} \mathbf{d}_{iB} \\ \mathbf{d}_{iN} \end{pmatrix} \; i = 1, 2, \quad \mathbf{A}\mathbf{d}_1 = \mathbf{A}\mathbf{d}_2 = \mathbf{0} \Rightarrow \mathbf{B}\mathbf{d}_{iB} + \mathbf{N}\mathbf{d}_{iN} = \mathbf{0} \Rightarrow$$

Hence, $\mathbf{B}\mathbf{d}_{iB} = -\mathbf{N}\mathbf{e}_j = -\mathbf{a}_j$ $\mathbf{d}_{iB} = \mathbf{B}^{-1}\mathbf{a}_j \Rightarrow \mathbf{d} = \mathbf{d}_1 = \mathbf{d}_2$, so $\mathbf{d}$ is an ED, and therefore, $\bar{\mathbf{d}}$ is an ED. Conversely: Suppose $\bar{\mathbf{d}}$ is an ED and derive a formula above. We may assume

$$\bar{\mathbf{d}} = (\bar{d}_1, \ldots, \bar{d}_k, 0, \ldots, 0, \bar{d}_j, 0, \ldots, 0)^\top$$

. We claim that $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly independent. We show it by a contradiction: $\Rightarrow \exists \lambda_1, \ldots, \lambda_k$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^\top \neq \mathbf{0} : \sum_{i=1}^k \lambda_i \mathbf{a}_i = \mathbf{0}$. We choose $\alpha > 0$ small that $\mathbf{d}_1 = \bar{\mathbf{d}} + \alpha \boldsymbol{\lambda}$, $\mathbf{d}_2 = \bar{\mathbf{d}} - \alpha \boldsymbol{\lambda}$

and $\mathbf{d}_1, \mathbf{d}_2 \geq \mathbf{0}$. Note that $\mathbf{A}\mathbf{d}_1 = \mathbf{A}\bar{\mathbf{d}} + \alpha \sum_i \lambda_i \mathbf{a}_i = \mathbf{0}$ and $\mathbf{A}\mathbf{d}_2 = \mathbf{0}$. Together they are distinct directions and $\bar{\mathbf{d}} = (\mathbf{d}_1 + \mathbf{d}_2)/2$, so it is a contradiction with $\bar{\mathbf{d}}$ is an ED. Therefore, $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly independent. As $r(\mathbf{A}) = m$, we know $k \leq m$. If $k < m$ then we may add $m - k$ vectors $\mathbf{a}_i$ such that $i \in \{k+1, \ldots, n\} \setminus \{j\}$ to form $\mathbf{B}$ matrix (with $r(\mathbf{B}) = m$ by a suitable selection of columns). So, $\mathbf{A}\bar{\mathbf{d}} = \mathbf{0} = \mathbf{B}\bar{\mathbf{d}}_B + \mathbf{a}_j \bar{d}_j$ and $\bar{\mathbf{d}}_B = -\bar{d}_j \mathbf{B}^{-1}\mathbf{a}_j$ and $\bar{\mathbf{d}} = \bar{d}_j \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix}$ and also with $\bar{\mathbf{d}} \geq \mathbf{0}$, $\bar{d}_j > 0 \Rightarrow -\mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0}$. □

Exercise 95.

For own LP with 4 variables and 2 equality constraints, compute one extreme direction. Hint: Draw a figure first!

Corollary 96 (Finite number of EDs)**.**

The number of extreme directions of S is finite and less or equal to $(n-m)\begin{pmatrix} n \\ m \end{pmatrix}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$.

Proof: For each of $\begin{pmatrix} n \\ m \end{pmatrix}$ choices of $\mathbf{B}$ from $\mathbf{A}$, there are $(n-m)$ ways to extract $\mathbf{a}_j$ from matrix $\mathbf{N}$. □

Theorem 97 (Existence of EDs)**.**

Let $S = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \neq \emptyset$, $r(\mathbf{A}) = m$. S has at least one extreme direction $\Leftrightarrow$ S is unbounded.

Proof: $\Rightarrow$: It is valid from the definition of ED.
$\Leftarrow$ Contradiction S is unbounded and there is no extreme direction. From the following theorem: $\forall \mathbf{x} \in S$: $\mathbf{x} = \sum_j \lambda_j \mathbf{x}_j + \sum_j \mu_j \mathbf{d}_j = \sum_j \lambda_j \mathbf{x}_j$ (because of no extreme direction). Then $\|\mathbf{x}\| = \|\sum_j \lambda_j \mathbf{x}_j\| \leq$ (Schwarz inequality) $\sum_j \|\lambda_j \mathbf{x}_j\| =$ (because $\lambda_j \geq 0$ and $\lambda_j \leq 1$) $\sum_{i=1}^{j} \lambda_j \|\mathbf{x}_j\| \leq \sum_{i=1}^{j} \|\mathbf{x}_j\| = K \in \mathbb{R}$. Therefore, $\exists K \in \mathbb{R} : \forall \mathbf{x} \in S : \|\mathbf{x}\| < K$. It leads to the contradiction with unboundedness of S assumption. Hence, S has at least one extreme direction. □
The following representation theorem says: Any point of LP standard form feasible set can be represented as a sum of a convex combination of EPs of S and nonnegative linear combination of EDs of S.

Theorem 98 (Representation theorem)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, $r(\mathbf{A}) = m$. There are $\mathbf{x}_1, \ldots, \mathbf{x}_k$ all EPs (extreme points) of S and $\mathbf{d}_1, \ldots, \mathbf{d}_l$ are all EDs (extreme directions) of S. Then:

$$\mathbf{x} \in S \Leftrightarrow \exists \mu_j \geq 0, j = 1, \ldots, l, \ \exists \lambda_j \geq 0, j = 1, \ldots, k :$$
$$\textstyle\sum_{j=1}^{k} \lambda_j = 1 \ \wedge \mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{l} \mu_j \mathbf{d}_j.$$

Exercise 99.

Draw a figure in $\mathbb{R}^2$ for unbounded polyhedral S to illustrate the theorem.

Proof of Theorem 98: Let $M := \{\sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{l} \mu_j \mathbf{d}_j \mid \mu_j \geq 0, j = 1, \ldots, l,\ \lambda_j \geq 0, j = 1, \ldots, k\}$. M is a closed (because of definition), convex (prove directly), and nonempty ($S \neq \emptyset \Rightarrow \exists$ EP by Theorem 88), and $M \subset S$ (S is polyhedral $\Rightarrow \cap$ halfspaces $\Rightarrow$ convex, so it contains convex combinations with nonnegative linear combinations of extreme directions from M by definition). We want to show $S \subset M$. By contradiction: $\mathbf{z} \in S \wedge \mathbf{z} \notin M$. Then it is possible to separate $\mathbf{z}$ from M (by Theorem 51) by a hyperplane: $\exists \alpha \in \mathbb{R}$, $\exists \mathbf{p} \neq \mathbf{0} : \mathbf{p}^\top \mathbf{z} > \alpha$ and $\mathbf{p}^\top (sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{l} \mu_j \mathbf{d}_j)$ for each element of M. Then, $\mathbf{d}_j$ may arbitrarily be enlarged, so with $\mu_j \geq 0$ and $\mathbf{p}^\top (sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{l} \mu_j \mathbf{d}_j) \leq \alpha$ requirement, we get: $\forall j : \mathbf{p}^\top \mathbf{d}_j \leq 0$. When λ_j, μ_j are set to zeroes and only one $\lambda_i = 1$ then we obtain $\forall j : \mathbf{p}^\top \mathbf{x}_j \leq \alpha < \mathbf{p}^\top \mathbf{z}$. We choose the "tightest" $\mathbf{x}_j$ by $\mathbf{p}^\top \bar{\mathbf{x}} = \max_{1 \leq j \leq k} \mathbf{p}^\top \mathbf{x}_j$, also $\mathbf{p}^\top \bar{\mathbf{x}} \leq \alpha < \mathbf{p}^\top \mathbf{z}$. $\bar{\mathbf{x}}$ is an EP: $\bar{\mathbf{x}} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} \geq \mathbf{0}$. Further, we assume $\mathbf{B}^{-1}\mathbf{b} > \mathbf{0}$ (nondegenerate case) without loss of generality. $\mathbf{z} \in S \Rightarrow \mathbf{Az} = \mathbf{b}$, $\mathbf{z} \geq \mathbf{0}$, $\mathbf{Bz}_B + \mathbf{Nz}_N = \mathbf{b}$ and so: $\mathbf{z}_B = \mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{Nz}_N$. We know $\mathbf{p}^\top \bar{\mathbf{x}} < \mathbf{p}^\top \mathbf{z}$, so $0 < \mathbf{p}^\top \mathbf{z} - \mathbf{p}^\top \bar{\mathbf{x}} = (\mathbf{p}_B^\top, \mathbf{p}_N^\top) \begin{pmatrix} \mathbf{z}_B \\ \mathbf{z}_N \end{pmatrix} - (\mathbf{p}_B^\top, \mathbf{p}_N^\top) \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} = \mathbf{p}_B^\top (\mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{Nz}_N) + \mathbf{p}_N^\top \mathbf{z}_N - \mathbf{p}_B^\top \mathbf{B}^{-1}\mathbf{b} = (\mathbf{p}_N^\top - \mathbf{p}_B^\top \mathbf{B}^{-1}\mathbf{N})\mathbf{z}_N \Rightarrow \exists j : z_j > 0$ and $p_j - \mathbf{p}_B^\top \mathbf{B}^{-1}\mathbf{a}_j > 0$. We check the sign of $\mathbf{B}^{-1}\mathbf{a}_j$ denoting $\mathbf{y}_j = \mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0}$. We assume $\mathbf{y}_j \leq \mathbf{0}$ then $\begin{pmatrix} -\mathbf{y}_j \\ \mathbf{e}_j \end{pmatrix} \geq \mathbf{0}$ is an ED of S by Theorem 94. Earlier, we noticed that $\mathbf{p}^\top \mathbf{d}_j \leq 0$, for EDs. So, $(\mathbf{p}_B^\top, \mathbf{p}_N^\top) \begin{pmatrix} -\mathbf{y}_j \\ \mathbf{e}_j \end{pmatrix} \leq \mathbf{0}$ and we get contradiction (previously $p_j - \mathbf{p}_B^\top \mathbf{B}^{-1}\mathbf{a}_j > 0$). So $\mathbf{y}_j \not\leq \mathbf{0}$ and we can construct:

$$\mathbf{x} = \begin{pmatrix} \bar{\mathbf{b}} \\ \mathbf{0} \end{pmatrix} + \lambda \begin{pmatrix} -\mathbf{y}_j \\ \mathbf{e}_j \end{pmatrix}, \bar{\mathbf{b}} = \mathbf{B}^{-1}\mathbf{b}$$

and $\lambda = \min_{1 \leq i \leq m} \{\frac{\bar{b}_i}{y_{ij}} \mid y_{ij} > 0\}$. So, $\lambda > 0$ (because of nondegeneracy $\bar{b}_i > 0$), $\lambda = \frac{\bar{b}_r}{y_{rj}}$ (rth component), $\mathbf{x} \geq \mathbf{0}$ by definition (has at most m positive components and rth drops to zero and jth is λ). Then $\mathbf{Ax} = \mathbf{B}(\mathbf{B}^{-1}\mathbf{b} - \lambda \mathbf{B}^{-1}\mathbf{a}_j) + \mathbf{N}\lambda \mathbf{e}_j = \mathbf{Ax} = \mathbf{B}(\mathbf{B}^{-1}\mathbf{b} - \lambda \mathbf{B}^{-1}\mathbf{a}_j) + \lambda \mathbf{a}_j = \mathbf{b} \Rightarrow \mathbf{x} \in S$. Then $\mathbf{a}_1, \ldots, \mathbf{a}_{r-1}, \mathbf{a}_{r+1}, \ldots, \mathbf{a}_m, \mathbf{a}_j$ are linearly independent. (As $\mathbf{By}_j = \mathbf{a}_j$ and we take out $y_{rj} \neq 0$, if $y_{rj} = 0$ then linear dependence occur), hence $\mathbf{x}$ is an EP. $\mathbf{p}^\top \mathbf{x} = (\mathbf{p}_B^\top, \mathbf{p}_N^\top) \begin{pmatrix} \bar{\mathbf{b}} - \lambda \mathbf{y}_j \\ \lambda \mathbf{e}_j \end{pmatrix} = \mathbf{p}_B^\top \bar{\mathbf{b}} - \lambda \mathbf{p}_B^\top \mathbf{y}_j + \lambda p_j = \mathbf{p}^\top \bar{\mathbf{x}} + \lambda(p_j - \mathbf{p}_B^\top \mathbf{B}^{-1}\mathbf{a}_j) > \mathbf{p}^\top \bar{\mathbf{x}}$ So, this is a contradiction with the assumption that we took the nearest EP, and hence, $z \in S$. □

Exercise 100.

For the given simple S and $\mathbf{x} \in S$ express $\mathbf{x}$ by Representation Theorem.

1.4 Linear programming — revisited

1.4.1 Assumptions, standard form, and full enumeration

Remark 101.

At the beginning, review important ideas related to linear programs:

- Modelling requirements: linearity (cf. NLP), certainty (cf. stochastic programming), and divisibility (cf. integer programming).
- Simple examples ($n = 2$) are solvable graphically. It gives ideas to the general case.
- Software is widely available: "hidden solvers" (e.g., MS Excel).
- There are successful applications of large-scale problems ($10^5 - -10^8$ variables).
- Mathematical formulations may differ but theory and algorithm is developed for LP in the standard form:

$$? \in \operatorname{argmin}\{\mathbf{c}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}.$$

  Additional assumptions further used are $r(\mathbf{A}) = m$ (nondegeneracy) and $S \neq \emptyset$.

Remark 102 (Standard form of LP)**.**

Any LP can be put in the standard form:

- Minimization:

$$\begin{aligned} \operatorname{argmin}\{-f(\mathbf{x}) \mid \mathbf{x} \in C\} &= \operatorname{argmax}\{f(\mathbf{x}) \mid \mathbf{x} \in C\} \quad \text{and} \\ -\min\{-f(\mathbf{x}) \mid \mathbf{x} \in C\} &= \max\{f(\mathbf{x}) \mid \mathbf{x} \in C\} \end{aligned}$$

- Slacks: $\bar{\mathbf{x}} \in \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \leq b\} \Leftrightarrow (\bar{\mathbf{x}}^\top, s)^\top \in \{(\mathbf{x}^\top, s)^\top \mid \mathbf{a}^\top \mathbf{x} + s = b, s \geq 0\}$ and $\bar{\mathbf{x}} \in \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \geq b\} \Leftrightarrow (\bar{\mathbf{x}}^\top, s)^\top \in \{(\mathbf{x}^\top, s)^\top \mid \mathbf{a}^\top \mathbf{x} - s = b, s \geq 0\}$
- $\bar{\mathbf{x}} \in C \Leftrightarrow \exists \bar{\mathbf{x}}^+, \bar{\mathbf{x}}^- \geq \mathbf{0} : \bar{\mathbf{x}} = \bar{\mathbf{x}}^+ - \bar{\mathbf{x}}^- \in C$

Algorithm 103 (Full enumeration)**.**

1. Check whether there is an ED $\mathbf{d}_j$ (Notice that it needs to check all EDs!) decreasing the value of the objective, i.e. $\exists j = 1, \ldots, l : \mathbf{c}^\top \mathbf{d}_j < 0$.
2. If "yes" then there is no optimal solution because of unboundedness of the objective function on the feasible region. Otherwise, compute and compare values of the objective for all EPs. Then the minimum $\mathbf{x}_{\min}$ is identified by the smallest value of $\mathbf{c}^\top \mathbf{x}_{\min}$.

Although this procedure works theoretically, it does not work practically because of a huge amount of EDs and EPs to be checked.

1.4.2 Linear programming theory

Theorem 104 (Optimality conditions in LP)**.**

We have a linear program (LP) in the standard form. We denote $S = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ and also $D = \text{argmin}_{\mathbf{x}}\{\mathbf{c}^\top\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$. We assume $r(\mathbf{A}) = m$, $S \neq \emptyset$, and that all EPs and EDs are listed: $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are all EPs of S and $\mathbf{d}_1 \ldots, \mathbf{d}_l$ are all EDs of S. Then: $D \neq \emptyset \Leftrightarrow \forall j \in \{1 \ldots, l\} : \mathbf{c}^\top\mathbf{d}_j \geq 0$. In addition if $D \neq \emptyset$ then $\exists \mathbf{x}_i \in D$ that is an EP of S.

Proof: We use Theorem 98: $\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \in S \Leftrightarrow \exists \lambda_j \geq 0, j = 1, \ldots, k, \sum_{j=1}^k \lambda_j = 1$ and $\exists \mu_j \geq 0, j = 1, \ldots, l$ such that $\mathbf{x} = \sum_{j=1}^k \lambda_j \mathbf{x}_j + \sum_{j=1}^l \mu_j \mathbf{d}_j$. So, we may rewrite the LP using this representation of $\mathbf{x} \in S$ as follows:

$$? \in \text{argmin}_{\lambda,\mu}\{\ \mathbf{c}^\top(\textstyle\sum_{j=1}^k \lambda_j \mathbf{x}_j + \sum_{j=1}^l \mu_j \mathbf{d}_j) \mid$$
$$\lambda_j \geq 0, j = 1, \ldots, k, \textstyle\sum_{j=1}^k \lambda_j = 1, \mu_j \geq 0, j = 1, \ldots, l\}.$$

If $\mathbf{c}^\top\mathbf{d}_j < 0$ for some j then with $\mu_j \longrightarrow \infty$ (and remaining μ_j equal 0) the objective function is unbounded from below. So, necessity and sufficiency is clear If $\forall j : \mathbf{c}^\top\mathbf{d}_j \geq 0$ then because we minimize we assign $\forall j : \mathbf{m}_j = 0$. Then we need to solve

$$? \in \text{argmin}_{\lambda,\mu}\{\ \mathbf{c}^\top \textstyle\sum_{j=1}^k \lambda_j \mathbf{x}_j \mid \lambda_j \geq 0, j = 1, \ldots, k, \sum_{j=1}^k \lambda_j = 1\}.$$

We can do it by setting $\lambda_i = 1$ for i such that $\min\{\mathbf{c}^\top\mathbf{x}_j \mid j = 1, \ldots, k\} = \mathbf{c}^\top\mathbf{x}_i$ (remaining λ_j are zeroes). □

The theorem explains the full enumeration algorithm but because of its inefficiency is useless. We need something more sophisticated (like “Go down and follow edges").

1.4.3 Simplex method

The idea is to use $\mathbf{x}_{n+1} := \mathbf{x}_n + \lambda_n \mathbf{d}_n$ where $\mathbf{x}_n$ is the approximate optimal solution after n iterations, λ_n is a n'th step size, and $\mathbf{d}_n$ is the n'th direction along that we move from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$. The step should be chosen such that it improves the objective function value, i.e. $\mathbf{c}^\top\mathbf{x}_{n+1} \leq \mathbf{c}^\top\mathbf{x}_n$.

Remark 105 (LP notation)**.**

We have $\mathbf{x} \in \mathbb{R}^n$ variable, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ constant vectors, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix. We have the linear program in a standard form: $D = \text{argmin}_{\mathbf{x}}\{\mathbf{c}^\top\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ We denote $S = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ a feasible set and $D = \text{argmin}_{\mathbf{x}}\{\mathbf{c}^\top\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is a set of optimal solutions.
We assume that $\mathbf{A} = (\mathbf{B}, \mathbf{N})$ and $\mathbf{B}^{-1}$ exists. We call $\mathbf{B}$ matrix a basis. Related EP $\begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} \geq \mathbf{0}$ is also called a Basic Feasible Solution (BFS). We denote J_B a set of indices of $\mathbf{B}$ columns in $\mathbf{A}$and J_N is a set of indices of $\mathbf{N}$ columns in $\mathbf{A}$.
Rewriting the LP using $\mathbf{B}$ and $\mathbf{N}$, we get: $D = \text{argmin}_{\mathbf{x}_B, \mathbf{x}_N}\{\mathbf{c}_B^\top\mathbf{x}_B + \mathbf{c}_N^\top\mathbf{x}_N \mid \mathbf{B}\mathbf{x}_B + \mathbf{N}\mathbf{x}_N = \mathbf{b}, \mathbf{x}_B \geq \mathbf{0}, \mathbf{x}_N \geq \mathbf{0}\}$.

Remark 106 (Related concepts).

Remember that $\mathbf{w}^\top = \mathbf{c}_B^\top \mathbf{B}^{-1}$ are so called simplex multipliers, cf. with dual problems of LP. Denoting $(z_j)_j = \mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{N}$, we have the condition in the form $\forall j \in J_N : z_j - c_j \leq 0$. Differences $c_j - z_j$ (sometimes $z_j - c_j$) are called reduced costs.

Theorem 107 (Optimality test).

Let $\bar{\mathbf{x}} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} \geq \mathbf{0}$ is the EP of S. Assume that nondegeneracy condition is satisfied, i.e. $\mathbf{B}^1\mathbf{b} > \mathbf{0}$. Then: $\bar{\mathbf{x}} \in D \Leftrightarrow -\mathbf{c}_N^\top + \mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{N} \leq \mathbf{0}^\top$.

Proof: Let $\bar{\mathbf{x}} = \begin{pmatrix} \bar{\mathbf{x}}_B \\ \bar{\mathbf{x}}_N \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{b}} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} \geq \mathbf{0}$ be an EP. Compute the objective function value: $\mathbf{c}^\top \bar{\mathbf{x}} = (\mathbf{c}_B^\top \mathbf{c}_N^\top) \begin{pmatrix} \bar{\mathbf{x}}_B \\ \bar{\mathbf{x}}_N \end{pmatrix} = \mathbf{c}^\top \mathbf{B}^{-1}\mathbf{b}$. Let $\mathbf{x} \in S$. So it satisfies: $\mathbf{x} = \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix} \geq \mathbf{0}$ and $\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}_B + \mathbf{N}\mathbf{x}_N = \mathbf{b}$. Hence $\mathbf{x}_B = \mathbf{B}^{-1}(\mathbf{b} - \mathbf{N}\mathbf{x}_N) = \bar{\mathbf{x}} - \mathbf{B}^{-1}\mathbf{N}\mathbf{x}_N$. We compute for the new solution x: $\mathbf{c}^\top \mathbf{x} = \mathbf{c}_B^\top \mathbf{x}_B + \mathbf{c}_N^\top \mathbf{x}_N = \mathbf{c}_B^\top \mathbf{B}^{-1}(\mathbf{b} - \mathbf{N}\mathbf{x}_N) + \mathbf{c}_N^\top \mathbf{x}_N = \mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{b} - (\mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{N} + \mathbf{c}_N^\top)\mathbf{x}_N = \mathbf{c}^\top \bar{\mathbf{x}} - (z_j - c_j)_j^\top \mathbf{x}_N$. So with $\forall j : z_j - c_j \leq 0$ the solution $\bar{\mathbf{x}}$ cannot be improved. □

Theorem 108 (Improving direction).

Let $\bar{\mathbf{x}} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} \geq \mathbf{0}$ is the given EP and $\mathbf{B}^1\mathbf{b} > \mathbf{0}$. We also assume that $\exists k \in J_N : z_k - c_k > 0$. Then either $\exists \mathbf{x}$ EP (improved BFS) such that $\mathbf{c}^\top \mathbf{x} < \mathbf{c}^\top \bar{\mathbf{x}}$ or $\forall x_k \geq 0 : \exists \mathbf{x} \in S$ with x_k as a j'th component satisfying $\mathbf{c}^\top \mathbf{x} < \mathbf{c}^\top \bar{\mathbf{x}}$ (unbounded case).

Proof: Let $k \in J_N : z_k - c_k > 0$, i.e. $\mathbf{c}^\top \mathbf{B}^{-1}\mathbf{N} - \mathbf{c}_N^\top \not\leq \mathbf{0}$. We see that $z_k = \mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{a}_k$. (Why? Hint: Visualize the multiplication of matrices.) We set $\forall j \in J_N \setminus \{k\} : x_j := 0$ (but $x_k \neq 0$), so $\mathbf{c}^\top \mathbf{x} = \mathbf{c}^\top \bar{\mathbf{x}} - (\mathbf{c}_B^\top \mathbf{B}^{-1}\mathbf{a}_k - c_k)x_k = \mathbf{c}^\top \bar{\mathbf{x}} - (z_k - c_k)x_k$. (How big x_k to choose?) We choose the biggest but feasible x_k to get the biggest improvement of the objective function value. Therefore, $\mathbf{x}_B = \mathbf{B}^{-1}(b - \mathbf{N}\mathbf{x}_N) \geq \mathbf{0}$ must be valid. So, we may rewrite the nonnegativity constraints as $\mathbf{x}_B = \bar{\mathbf{x}}_B - \mathbf{B}^{-1}\mathbf{N}\mathbf{x}_N = \bar{\mathbf{b}} - \mathbf{B}^{-1}\mathbf{a}_k x_k \geq \mathbf{0}$ and we search for: $\max\{x_k \mid \mathbf{B}^{-1}\mathbf{a}_k x_k \leq \bar{\mathbf{b}}\}$. We denote $\mathbf{y}_k = \mathbf{B}^{-1}\mathbf{a}_k$ for simplicity. Then we find the maximum as follows: If $\mathbf{y}_k < \mathbf{0}$ then $x_k \geq \frac{\bar{b}_i}{y_{ik}}, i = 1, \ldots, m$, and hence, $x_k \longrightarrow \infty$ and the unbounded case is proven. Otherwise: $\mathbf{y}_k \not\leq \mathbf{0} \Rightarrow \lambda = \frac{\bar{b}_r}{y_{rk}} = \min\{\frac{\bar{b}_i}{y_{ik}} : y_{ik} > 0\}$ and $x_{k,\max} = \lambda$. Then (for local index i related to the column of $\mathbf{B}$ and corresponding to index j of the column of the original $\mathbf{A}$): $\forall j \in J_B : x_j = x_{Bi} = \bar{b}_i - \frac{\bar{b}_r}{y_{rk}} y_{ik}, i = 1, \ldots, m$ (and $x_{Bi} = 0$ for $i = r$), $x_k = \frac{\bar{b}_r}{y_{rk}}$ and $x_j = 0$ for $j \in J_N \setminus \{k\}$ and the new solution is an EP formed from the updated $\mathbf{B}$ that has linearly independent columns $\mathbf{a}_j, j \in J_N \setminus \{r\}$ and the column $\mathbf{a}_k$. It is the adjacent basis and EP to the previous one. The requirement $\mathbf{c}^\top \mathbf{x} < \mathbf{c}^\top \bar{\mathbf{x}}$ is also satisfied. □

Corollary 109 (Finite convergence).

Assuming LP nondegeneracy, the simplex algorithm described in the proof converges in a finite number of steps.

Proof Number of EPs is finite and the strict decrease of the objective function value excludes the possibility to pass twice through any EP. □

Algorithm 110 (Simplex algorithm — matrix form)**.**

0. Initialization: Solve

$$? \in \operatorname{argmin}_{\mathbf{x}}\{\mathbf{c}^\top\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}, r(\mathbf{A}) = m.$$

Denote $J = \{1, \ldots, n\}$. Assume that the initial basis is given $J = J_B \cup J_N$, $J_B \cap J_N = \emptyset$, $\mid J_B \mid = m$, $\mid J_N \mid = n - m$. Therefore, we have $\mathbf{B} = (\mathbf{a}_j)_{j \in J_B}$, $r(\mathbf{B}) = m$, $\mathbf{N} = (\mathbf{a}_j)_{j \in J_N}$, and $\mathbf{c}_B^\top$, $\mathbf{c}_N^\top$.

1. Solve system of linear equations $\mathbf{B}\mathbf{x}_B = \mathbf{b}$ efficiently to get $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b}$. Set $\mathbf{x}_N = \mathbf{0}$, and compute $z = \mathbf{c}_B^\top \mathbf{x}_B$.
2. Solve $\mathbf{w}^\top\mathbf{B} = \mathbf{c}_B^\top$ efficiently to get $\mathbf{w}^\top = \mathbf{c}_B^\top\mathbf{B}^{-1}$. Compute $\forall j \in J_N : z_j - c_j = \mathbf{w}^\top\mathbf{a}_j - c_j$. Search $z_k - c_k = \max\{z_j - c_j \mid j \in J \in J_N\}$. If $z_k - c_k \leq 0$ then STOP because the optimum was reached.
3. Solve $\mathbf{B}\mathbf{y}_k = \mathbf{a}_k$ efficiently to get $\mathbf{y}_k = \mathbf{B}^{-1}\mathbf{a}_k$. If $\mathbf{y}_k \leq \mathbf{0}$ then STOP because of the unboundedness and recession direction is specified by $\begin{pmatrix} -\mathbf{y}_k \\ \mathbf{e}_k \end{pmatrix}$.
4. Solve $\frac{\bar{b}_r}{y_{rk}} = \min_{i \in J_B}\{\frac{\bar{b}_i}{y_{ik}} : y_{ik} > 0\}$. Then: $J_B := J_B \setminus \{r\} \cup \{k\}$, $J_N := J_N \setminus \{k\} \cup \{r\}$, and update $\mathbf{B} := (\mathbf{a}_j)_{j \in J_B}$, $\mathbf{N} := (\mathbf{a}_j)_{j \in J_N}$ and GO TO 1 .

Exercise 111.

What is the most important numerical procedure for the simplex method?

Remark 112.

We expressed new $\mathbf{x}$ using $\bar{\mathbf{x}}$, λ, and $\mathbf{d}$. So, $\mathbf{x}_{n+1} := \mathbf{x}_n + \lambda_n\mathbf{d}_n$ where:

$$\begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix} := \begin{pmatrix} \bar{\mathbf{b}} \\ \mathbf{0} \end{pmatrix} + x_k \begin{pmatrix} -\mathbf{y}_k \\ \mathbf{e}_k \end{pmatrix} = \begin{pmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{pmatrix} + x_k \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_k \\ \mathbf{e}_k \end{pmatrix}$$

Remember that $\mathbf{x}_n$ is feasible, so $\bar{\mathbf{b}} \geq \mathbf{0}$ (or $\bar{\mathbf{b}} > \mathbf{0}$ for the nondegenerate case). Further $x_k = \frac{\bar{b}_r}{y_{rk}} > 0$. If $\mathbf{y}_k \leq \mathbf{0}$ then we get a descent ED. For the objective function, we get:

$$\mathbf{c}^\top\mathbf{x} = \mathbf{c}^\top\bar{\mathbf{x}} + x_k(\mathbf{c}_B^\top, \mathbf{c}_N^\top) \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_k & \mathbf{e}_k \end{pmatrix}.$$

To decrease the objective function value $(\mathbf{c}_B^\top, \mathbf{c}_N^\top) \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_k & \mathbf{e}_k \end{pmatrix} < 0$ must be satisfied.

Remark 113 (Simplex Tableau)**.**

What is behind the simplex tableau? We rewrite the LP as follows:

$$? \in \operatorname{argmin}\{z \mid z - \mathbf{c}^\top \mathbf{x} = 0, \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

Then we may use the tabular form often used for the Gauss elimination method solving systems of linear equations. The sequence of simplex tables derived by basic equivalent transformations applied to matrix blocks follows:

$$\left(\begin{array}{ccc|c} 1 & -\mathbf{c}^\top & & 0 \\ \mathbf{0} & \mathbf{A} & & \mathbf{b} \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -\mathbf{c}_B^\top & -\mathbf{c}_N^\top & 0 \\ \mathbf{0} & \mathbf{B} & \mathbf{N} & \mathbf{b} \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -\mathbf{c}_B^\top & -\mathbf{c}_N^\top & 0 \\ \mathbf{0} & \mathbf{B}^{-1}\mathbf{B} & \mathbf{B}^{-1}\mathbf{N} & \mathbf{B}^{-1}\mathbf{b} \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & -\mathbf{c}_B^\top & -\mathbf{c}_N^\top & 0 \\ \mathbf{0} & \mathbf{I} & \mathbf{B}^{-1}\mathbf{N} & \mathbf{B}^{-1}\mathbf{b} \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & \mathbf{c}_B^\top - \mathbf{c}_B^\top & \mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{N} - \mathbf{c}_N^\top & \mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{I} & \mathbf{B}^{-1}\mathbf{N} & \mathbf{B}^{-1}\mathbf{b} \end{array}\right) \sim$$

$$\left(\begin{array}{ccc|c} 1 & \mathbf{0}^\top & \mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{N} - \mathbf{c}_N^\top & \mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{I} & \mathbf{B}^{-1}\mathbf{N} & \mathbf{B}^{-1}\mathbf{b} \end{array}\right)$$

We see that the final table contains the information suitable for the simplex algorithm and the choice of k (see the first row) and r (see the last and k'th column containing $\mathbf{y}_k$). In the last column, we find the information about the last $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b}$ as we know that $\mathbf{x}_B = \bar{\mathbf{x}}_B - \mathbf{B}^{-1}\mathbf{N}\mathbf{x}_N = \bar{\mathbf{b}} - (\mathbf{y}_j)_{j\in J_N}\mathbf{x}_N$ and $\mathbf{x}_N = \mathbf{0}$. In the first row, we find the optimality condition term, and the objective function value $\mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{b}$ as $z = \mathbf{c}^\top\mathbf{x} = \mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{b} - (\mathbf{c}_B^\top\mathbf{B}^{-1}\mathbf{N} - \mathbf{c}_N^\top)\mathbf{x}_N = \bar{z} - (z_j - c_j)_j^\top\mathbf{x}_N$.

Remark 114 (Important ideas).

- Simple lower and upper bounds $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$ can be implemented in the special way and need not be considered as usual constraints.
- Initial basic feasible solution can be found by Two Phase Method that begins with the solution of the modified first phase program ($\mathbf{b} \geq \mathbf{0}$): $? \in \operatorname{argmin}\{\mathbf{1}^\top \mathbf{x} \mid \mathbf{A}\mathbf{x} + \mathbf{x}_s = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \mathbf{x}_s \geq \mathbf{0}\}$. When the first phase program is solved the original one is either detected as infeasible (for $z_{\min} > 0$) or (for $z_{\min} = 0$) the initial basis and J_B is obtained for the original program.
- The identification of all optimal solutions needs to find all EPs and EDs of the polyhedral set D. There are backtracking procedures based on the spanning tree search modified for LP.
- Instead of Dantzig's rule (used in Algorithm 110), identifying k index by $z_k - c_k = \max\{z_j - c_j \mid j \in J_N\}$, another rules for the choice of the 'key column' can be used (see also Deveaux rule and column generation techniques for large-scale problems).
- Cycling (stalling) may be caused by degeneracy. There are theoretical lexicographical and Bland's rules avoiding it. In practice, rounding errors may help. For integer network flow programs, the network simplex method is developed.
- Regarding memory requirements: For sparse matrix structures, the revised simplex method (also in the product matrix form) has been developed.
- The number of required operations per one iteration can be expressed by a polynomial function of the program size.
- For special structures decomposition approaches may help.
- There is a difference between theoretical (the worst case) overall computational complexity that is exponential (the number of iterations required for the problem with n variables might be 2^n) in comparison with a practical experience showing that for certain programs even $2m - -3m$ iterations are enough (m is a number of constraints).
- There are also polynomial algorithms for linear programs. The first polynomial Khachian's algorithm had the theoretical importance. The next Karmarkar's algorithm led to the development of the class of the efficient interior point methods for the huge linear programs.
- Previous formulas lead to basic sensitivity results

$$\frac{\partial x_{Bl}}{\partial x_{Nj}} = -y_{lj}, \ \frac{\partial z}{\partial x_{Nj}} = c_j - z_j, \ \frac{\partial x_{Bl}}{\partial b_i} = B^{li}, \ \frac{\partial z}{\partial b_i} = w_i$$

  where B^{li} represents the element of $\mathbf{B}^{-1}$.
- LP duality will be revised with NLP later.

Exercise 115.

Compute typical examples from linear programming reviewing: the standard form, two-phase method, and simplex tableau computations.

1.5 Convex functions

1.5.1 Basic properties

Definition 116 (Convex function)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set, $f : S \longrightarrow \mathbb{R}$.
The function f is said to be convex on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S$, $\forall \lambda \in (0,1) : f(\lambda \mathbf{x}_1 + (1-\lambda) f(\mathbf{x}_2)) \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$.
The function f is said to be strictly convex on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S$, $\mathbf{x}_1 \neq \mathbf{x}_2$, $\forall \lambda \in (0,1) : f(\lambda \mathbf{x}_1 + (1-\lambda) f(\mathbf{x}_2)) < \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$.
The function f is said to be concave (strictly concave) on $S \Leftrightarrow -f$ is convex (strictly convex) on S.

Exercise 117.

How Definition 116 changes when $\lambda \in (0,1)$ is replaced by $\lambda \in [0,1]$?

Exercise 118.

Draw figures of different real functions of one real variable (convex, concave, etc.).

Remark 119 (Geometric interpretation)**.**

For a convex function f, the value of f at points on line segment $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is less or equal to the height of the chord joining $[\mathbf{x}_1, f(\mathbf{x}_1)]$ and $[\mathbf{x}_2, f(\mathbf{x}_2)]$.

Remark 120.

If f is both convex and concave $\Leftrightarrow$ f is affine (linear). f may be non-convex over S but can be convex over a convex subset of S.

Exercise 121.

Why a convex function must be defined on a convex set? Explain, using a simple example!

Definition 122 (Level set)**.**

$S_\alpha = \{\mathbf{x} \in S \mid f(\mathbf{x}) \leq \alpha\}$ is called an α-level set.

Theorem 123 (Convexity of a level set)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set, $f : S \longrightarrow \mathbb{R} \Rightarrow \forall \alpha \in \mathbb{R} : S_\alpha$ is a convex set.

Proof: Let $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha \Rightarrow \mathbf{x}_1, \mathbf{x}_2 \in S$ as $S_\alpha \subset S$. By Definition 122 $f(\mathbf{x}_1) \leq \alpha$, $f(\mathbf{x}_2) \leq \alpha$. We choose $\lambda \in (0,1)$ and we have $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$. Because of convexity of f: $f(\mathbf{x}) \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \leq \lambda\alpha + (1-\lambda)\alpha = \alpha \Rightarrow f(\mathbf{x}) \leq \alpha \Rightarrow \mathbf{x} \in S_\alpha$. □

Exercise 124.

Why Theorem 123 is very important for nonlinear programming? Hint: Think about the NLP constraint $g_i(\mathbf{x}) \leq 0$ where $g_i : S \longrightarrow \mathbb{R}$, and then $C = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m\} = \cap_i C_i = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \leq 0\}$.

Theorem 125 (Continuity of a convex function)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set, $f : S \longrightarrow \mathbb{R}$ is a convex function $\Rightarrow f$ is a continuous function on $\operatorname{int} S$.

Proof: Let $\bar{\mathbf{x}} \in \operatorname{int} S$, we need to show $\forall \varepsilon > 0\ \exists \delta > 0 : \|\mathbf{x} - \bar{\mathbf{x}}\| \leq \delta \Rightarrow |f(\mathbf{x}) - f(\bar{\mathbf{x}})| \leq \varepsilon$. Because $\bar{\mathbf{x}} \in \operatorname{int} S \Rightarrow$ we set $\varepsilon > 0$. $\exists \delta' : \|\mathbf{x} - \bar{\mathbf{x}}\| \leq \delta' \Rightarrow \mathbf{x} \in S$. We define θ as $\theta = \max_{1 \leq i \leq n}\{\max\{f(\bar{\mathbf{x}} + \delta'\mathbf{e}_i) - f(\bar{\mathbf{x}}), f(\bar{\mathbf{x}} - \delta'\mathbf{e}_i) - f(\bar{\mathbf{x}})\}\}$, i.e. maximum increase by a coordinate direction. Because of convexity: $0 \leq \theta < \infty$ (finite and nonnegative). $\delta := \min(\frac{\delta'}{n}, \frac{\varepsilon\delta'}{n\theta})$. Choose $\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\| \leq \delta$. If $x_i - \bar{x}_i \geq 0$ let $\mathbf{z}_i = \delta'\mathbf{e}_i$ otherwise $\mathbf{z}_i = -\delta'\mathbf{e}_i$. Then $\mathbf{x} - \bar{\mathbf{x}} = \sum_{i=1}^n \alpha_i \mathbf{z}_i$ where $\forall i : \alpha_i \geq 0$. Furthermore $\|\mathbf{x} - \bar{\mathbf{x}}\| = \delta'(\sum_{i=1}^n \alpha_i^2)^{\frac{1}{2}}$. Because $\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \delta \Rightarrow \delta'(\sum_{i=1}^n \alpha_i^2)^{\frac{1}{2}} \leq \delta = \min\{\frac{\delta'}{n}, \frac{\delta'}{n}\} \Rightarrow (\sum_{i=1}^n \alpha_i^2)^{\frac{1}{2}} \leq \min\{\frac{1}{n}, \frac{\varepsilon}{\theta}\frac{1}{n}\} \leq \frac{1}{n}$ $\Rightarrow$ if only one $\alpha_i > 0 \Rightarrow |\alpha_i| = \alpha_i \leq \frac{1}{n}$ and $0 \leq n\alpha_i \leq 1$. Then $f(\mathbf{x}) = f(\bar{\mathbf{x}} + \sum_{i=1}^i \alpha_i \mathbf{z}_i) = f(\frac{1}{n}\sum_{i=1}^n (\bar{\mathbf{x}} + n\alpha_i \mathbf{z}_i)) \leq$ (convexity) $\frac{1}{n}\sum_{i=1}^n f(\bar{\mathbf{x}} + n\alpha_i \mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^n f((1 - n\alpha_i)\bar{\mathbf{x}} + n\alpha_i(\bar{\mathbf{x}} + \mathbf{z}_i)) \leq$ (convexity) $\frac{1}{n}\sum_{i=1}^n ((1 - n\alpha_i) f(\bar{\mathbf{x}}) + n\alpha_i f(\bar{\mathbf{x}} + \mathbf{z}_i))$. Then $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}) + \sum_{i=1}^n (\alpha_i f(\bar{\mathbf{x}} + \mathbf{z}_i) - \alpha_i f(\bar{\mathbf{x}})) - f(\bar{\mathbf{x}})$. So, $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leq \sum_{i=1}^n \alpha_i (f(\bar{\mathbf{x}} + \mathbf{z}_i) - f(\bar{\mathbf{x}}) \leq$ (see above) $\theta \sum_{i=1}^n \alpha_i$ (that is $\leq \frac{1}{n}$ and $\frac{\varepsilon}{n\theta}$), so $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leq \theta \sum_{i=1}^n \frac{\varepsilon}{n\theta} = \varepsilon$ and upper semicontinuity is proven. We need $f(\bar{\mathbf{x}}) - f(\mathbf{x}) \leq \varepsilon$ to have $|f(\mathbf{x}) - f(\bar{\mathbf{x}})| \leq \varepsilon$. But $\bar{\mathbf{x}} = \frac{\mathbf{y}+\mathbf{x}}{2}$ and $f(\bar{\mathbf{x}}) \leq \frac{f(\mathbf{y}) + f(\mathbf{x})}{2}$ (convexity). Then $f(\mathbf{y}) \geq 2f(\bar{\mathbf{x}}) - f(\mathbf{x}) \Rightarrow \varepsilon \geq 2f(\bar{\mathbf{x}}) - f(\mathbf{x}) - f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) - f(\mathbf{x})$. □

Exercise 126.

Why Theorem 125 is useful for NLP? Hint: Think about Weierstrass' Theorem 12 assumptions!

Exercise 127.

Illustrate Theorem 125 to show that discontinuity on boundary is allowed. Hint: For example, use $y = x^2$ on $[-1, 1]$.

Definition 128 (Directional derivative)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : S \longrightarrow \mathbb{R}$, $\bar{\mathbf{x}} \in S$, $\mathbf{d} \neq \mathbf{0} : \exists \lambda_0 > 0 \; \forall \lambda < \lambda_0 : \bar{\mathbf{x}} + \lambda \mathbf{d} \in S$. Then, we define the directional derivative of f at $\bar{\mathbf{x}}$ along $\mathbf{d}$ as:

$$f'(\bar{\mathbf{x}}; \mathbf{d}) := \lim_{\lambda \longrightarrow 0^+} \frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d} - f(\bar{\mathbf{x}}))}{\lambda}.$$

Lemma 129 (Existence of a directional derivative)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then $\forall \mathbf{x} \in \mathbb{R}^n \; \forall \mathbf{d} \in \mathbb{R}^n \; \mathbf{d} \neq \mathbf{0} : \exists f'(\bar{\mathbf{x}}; \mathbf{d})$.

Proof: Let $\lambda_2 > \lambda_1 > 0 : f(\bar{\mathbf{x}} + \lambda_1 \mathbf{d}) = f(\frac{\lambda_1}{\lambda_2}(\bar{\mathbf{x}} + \lambda_2 \mathbf{d}) + (1 - \frac{\lambda_1}{\lambda_2})\bar{\mathbf{x}}) \leq \frac{\lambda_1}{\lambda_2} f(\bar{\mathbf{x}} + \lambda_2 \mathbf{d}) + (1 - \frac{\lambda_1}{\lambda_2}) f(\bar{\mathbf{x}})$ $\Rightarrow \frac{f(\bar{\mathbf{x}} + \lambda_1 \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda_1} \leq \frac{1}{\lambda_1}(\frac{\lambda_1}{\lambda_2} f(\bar{\mathbf{x}} + \lambda_2 \mathbf{d}) + (1 - \frac{\lambda_1}{\lambda_2}) f(\bar{\mathbf{x}})) = \frac{f(\bar{\mathbf{x}} + \lambda_2 \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda_2}$. So the difference quotient is monotone decreasing (non-increasing) for $\lambda \longrightarrow 0^+$. So, $\forall \lambda > 0 : f(\bar{\mathbf{x}}) = f(\frac{\lambda}{1+\lambda}(\bar{\mathbf{x}} - \mathbf{d}) + \frac{1}{1+\lambda}(\bar{\mathbf{x}} + \lambda \mathbf{d}))$ (because $\bar{\mathbf{x}} = \frac{1+\lambda}{1+\lambda}(\bar{\mathbf{x}} - \mathbf{d} + \mathbf{d}) = \frac{\lambda}{1+\lambda}(\bar{\mathbf{x}} - \mathbf{d}) + \frac{1}{1+\lambda}(\lambda \mathbf{d} + \bar{\mathbf{x}} - \mathbf{d} + \mathbf{d})) \leq \frac{\lambda}{1+\lambda} f(\bar{\mathbf{x}} - \mathbf{d}) + \frac{1}{1+\lambda} f(\bar{\mathbf{x}} + \lambda \mathbf{d})$ $\Rightarrow (1 + \lambda) f(\bar{\mathbf{x}}) \leq \lambda f(\bar{\mathbf{x}} - \mathbf{d}) + f(\bar{\mathbf{x}} + \lambda \mathbf{d}) \Rightarrow f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}}) \geq \lambda f(\bar{\mathbf{x}}) + \lambda f(\bar{\mathbf{x}} - \mathbf{d}) \Rightarrow \frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} \geq f(\bar{\mathbf{x}}) + f(\bar{\mathbf{x}} - \mathbf{d})$. For the given $\bar{\mathbf{x}}$, $\mathbf{d}$, the RHS is constant and the LHS is bounded from below. From theorems of mathematical analysis, we conclude that the limit in the theorem exists and:

$$\lim_{\lambda \longrightarrow 0^+} \frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} = \inf_{\lambda > 0} \frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}. \quad \square$$

Remark 130.

Let f convex on convex S, $\mathbf{d}$ by Definition 128, $\bar{\mathbf{x}} \in \operatorname{int} S \Rightarrow \exists f'(\mathbf{x}; \mathbf{d})$. If $\bar{\mathbf{x}} \in \partial S$: $f'(\mathbf{x}; \mathbf{d})$ might be $-\infty$ (also for continuous f — illustrate by a figure in $\mathbb{R}^2$).

Definition 131 (Graph)**.**

Let $S \subset \mathbb{R}^n$, $f : S \longrightarrow \mathbb{R}$. Then $\{(\mathbf{x}^\top, f(\mathbf{x}))^\top \mid \mathbf{x} \in S\} \subset \mathbb{R}^{n+1}$ is called a graph of f.

Definition 132 (Epigraph)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : S \longrightarrow \mathbb{R}$. Then $\operatorname{epi} f := \{(\mathbf{x}^\top, y)^\top \mid \mathbf{x} \in S, y \in \mathbb{R}, y \geq f(\mathbf{x})\}$ is called an epigraph of f.

Definition 133 (Hypograph)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : S \longrightarrow \mathbb{R}$. Then $\operatorname{hyp} f := \{(\mathbf{x}^\top, y)^\top \mid \mathbf{x} \in S, y \in \mathbb{R}, y \leq f(\mathbf{x})\}$ is called a hypograph of f.

Theorem 134 (Convexity of an epigraph)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ is a convex set. Then $f : S \longrightarrow \mathbb{R}$ is a convex function $\Leftrightarrow$ epi$f$ is a convex set.

Remark 135.

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ is a convex set, $f : S \longrightarrow \mathbb{R}$ is a concave function $\Leftrightarrow$ hyp$f$ is a convex set.

Proof of Theorem 134: $\Rightarrow$: f convex, $(\mathbf{x}_1^\top, y_1)^\top$, $(\mathbf{x}_2^\top, y_2)^\top \in$ epi$f$ as $(\mathbf{x}_1, \mathbf{x}_2 \in S)$, $y_1 \geq f(\mathbf{x}_1)$, $y_2 \geq f(\mathbf{x}_2)$. Take $\lambda \in (0,1)$ arbitrarily: $\lambda y_1 + (1-\lambda)y_2 \geq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) \geq f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)$. So, $y \geq f(\mathbf{x})$ and $(\mathbf{x}^\top, y)^\top \in$ epi$f$ and epi$f$ is convex.
$\Leftarrow$: Conversely, assume epi$f$ is convex and S convex. Let $\mathbf{x}_1, \mathbf{x}_2 \in S$ then $(\mathbf{x}_1^\top, f(\mathbf{x}_1))^\top$, $(\mathbf{x}_2^\top, f(\mathbf{x}_2))^\top \in$ epi$f$. Then by Definition 132 and by the assumption of Theorem 134: $\forall\lambda \in (0,1) : (\lambda\mathbf{x}_1^\top + (1-\lambda)\mathbf{x}_2^\top, \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2))^\top \in$ epi$f$. Since $\mathbf{x} \in S$ then $(\mathbf{x}^\top, f(\mathbf{x}))^\top \in$ epi$f$ by Definition 132 and $\lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) = y \geq f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)$, so f is convex. □

Remark 136 (Usefulness of epigraph)**.**

When $f(\mathbf{x})$ is nondifferentiable then the solution of $? \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ can be obtained by solving of a modified program $? \in \operatorname{argmin}\{z \mid z \geq f(\mathbf{x}), \mathbf{x} \in S\}$ where the epigraph of f is considered in the feasibility region description.

Exercise 137.

Solve the following regression problem: For points $(x_j, y_j), j = 1,2,3$: $(1,1), (2,4), (3,2)$ find a straight line formula $y = \beta x$ minimizing $\max\{|y_j - \beta x_j| : j = 1,2,3\}$. Hint: Use the criterion epigraph to reformulate the problem as an LP. Draw a figure!

Definition 138 (Subgradient)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : S \longrightarrow \mathbb{R}$ is convex. Then $\boldsymbol{\xi} \in \mathbb{R}^n$ is called a subgradient of f at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S : f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})$.

Subgradient for a concave function is defined using $\leq$ in Definition 138.

Definition 139 (Subdifferential)**.**

The set of all subgradients of f at $\bar{\mathbf{x}}$ is called a subdifferential of f at $\bar{\mathbf{x}}$.

Exercise 140.

Draw subgradient for $S \subset \mathbb{R}$ and $S \subset \mathbb{R}^2$, and $f : S \longrightarrow \mathbb{R}$.

Remark 141.

Recognize that as epi$f$ is convex set then $\exists$ a supporting hyperplane, which is described by a subgradient. $f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})$ corresponds to a supporting hyperplane of epi$f$ at $(\bar{\mathbf{x}}^\top, f(\bar{\mathbf{x}}))^\top$. Then $\boldsymbol{\xi}$identifies its slope.

Exercise 142.

Find a subgradient and draw a figure for $y = |x|$, $\bar{x} = 0$. Hint: Find that $\xi \in [-1, 1]$ as $y = 0 + \xi(x - 0)$.

Exercise 143.

Develop own examples in RR^2 and draw contour graphs for $\bar{\mathbf{x}}, \mathbf{x} \in S$ and $\boldsymbol{\xi} \in \mathbb{R}^2$.

Theorem 144 (Existence of subdifferential)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be convex, $f : S \longrightarrow \mathbb{R}$ be convex $\Rightarrow \forall \bar{\mathbf{x}} \in \operatorname{int} S\ \exists \boldsymbol{\xi} : H = \{(\mathbf{x}^\top, y)^\top \mid y = f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})\}$ supports epi$f$ at $(\mathbf{x}^\top, f(\bar{\mathbf{x}}))^\top$. In particular, $\forall \mathbf{x} \in S : f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})$, so $\boldsymbol{\xi}$ is a subgradient of f at $\bar{\mathbf{x}}$, i.e. subdifferential of f is a nonempty set for interior points of S.

Proof: By theorem epi$f$ is convex. Let $(\bar{\mathbf{x}}^\top, f(\bar{\mathbf{x}}))^\top \in \partial \text{epi} f$. So $\exists (\boldsymbol{\xi}_0^\top, \mu)^\top \in \mathbb{R}^{n+1}$ and $(\boldsymbol{\xi}_0^\top, \mu)^\top \neq \mathbf{0}$ such that $\forall (\mathbf{x}^\top, y)^\top \in \text{epi} f : \boldsymbol{\xi}_0^\top(\mathbf{x} - \bar{\mathbf{x}}) + \mu(y - f(\bar{\mathbf{x}})) \leq 0$ (by Theorem 62 about supporting hyperplane). If $\mu > 0 \Rightarrow y$ chosen enough large and we obtain a contradiction. If $\mu > 0 \Rightarrow \forall \mathbf{x} \in S : \boldsymbol{\xi}_0^\top(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$ Since $\bar{\mathbf{x}} \in \operatorname{int} S\ \exists \lambda > 0$ such that $\bar{\mathbf{x}} + \lambda \boldsymbol{\xi}_0 \in S$. Therefore, we may take arbitrary direction, even $\boldsymbol{\xi}$. So, $\boldsymbol{\xi}_0^\top(\bar{\mathbf{x}} + \lambda \boldsymbol{\xi}_0 - \bar{\mathbf{x}}) = \lambda \boldsymbol{\xi}_0^\top \boldsymbol{\xi}_0 = \lambda \|\boldsymbol{\xi}_0\|^2 \leq 0$. Because $\lambda > 0$ then $\boldsymbol{\xi}_0 = \mathbf{0}$ that is a contradiction with $\mu > 0$. Then $\mu < 0$ and we have $\frac{\boldsymbol{\xi}_0^\top}{|\mu|}(\mathbf{x} - \bar{\mathbf{x}}) + \frac{\mu}{|\mu|}(y - f(\bar{\mathbf{x}})) = \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}}) + f(\bar{\mathbf{x}}) - y \leq 0$. By letting $y = f(\mathbf{x})$ and $\boldsymbol{\xi} = \frac{\boldsymbol{\xi}_0^\top}{|\mu|}$ we get the conclusion of theorem. □

Corollary 145.

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be convex, $f : S \longrightarrow \mathbb{R}$ strictly convex function. Then $\forall \bar{\mathbf{x}} \in \operatorname{int} S : \exists \boldsymbol{\xi} \in \mathbb{R}^n : \forall \mathbf{x} \in S\ \mathbf{x} \neq \bar{\mathbf{x}} : f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})$.

Proof: By theorem $\exists \boldsymbol{\xi}\ \forall \mathbf{x} \in S : f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}})$. By contradiction $\exists \hat{\mathbf{x}} \neq \bar{\mathbf{x}} : f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\hat{\mathbf{x}} - \bar{\mathbf{x}})$ and by strict convexity of f for $\lambda \in (0, 1) : f(\lambda \bar{\mathbf{x}} + (1 - \lambda)\hat{\mathbf{x}}) < \lambda f(\bar{\mathbf{x}}) + (1 - \lambda) f(\hat{\mathbf{x}}) = \lambda f(\bar{\mathbf{x}}) + (1 - \lambda)(f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\hat{\mathbf{x}} - \bar{\mathbf{x}})) = f(\bar{\mathbf{x}}) + (1 - \lambda)\boldsymbol{\xi}^\top(\hat{\mathbf{x}} - \bar{\mathbf{x}})$. For $\mathbf{x} = \lambda \bar{\mathbf{x}} + (1 - \lambda)\hat{\mathbf{x}}$ by theorem $f(\lambda \bar{\mathbf{x}} + (1 - \lambda)\hat{\mathbf{x}}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\lambda \bar{\mathbf{x}} + (1 - \lambda)\hat{\mathbf{x}} - \bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) + (1 - \lambda)(\hat{\mathbf{x}} - \bar{\mathbf{x}})$. □

Remark 146.

The converse ($\Leftarrow$) is not true in general, so $\exists$ subgradients $\forall \mathbf{x} \in \operatorname{int} S \not\Rightarrow f$ is convex on S (however, it is true on int S).

Theorem 147 (Convexity implied by subgradient)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, convex, $f : S \longrightarrow \mathbb{R}$, $\forall \bar{\mathbf{x}} \in \operatorname{int} S : \exists \boldsymbol{\xi}\ \forall \mathbf{x} \in S : f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top (\mathbf{x} - \bar{\mathbf{x}})$ $\Rightarrow$ f is convex on $\operatorname{int} S$.

Proof: $\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{int} S$, $\lambda \in (0,1)$, $\operatorname{int} S$ of convex set is a convex set, so $\bar{\mathbf{x}} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \operatorname{int} S \subset S \Rightarrow$ $\exists \boldsymbol{\xi}$ of f at $\bar{\mathbf{x}}$, so: For $\mathbf{x}_1$: $f(\mathbf{x}_1) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top (\mathbf{x}_1 - \lambda \mathbf{x}_1 - (1-\lambda)\mathbf{x}_2) = f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + (1-\lambda)\boldsymbol{\xi}^\top (\mathbf{x}_1 - \mathbf{x}_2)$. For $\mathbf{x}_2$: $f(\mathbf{x}_2) \geq f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + \boldsymbol{\xi}^\top (\mathbf{x}_2 - \lambda \mathbf{x}_1 - (1-\lambda)\mathbf{x}_2) = f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + \lambda \boldsymbol{\xi}^\top (\mathbf{x}_2 - \mathbf{x}_1)$. Then, multiplying the inequalities by λ and $1-\lambda$ and summing them, we obtain $\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \geq (\lambda + 1 - \lambda) f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + 0$. So, f is convex on $\operatorname{int} S$. $\square$

1.5.2 Differentiable convex functions

Definition 148 (Gradient, differentiable function at point)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : S \longrightarrow \mathbb{R}$. Then f is said to be differentiable at $\bar{\mathbf{x}} \in \operatorname{int} S$ $\Leftrightarrow$ $\exists \nabla f(\bar{\mathbf{x}}) \in \mathbb{R}^n$ (gradient of f at $\bar{\mathbf{x}}$) and function $\alpha(\bar{\mathbf{x}}; \cdot - \bar{\mathbf{x}}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $\forall \mathbf{x} \in S : f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\| \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$ where $\lim_{\mathbf{x} \longrightarrow \bar{\mathbf{x}}} \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}) = 0$.

Compare the expression with the first order Taylor expansion, the first order Taylor approximation, and the reminder term $(\alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}))$.

Exercise 149.

Draw contour graphs in $\mathbb{R}^2$ illustrating the gradient concept.

Definition 150 (Differentiable function on set)**.**

f is said to be differentiable on the open set $S' \subset S$ if f is differentiable at $\mathbf{x}$, $\forall \bar{\mathbf{x}} \in S'$.

Remark 151.

If f is differentiable at $\bar{\mathbf{x}} \Rightarrow \exists!\ \nabla f(\bar{\mathbf{x}}) = (\frac{\partial f(\bar{\mathbf{x}})}{\partial x_i})^\top_{i=1,\ldots,n} = (f'_{x_i}(\bar{\mathbf{x}}))^\top_{i=1,\ldots,n}$ vector of partial derivatives of f at $\bar{\mathbf{x}}$.

Lemma 152 (Subgradient and gradient)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be convex, $f : S \longrightarrow \mathbb{R}$ is convex. Suppose that f is differentiable $\forall \bar{\mathbf{x}} \in \operatorname{int} S$. $\Rightarrow \forall \bar{\mathbf{x}} \in \operatorname{int} S : |\Xi| = |\{\boldsymbol{\xi} \mid \boldsymbol{\xi} \text{ is a subdifferential of } f \text{at} \bar{\mathbf{x}}\}| = 1$ and $\boldsymbol{\xi} = \nabla f(\bar{\mathbf{x}})$.

Proof: $\Xi \neq \emptyset$ by theorem, so $\boldsymbol{\xi} \in \Xi$ then $\forall \mathbf{d}\ \exists \lambda_0 > 0\ \forall \lambda < \lambda_0,\ \lambda > 0$: $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) \geq (\bar{\mathbf{x}}) + \lambda \boldsymbol{\xi}^\top \mathbf{d}$ and by differentiability $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} + \lambda \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d})$. Together $0 \geq \lambda(\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}}))^\top \mathbf{d} - \lambda \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d})$ (multiply by $\frac{1}{\lambda}$). Then $0 \geq (\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}}))^\top \mathbf{d} - \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d})$. With $\lambda \longrightarrow 0^+$ then $\alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}) \longrightarrow 0$, so $0 \geq (\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}}))^\top \mathbf{d}$ for any $\mathbf{d}$ even for $\mathbf{d} = (\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}}))$. So, $0 \geq \|\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}})\|^2 \Rightarrow \boldsymbol{\xi} = \nabla f(\bar{\mathbf{x}})$. □

Theorem 153 (Convexity and existence of gradient)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be an open convex set, $f : S \longrightarrow \mathbb{R}$ differentiable on S. Then: f is convex (strictly convex) on $S \Leftrightarrow \forall \bar{\mathbf{x}} \in S$, $\forall \mathbf{x} \in S$ (for strictly convex add $\mathbf{x} \neq \bar{\mathbf{x}}$): $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})$ (for strictly convex $>$).

Proof: Directly follows from the subgradient theorem 144. □

Remark 154 (Geometrical interpretations)**.**

1. For $\min\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \geq \min\{f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \mid \mathbf{x} \in S\}$ gives a lower bound on the optimum objective function value of the original program. This fact is useful for some optimization algorithms.
2. We may construct outer linearization of $S = \{\mathbf{x} \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}$ convex as $S' = \{\mathbf{x} \mid g_i(\bar{\mathbf{x}}) + \nabla g_i(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0, i = 1, \ldots, m\}$. So, $S \subset S'$.
3. For $f : \mathbb{R} \longrightarrow \mathbb{R}$ (differentiable) non-decreasing slope $\Leftrightarrow$ f convex.

Exercise 155.

Illustrate Remark 154 by figures in $\mathbb{R}^2$.

Theorem 156.

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex open set, $f : S \longrightarrow \mathbb{R}$ differentiable on S. Then f is convex (strictly convex) on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S$ ($\mathbf{x}_1 \neq \mathbf{x}_2$ for strictly convex f): $(\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1))^\top (\mathbf{x}_2 - \mathbf{x}_1) \geq 0$ (for strictly convex: $>$).

Proof: Assume f convex, $\mathbf{x}_1, \mathbf{x}_2 \in S$ then $f(\mathbf{x}_1) \geq f(\mathbf{x}_2) + \nabla f(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2)$ and $f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_1 - \mathbf{x}_2)$. The sum of inequalities leads to: $f(\mathbf{x}_1) + f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + f(\mathbf{x}_2) + (\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1))^\top (\mathbf{x}_1 - \mathbf{x}_2)$. Hence, $0 \leq (\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1))^\top (\mathbf{x}_2 - \mathbf{x}_1)$.
Conversely: $\mathbf{x}_1, \mathbf{x}_2 \in S$ then we get by the multivariate mean value theorem (f is differentiable, and so continuous) $\exists \mathbf{x} \in (\mathbf{x}_1, \mathbf{x}_2)$, $\nabla f(\mathbf{x})$, $\exists \lambda \in (0, 1) : \mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$ and $f(\mathbf{x}_2) - f(\mathbf{x}_1) = \nabla f(\mathbf{x})(\mathbf{x}_2 - \mathbf{x}_1)$. Assumption is $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_1))^\top (\mathbf{x} - \mathbf{x}_1) \geq 0 \Rightarrow (1 - \lambda)(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_1))^\top (\mathbf{x}_2 - \mathbf{x}_1) \geq 0$ because $\mathbf{x} = \lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2$. Therefore, $-(1 - \lambda) \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) + (1 - \lambda) \nabla f(\mathbf{x})^\top (\mathbf{x}_2 - \mathbf{x}_1) \geq 0$ and $-(1 - \lambda) \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) + (1 - \lambda)(f(\mathbf{x}_2) - f(\mathbf{x}_1)) \geq 0$. Then (multiplying by $\frac{1}{1-\lambda}$) $f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1)$ and so f is convex (similarly for strictly convex). □

All these results are useful for algorithms but they are not helpful for the function f convexity check.

1.5.3 Twice differentiable convex functions

Definition 157 (Twice differentiable function at point)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : S \longrightarrow \mathbb{R}$. Then f is said to be twice differentiable at $\bar{\mathbf{x}} \in \operatorname{int} S \Leftrightarrow \exists \nabla f(\bar{\mathbf{x}}) \in \mathbb{R}^n$ and $n \times n$ symmetric matrix $\mathbf{H}(\bar{\mathbf{x}})$ (Hessian) and $\alpha(\bar{\mathbf{x}}; \cdot - \bar{\mathbf{x}}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $\forall \mathbf{x} \in S : f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\top \mathbf{H}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$ where $\lim_{\mathbf{x} \longrightarrow \bar{\mathbf{x}}} \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}) = 0$.

Definition 158 (Twice differentiable function on set)**.**

f is said to be twice differentiable on the open set $S' \subset S$ if f is twice differentiable at $\mathbf{x}$, $\forall \bar{\mathbf{x}} \in S'$.

Remark 159.

If f is differentiable at $\bar{\mathbf{x}} \Rightarrow \exists!\ \mathbf{H}(\bar{\mathbf{x}}) = (\frac{\partial^2 f(\bar{\mathbf{x}})}{\partial x_i \partial x_j})^\top_{i,j=1,\ldots,n} = (f''_{x_i,x_j}(\bar{\mathbf{x}}))^\top_{i=1,\ldots,n}$ a matrix of the second partial derivatives of f at $\bar{\mathbf{x}}$. And above mentioned the second order Taylor expansion can be rewritten in the summation-index form: $f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \sum_{j=1}^n f'_{x_j}(\bar{\mathbf{x}})(x_j - \bar{x}_j) + \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (x_i - \bar{x}_i) f''_{x_i,x_j}(\bar{\mathbf{x}})(x_j - \bar{x}_j) + \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$.

Definition 160 (Definiteness of matrix)**.**

Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be an $n \times n$ symmetric matrix. Then $\mathbf{D}$ is said to be

positive definite (PD) $\Leftrightarrow \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0} : \mathbf{x}^\top \mathbf{D} \mathbf{x} > 0$,
positive semidefinite (PSD) $\Leftrightarrow \forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{D} \mathbf{x} \geq 0$,
negative definite (ND) $\Leftrightarrow \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0} : \mathbf{x}^\top \mathbf{D} \mathbf{x} < 0$, and
negative semidefinite (NSD) $\Leftrightarrow \forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{D} \mathbf{x} \leq 0$.

Remark 161.

From linear algebra is known that $\mathbf{x}^\top \mathbf{D} \mathbf{x} = \sum_{i=1}^n \lambda_i y_i^2$ where all eigenvalues λ_i solve $|\mathbf{D} - \lambda \mathbf{I}| = 0$ equation and $\mathbf{x} = \mathbf{B}\mathbf{y}$ where columns (eigenvectors $\mathbf{b}_j \neq \mathbf{0}$) of $\mathbf{B}$ and $\forall j : \mathbf{b}_j$ solve $(\mathbf{D} - \lambda_j \mathbf{I})\mathbf{b}_j = \mathbf{0}$.

Theorem 162 (Definiteness and eigenvalues)**.**

$\mathbf{D}$ is positive definite $\Leftrightarrow \forall j : \lambda_j > 0$,
$\mathbf{D}$ is positive semidefinite $\Leftrightarrow \forall j : \lambda_j \geq 0$,
$\mathbf{D}$ is negative definite $\Leftrightarrow \forall j : \lambda_j < 0$, and
$\mathbf{D}$ is negative semidefinite $\Leftrightarrow \forall j : \lambda_j \leq 0$.

Exercise 163.

Use software (e.g., Matlab) to compute eigenvalues and check definiteness of some matrices.

Theorem 164 (Convexity and definiteness)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ open convex, $f : S \longrightarrow \mathbb{R}$ twice differentiable on S. Then:

f is convex on $S \Leftrightarrow$ its $\mathbf{H}(\bar{\mathbf{x}})$ is PSD $\forall \bar{\mathbf{x}} \in S$.
f is strictly convex on $S \Leftarrow$ its $\mathbf{H}(\bar{\mathbf{x}})$ is PD $\forall \bar{\mathbf{x}} \in S$.
f is strictly convex on $S \Rightarrow$ its $\mathbf{H}(\bar{\mathbf{x}})$ is PSD $\forall \bar{\mathbf{x}} \in S$.
f is quadratic and strictly convex on $S \Rightarrow$ its $\mathbf{H}(\bar{\mathbf{x}})$ is PD $\forall \bar{\mathbf{x}} \in S$.

Proof: $\Rightarrow$: f convex, $\bar{\mathbf{x}} \in S$. We need to show $\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{x} \geq 0$. Since S is open $\Rightarrow \exists \lambda_0 \; \forall \mathbf{x} \in \mathbb{R}^n$ $\forall \lambda \in [-\lambda_0, \lambda_0], \lambda \neq 0 : \bar{\mathbf{x}} + \lambda \mathbf{x} \in S$. We already know: $f(\bar{\mathbf{x}} + \lambda \mathbf{x}) = f(\bar{\mathbf{x}}) + \lambda \nabla f(\mathbf{x})\mathbf{x} + \frac{1}{2}\lambda^2 \mathbf{x}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{x} + \lambda^2 \|\mathbf{x}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$. By subtraction, we have $0 \leq \frac{1}{2}\lambda^2 \mathbf{x}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{x} + \lambda^2 \|\mathbf{x}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$. We multiply the inequality by $\frac{1}{\lambda^2}$ and with $\lambda \longrightarrow 0$ we have PSD $\mathbf{H}(\bar{\mathbf{x}})$.
$\Leftarrow$: $\forall \bar{\mathbf{x}} \in S : \mathbf{H}(\bar{\mathbf{x}})$ is PSD. Then for $\mathbf{x}$ and $\bar{\mathbf{x}}$ by the mean value theorem, we have: $f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\top \mathbf{H}(\hat{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})$ where $\hat{\mathbf{x}} = \lambda \bar{\mathbf{x}} + (1 - \lambda)\mathbf{x}$ for some $\lambda \in (0, 1) \Rightarrow \hat{\mathbf{x}} \in S$. By assumption $\mathbf{H}(\hat{\mathbf{x}})$ is PSD, so $(\mathbf{x} - \bar{\mathbf{x}})^\top \mathbf{H}(\hat{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) > 0$ and $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})$ then f is convex. □

Remark 165.

Theorem is useful for a quadratic functions because $\mathbf{H}(\bar{\mathbf{x}})$ is constant in this case and definiteness may be easily checked (cf. use of Matlab computations of eigenvalues with symbolic computations in the general case).

Theorem 166 (Univariate function convexity)**.**

$S \subset \mathbb{R}$, $S \neq \emptyset$ open convex, $f : S \longrightarrow \mathbb{R}$ infinitely differentiable. Then f is strictly convex on $S \Leftrightarrow \forall \bar{x} \in S \; \exists k \in \mathbb{N} : f^{(2k)}(\bar{x}) > 0$ while $f^{(j)}(\bar{x}) = 0$, $\forall j \in \{2, \ldots, 2k - 1\}$.

Proof: See calculus, mathematical analysis. □

Theorem 167.

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\bar{\mathbf{x}} \in \mathbb{R}^n$, $\mathbf{d} \in \mathbb{R}^n, \mathbf{d} \neq \mathbf{0}$ and define $F_{\bar{\mathbf{x}},\mathbf{d}}(\lambda) := f(\bar{\mathbf{x}} + \lambda \mathbf{d})$. Then f is (strictly) convex $\Leftrightarrow F_{\bar{\mathbf{x}},\mathbf{d}}$ is (strictly) convex $\forall \bar{\mathbf{x}} \in \mathbb{R}^n, \mathbf{d} \in \mathbb{R}^n, \mathbf{d} \neq \mathbf{0}$.

Proof is omitted. Theorem may be used (examining f via univariate cross sections $F_{\bar{\mathbf{x}},\mathbf{d}}$) in algorithms.

Lemma 168.

$\mathbf{H} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is PSD $\Leftrightarrow a \geq 0,\ c \geq 0,\ ac - b^2 \geq 0.$

Proof: $\Rightarrow$: By Definition 160 of PSD: $\mathbf{d}^\top \mathbf{H}\mathbf{d} \geq 0,\ \forall \mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$ so $a{d_1}^2 + 2bd_1d_2 + c{d_2}^2 \geq 0 \Rightarrow a \geq 0, c \geq 0$ otherwise set either $d_1 = 0$ or $d_2 = 0$ and the remaining d_1 or d_2 large and < 0. Also $a = 0 \Rightarrow b = 0$ (or $|d_1|$ large and < 0) $\Rightarrow ac - b^2 = 0$. With $a > 0$ completing the squares, we get: $\mathbf{d}^\top \mathbf{H}\mathbf{d} = a(d_1^2 + \frac{2bd_1d_2}{a} + \frac{b^2}{a^2}d_2^2) + d_2^2(c - \frac{b^2}{a}) = a(d_1 + \frac{b}{a}d_2)^2 + d_2^2(\frac{ac-b^2}{a}) \Rightarrow ac - b^2 \geq 0$. Otherwise with $d_1 = -\frac{b}{a}d_2 \Rightarrow < 0$.
$\Leftarrow$: $a \geq 0,\ c \geq 0,\ ac - b^2 \geq 0$ assuming $a = 0 \Rightarrow b = 0 \Rightarrow \mathbf{d}^\top \mathbf{H}\mathbf{d} = cd_2^2 \geq 0$. If $a > 0$ then $\mathbf{d}^\top \mathbf{H}\mathbf{d} = a(d_1 + \frac{b}{a}d_2)^2 + d_2^2(\frac{ac-b^2}{a}) \geq 0$. PD similarly. □

Exercise 169.

From Lemma 168, we get for ND $a < 0,\ c < 0,\ ac - b^2 \geq 0$. Why? Hint: Matrix $\mathbf{D}$ is NSD (ND) $\Leftrightarrow -\mathbf{D}$ is PSD (PD).

Exercise 170.

How to deal with non-symmetric $\mathbf{H}$ for $\mathbf{d}^\top \mathbf{H}\mathbf{d}$? Hint: Consider that we may write $\mathbf{d}^\top \mathbf{H}\mathbf{d} = \mathbf{d}^\top \mathbf{H}^\top \mathbf{d} = \frac{1}{2}\mathbf{d}^\top(\mathbf{H} + \mathbf{H}^\top)\mathbf{d}$.

Theorem 171 (Checking definiteness).

Let $\mathbf{H} = (h_{ij})$ be a symmetric $m \times n$ matrix:

(a) $\exists i \in \{1, \ldots, n\} : h_{ii} \leq 0 \Rightarrow \mathbf{H}$ is not PD. If $h_{ii} < 0 \Rightarrow$ is not PSD.
(b) $\exists i \in \{1, \ldots, n\} : h_{ii} = 0 \Rightarrow \forall j \in \{1, \ldots, n\} : h_{ij} = h_{ji} = 0$ or $\mathbf{H}$ is not PSD.
(c) If $n = 1$: $\mathbf{H}$ is PSD (PD) $\Leftrightarrow h_{11} \geq 0$ ($h_{11} > 0$) otherwise $n \geq 2$ let $\begin{pmatrix} h_{11} & \mathbf{q}^\top \\ \mathbf{q} & \mathbf{G} \end{pmatrix}$)
(by (b) $h_{11} = 0 \Rightarrow \mathbf{q} = \mathbf{0}$ to keep PSD). So for $h_{11} > 0$ as a pivot, perform elementary Gauss-Jordan operation with the first row obtaining: $\begin{pmatrix} h_{11} & \mathbf{q}^\top \\ \mathbf{0} & \mathbf{G}_{\text{new}} \end{pmatrix}$)
where $\mathbf{G}_{\text{new}}$ is a symmetric $(n-1) \times (n-1)$ and $\mathbf{H}$ is PSD $\Leftrightarrow \mathbf{G}_{\text{new}}$ is PSD. (Moreover for $h_{11} > 0$: $\mathbf{H}$ is PD $\Leftrightarrow \mathbf{G}_{\text{new}}$ is PD).

Proof: (a) $\mathbf{d}^\top \mathbf{H}\mathbf{d} = d_i^2 h_{ii}$ for $\mathbf{d} = (0, \ldots, 0, d_i, 0, \ldots, 0)^\top$ and proved.
(b) $h_{ii} = 0,\ h_{ij} \neq 0 \Rightarrow$ choose $d_k = 0$ for $k \neq i,\ k \neq j \Rightarrow \mathbf{d}^\top \mathbf{H}\mathbf{d} = 2h_{ij}d_id_j + d_{jj}^2 h_j \Rightarrow d_i$ large and then contradiction.
(c) $n = 1$: trivial, $n \geq 2 : \mathbf{d}^\top = (d_1, \boldsymbol{\delta}^\top)$. If $h_{11} = 0$ then $\mathbf{q} = \mathbf{0}$ $\mathbf{G} = \mathbf{G}_{\text{new}}$ and $\mathbf{d}^\top \mathbf{H}\mathbf{d} = \boldsymbol{\delta}^\top \mathbf{H}\boldsymbol{\delta}$ and theorem is valid. If $h_{11} > 0$ then $\mathbf{d}^\top \mathbf{H}\mathbf{d} = d_1^2 h_{11} + 2d_1(\mathbf{q}^\top \boldsymbol{\delta}) + \boldsymbol{\delta}^\top \mathbf{G}\boldsymbol{\delta}$. By Gauss-Jordan, we have: $\mathbf{G}_{\text{new}} = \mathbf{G} - \frac{1}{h_{11}} \begin{pmatrix} q_1\mathbf{q}^\top \\ \vdots \\ q_n\mathbf{q}^\top \end{pmatrix} = \mathbf{G} - \frac{1}{h_{11}}\mathbf{q}\mathbf{q}^\top$. It is a symmetric matrix because it is a difference of symmetric

matrices. Therefore, $\mathbf{d}^\top\mathbf{H}\mathbf{d} = d_1^2 h_{11} + 2d_1(\mathbf{q}^\top\boldsymbol{\delta}) + \boldsymbol{\delta}^\top(\mathbf{G}_{\text{new}} + \frac{1}{h_{11}}\mathbf{q}\mathbf{q}^\top)\boldsymbol{\delta} = \boldsymbol{\delta}^\top\mathbf{G}_{\text{new}}\boldsymbol{\delta} + h_{11}(d_1 + \frac{\mathbf{q}^\top\boldsymbol{\delta}}{h_{11}})^2$ and theorem is valid because of the second term nonnegativity in the sum. Similarly PD. □

Algorithm 172 (Gauss-Jordan)**.**

1. Scan diagonal elements of $\mathbf{H}$ (cf. (a) and (b) of Theorem 171).
2. Perform Gauss-Jordan reduction to obtain $\mathbf{G}_{\text{new}}$ by (c).
3. Take $\mathbf{G}_{\text{new}}$ instead of $\mathbf{H}$ and GOTO 1. till $n = 1$ (if $n = 1$ then (c)).

Remark 173.

Each step (1)–(3) takes m arithmetic operations and comparisons where $m \leq Kn^2$ for certain K and there is at most n steps and then the total number of operations $M \leq Kn^3$ (is of $O(n^3)$). The algorithm has a polynomial complexity.

Exercise 174.

Compute own numerical examples.

Corollary 175.

$\mathbf{H}$ is an $n \times n$ symmetric matrix: $\mathbf{H}$ is PD $\Leftrightarrow$ $\mathbf{H}$ is PSD and nonsingular ($\exists\mathbf{H}^{-1}$, $r(\mathbf{H}) = n$).

Proof: $\Rightarrow$: $\mathbf{H}$ is PD $\Rightarrow$ Gauss-Jordan reduction gives the upper triangular matrix with > 0 in diagonal $\Rightarrow r(\mathbf{H}) = n$.
$\Leftarrow$: $\mathbf{H}$ is PSD and nonsingular, so $r(\mathbf{H}) = n \Rightarrow$ Gauss-Jordan gives the upper triangular matrix, diagonal ≥ 0 but > 0 because of the full rank. □

Remark 176 (Convexity of composed functions)**.**

1. $f_1, \ldots, f_k : \mathbb{R}^n \longrightarrow \mathbb{R}$ are convex. Then (a) $\forall \alpha_j > 0, j = 1, \ldots, k : \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x})$ is convex. (b) $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ is convex.
2. Let $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a concave function, $S = \{\mathbf{x} \mid g(\mathbf{x}) > 0\}$ is convex. We define $f : S \longrightarrow \mathbb{R}$ as $f(\mathbf{x}) = \frac{1}{g(\mathbf{x})}$ is convex.
3. Let $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ be a linear function where $\mathbf{h} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \longrightarrow \mathbb{R}$ convex then $f(\mathbf{x}) = g(\mathbf{h}(\mathbf{x}))$ is a convex function.
4. Let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a non-decreasing (univariate) convex and let $\mathbf{h} : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function then $f(\mathbf{x}) = g(\mathbf{h}(\mathbf{x}))$ is convex.

Let $h_1 : \mathbb{R} \longrightarrow \mathbb{R}$ and $h_2 : \mathbb{R} \longrightarrow \mathbb{R}$ be univariate functions. Then, the composed function $h_2(h_1) : \mathbb{R} \longrightarrow \mathbb{R}$ inherits the convexity (concavity) from functions h_1 and h_2 by Figure 1.

Exercise 177.

Explain the conclusions of Remark 176. Try to prove them.

1.5.4 Minima and maxima of convex functions

Exercise 178.

Review previously introduced concepts: feasible solution; infimum, minimum; optimal solution (minimum, maximum); local or global optimal solution; alternative optimal solutions; strict local (global) optimal solution; isolated local (global) optimal solution. Draw own figures in $\mathbb{R}^2$ to illustrate those various concepts.

Exercise 179.

For further piece-wise defined function $f(x)$ on $[0, 5)$ identify all optimal solutions. $f(x) = 1 - x$ for $x \in [0, 1]$, $f(x) = x - 1$ for $x \in (1, 2]$, $f(x) = 1$ for $x \in (2, 3]$, $f(x) = 4 - x$ for $x \in (3, 4)$, and $f(x) = x - 3$ for $x \in [4, 5)$.

Theorem 180 (Convexity and global minimum)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be convex and $f : S \longrightarrow \mathbb{R}$ convex on S. Let $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ then $\bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. If $\bar{\mathbf{x}}$ is a strict local minimum ($\bar{\mathbf{x}} \in \operatorname{argstrictlocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$) or f is strictly convex on S then $\bar{\mathbf{x}}$ is a unique global minimum of f on S (and an isolated local minimum).

Proof: Let $\bar{\mathbf{x}} \in \text{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. Then $\exists \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) : \forall \mathbf{x} \in S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) : f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$. By contradiction suppose $\exists \hat{\mathbf{x}} \in S : f(\hat{\mathbf{x}}) < f(\bar{\mathbf{x}})$. By convexity of f: $\forall \lambda \in (0,1) : f(\lambda\hat{\mathbf{x}}) + (1-\lambda)\bar{\mathbf{x}} \leq \lambda f(\hat{\mathbf{x}}) + (1-\lambda)f(\bar{\mathbf{x}}) < \lambda f(\bar{\mathbf{x}}) + (1-\lambda)f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}})$. But for $\lambda > 0$ small $\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}} \in S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$, and so $f(\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}) \geq f(\bar{\mathbf{x}})$ and we obtain contradiction. So, $\bar{\mathbf{x}} \in \text{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.
Suppose $\bar{\mathbf{x}}$ is a strict local minimum. By contradiction assume that is not unique: $\exists \hat{\mathbf{x}} \in S : f(\bar{\mathbf{x}}) = f(\hat{\mathbf{x}})$ and $\forall \mathbf{x}_\lambda \in S : \mathbf{x}_\lambda = \lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}$ by convexity of f and S: $\mathbf{x}_\lambda \in S$ and $f(\mathbf{x}_\lambda) = f(\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}) \leq \lambda f(\hat{\mathbf{x}}) + (1-\lambda)f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}})$. For $\lambda \longrightarrow 0^+ : \mathbf{x}_\lambda \in S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ and contradiction, so $\bar{\mathbf{x}}$ is a unique global and isolated minimum.
Finally, assume f is strictly convex and $\bar{\mathbf{x}} \in \text{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. Because strict convexity of f on $S \Rightarrow$ convexity of f on S then $\bar{\mathbf{x}} \in \text{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ as above. By contradiction uniqueness $\exists \mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}} : f(\mathbf{x}) = f(\bar{\mathbf{x}})$, so $f(\frac{1}{2}\mathbf{x} + \frac{1}{2}\bar{\mathbf{x}}) <$ (strict convexity) $\frac{1}{2}f(\mathbf{x}) + \frac{1}{2}f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}})$ that is the contradiction with $\bar{\mathbf{x}} \in \text{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. So, it is unique and isolated. □

Theorem 181 (Convexity and subgradient)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then $\bar{\mathbf{x}} \in \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \exists \boldsymbol{\xi}$ subgradient of f at $\bar{\mathbf{x}}$ such that $\forall \mathbf{x} \in S : \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$.

Proof: $\Leftarrow$: $\forall \mathbf{x} \in S : \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$. For $\boldsymbol{\xi}$ subgradient $\forall \mathbf{x} \in S : f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}}) \leq f(\mathbf{x})$, so $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$, and hence $\bar{\mathbf{x}}$ is a minimum.
$\Rightarrow$ (Main idea): Let $\bar{\mathbf{x}}$ be a minimum, define: $M_1 = \{(\mathbf{x}^\top - \bar{\mathbf{x}}^\top, y)^\top \mid \mathbf{x} \in \mathbb{R}^n, y > f(\mathbf{x}) - f(\bar{\mathbf{x}})\}$ is a convex set and $M_2 = \{(\mathbf{x}^\top - \bar{\mathbf{x}}^\top, y)^\top \mid \mathbf{x} \in S, y \leq 0\}$ is also a convex set. By definitions of M_1 and M_2, we have $M_1 \cap M_2 = \emptyset$. Otherwise, $\exists (\mathbf{x}^\top - \bar{\mathbf{x}}^\top, y)^\top : \mathbf{x} \in S : 0 \geq y > f(\mathbf{x}) - f(\bar{\mathbf{x}})$ (contradiction with minimization requirement). By theorem about separation of convex sets, we may find a separating (even supporting) hyperplane for M_1 and M_2. The normal vector of this hyperplane specifies a subgradient $\boldsymbol{\xi}$ with the required property. □

Corollary 182.

If assumptions of Theorem 181 are valid and in addition S is an open set then: $\bar{\mathbf{x}} \in \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \exists \boldsymbol{\xi} = \mathbf{0}$.

Proof: Because S is open then $\exists \lambda_0 > 0$ then $\forall \lambda < \lambda_0 : \bar{\mathbf{x}} - \lambda\boldsymbol{\xi} \in S \Rightarrow \boldsymbol{\xi}^\top(\bar{\mathbf{x}} - \lambda\boldsymbol{\xi} - \bar{\mathbf{x}}) \geq 0 \Rightarrow \boldsymbol{\xi} = \mathbf{0}$. □

Corollary 183.

If assumptions of Theorem 181 are valid and f is differentiable then: $\bar{\mathbf{x}} \in \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \forall \mathbf{x} : \nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$.

Corollary 184.

If assumptions of Theorem 181 are valid and f is differentiable and S is open then: $\bar{\mathbf{x}} \in \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Remark 185.

1. Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be open convex and $f : S \longrightarrow \mathbb{R}$ convex differentiable. Then: $\bar{\mathbf{x}} \in \text{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ (Necessary and sufficient condition).
2. If $\mathbf{x}$ is non-optimal then $\nabla f(\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}}) < 0$ and we may improve it using $\mathbf{d} = \mathbf{x}-\bar{\mathbf{x}}$ direction and step $\lambda_0 = \min_\lambda\{f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \mid \lambda \geq 0, \bar{\mathbf{x}} + \lambda\mathbf{d} \in S\}$ (This idea will introduce feasible direction methods).
3. Let $\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})$ be a feasible solution, so $\mathbf{d} = \mathbf{x} - \bar{\mathbf{x}}$ is a feasible direction and $-\nabla f(\bar{\mathbf{x}})$ is a recessive direction. If the angle between $-\nabla f(\bar{\mathbf{x}})$ and $\mathbf{d}$ is less or equal then the right angle then we may improve $\bar{\mathbf{x}}$ to get a smaller objective function value.
4. If f differentiable: $f'(\bar{\mathbf{x}}; \mathbf{d}) = \nabla f(\bar{\mathbf{x}})^\top \mathbf{d}$ and $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}})$. If $\nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) < 0$ then $\bar{\mathbf{x}}$ may be improved.
5. When $\bar{\mathbf{x}} \in \text{int}\, S : f'(\bar{\mathbf{x}}; \mathbf{d}) \geq 0 \Rightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Exercise 186.

Illustrate Remark 185 by figures using contour graphs.

Theorem 187 (Alternative optimal solutions)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : S \longrightarrow \mathbb{R}$ convex twice differentiable, denote $S^* = \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ then: $S^* = \{\mathbf{x} \in S \mid \nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) \leq 0, \nabla f(\mathbf{x}) = \nabla f(\bar{\mathbf{x}})\}$.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty pages 104-105). □

Corollary 188 (Properties of S^*)**.**

1. $S^* = \{\mathbf{x} \in S \mid \nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) = 0, \nabla f(\mathbf{x}) = \nabla f(\bar{\mathbf{x}})\}$.
2. In addition, suppose f is a quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x} + \mathbf{c}^\top\mathbf{x}$ and S is a polyhedral set. Then S^* is polyhedral and $S^* = \{\mathbf{x} \in S : \mathbf{c}^\top(\mathbf{x} - \bar{\mathbf{x}}) = 0, \mathbf{H}(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{0}\}$.

Theorem 189 (Convexity and maximum)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convex, $\bar{\mathbf{x}} \in \text{arglocmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow$ $\forall \mathbf{x} \in S$, $\forall \boldsymbol{\xi}$ subgradients of f at $\bar{\mathbf{x}}$: $\boldsymbol{\xi}^\top(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$.

Proof: Let $\bar{\mathbf{x}} \in \operatorname{arglocmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ then $\exists \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) \cap S$ and $\mathbf{x} \in \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) : f(\mathbf{x}) \leq f(\bar{\mathbf{x}})$ then for $\lambda > 0$ small enough $f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) \leq f(\bar{\mathbf{x}})$ and $\boldsymbol{\xi}$ subgradient $f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) - f(\bar{\mathbf{x}}) \geq \lambda \boldsymbol{\xi}^\top (\mathbf{x} - \bar{\mathbf{x}})$ from convexity. So, $0 \geq \lambda \boldsymbol{\xi}^\top (\mathbf{x} - \bar{\mathbf{x}})$ for $\lambda > 0$ and theorem is proven. □

Corollary 190.

In addition if f is differentiable and $\bar{\mathbf{x}} \in \operatorname{arglocmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \forall \mathbf{x} \in S : \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0$.

Theorem 191.

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ compact, polyhedral and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convex. Then: $\exists \bar{\mathbf{x}} \in \operatorname{arglocmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \neq \emptyset$ and $\bar{\mathbf{x}}$ is EP of S.

Proof: Use Weierstrass' Theorem, convexity properties, and extreme point properties. □
Think about the reversed umbrella!

1.5.5 Generalizations of convex functions

Definition 192 (Quasiconvex function).

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set. $f : S \longrightarrow \mathbb{R}$ is said to be quasiconvex function on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S \; \forall \lambda \in (0,1) : f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. f is called quasiconcave $\Leftrightarrow -f$ is quasiconvex.

Univariate cross section of quasiconvex function is either monotone or unimodal. Function that is quasiconvex and quasiconcave is quasimonotone. Strictly quasiconvex function is also alternatively called a semistrictly, explicitly, functionally quasiconvex function.

Exercise 193.

Illustrate concepts above by figures in $\mathbb{R}^2$.

Theorem 194 (Quasiconvexity characterization).

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex. Then $f : S \longrightarrow \mathbb{R}$ is quasiconvex $\Leftrightarrow \forall \alpha \in \mathbb{R} : S_\alpha$ is a convex set.

Proof: $\Rightarrow$: $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha \Rightarrow \mathbf{x}_1, \mathbf{x}_2 \in S \wedge \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \leq \alpha$. Choose $\lambda \in (0,1)$, so $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ $\Rightarrow \mathbf{x} \in S$ by convexity of S. By quasiconvexity of f: $f(\mathbf{x}) \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \leq \alpha \Rightarrow \mathbf{x} \in S_\alpha$.
$\Leftarrow$: S_α, $\mathbf{x}_1, \mathbf{x}_2 \in S$, $\lambda \in (0,1)$: $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. So, $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha$ for $\alpha = \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$ by assumption $\mathbf{x} \in S_\alpha$ (convex combination), and so $f(\mathbf{x}) \leq \alpha = \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. □

Remark 195.

f quasiconcave $\Leftrightarrow S_\alpha^u = \{\mathbf{x} \mid f(\mathbf{x}) \geq \alpha\}$ is convex. f quasimonotone $\Leftrightarrow S_\alpha^= = \{\mathbf{x} \mid f(\mathbf{x}) = \alpha\}$ is convex.

Theorem 196 (Quasiconvexity and maximization)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ polyhedral and compact, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ quasiconvex and continuous: $\exists \bar{\mathbf{x}} \in \operatorname{argmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \wedge \bar{\mathbf{x}}$ is EP of S.

Proof: By contradiction: $\mathbf{x}' \in \operatorname{argmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ is not an EP. Because of compactness $\mathbf{x}'$ belongs to a convex hull of extreme points of S and $f(\mathbf{x}')$ is greater than the objective function value at any EP $\mathbf{x}_j$. We take $\max f(\mathbf{x}_j) = f(\mathbf{x}_k) = \alpha$. We consider $S_\alpha \Rightarrow \forall j : \mathbf{x}_j \in S_\alpha \Rightarrow \mathbf{x}' \in S_\alpha \Rightarrow f(\mathbf{x}') \leq \alpha$ and contradiction. □

Theorem 197 (Differentiable quasiconvex function characterization)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be an open convex set, $f : S \longrightarrow \mathbb{R}$ differentiable on S. Then f is quasiconvex $\Leftrightarrow$ one of the following equivalent statements holds:

1. $\forall \mathbf{x}_1, \mathbf{x}_2 \in S : f(\mathbf{x}_1) \leq f(\mathbf{x}_2) \Rightarrow \nabla f(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2) \leq 0$.
2. $\forall \mathbf{x}_1, \mathbf{x}_2 \in S : \nabla f(\mathbf{x}_2)^\top (\mathbf{x}_1 - \mathbf{x}_2) > 0 \Rightarrow f(\mathbf{x}_1) > f(\mathbf{x}_2)$.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty). □

Definition 198 (Strictly quasiconvex function)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set. $f : S \longrightarrow \mathbb{R}$ is said to be strictly quasiconvex function on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S$ with $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$: $\forall \lambda \in (0,1) : f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. f is called strictly quasiconcave $\Leftrightarrow -f$ is strictly quasiconvex.

Exercise 199.

Explain that f convex $\Rightarrow f$ strictly quasiconvex $\not\Rightarrow$ quasiconvex by figures. Hint: "Flat spots" may occur in extremizing points (only!). Think about $y = 0$ for $x \in \mathbb{R} \setminus \{0\}$ and $y = 1$ for $x = 0$.

Theorem 200 (Local minima for strictly quasiconvex)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be strictly quasiconvex, and $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty page 111, by contradiction). □

Lemma 201 (Semicontinuous and strictly quasiconvex function)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : S \longrightarrow \mathbb{R}$ strictly quasiconvex and lower semicontinuous $\Rightarrow$ f is quasiconvex.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty page 112). □

Definition 202 (Strongly quasiconvex function)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set. $f : S \longrightarrow \mathbb{R}$ is said to be strongly quasiconvex function on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S$ with $\mathbf{x}_1 \neq \mathbf{x}_2$: $\forall \lambda \in (0,1) : f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. f is called strongly quasiconcave $\Leftrightarrow -f$ is strongly quasiconvex.

Alternatively, some authors replace 'strongly quasiconvex' with 'strictly quasiconvex'. Note that strong quasiconvexity enforces unimodality:

Theorem 203 (Unimodality)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ strongly quasiconvex on S. Then $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \{\bar{\mathbf{x}}\} = \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty page 113). □

Remark 204.

However, differentiable and strongly quasiconvex function does not have the property $\nabla f(\bar{\mathbf{x}}) = \mathbf{0} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$. Therefore, we introduce:

Definition 205 (Pseudoconvex function)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex, $f : S \longrightarrow \mathbb{R}$ differentiable on S. f is said pseudoconvex on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S : \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) \geq 0 \Rightarrow f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$. f is pseudoconcave $\Leftrightarrow -f$ is pseudoconvex.

Or equivalently $f(\mathbf{x}_2) < f(x_1) \Rightarrow \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) < 0$. f is said strictly pseudoconvex on $S \Leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in S : \nabla f(\mathbf{x}_1)^\top (\mathbf{x}_2 - \mathbf{x}_1) \geq 0 \Rightarrow f(\mathbf{x}_2) > f(\mathbf{x}_1)$.

Exercise 206.

Draw a figure explaining the concept of pseudoconvex function f on S.

Theorem 207 (Relations among different convex functions)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$ be an open and convex set and $f : S \longrightarrow \mathbb{R}$:

1. f is strictly convex on $S \Rightarrow f$ is convex on S.
2. f is strictly convex on $S \Rightarrow f$ is strongly quasiconvex on S.
3. f is convex on $S \Rightarrow f$ is (strictly) quasiconvex on S.
4. f is differentiable and (strictly) convex on $S \Rightarrow f$ is (strictly) pseudoconvex on S.
5. f is strictly pseudoconvex on $S \Rightarrow f$ is pseudoconvex on S.
6. f is strictly pseudoconvex on $S \Rightarrow f$ is strongly quasiconvex on S.
7. f is pseudoconvex on $S \Rightarrow f$ is strictly quasiconvex on S.
8. f is strongly quasiconvex on $S \Rightarrow f$ is (strictly) quasiconvex on S.
9. f is strictly quasiconvex and lower semicontinuous on $S \Rightarrow f$ is quasiconvex.

Proof: Not required (see, e.g., Bazaraa-Sherali-Shetty page 114–115). □

Definition 208 (Convex function at point)**.**

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ be a convex set. $f : S \longrightarrow \mathbb{R}$. Then:

1. f is said to be a convex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S \; \forall \lambda \in (0,1) : f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}) \leq \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x})$.
2. f is said to be a strictly convex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$, $\forall \lambda \in (0,1) : f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}) < \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x})$.
3. f is said to be a quasiconvex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$, $\forall \lambda \in (0,1) : f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}) \leq \max\{f(\bar{\mathbf{x}}), f(\mathbf{x})\}$.
4. f is said to be a strictly quasiconvex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$, $f(\mathbf{x}) \neq f(\bar{\mathbf{x}})$, $\forall \lambda \in (0,1) : f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}) < \max\{f(\bar{\mathbf{x}}), f(\mathbf{x})\}$.
5. f is said to be a strongly quasiconvex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$, $\forall \lambda \in (0,1) : f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}) < \max\{f(\bar{\mathbf{x}}), f(\mathbf{x})\}$.
6. f is said to be a pseudoconvex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$: $\nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \geq 0 \Rightarrow f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$.
7. f is said to be a strictly pseudoconvex function at $\bar{\mathbf{x}} \in S \Leftrightarrow \forall \mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$: $\nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) > 0 \Rightarrow f(\mathbf{x}) > f(\bar{\mathbf{x}})$.

Remark 209 (Properties).

$S \subset \mathbb{R}^n$, $S \neq \emptyset$ convex and $f : S \longrightarrow \mathbb{R}$. Then:

1. f convex and differentiable at $\bar{\mathbf{x}} \Rightarrow f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})$.
2. f convex and twice differentiable at $\bar{\mathbf{x}} \Rightarrow \mathbf{H}(\bar{\mathbf{x}})$ is PSD.
3. f convex at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.
4. f convex and differentiable at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}} \in \operatorname{int} S : \bar{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$. Otherwise, $\bar{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow \forall \mathbf{x} \in S : \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \geq 0$.
5. f convex and differentiable at $\bar{\mathbf{x}}$. $\bar{\mathbf{x}} \in \operatorname{arglocmax}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \forall \mathbf{x} \in S : \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0$.
6. f strictly quasiconvex at $\bar{\mathbf{x}}$: $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.
7. f strongly quasiconvex at $\bar{\mathbf{x}}$: $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ and $\bar{\mathbf{x}}$ is unique.
8. f pseudoconvex at $\bar{\mathbf{x}}$: $\nabla f(\bar{\mathbf{x}}) = \mathbf{0} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.
9. f strictly pseudoconvex at $\bar{\mathbf{x}}$: $\nabla f(\bar{\mathbf{x}}) = \mathbf{0} \Rightarrow \bar{\mathbf{x}} \in \operatorname{argglobmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ and $\bar{\mathbf{x}}$ is unique.

1.6 Theory of unconstrained optimization

Exercise 210.

Review concepts: local minimum, global minimum.

Theorem 211 (Existence of a descent direction).

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be differentiable at $\bar{\mathbf{x}}$. $\exists \mathbf{d} \in \mathbb{R}^n : \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} < 0 \Rightarrow \exists \delta > 0$ $\forall \lambda \in (0, \delta) : f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})$. So $\mathbf{d}$ is a descent direction of f at $\bar{\mathbf{x}}$.

Proof: By differentiability: $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} + \lambda \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}, \lambda \mathbf{d})$ where $\alpha(\cdot, \cdot) \longrightarrow 0$ for $\lambda \longrightarrow 0$. We may rewrite it as: $\frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} = \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} + \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}, \lambda \mathbf{d})$. Since $\nabla f(\bar{\mathbf{x}})^\top \mathbf{d} < 0$ by assumption and $\alpha(\cdot, \cdot) \longrightarrow 0$ for $\lambda \longrightarrow 0 \Rightarrow \exists \delta > 0 \; \forall \lambda \in (0, \delta) : \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} + \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}, \lambda \mathbf{d}) < 0$, and hence $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})$. □

Corollary 212 (First-order necessary optimality condition).

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be differentiable at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}} \in \operatorname{arglocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\} \Rightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Proof: By contradiction: $\nabla f(\bar{\mathbf{x}}) \neq \mathbf{0}$. Set $\mathbf{d} = -\nabla f(\bar{\mathbf{x}})$. Then $\nabla f(\bar{\mathbf{x}})^\top \mathbf{d} = -\|\nabla f(\bar{\mathbf{x}})\|^2 < 0$ and by Theorem 211 we have found a descent direction of f at $\bar{\mathbf{x}}$. □

Theorem 213 (Second-order necessary optimality condition)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be twice differentiable at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}} \in \text{arglocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\} \Rightarrow \mathbf{H}(\bar{\mathbf{x}})$ is PSD.

Proof: Let $\mathbf{d}$ be an arbitrarily chosen direction. By differentiability of f: $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top \lambda\mathbf{d} + \frac{1}{2}\lambda^2 \mathbf{d}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{d} + \lambda^2\|\mathbf{d}\|^2\alpha(\bar{\mathbf{x}}, \lambda\mathbf{d})$ where for $\lambda \longrightarrow 0 \Rightarrow \alpha(\cdot,\cdot) \longrightarrow 0$. By Corollary 212: $\bar{\mathbf{x}} \in \text{arglocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\} \Rightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$. Then $\frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda^2} = \frac{1}{2}\mathbf{d}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{d} + \|\mathbf{d}\|^2\alpha(\bar{\mathbf{x}}, \lambda\mathbf{d})$. By Corollary 212: Left-hand-side ≥ 0 ($\bar{\mathbf{x}}$ is minimum). It is true $\forall\lambda > 0$ then also right-hand-side ≥ 0. Then taking a limit of both sides we have $\alpha(\cdot,\cdot) \longrightarrow 0$, and so (because differentiability implies continuity) $\frac{1}{2}\mathbf{d}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{d} \geq 0$ so $\mathbf{H}(\bar{\mathbf{x}})$ is PSD. □

Theorem 214 (Sufficient differentiability condition)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be twice differentiable at $\bar{\mathbf{x}}$. If $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ and $\mathbf{H}(\bar{\mathbf{x}})$ is PD $\Rightarrow$ $\bar{\mathbf{x}} \in \text{argstrictlocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$ (and the isolated minimum).

Proof: By twice differentiability: $\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{f}(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\top \mathbf{H}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\|^2\alpha(\bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}})$ where $\mathbf{x} \longrightarrow \bar{\mathbf{x}} \Rightarrow \alpha \longrightarrow 0$. By contradiction: $\bar{\mathbf{x}} \notin \text{argstrictlocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$: $\exists\{\mathbf{x}_k\} : \mathbf{x}_k \longrightarrow \bar{\mathbf{x}}$, $\forall k : f(\mathbf{x}_k) \leq f(\bar{\mathbf{x}})$, $\mathbf{x}_k \neq \bar{\mathbf{x}}$. So, we know $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, we denote $\mathbf{d}_k = \frac{\mathbf{x}_k - \bar{\mathbf{x}}}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}$ and we obtain: $\forall k : \frac{1}{2}\mathbf{d}_k^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{d}_k + \alpha(\bar{\mathbf{x}}, \mathbf{x}_k - \bar{\mathbf{x}}) \leq 0$ and $\|\mathbf{d}_k\| = 1$ (a bounded sequence). Therefore, $\exists\mathcal{K} : \{\mathbf{d}_k\}_{k\in\mathcal{K}}$ converges to $\mathbf{d}$ such that $\|\mathbf{d}\| = 1$. For $k \longrightarrow \infty$, $\alpha \longrightarrow 0$ so $\mathbf{d}^\top \mathbf{H}(\bar{\mathbf{x}})\mathbf{d} \leq 0$ and contradiction with PD. □

Remark 215.

Because $\mathbf{H}(\bar{\mathbf{x}})$ is PD $\Rightarrow$ $\mathbf{H}(\mathbf{x})$ is PD on some $\mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ and f is strictly convex there, so $\bar{\mathbf{x}}$ is an isolated minimum, too.

Theorem 216 (Sufficient convexity condition)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be pseudoconvex at $\bar{\mathbf{x}}$. $\bar{\mathbf{x}} \in \text{argglobmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\} \Leftrightarrow \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Proof: $\Rightarrow$: $\bar{\mathbf{x}}$ is a global minimum, it is also a local minimum and by corollary $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.
$\Leftarrow$: $\nabla f(\bar{\mathbf{x}}) = \mathbf{0} \Rightarrow \forall \mathbf{x} \in \mathbb{R}^n : \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) = 0$. By pseudoconvexity: Left-hand-side $\geq 0 \Rightarrow \forall \mathbf{x}$ $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$, so $\bar{\mathbf{x}}$ is a global minimum. □

Example 217 (Useful for regression analysis)**.**

We have n measurements $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of independent variables $\mathbf{x} = (x_1, \ldots, x_k)^\top$ (so $\mathbf{x}_j = (x_{j1}, \ldots, x_{jk})^\top$) and related $y_1, \ldots, y_n$ (denoting $\mathbf{y} = (y_1, \ldots, y_n)^\top$) of dependent variable y. We search for $y = \varphi(\mathbf{x}; \boldsymbol{\beta})$ (φ known, $\boldsymbol{\beta}$ unknown). We know that measurements were obtained from $\eta = \varphi(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon(\mathbf{x})$ where $\varepsilon(\mathbf{x}_i)$ are *i.i.d.* (identically independent distributed) and $\varepsilon(\mathbf{x}_i) \sim N(0, \sigma^2)$. We assume known σ^2 and we look for point estimate $\mathbf{b}$ of unknown $\boldsymbol{\beta}$ using maximization of the likelihood function (h is a density function of η):

$$\max_{\mathbf{b}} \prod_{j=1}^{n} h(y_j, \mathbf{b}, \sigma) \quad = \quad \max_{\mathbf{b}} \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \; e^{-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \varphi(\mathbf{x}_j; \mathbf{b}))^2}.$$

It is equivalent with LSQ criterion:

$$\min_{\mathbf{b}} S^*(\mathbf{b}) = \min_{\mathbf{b}} \sum_{j=1}^{n} (y_j - \varphi(\mathbf{x}_j; \mathbf{b}))^2.$$

We put $\nabla S^*(\mathbf{b}) = \mathbf{0}$. Then:

$$\begin{aligned} \frac{\partial}{\partial b_i} s^*(\mathbf{b}) &= 0 && i = 1, \ldots, m \\ \sum_{j=1}^{n} 2 \cdot (y_j - \varphi(\mathbf{x}_j, \mathbf{b})) \cdot (-\frac{\partial}{\partial b_i} \varphi(\mathbf{x}_j, \mathbf{b})) &= 0 && i = 1, \ldots, m. \end{aligned}$$

Generally, we have a system of m nonlinear equations of m unknown variables b_i. The global minimum $\mathbf{b}_0$ is among roots of this system (solvable by further numerical methods). We denote $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^\top$ and assume the case of a linear regression function $\varphi(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} = \sum_{i=1}^{m} \beta_i f_i(\mathbf{x})$. So, we solve:

$$\min_{\mathbf{b}} \sum_{j=1}^{n} (y_j - \mathbf{f}(\mathbf{x}_j)^\top \mathbf{b})^2 \quad = \quad \min_{\mathbf{b}} \sum_{j=1}^{n} (y_j - \sum_{l=1}^{m} f_l(\mathbf{x}_j) \cdot b_l)^2,$$

and then by Theorems:

$$\begin{aligned} \sum_{j=1}^{n} [2 \cdot (y_j - \sum_{l=1}^{m} f_l(\mathbf{x}_j) \cdot b_l) \cdot (-\frac{\partial}{\partial b_i} \sum_{l=1}^{m} f_l(\mathbf{x}_j) \cdot b_l)] &= 0 && i = 1, \ldots, m \\ \sum_{j=1}^{n} [(y_j - \sum_{l=1}^{m} f_l(\mathbf{x}_j) \cdot b_l) \cdot f_i(\mathbf{x}_j)] &= 0 && i = 1, \ldots, m \\ \sum_{j=1}^{n} \sum_{l=1}^{m} f_l(\mathbf{x}_j) \cdot f_i(\mathbf{x}_j) \cdot b_l - \sum_{j=1}^{n} y_j \cdot f_i(\mathbf{x}_j) &= 0 && i = 1, \ldots, m \\ \sum_{l=1}^{m} (\sum_{j=1}^{n} f_i(\mathbf{x}_j) \cdot f_l(\mathbf{x}_j)) \cdot b_l &= \sum_{j=1}^{n} f_i(\mathbf{x}_j) \cdot y_j && i = 1, \ldots, m. \end{aligned}$$

Because of the linear regression function we have a system of linear equations. To simplify a notation, we introduce matrix $\mathbf{F}$ of type $m \times n$ and define it as $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n)) = (f_i(\mathbf{x}_j))_{ij}$. Then we have a system of normal equations in the form $\mathbf{F}\mathbf{F}^\top \mathbf{b} = \mathbf{F}\mathbf{y}$ that is solvable by the Gauss elimination. In the case of the unique solution $\mathbf{b}_0$ ($\mathbf{F}$ of full rank, $r(\mathbf{F}) = m$), we may derive the explicit form using a matrix inverse:

$$\mathbf{b}_0 = (\mathbf{F}\mathbf{F}^\top)^{-1} \mathbf{F}\mathbf{y}.$$

To check whether $\mathbf{b}_0$ represents a minimum we have to check convexity of $S^*(\mathbf{b})$ at $\mathbf{b}_0$. For linear regression function, Hessian of $S^*(\mathbf{b})$ is $\mathbf{H}(\mathbf{b}) = 2\mathbf{F}\mathbf{F}^\top$. By linear algebra theory, we have $\mathbf{F}\mathbf{F}^\top = \mathbf{F}\mathbf{I}\mathbf{F}^\top$. Because $\mathbf{I}$ is PD and $\mathbf{F}$ is of the full rank then $\mathbf{H}(\mathbf{b})$ is also PD, and hence, $S^*(\mathbf{b})$ is strictly convex and $\mathbf{b}_0$ is a global minimum.

Theorem 218 (One dimensional case)**.**

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be an infinitely differentiable function ($f \in \mathcal{C}^{\infty}(\mathbb{R})$). Then $\bar{x} \in \mathbb{R}$ satisfies $\bar{x} \in \operatorname{arglocmin}\{f(x) \mid x \in \mathbb{R}\} \Leftrightarrow$ Either $\forall j = 1, 2, \ldots\ f^{(j)}(\bar{x}) = 0$ or $\exists n = 2k$ (even) $f^{(n)(\bar{x})} > 0$ and $\forall j = 1, \ldots, n-1 : f^{(j)}(\bar{x}) = 0$.

Proof: It is not required. □

There is $f^{(n)}$ n'th order derivative of f. The result essentially asserts that for discussed case $\bar{x} \in \operatorname{arglocmin}\{f(x) \mid x \in \mathbb{R}\} \Leftrightarrow f$ is locally convex about $\bar{x}$. For comparison with multidimensional case, read the discussion by Bazaraa, Sherali, and Shetty, pages 135–137.

1.7 Algorithms – general viewpoint

Remark 219 (Solution sets)**.**

For NLP $? \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$, we need an algorithm generating a sequence of points converging to a global optimal solution. Usually, we obtain less. We often obtain only the limit point belonging to some solution set Ω. There are the following typical sets Ω:

$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \operatorname{arglocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}\}$.
$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in S \wedge f(\bar{\mathbf{x}}) \leq b\}$ where b is an acceptable level of the objective function value.
$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in S \wedge f(\bar{\mathbf{x}}) < \mathrm{LB} + \varepsilon\}$where LB is a lower bound and ε is a tolerance.
$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in S \wedge f(\bar{\mathbf{x}}) - f_{\min} < \varepsilon\}$ where only optimal objective function value is available.
$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}}$ satisfies KKT conditions$\}$.
$\Omega = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}}$ satisfies FJ conditions$\}$.

Definition 220 (Algorithmic map)**.**

Let $X \subset \mathbb{R}^n$, $\mathcal{A} : X \longrightarrow 2^X$ point-set mapping, and $\mathbf{x}_1 \in X$. We call $\mathcal{A}$ an algorithmic map (mapping).

Remark 221 (Convergence of algorithm)**.**

By $\mathcal{A}$, we may generate an aforementioned sequence $\mathbf{x}_1, \ldots, x_k, \ldots$ approximating solutions of NLP satisfying $\mathbf{x}_{k+1} \in \mathcal{A}(\mathbf{x}_k)$ for $k = 1, 2, \ldots$. Convergence of NLP algorithms will be related to properties of $\mathcal{A}$ with respect to Ω.

Definition 222 (Convergence of algorithmic map)**.**

Algorithmic map $\mathcal{A} : X \longrightarrow 2^X$ is said to converge over $Y \subset X$ if $\forall \mathbf{x}_1 \in Y$ the limit of any convergent subsequence of the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ generated by the algorithm belongs to the solution set Ω.

Exercise 223.

Consider $\operatorname{argmin}\{x^2 \mid x \geq 1\}$. Draw a figure explaining that the solution set $\Omega = \{1\}$.

1. Analyze graphically $\mathcal{A}(x) = \{\frac{x+1}{2}\}$.
2. Analyze graphically $\mathcal{A}(x) = [1, \frac{1}{2}(x+1)]$ for $x \geq 1$ and $\mathcal{A}(x) = [\frac{1}{2}(x+1), 1]$ for $x < 1$.
3. Analyze graphically $\mathcal{A}(x) = [\frac{3}{2} + \frac{1}{4}x, 1 + \frac{1}{2}x]$ for $x \geq 2$ and $\mathcal{A}(x) = \{\frac{1}{2}(x+1)\}$ for $x < 2$. Hint: If $x_1 \geq 2$ then $x_k \not\longrightarrow x \in \Omega$. Why?

Exercise 224.

Check graphically that all three previous cases may satisfy: $x_k \in S \Rightarrow x_{k+1} \in S$ (keeps feasibility), $x_k \in S, x_k \notin \Omega \Rightarrow f(x_k) > f(x_{k+1})$ (decreasing objective), and $x_k \in \Omega \Rightarrow x_{k+1} \in \Omega$ (keeps optimality).

Definition 225 (Closed algorithmic map)**.**

Let $X \subset \mathbb{R}^p$, $Y \subset \mathbb{R}^q$, both nonempty and closed. $\mathcal{A} : X \longrightarrow 2^Y$. Then $\mathcal{A}$ is said to be closed at $\mathbf{x} \in X \Leftrightarrow (\mathbf{x}_k \in X, \mathbf{x}_k \longrightarrow \mathbf{x}, \mathbf{y}_k \in \mathcal{A}(\mathbf{x}_k), \mathbf{y}_k \longrightarrow \mathbf{y} \Rightarrow \mathbf{y} \in \mathcal{A}(\mathbf{x}))$. The map $\mathcal{A}$ is said to be closed on $Z \subset X$ if it is closed $\forall \mathbf{x} \in Z$.

Theorem 226 (About convergence)**.**

Let $X \subset \mathbb{R}^n$, $X \neq \emptyset$, closed and Ω be a solution set, $\mathcal{A} : X \longrightarrow 2^X$, $\mathbf{x}_1 \in X$, $\{\mathbf{x}_k\}$ is defined $\forall k \in \mathbb{N}$ and for $\mathbf{x}_k \notin \Omega : \mathbf{x}_{k+1} := \mathcal{A}(\mathbf{x}_k)$. If $\mathbf{x}_k \in \Omega$ then stop (or $\mathbf{x}_{k+1} := \mathbf{x}_k$). We suppose that $\{\mathbf{x}_k\}$ is contained in a compact subset of X and that there exists $\alpha : X \longrightarrow \mathbb{R}$ (e.g., $f(\mathbf{x})$ or $\|\nabla f(\mathbf{x})\|$) a descent function continuous such that $\alpha(\mathbf{y}) < \alpha(\mathbf{x})$ if $\mathbf{x} \notin \Omega$ and $\mathbf{y} \in \mathcal{A}(\mathbf{x})$ if $\mathcal{A}$ is closed over $X \setminus \Omega$ then either the algorithm stops in a finite number of steps with point in Ω or it generates the infinite sequence $\{\mathbf{x}_k\}$ such that:

1. every convergent subsequence of $\{\mathbf{x}_k\}$ has a limit in Ω.
2. $\alpha(\mathbf{x}_k) \longrightarrow \alpha(\mathbf{x})$ for some $\mathbf{x} \in \Omega$.

Proof: Let $\mathbf{x}_k \in \Omega$ then stop. Otherwise, we have $\mathbf{x}_1, \mathbf{x}_2, \ldots$, i.e. $\{\mathbf{x}_k\}$. Let $\{\mathbf{x}_k\}_{k \in \mathcal{K}}$ be a convergent subsequence with limit $\mathbf{x} \in X$. By continuity of α: $k \in \mathcal{K}$ $\alpha(\mathbf{x}_k) \longrightarrow \alpha(\mathbf{x})$ and $\forall \varepsilon > 0$ $\exists K \in \mathcal{K} : \forall k \in \mathcal{K}$,

$k \geq K$ $\alpha(\mathbf{x}_k) - \alpha(\mathbf{x}) < \varepsilon$. Because $\alpha(\mathbf{x}_K) - \alpha(\mathbf{x}) < \varepsilon$ then for $k > K$ (because of descent function α) $\alpha(\mathbf{x}_k) < \alpha(\mathbf{x}_K)$ and $\alpha(\mathbf{x}_k) - \alpha(\mathbf{x}) = \alpha(\mathbf{x}_k) - \alpha(\mathbf{x}_K) + \alpha(\mathbf{x}_K) - \alpha(\mathbf{x}) < 0 + \varepsilon$, so $\lim_{k \longrightarrow \infty} \alpha(\mathbf{x}_k) = \alpha(\mathbf{x})$. By contradiction, we show that $\mathbf{x} \in \Omega$. Assume $\mathbf{x} \notin \Omega$: $\{\mathbf{x}_{k+1}\}_{\mathcal{K}}$ is contained in a compact subset of X and has a convergent subsequence $\{\mathbf{x}_{k+1}\}_{\bar{\mathcal{K}}}$ with limit $\bar{\mathbf{x}} \in X$. Then $\lim_{k \longrightarrow \infty} \alpha(\mathbf{x}_k) = \alpha(\mathbf{x}) \Rightarrow \alpha(\mathbf{x}) = \alpha(\bar{\mathbf{x}})$. Because $\mathcal{A}$ is closed at $\mathbf{x} \Rightarrow k \in \bar{\mathcal{K}} \Rightarrow \mathbf{x}_k \longrightarrow \mathbf{x}$, $\mathbf{x}_{k+1} \in \mathcal{A}(\mathbf{x}_k)$, $\mathbf{x}_{k+1} \longrightarrow \bar{\mathbf{x}}$, $\bar{\mathbf{x}} \in \mathcal{A}(\mathbf{x})$ then $\alpha(\bar{\mathbf{x}}) < \alpha(\mathbf{x})$ and contradiction. □

Corollary 227.

If Theorem's assumptions are valid and $\Omega = \{\bar{\mathbf{x}}\} \Rightarrow \mathbf{x}_k \longrightarrow \bar{\mathbf{x}}$.

Definition 228 (Composite map)**.**

Let $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^p$, $Z \subset \mathbb{R}^q$ nonempty closed sets. Let $\mathcal{B} : X \longrightarrow 2^Y$ and $\mathcal{C} : Y \longrightarrow 2^Z$. Then, the composite map $\mathcal{A} = \mathcal{C}\mathcal{B} : X \longrightarrow 2^Z$ ($\mathcal{C}$ is applied after $\mathcal{B}$) where $\mathcal{A}(\mathbf{x}) = \cup\{\mathcal{C}(\mathbf{y}) \mid \mathbf{y} \in \mathcal{B}(\mathbf{x})\}$.

The definition of composite map is useful for optimization algorithms because they are also composed.

Theorem 229 (Closed composite map)**.**

Let $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^p$, $Z \subset \mathbb{R}^q$ nonempty closed sets. Let $\mathcal{B} : X \longrightarrow 2^Y$, $\mathcal{C} : Y \longrightarrow 2^Z$, and $\mathcal{A} = \mathcal{C}\mathcal{B}$. We suppose that $\mathcal{B}$ is closed at $\mathbf{x}$ and $\mathcal{C}$ is closed on $\mathcal{B}(\mathbf{x})$. Furthermore, we suppose that $\mathbf{x}_k \longrightarrow \mathbf{x}$ and $\mathbf{y}_k \in \mathcal{B}(\mathbf{x}_k) \Rightarrow \exists$ convergent subsequence $\{\mathbf{y}_k\}$. Then $\mathcal{A}$ is closed at $\mathbf{x}$.

Proof: $\mathbf{x}_k \longrightarrow \mathbf{x}$, $\mathbf{z}_k \in \mathcal{A}(\mathbf{x}_k)$ and $\mathbf{z}_k \longrightarrow \mathbf{z}$. We need to show $\mathbf{z} \in \mathcal{A}(\mathbf{x})$ to prove Theorem 229. By definition: $\forall k : \exists \mathbf{y}_k \in \mathcal{B}(\mathbf{x}_k)$ such that $\mathbf{z}_k \in \mathcal{C}(\mathbf{y}_k)$. By assumption $\exists\{\mathbf{y}_k\}_{\mathcal{K}}$ convergent subsequence with limit $\mathbf{y}$. $\mathcal{B}$ is closed $\Rightarrow \mathbf{y} \in \mathcal{B}(\mathbf{x})$, $\mathcal{C}$ is closed on $\mathcal{B}(\mathbf{x})$ and at point, e.g., $\mathbf{y}$. Therefore, $\mathbf{z} \in \mathcal{C}(\mathbf{y})$, $\mathbf{z} \in \mathcal{C}\mathcal{B}(\mathbf{x}) = \mathcal{A}(\mathbf{x})$. □

Corollary 230.

Let $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^p$, $Z \subset \mathbb{R}^q$ nonempty closed sets. Let $\mathcal{B} : X \longrightarrow 2^Y$ and $\mathcal{C} : Y \longrightarrow 2^Z$, $\mathcal{B}$ closed at $\mathbf{x}$, $\mathcal{C}$ closed on $\mathcal{B}(\mathbf{x})$, and Y is a compact set. Then $\mathcal{A} = \mathcal{C}\mathcal{B}$ is closed at $\mathbf{x}$.

Corollary 231.

Let $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^p$, $Z \subset \mathbb{R}^q$ nonempty closed sets. Let $\mathcal{B} : X \longrightarrow Y$ (point-to-point function, i.e. point-to-one-point-set mapping), $\mathcal{C} : Y \longrightarrow 2^Z$, $\mathcal{B}$ continuous at $\mathbf{x}$, and $\mathcal{C}$ closed on $\mathcal{B}(\mathbf{x})$. Then $\mathcal{A} = \mathcal{C}\mathcal{B}$ is closed at $\mathbf{x}$.

Theorem 232 (More about convergence)**.**

$X \subset \mathbb{R}^n$, $X \neq \emptyset$ is a closed set. $\Omega \subset X$, $\Omega \neq \emptyset$ is a solution set. $\alpha : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous mapping. $\mathcal{C} : X \longrightarrow 2^X$ mapping satisfies $\forall \mathbf{x} \in X$ and $\mathbf{y} \in \mathcal{C}(\mathbf{x})$: $\alpha(\mathbf{y}) \leq \alpha(\mathbf{x})$. $\mathcal{B} : X \longrightarrow 2^X$ is a closed map over $X \setminus \Omega$ and $\forall \mathbf{x} \notin \Omega$, $\forall \mathbf{y} \in \mathcal{B}(\mathbf{x})$: $\alpha(\mathbf{y}) < \alpha(\mathbf{x})$. Assume that $\mathcal{A} = \mathcal{C}\mathcal{B}$ generates for $\mathbf{x}_1$ the sequence $\{\mathbf{x}_k\}$ as follows: Either $\mathbf{x}_k \in \Omega$ (and stop) or $\mathbf{x}_k \notin \Omega : \mathbf{x}_{k+1} \in \mathcal{A}(\mathbf{x}_k)$. If $\Lambda = \{\mathbf{x} \mid \alpha(\mathbf{x}) \leq \alpha(\mathbf{x}_1)\}$ is a compact set then: Either the algorithm stops in a finite number of steps with a point in Ω or all accumulation points of $\{\mathbf{x}_k\}$ belong to Ω.

Proof: $\mathbf{x}_k \in \Omega \Rightarrow$ stop by assumptions. Let $\{\mathbf{x}_k\}$ be generated by $\mathcal{A}$ and $\{\mathbf{x}_k\}_\mathcal{K}$ be a convergent subsequence with limit $\mathbf{x}$. Thus $\alpha(\mathbf{x}_k) \longrightarrow \alpha(\mathbf{x})$ for $k \in \mathcal{K}$ by continuity of α. By monotonicity of α as before $\lim_{k \longrightarrow \infty} \alpha(\mathbf{x}_k) = \alpha(\mathbf{x})$. We need to show $\mathbf{x} \in \Omega$. by contradiction assume $\mathbf{x} \notin \Omega$ and consider the sequence $\{\mathbf{x}_{k+1}\}_\mathcal{K}$. By definition $\mathcal{A}$, we have $\mathbf{x}_{k+1} \in \mathcal{C}(\mathbf{y}_k)$ for $\mathbf{y}_k \in \mathcal{B}(\mathbf{x}_k)$ and $\mathbf{y}_k, \mathbf{x}_{k+1} \in \Lambda$. By compactness of Λ $\exists$ another subsequence $\{\mathbf{x}_{k+1}\}_{\mathcal{K}'}$, $\{\mathbf{y}_k\}_{\mathcal{K}'}$: $\mathbf{x}_{k+1} \longrightarrow \mathbf{x}'$, $\mathbf{y}_k \longrightarrow \mathbf{y}$. $\mathcal{B}$ closed at $\mathbf{x} \Rightarrow$ $\mathbf{y} \in \mathcal{B}(\mathbf{x})$ and $\alpha(\mathbf{y}) < \alpha(\mathbf{x})$ (as $\mathbf{x}$ is not a limit). $\mathbf{x}_{k+1} \in \mathcal{C}(\mathbf{y}_k) \Rightarrow \alpha(\mathbf{x}_{k+1}) \leq \alpha(\mathbf{y}_k)$ and $\alpha(\mathbf{x}') \leq \alpha(\mathbf{y})$ that is a contradiction. □

Theorem 233 (Minimization along independent directions)**.**

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function. $? \in \operatorname{arglocmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$ is a NLP (unconstrained). $\mathcal{A}$ is defined as follows: $\mathbf{y} \in \mathcal{A}(\mathbf{x})$ is obtained by n times (sequential) minimization of f along directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$ (starting from $\mathbf{x}$) where $\|\mathbf{d}_j\| = 1, j = 1, \ldots, n$. Denote matrix $\mathbf{D}(\mathbf{x}) = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$ (for given $\mathbf{x}$), assume that determinant $\det(\mathbf{D}(\mathbf{x})) \geq \varepsilon > 0$ for some ε and $\forall \mathbf{x} \in \mathbb{R}^n$. Suppose that minimum of f along any line in $\mathbb{R}^n$ is unique. For $\mathbf{x}_1$, $\{\mathbf{x}_k\}$ is generated as follows: If $\nabla f(\mathbf{x}_k) = \mathbf{0}$ then stop in $\mathbf{x}_k$. Otherwise, $\mathbf{x}_{k+1} \in \mathcal{A}(\mathbf{x}_k)$ (and $k := k + 1$) and repeat. If $\{\mathbf{x}_k\}$ is contained in a compact set then each accumulation point $\mathbf{x}$ of $\{\mathbf{x}_k\}$ satisfies $\nabla f(\mathbf{x}_k) = \mathbf{0}$.

Remark 234.

1. The discussed general attempt (and also last Theorem 233) to convergence is suitable for many optimization algorithms.
2. No additional assumption about closedness or continuity is needed for Theorem 233. However, all search directions should be linearly independent (e.g., it is valid for coordinate or orthogonal directions).

Exercise 235.

Why uniqueness in Theorem 233 is required? Hint: Draw a figure for $z = x_2(1 - x_1)$ and use that $\nabla f(\mathbf{x}) \neq \mathbf{0} \Rightarrow f(\mathbf{x}_2) < f(\mathbf{x}_1)$.

Remark 236 (Termination of algorithm)**.**

In the case of convergence to Ω in a limit sense, we need stopping rules (assume that $\varepsilon > 0$ and $N \in \mathbb{N}$):

1. $\|\mathbf{x}_{k+N} - \mathbf{x}_k\| < \varepsilon$ (small movement).
2. $\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|}{\|\mathbf{x}_k\|} < \varepsilon$ (small relative distance change).
3. $\alpha(\mathbf{x}_k) - \alpha(\mathbf{x}_{k+N}) < \varepsilon$ (descent function does not improve).
4. $\frac{\alpha(\mathbf{x}_k) - \alpha(\mathbf{x}_{k+1})}{|\alpha(\mathbf{x}_k)|} < \varepsilon$ (small relative improvement).
5. $\alpha(\mathbf{x}_k) - \alpha(\bar{\mathbf{x}}) < \varepsilon$ (e.g., $\Omega = \{\bar{\mathbf{x}} \mid \nabla f(\bar{\mathbf{x}}) = \mathbf{0}\} \wedge \alpha(\bar{\mathbf{x}}) = \|\nabla f(\bar{\mathbf{x}})\|$).

Remark 237 (Comparison of algorithms)**.**

They can be compared using various features:

Generality is given by required assumptions, solvable problems.

Reliability (robustness) is defined by achievable accuracy for solvable problems.

Precision is specified by quality of points after number of iterations.

Sensitivity relates to setup of algorithm parameters and problem's input data.

Preparations required identify preprocessing steps, e.g., computations of gradient, Hessian.

Computational effort depends on computing time, number of iterations, elementary operations required, functional evaluations, etc.

Remark 238 (Convergence)**.**

It is important to make a difference between theoretical and practical convergence (e.g., average rate based on statistical observations). In theory, we have $\mathbf{x}_k \longrightarrow \bar{\mathbf{x}}$ $(\forall k : \mathbf{x}_k \neq \bar{\mathbf{x}})$. The order of convergence is defined by

$$\sup\{p \mid p \geq 0 \wedge \exists \beta : \limsup_{k \longrightarrow \infty} \frac{\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}\|}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|^p} = \beta < \infty\}.$$

Alternatively the fraction can be given as $\frac{|\alpha(\mathbf{x}_{k+1}) - \alpha(\bar{\mathbf{x}})|}{|\alpha(\mathbf{x}_k) - \alpha(\bar{\mathbf{x}})|^p}$. If $p = 1$ and $\beta < 1$ then we call convergence linear (geometric -why?). If $p > 1$ (or $p = 1$ and $\beta = 0$) then convergence is said superlinear. If $p = 2$ and $\beta < \infty$ then convergence is called quadratic (second order). Notice that with greater p theoretical convergence is faster.

1.8 Line search methods revisited

Remark 239 (Motivation)**.**

Efficient line search algorithms are very important (Why? Give examples!) because they are repeatedly used by multidimensional algorithms solving NLP problems. Many multidimensional NLP algorithms use the following iteration formula (Explain by own example!):

$$\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k,$$

where $\mathbf{x}_k$ is a point known from the previous iteration, $\mathbf{d}_k$ is a direction specified by multidimensional algorithm, and λ_k is a step size. For given $\mathbf{x}_k$ and $\mathbf{d}_k$, step size λ_k (also denoted as $\lambda_{\min}$) is computed as follows:

$$\lambda_k \in \operatorname*{argmin}_{\lambda \in \mathcal{L}} \theta(\lambda) = \operatorname{argmin}\{f(\mathbf{x}_k + \lambda \mathbf{d}_k) \mid \lambda \in \mathcal{L}\},$$

which is a univariate minimization problem with unknown λ (Does it change with iterations?). We will introduce line search algorithms for this problem.

Remark 240 (Interval of uncertainty)**.**

A feasible set $\mathcal{L}$ may be defined in different ways (e.g., $\mathcal{L} = \mathbb{R}$, $\mathcal{L} = [0; \infty)$,$\mathcal{L} = [a; b]$). We will search for $\lambda_{\min} \in \mathcal{L} = [a; b]$. It is so called *interval of uncertainty* since the location of $\lambda_{\min}$ over $[a; b]$ is not known. The goal is to reduce the interval of uncertainty by the line search excluding parts of it.

Remark 241 (Reduction for strictly quasiconvex function)**.**

If we assume that $\lambda_{\min} \in \mathcal{L}$ and θ is strictly quasiconvex (i.e. any local minimum is also global minimum) the important reduction idea is given by the following theorem (Explain by figure!):

Theorem 242.

Let $\theta : \mathbb{R} \longrightarrow \mathbb{R}$ be strictly quasiconvex over the interval $[a; b]$. Let $\lambda, \mu \in [a; b]$ and $\lambda < \mu$. If $\theta(\lambda) > \theta(\mu)$ then $\forall z \in [a; \lambda) : \theta(z) \geq \theta(\mu)$. If $\theta(\lambda) \leq \theta(\mu)$ then $\forall z \in (\mu; b] : \theta(z) \geq \theta(\lambda)$.

Proof: Suppose that $\theta(\lambda) > \theta(\mu)$ and let $z \in [a; \lambda)$. By contradiction, suppose that $\theta(z) < \theta(\mu)$. Since λ can be written as a convex combination of z and μ then by strict convexity definition, we obtain $\theta(\lambda) < \max\{\theta(z), \theta(\mu)\} = \theta(\mu)$. This is a contradiction with $\theta(\lambda) > \theta(\mu)$. So $\theta(z) \geq \theta(\mu)$. The rest of theorem is proven in the similar way. □

Therefore, in the case of strict quasiconvexity we have the following simple rules. (Explain for the forthcoming algorithm!):

- If $\theta(\lambda) > \theta(\mu)$ then the new interval of uncertainty is $[\lambda; b]$

- If $\theta(\lambda) \leq \theta(\mu)$ then the new interval of uncertainty is $[a; \mu]$.

Remark 243 (Unimodality).

A reader may also be interested in the relation between different types of generalized convexity introduced earlier and different concepts of unimodality often used in the books on numerical algorithms. So, we give a brief overview (Draw figures!):

Definition 244.

Let $\theta : \mathcal{L} \longrightarrow \mathbb{R}$ be a univariate function and $\mathcal{L} \subset \mathbb{R}$ is an interval on $\mathbb{R}$. θ is *unimodal* on $\mathcal{L}$ if there exists $\lambda_{\min} \in \operatorname{argmin}\{\theta(\lambda)|\lambda \in \mathcal{L}\}$ and θ is nondecreasing on the interval $\{\lambda \in \mathcal{L}|\lambda \geq \lambda_{\min}\}$ and nonincreasing on the interval $\{\lambda \in \mathcal{L}|\lambda \leq \lambda_{\min}\}$.

Assuming that θ attains a minimum on $\mathcal{L}$ (Why? Figure! avoiding cases as, e.g., $\theta(\lambda) = \lambda^3$) it can be shown that θ is quasiconvex if and only if it is unimodal on $\mathcal{L}$.

Definition 245.

A function $\theta : \mathbb{R} \longrightarrow \mathbb{R}$ to be minimized is said to be *strictly unimodal* over $[a; b]$ if there exists $\lambda_{\min} \in \operatorname{argmin}\{\theta(\lambda)|\lambda \in [a; b]\}$ and $\forall \lambda_1, \lambda_2 \in [a; b]$ such that $\theta(\lambda_1) \neq \theta(\lambda_{\min})$ and $\theta(\lambda_2) \neq \theta(\lambda_{\min})$ and $\lambda_1 < \lambda_2$ we have $(\lambda_2 \leq \lambda_{\min}) \Rightarrow (\theta(\lambda_1) > \theta(\lambda_2))$ and $(\lambda_1 \geq \lambda_{\min}) \Rightarrow (\theta(\lambda_1) < \theta(\lambda_2))$.

It may be shown that if θ is strictly unimodal and continuous over$[a; b]$ then θ is strictly quasiconvex over $[a; b]$. Conversely if θ is strictly quasiconvex over $[a; b]$ and has a minimum in this interval (Why? Figure!) then it is strictly unimodal over this interval.

Definition 246.

A function $\theta : \mathbb{R} \longrightarrow \mathbb{R}$ to be minimized is said to be *strongly unimodal* over $[a; b]$ if there exists $\lambda_{\min} \in \operatorname{argmin}\{\theta(\lambda)|\lambda \in [a; b]\}$ and $\forall \lambda_1, \lambda_2 \in [a; b]$ such that $\lambda_1 < \lambda_2$ we have $(\lambda_2 \leq \lambda_{\min}) \Rightarrow (\theta(\lambda_1) > \theta(\lambda_2))$ and $(\lambda_1 \geq \lambda_{\min}) \Rightarrow (\theta(\lambda_1) < \theta(\lambda_2))$.

It can be shown that if θ is strongly unimodal over $[a; b]$ then θ is strongly quasiconvex over $[a; b]$. Conversely, if θ is strongly quasiconvex over $[a; b]$ and has a minimum in this interval (Why? Figure!), then it is strongly unimodal over the interval.

Definition 247.

Let $f : S \longrightarrow \mathbb{R}$, where f is lower-semicontinuous and $S \subset \mathbb{R}^n$ is a convex set. f is *strongly unimodal* on S if $\forall \mathbf{x}_1, \mathbf{x}_2 \in S$ such that the function $F(\lambda) = f(\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)), \lambda \in [0; 1]$, attains a minimum $\lambda_{\min} > 0$ we have $\forall \lambda \in (0, \lambda_{\min}) : F(0) > F(\lambda) > F(\lambda_{\min})$.

It can be shown that f is strongly quasiconvex on S if and only if f is strongly unimodal on S.

Classification of line search algorithms:

- LS Without using derivatives: uniform search, dichotomous search, golden section, Fibonacci.
- LS Using derivatives: bisection search, Newton's method.
- Approximation-based LS: quadratic (cubic) approximations.
- Inexact LS: Armijo's rule.

1.8.1 Line search without using derivatives

Uniform search. We know n grid points in advance at which functional evaluations ($\theta(\lambda)$) are to be made. The big disadvantage is that the information obtained during computations is not further used (Why?). So, for k'th step and step size δ we assign $\lambda_k := a + k\delta,\ k\ = 1, \ldots, n$ and $\theta(\lambda_k)$. We search for r such that $\theta(\lambda_r) = \min_k \theta(\lambda_k)$. The uncertainty interval is then reduced to $[\lambda_r - \delta; \lambda_r + \delta]$. This technique may be used for the initial bracketing of $\lambda_{\min}$.

Dichotomous search. It already uses information taken from previous iterations. Again, we assume that θ is strictly quasiconvex on $[a; b]$. The idea is to place λ and μ near to the midpoint to guard against the worst possible outcome in reduction of the uncertainty interval.

Algorithm 248.

Choose the constant δ>0 and the allowable final length of uncertainty interval $\varepsilon > 0$. Let $[a_1; b_1]$ be the initial interval of uncertainty, let $k := 1$, and continue.

1. If $b_k - a_k < \varepsilon$ then STOP the minimum lies in the interval $[a_k; b_k]$. Otherwise:

$$\lambda_k := \frac{a_k + b_k}{2} - \delta \qquad \mu_k := \frac{a_k + b_k}{2} + \delta$$

2. If $\theta(\lambda_k) < \theta(\mu_k)$, let $a_{k+1} := a_k$ and $b_{k+1} := \mu_k$. Otherwise, let $a_{k+1} := \lambda_k$ a $b_{k+1} := b_k$. In both cases $k := k\ + 1$ and GOTO 1.

(Be able to use it!) We have to notice that each iteration requires two new computations of θ function values.

Golden section method. To compare various line search methods, usually we consider the ratio of the uncertainty interval length after k'th iteration and before any iteration taken. It is clear that more efficient algorithm has the smaller ratio. This seems to be an argument for dichotomous search. For it (and also for the developed golden section method) λ_k and μ_k are chosen in such a way that $\mu_k - a_k = b_k - \lambda_k = b_{k+1} - a_{k+1}$ because we do not know our preference for either μ_k or λ_k in advance. It may be also acceptable to require

$$\frac{b_{k+1} - a_{k+1}}{b_k - a_k} = \alpha, \tag{1.3}$$

and so, reduction ratio $\alpha \in (0; 1)$ is a constant. We may think about another possibility to save computational time. We may find that our previous idea to count iterations is imprecise because we are not taking into account needed functional evaluations (There are 2 per iteration

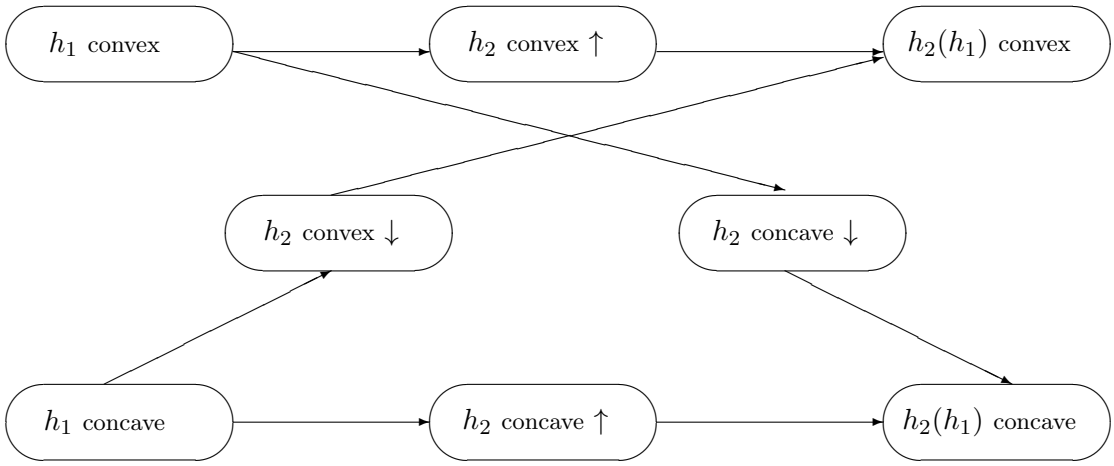

Figure 1.1: Convexity and concavity of composed functions.

for dichotomous search). So, we try to utilize already computed function value for the next iteration (This is the main idea of the golden section method! Could you utilize the following figure for the computation of α?). Therefore, for the case $a_{k+1} := \lambda_k$ a $b_{k+1} := b_k$ we have to assign $\lambda_{k+1} = \mu_k$. By Figure 2, two different descriptions of μ_k utilizing constant α may be used:

$$a_k + \alpha(b_k - a_k) = \lambda_k + (1 - \alpha)(b_k - \lambda_k).$$

We may replace λ_k with $a_k + (1 - \alpha)(b_k - a_k)$ and after the following simplification:

$$\begin{aligned} a_k + \alpha(b_k - a_k) &= a_k + (1 - \alpha)(b_k - a_k) + (1 - \alpha)(b_k - (a_k + (1 - \alpha)(b_k - a_k))) \\ \alpha(b_k - a_k) &= 2(1 - \alpha)(b_k - a_k) - (1 - \alpha)^2(b_k - a_k))) \\ \alpha &= 2(1 - \alpha) - (1 - \alpha)^2 \end{aligned}$$

we obtain a quadratic equation $\alpha^2 + \alpha - 1 = 0$ (it is also the same for the case $a_{k+1} := a_k$ a $b_{k+1} := \mu_k$) We search for $\alpha > 0$ and that is $\alpha = -1/2 + \sqrt{5}/2 = 0.618$ the golden section ratio.

For strictly quasiconvex function θ on $[a; b]$ we form the following algorithm (Be able to use it! - with help):

Algorithm 249.

Choose $\varepsilon > 0$, $[a_1; b_1]$, $\alpha := 0.618$, compute $\lambda_1 := a_1 + (1 - \alpha)(b_1 - a_1)$, $\mu_1 := a_1 + \alpha(b_1 - a_1)$ together with $\theta(\lambda_1)$ a $\theta(\mu_1)$, and assign $k := 1$.

1. If $b_k - a_k < \varepsilon$ STOP, the minimum lies in $[a_k; b_k]$. Otherwise, if $\theta(\lambda_k) > \theta(\mu_k)$ then GOTO 2.; and if $\theta(\lambda_k) \leq \theta(\mu_k)$ then GOTO 3.

2. Assign $a_{k+1} := \lambda_k$ and $b_{k+1} := b_k$. Furthermore, $\lambda_{k+1} := \mu_k$ and $\mu_{k+1} := a_{k+1} + \alpha(b_{k+1} - a_{k+1})$. Evaluate $\theta(\mu_{k+1})$ and GOTO 4.

3. Assign $a_{k+1} := a_k$ and $b_{k+1} := \mu_k$. Furthermore, $\mu_{k+1} := \lambda_k$ and $\lambda_{k+1} := a_{k+1} + (1 - \alpha)(b_{k+1} - a_{k+1})$. Evaluate $\theta(\lambda_{k+1})$ and GOTO 4.

4. $k := k + 1$, GOTO 1.

Example 250.

Solve the problem $\min\{\lambda^2 + 2\lambda \mid -3 \leq \lambda \leq 5\}$ using the golden section method. The results of 4 iterations are contained in the table (cf. with the true minimum $\lambda_{\min} = -1$).

Iteration k	a_k	b_k	λ_k	μ_k	$\theta(\lambda_k)$	$\theta(\mu_k)$
1	−3.000	5.000	0.056	1.944	0.115	7.667
2	−3.000	1.944	−1.112	0.056	−0.987	0.115
3	−3.000	0.056	−1.832	−1.112	−0.308	−0.987
4	−1.832	0.056	−1.112	−0.664	−0.987	−0.887

The Fibonacci search. It is useful in the cases when the number of iterations N to be realized is known. The Fibonacci sequence of $N+1$ numbers F_k is defined by the assignment $F_{k+1} := F_k + F_{k-1},\ k = 1, \ldots, N-1$, where $F_0 = F_1 = 1$. The basic idea is common with the golden section method which is the limiting case of the Fibonacci method (think about the limit of sequence of ratios used instead of α). Notice that the number of iterations must be known in advance because F_N is used at the beginning.

Algorithm 251 (Fibonacci)**.**

Choose $\varepsilon > 0$, $\delta > 0$, $[a_1; b_1]$, N such that $F_N > (b_1 - a_1)/\varepsilon$. Then $\lambda_1 := a_1 + (F_{N-2}/F_N)(b_1 - a_1)$, $\mu_1 := a_1 + (F_{N-1}/F_N)(b_1 - a_1)$, $\theta(\lambda_1)$, and $\theta(\mu_1)$ a $k := 1$.

1. If $\theta(\lambda_k) > \theta(\mu_k)$ then GOTO 2., otherwise, if $\theta(\lambda_k) \leq \theta(\mu_k)$ then GOTO 3.

2. $a_{k+1} := \lambda_k$ and $b_{k+1} := b_k$. Furthermore, $\lambda_{k+1} := \mu_k$ and $\mu_{k+1} := a_{k+1} + (F_{N-k-1}/F_{N-k})(b_{k+1} - a_{k+1})$. If $k = N-2$ then GOTO 5., otherwise compute $\theta(\mu_{k+1})$ and GOTO 4.

3. $a_{k+1} := a_k$ and $b_{k+1} := \mu_k$. Furthermore, $\mu_{k+1} := \lambda_k$ and $\lambda_{k+1} := a_{k+1} + (F_{N-k-2}/F_{N-k})(b_{k+1} - a_{k+1})$. If $k = N-2$ then GOTO 5., otherwise compute $\theta(\lambda_{k+1})$ and GOTO 4.

4. $k := k + 1$, GOTO 1.

5. Set $\lambda_N := \lambda_{N-1}$ and $\mu_N := \lambda_{N-1} + \delta$. If $\theta(\lambda_N) > \theta(\mu_N)$ then $a_N := \lambda_N$ and $b_N := b_{N-1}$. If $\theta(\lambda_N) \leq \theta(\mu_N)$ then $a_N := a_{N-1}$ and $b_N := \lambda_N$. STOP and minimum lies in interval $[a_N; b_N]$.

Example 252.

Solve the problem $\min\{\lambda^2 + 2\lambda \mid -3 \leq \lambda \leq 5\}$ using the Fibonacci search for $N = 9$. Results of 4 iterations are contained in the table.

Iteration k	a_k	b_k	λ_k	μ_k	$\theta(\lambda_k)$	$\theta(\mu_k)$
1	−3.000	5.000	0.054	1.945	0.112	7.675
2	−3.000	1.945	−1.109	0.054	−0.988	0.112
3	−3.000	0.054	−1.836	−1.109	−0.300	−0.988
4	−1.836	0.054	−1.109	−0.672	−0.988	−0.892

Comparison of derivative-free line search methods. For the given precision $\varepsilon > 0$, the required number of iterations $n \in \mathbb{N}$ must satisfy the following inequalities:
Do you have any idea how they were derived?

Uniform search method: $n \geq (b_1 - a_1)/(\varepsilon/2) - 1$

Dichotomous search method: $(0.5)^{n/2} \geq \varepsilon/(b_1 - a_1)$

Golden section method: $(0.618)^{n-1} \geq \varepsilon/(b_1 - a_1)$

Fibonacci search method: $F_n \geq (b_1 - a_1)/\varepsilon$

1.8.2 Line search using derivatives

Using calculus. We may try to compute $\theta'(\lambda) = 0$ then we obtain:

$$0 = \theta'(\lambda) = \mathbf{d}_k^\top \nabla f(\mathbf{x}_k + \lambda \mathbf{d}_k)$$

and it is usually a nonlinear equation. Be able to use it during examination! If f is not differentiable then also θ is not differentiable and previous methods are used. With differentiable θ we may use the following algorithms.

Bisection search method. It converges for pseudoconvex function θ Compare with the root search bisection! Be able to use it for computations.

Algorithm 253.

Identify $[a_1; b_1]$, $\varepsilon > 0$, and $N \in \mathbb{N}$ satisfying $0.5^N \leq \varepsilon/(b_1 - a_1)$, $k := 1$.

1. $\lambda_k := \frac{1}{2}(a_k + b_k)$, compute $\theta'(\lambda_k)$. If $\theta'(\lambda_k) = 0$ then STOP and λ_k is a minimum. For $\theta'(\lambda_k) > 0$ GOTO 2. and for $\theta'(\lambda_k) < 0$ GOTO 3.
2. $a_{k+1} := a_k$ and $b_{k+1} := \lambda_k$, GOTO 4.
3. $a_{k+1} := \lambda_k$ and $b_{k+1} := b_k$, GOTO 4.
4. If $k = N$, STOP, and the minimum lies in $[a_{N+1}; b_{N+1}]$. Otherwise $k := k + 1$ and GOTO 1.

Example 254.

Solve the problem $\min\{\lambda^2 + 2\lambda \mid -3 \leq \lambda \leq 6\}$ by the bisection search method.

Iteration k	a_k	b_k	λ_k	$\theta'(\lambda_k)$
1	−3.000	6.000	1.500	5.000
2	−3.000	1.500	−0.750	0.500
3	−3.000	−0.750	−1.875	−1.750
4	−1.875	−0.750	−1.312	−0.625

Newton's method. It searches roots of equation $f(x) = 0$. So, we may use it for our equation $\theta'(\lambda) = 0$. By the 2nd order Taylor series approximation of the function θ at λ_k we obtain:

$$T_2(\lambda) = \theta(\lambda_k) + \theta'(\lambda_k)(\lambda - \lambda_k) + \frac{1}{2}\theta''(\lambda_k)(\lambda - \lambda_k)^2,$$

and the following iteration formula (Be able to derive it!):

$$\lambda_{k+1} = \lambda_k - \frac{\theta'(\lambda_k)}{\theta''(\lambda_k)}$$

If $|\lambda_{k+1} - \lambda_k| < \varepsilon$ or $|\theta'(\lambda_k)| < \varepsilon$ then the iteration process is stopped. The use of this procedure requires the existence of non-zero 2nd order derivatives of θ at $\lambda_k, k = 1, 2, \ldots$. This method does not converge to the stationary point for all initial points λ_1. See Bazaraa-Shetty for theoretical sufficient convergence conditions ("the initial point λ_1 should be chosen near to the minimum").

1.8.3 Approximating line search

Motivating idea. Previous methods do not accelerate the computational process using information about the shape of function θ (with exception of non-globally convergent Newton's method and its $\theta'\prime$ use). Therefore, quadratic fit line search has been developed (even cubic also exists). During each iteration, the original function θ is replaced by parabola (Explain algorithm by figures). Its minimum approximates λ_k. We assume that θ is strictly quasiconvex and continuous and $\lambda > 0$.

Algorithm 255.

We suppose that three points are given $0 \leq \lambda_1 < \lambda_2 < \lambda_3$ such that they satisfy $\theta_1 \geq \theta_2$ and $\theta_2 \leq \theta_3$, where $\theta_j = \theta(\lambda_j)$ for $j = 1, 2, 3$. Choose $\varepsilon > 0$.

1. Three points $(\lambda_j; \theta_j), j = 1, \ldots, 3$ determine the parabola (How?):

$$q(\lambda) = \frac{\theta_1(\lambda - \lambda_2)(\lambda - \lambda_3)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)} + \frac{\theta_2(\lambda - \lambda_1)(\lambda - \lambda_3)}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} + \frac{\theta_3(\lambda - \lambda_1)(\lambda - \lambda_2)}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)}$$

and from $q'(\lambda) = 0$, we obtain:

$$\lambda^* = \frac{1}{2} \frac{b_{23}\theta_1 + b_{31}\theta_2 + b_{12}\theta_3}{a_{23}\theta_1 + a_{31}\theta_2 + a_{12}\theta_3},$$

where $a_{ij} = \lambda_i - \lambda_j$ and $b_{ij} = \lambda_i^2 - \lambda_j^2$. By substitution we obtain $\theta^* = \theta(\lambda^*)$.

2. If $\lambda^* > \lambda_2$ and $\theta^* \geq \theta_2$ then new three points are defined as follows $\lambda_1, \lambda_2, \lambda^*$. If $\lambda^* > \lambda_2$ a $\theta^* < \theta_2$ then new three points are specified as $\lambda_2, \lambda^*, \lambda_3$. GOTO 5.

3. if $\lambda^* < \lambda_2$ and $\theta^* \geq \theta_2$ then new three points are given as $\lambda^*, \lambda_2, \lambda_3$. If $\lambda^* < \lambda_2$ and $\theta^* < \theta_2$ then new three points are $\lambda_1, \lambda^*, \lambda_2$. GOTO 5.

4. If $\lambda^* = \lambda_2$, then no new point is derived and $0 < \delta < \lambda_3 - \lambda_1$ is chosen to move λ^* from λ_2 by $\delta/2$ (a direction does not matter) and GOTO 2.

5. Test whether $\theta_1 = \theta_2 = \theta_3$ or $\lambda_3 - \lambda_1 < \varepsilon$. If termination condition is not fulfilled then update λ_1, λ_2, and λ_3 by previous rules and GOTO 1.

1.8.4 Inexact line search

Armijo's rule. There are many rules that do not try to minimize $\theta(\lambda)$ but they search for the improved value of $\theta(\lambda)$ in such a way that an outer multivariate algorithm still converges. (Explain by figure!).

Algorithm 256.

Choose $0 < \delta < 1$ (0.2 is recommended) a $\alpha > 1$ (2 is recommended). The computational procedure is formed: $\theta^*(\lambda) = \theta(0) + \lambda\delta\theta'(0)$. The initial iteration λ^* is given.

1. If $\theta(\lambda^*) \leq \theta^*(\lambda^*)$ then multiply λ^* by coefficient α repeatedly until t is found such that satisfies $\theta(\alpha^t\lambda^*) \leq \theta^*(\alpha^t\lambda^*)$ and in addition $\theta(\alpha^{t+1}\lambda^*) > \theta^*(\alpha^{t+1}\lambda^*)$. Then $\lambda^* := \alpha^t\lambda^*$.

2. If $\theta(\lambda^*) > \theta^*(\lambda^*)$ then divide λ^* by coefficient α repeatedly until t is found such that satisfies $\theta(\lambda^*/\alpha^t) \leq \theta^*(\lambda^*/\alpha^t)$ and in addition $\theta(\lambda^*/\alpha^{t-1}) > \theta^*(\lambda^*/\alpha^{t-1})$. Then $\lambda^* := \lambda^*/\alpha^t$.

1.8.5 Convergence

Line search algorithmic map. Since the line search is a basic component of most NLP algorithms we will analyse its convergence by the concept of closed map (Do not memorize! But understand!). We introduce

$$\mathcal{M}(\mathbf{x};\mathbf{d}) = \{\mathbf{y} \mid \exists \lambda_{\min} \in \mathcal{L} : \mathbf{y} = \mathbf{x} + \lambda_{\min}\mathbf{d}, \forall \lambda \in \mathcal{L} : f(\mathbf{y}) \leq f(\mathbf{x} + \lambda \mathbf{d})\}.$$

$\mathcal{M}$ is generally a point-set mapping (Why? Think about uniqueness of $\lambda_{\min}$).

Theorem 257.

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be continuous at $\mathbf{x}$, $\mathcal{L}$ is a closed interval in $\mathbb{R}$, and $\mathbf{d} \neq \mathbf{0}$. Then above introduced $\mathcal{M} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is closed at $(\mathbf{x};\mathbf{d})$.

Proof: We want to show that $\mathbf{y} \in \mathcal{M}(\mathbf{x},\mathbf{d})$ for $(\mathbf{x}_k,\mathbf{d}_k) \longrightarrow (\mathbf{x},\mathbf{d})$ and $\mathbf{y}_k \longrightarrow \mathbf{y}$ where $\mathbf{y}_k \in \mathcal{M}(\mathbf{x}_k;\mathbf{d}_k)$. From $\mathbf{y}_k = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ and $\mathbf{d} \neq \mathbf{0} \Rightarrow \mathbf{d}_k \neq \mathbf{0}$ (for k large enough) we obtain $\lambda_k = \| \mathbf{y}_k - \mathbf{x}_k \| / \| \mathbf{d}_k \|$. As the $k \longrightarrow \infty$ then $\lambda_k \longrightarrow \lambda_{\min}$ and also the RHS: $\| \mathbf{y}_k - \mathbf{x}_k \| / \| \mathbf{d}_k \| \longrightarrow \| \mathbf{y} - \mathbf{x} \| / \| \mathbf{d} \|$ and so $\mathbf{y} = \mathbf{x} + \lambda_{\min}\mathbf{d}$. Furthermore, by closedness of $\mathcal{L}$: $\lambda_k \in \mathcal{L} \Rightarrow \lambda_{\min} \in \mathcal{L}$ and by continuity of f from $f(\mathbf{y}_k) \leq f(\mathbf{x}_k + \lambda \mathbf{d}_k), \forall \lambda \in \mathcal{L}, \forall k$ we may conclude that $f(\mathbf{y}) \leq f(\mathbf{x} + \lambda \mathbf{d}), \forall \lambda \in \mathcal{L}$. Thus $\mathbf{y} \in \mathcal{M}(\mathbf{x},\mathbf{d})$ and the proof is complete. □

Remark 258.

In addition, if $\mathcal{D}$ mapping identifying the direction $\mathbf{d}$ is also closed then by general algorithm-related theorems, the composed mapping $\mathcal{A} = \mathcal{M}\mathcal{D}$ is closed. See also Bazaraa-Shetty page 283 for an example where $\mathbf{d} = \mathbf{0}$ causes that $\mathcal{M}$ is not closed.

1.9 Multidimensional Search Without Using Derivatives

Methods that do not use derivatives are popular because they are robust and easy implementable (Do you agree?). They are suitable for modest size engineering problems that are not solved too often. They are also advantageous in the cases when the derivatives (or their approximations are hardly available). There are complex computational systems (e.g., Fluent, Ansys) that for the given input vector $\mathbf{x} \in \mathbb{R}^n$ generate the output $\mathbf{y} \in \mathbb{R}^m$ by transformation $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$,which is not explicitly known. For instance, it may be hidden in the program. In such case, when y_i is minimized and other variables are free, one of proposed methods in this section may be employed.

1.9.1 The Cyclic Coordinate Method

Basic idea. It is a trivial and slowly convergent method (You have to know it!). The main idea is to repeatedly realize line searches by axis directions $\mathbf{d}_j$. It often results in zig-zag behaviour, see Figure 3.

Algorithm 259.

Choose $\varepsilon > 0$ and directions $\mathbf{d}_j := \mathbf{e}_j$ pro $j = 1, \ldots, n$, where $\mathbf{e}_j$ are coordinate directions. Choose an initial point $\mathbf{x}_1$, set $\mathbf{y}_1 := \mathbf{x}_1$ and $j := 1$, $k := 1$.

1. Let λ_j be an optimal solution of $? \in \operatorname{argmin}_{\lambda \in \mathbb{R}} f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ and let $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$ then $j := j + 1$ and GOTO 1. If $j = n$ then continue.

2. Let $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$ then STOP. Otherwise $\mathbf{y}_1 := \mathbf{x}_{k+1}$, $j := 1$, $k := k + 1$ and GOTO 1.

Method convergence. It follows immediately from the theorem about linearly independent directions under the following assumptions (Why? What is **D**?):

1. The minimum of f along any line in $\mathbb{R}^n$ is unique.
2. The sequence of points generated by algorithm is contained in a compact subset of $\mathbb{R}^n$

Example 260.

Solve the problem $\min [(x_1 - 2)^4 + (x_1 - 2x_2)^2]$ by Algorithm 259. Remember that results in tables are rounded or truncated, transposition is very often omitted.

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{d}_j$	$\mathbf{y}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0,00; 3,00)	1	(1,0; 0,0)	(0,00; 3,00)	3,13	(3,13; 3,00)
	52,00	2	(0,0; 1,0)	(3,13; 3,00)	−1,44	(3,13; 1,56)
2	(3,13; 1,56)	1	(1,0; 0,0)	(3,13; 1,56)	−0,50	(2,63; 1,56)
	1,63	2	(0,0; 1,0)	(2,63; 1,56)	−0,25	(2,63; 1,31)
3	(2,63; 1,31)	1	(1,0; 0,0)	(2,63; 1,31)	−0,19	(2,44; 1,31)
	0,16	2	(0,0; 1,0)	(2,44; 1,31)	−0,09	(2,44; 1,22)
4	(2,44; 1,22)	1	(1,0; 0,0)	(2,44; 1,22)	−0,09	(2,35; 1,22)
	0,04	2	(0,0; 1,0)	(2,35; 1,22)	−0,05	(2,35; 1,17)

Figure 1.2: Illustration of the golden section rule.

1.9.2 The Method of Hooke and Jeeves

Acceleration step. The cyclic coordinate method is accelerated (and also zig-zag appears not so often) by including additional accelerating steps using direction $\mathbf{x}_{k+1} - \mathbf{x}_k$.
See Figure 4(And use Figure!).

Algorithm 261 (Hooke-Jeeves).

Choose $\varepsilon > 0$, directions $\mathbf{d}_j := \mathbf{e}_j$ for $j = 1, \ldots, n$, initial point $\mathbf{x}_1$, and assign $\mathbf{y}_1 := \mathbf{x}_1$ a $j := 1$, $k := 1$.

1. Let λ_j be an optimal solution of $? \in \operatorname{argmin}_{\lambda \in \mathbb{R}} f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ and set $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$ then $j := j + 1$ and GOTO 1. If $j = n$ then $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}$ and for $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$ STOP, otherwise continue.

2. Let $\mathbf{d} := \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\hat{\lambda}$ solves $? \in \operatorname{argmin}_{\lambda \in \mathbb{R}} f(\mathbf{x}_{k+1} + \lambda \mathbf{d})$. Determine $\mathbf{y}_1 := \mathbf{x}_{k+1} + \hat{\lambda}\mathbf{d}$, and so $j := 1$, $k := k + 1$ and GOTO 1.

Method convergence. The idea of convergence proof lies in the decomposition of the method into two maps: $\mathcal{B}$ denotes cyclic coordinate method already discussed and accelerating step by map $\mathcal{C}$. With additional assumptions: the unique minimum of f (differentiable) along any line, $\alpha = f \Rightarrow \forall \mathbf{x} \in \mathbb{R}^n \setminus \Omega : \alpha(\mathbf{y}) < \alpha(\mathbf{x})$, $\forall \mathbf{z} \in \mathcal{C}(\mathbf{y}) : \alpha(\mathbf{z}) \leq \alpha(\mathbf{y})$, $\Omega = \{\mathbf{x} | \nabla f(\mathbf{x}) = \mathbf{0}\}$, and $\{\mathbf{x} | f(\mathbf{x}) \leq f(\mathbf{x}_1)\}$ is compact for starting point $\mathbf{x}_1$ the convergence is established by the general convergence theorem for composed maps.

Example 262.

Solve the problem $\min [(x_1 - 2)^4 + (x_1 - 2x_2)^2]$ by Algorithm 261.

k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{d}_j$	$\mathbf{y}_j$	λ_j	$\mathbf{y}_{j+1}$	$\mathbf{d}$	$\hat{\lambda}$	$\mathbf{y}_{n+1} + \hat{\lambda}\mathbf{d}$
1	$(0,00; 3,00)$	1	$(1,0; 0,0)$	$(0,00; 3,00)$	$3,13$	$(3,13; 3,00)$			
	$52,00$	2	$(0,0; 1,0)$	$(3,13; 3,00)$	$-1,44$	$(3,13; 1,56)$	$(3,13; 1,44)$	$-0,10$	$(2,82; 1,70)$
2	$(3,13; 1,56)$	1	$(1,0; 0,0)$	$(2,82; 1,70)$	$-0,12$	$(2,70; 1,70)$			
	$1,63$	2	$(0,0; 1,0)$	$(2,70; 1,70)$	$-0,35$	$(2,70; 1,35)$	$(-0,43; -0,21)$	$1,50$	$(2,06; 1,04)$
3	$(2,70; 1,35)$	1	$(1,0; 0,0)$	$(2,06; 1,04)$	$-0,02$	$(2,04; 1,04)$			
	$0,24$	2	$(0,0; 1,0)$	$(2,04; 1,04)$	$-0,02$	$(2,04; 1,02)$	$(-0,66; -0,33)$	$0,06$	$(2,00; 1,00)$
4	$(2,04; 1,02)$	1	$(1,0; 0,0)$	$(2,00; 1,00)$	$-0,01$	$(2,00; 1,00)$			
	$0,000003$	2	$(0,0; 1,0)$	$(2,00; 1,00)$	$-0,01$	$(2,00; 1,00)$			

It is important to note that univariate minimization can be replaced by imprecise algorithms (e.g., discrete steps, Armijo's rule). Other methods without derivatives may be found, e.g., in Bazaraa-Shetty. They use different sets of directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$.

Figure 1.3: Illustration of the cyclic coordinate method.

1.9.3 The method of Nelder and Mead

Moving simplex. The method of moving simplex (do not mix with LP!) is often used by engineers. Its main idea is to choose points $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ to form a simplex. The algorithm moves and deform the simplex to achieve the situation when the searched minimum becomes the simplex interior point. By biological analogy this method is often implemented as AMOEBA (Be able to explain certain algorithm steps by figures). The method is robust but slow, so suitable for $n \leq 10$.

Algorithm 263.

Choose points $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1} \in \mathbb{R}^n$ to form a simplex. Choose coefficient of reflection $\alpha > 0$, coefficient of contraction $\beta < 1$ and coefficient of expansion $\gamma > 1$.

1. Search r and s such that $f(\mathbf{x}_r) = \min_{1 \leq j \leq n+1} f(\mathbf{x}_j)$ and $f(\mathbf{x}_s) = \max_{1 \leq j \leq n+1} f(\mathbf{x}_j)$. If points are near enough then STOP and the minimum is found. Otherwise, compute $\bar{\mathbf{x}} := \frac{1}{n} \sum_{j=1, j \neq s}^{n+1} \mathbf{x}_j$

2. Let $\hat{\mathbf{x}} := \bar{\mathbf{x}} + \alpha(\bar{\mathbf{x}} - \mathbf{x}_s)$. If $f(\mathbf{x}_r) > f(\hat{\mathbf{x}})$ then $\mathbf{x}_e := \bar{\mathbf{x}} + \gamma(\hat{\mathbf{x}} - \bar{\mathbf{x}})$ and GOTO 3. If $f(\mathbf{x}_r) \leq f(\hat{\mathbf{x}})$, then GOTO 4.

3. If $f(\hat{\mathbf{x}}) > f(\mathbf{x}_e)$ then $\mathbf{x}_s := \mathbf{x}_e$. If $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}_e)$ then $\mathbf{x}_s := \hat{\mathbf{x}}$. In both cases GOTO 1.

4. If $\max_{1 \leq j \leq n+1} \{f(\mathbf{x}_j) \mid j \neq s\} \geq f(\hat{\mathbf{x}})$ then $\mathbf{x}_s := \hat{\mathbf{x}}$ and GOTO 1. Otherwise GOTO 5.

5. Compute $\mathbf{x}'$ using $f(\mathbf{x}') = \min\{f(\hat{\mathbf{x}}); f(\mathbf{x}_s)\}$ and $\mathbf{x}'' := \bar{\mathbf{x}} + \beta(\mathbf{x}' - \bar{\mathbf{x}})$. If $f(\mathbf{x}'') > f(\mathbf{x}')$ then $\mathbf{x}_j := \mathbf{x}_j + \frac{1}{2}(\mathbf{x}_r - \mathbf{x}_j)$ for $j = 1, \ldots, n+1$ and GOTO 1. If $f(\mathbf{x}'') \leq f(\mathbf{x}')$ then $\mathbf{x}_s := \mathbf{x}''$ and GOTO 1.

1.10 Multidimensional Search Using Derivatives

1.10.1 Steepest descent (gradient) method

Remark 264 (Main idea).

It is a classical method (Gauss) for the multivariate differentiable function minimization. $\mathbf{d}$ is a descent direction (feasible because $S = \mathbb{R}^n$) of f at $\mathbf{x} \Leftrightarrow \exists \delta > 0 : \forall \lambda \in (0, \delta)$ $f(\mathbf{x} + \lambda \mathbf{d}) < f(\mathbf{x})$. In particular: If $\lim_{\lambda \longrightarrow 0^+} \frac{f(\mathbf{x}+\lambda \mathbf{d}) - f(\mathbf{x}))}{\lambda} < 0$ then $\mathbf{d}$ is a descent direction. To find a descent direction explicitly we should minimize a directional derivative (see limit above) under condition $\|\mathbf{d}\| = 1$.

Lemma 265 (Steepest descent direction).

Let $f : S \longrightarrow \mathbb{R}$ be a differentiable function at $\mathbf{x}$, $f(\mathbf{x}) \neq \mathbf{0}$. Then the steepest descent direction $\bar{\mathbf{d}}$ satisfies $\bar{\mathbf{d}} = \frac{-\nabla f(x)}{\|\nabla f(x)\|} \in \operatorname{argmin}\{f'(\mathbf{x}, \mathbf{d}) \mid \|\mathbf{d}\| \leq 1\}$.

Proof: By differentiability and definitions $f'(\mathbf{x}, \mathbf{d}) = \lim_{\lambda \longrightarrow 0^+} \frac{f(\mathbf{x}+\lambda \mathbf{d}) - f(\mathbf{x}))}{\lambda} = \nabla f(\mathbf{x})^\top \mathbf{d}$ (Taylor formula in use). So $\operatorname{argmin}_{\mathbf{d}} \{\nabla f(\mathbf{x})^\top \mathbf{d} \mid \|\mathbf{d}\| \leq 1\}$ have to be solved. We know that $\nabla f(\mathbf{x})^\top \mathbf{d} \geq -\|\nabla f(x)\| \|\mathbf{d}\| \geq -\|\nabla f(x)\|^2$. Hint: See the formula for the cosine of to vectors $\nabla f(x)$ and $\mathbf{d}$ and when it is equal to -1. □

So, $-\nabla f(\mathbf{x})$ is a direction of the steepest descent of f at $\mathbf{x}$, and hence, the method is also called the gradient method.

Algorithm 266 (Gradient method)**.**

We choose $\varepsilon > 0$, point $\mathbf{x}_1$, $k := 1$.

1. If $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$ then STOP. Otherwise we assign $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$. We get a solution λ_k of $\min\{f(\mathbf{x}_k + \lambda \mathbf{d}_k) \mid \lambda \geq 0\}$, and then we define $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$ and $k := k + 1$, GOTO 1.

Example 267.

Solve the program $\min\left[(x_1 - 2)^4 + (x_1 - 2x_2)^2\right]$ by the gradient method – Algorithm 266.

k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	$\nabla f(\mathbf{x}_k)$	$\|\nabla f(\mathbf{x}_k)\|$	$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$	λ_k	$\mathbf{x}_{k+1}$
1	$(0,00; 3,00)$ $52,00$	$(-44,00; 24,00)$	$50,12$	$(44,00; -24,00)$	$0,062$	$(3,13; 3,00)$
2	$(2,70; 1,51)$ $0,34$	$(0,73; 1,28)$	$1,47$	$(-0,73; -1,28)$	$0,24$	$(2,52; 1,20)$
3	$(2,52; 1,20)$ $0,09$	$(0,80; -0,48)$	$0,93$	$(-0,80; 0,48)$	$0,11$	$(2,43; 1,25)$
4	$(2,43; 1,25)$ $0,04$	$(0,18; 0,28)$	$0,33$	$(-0,18; -0,28)$	$0,31$	$(2,37; 1,16)$

Remark 268 (Convergence of the steepest descent method (Optional))**.**

Let set $\Omega = \{\bar{\mathbf{x}} \mid \nabla f(\bar{\mathbf{x}}) = \mathbf{0}\}$ be a solution set, $\alpha(\mathbf{x}) = f(\mathbf{x})$ is a descent function, $\mathcal{A} = \mathcal{M}\mathcal{D}$ is a composed algorithmic map, where $\mathcal{D} : \mathbf{x} \longmapsto [\mathbf{x}, \nabla f(\mathbf{x})]$ (for the given point, a new point and descent direction are generated). Therefore, $\mathcal{D}$ is a function. If f is continuously differentiable then $\mathcal{D}$ is continuous. $\mathcal{M}$ was already discussed with line search methods (check it!). There is $\mathcal{M} : [\mathbf{x}, \mathbf{d}] \longmapsto \{\mathbf{y} \mid \mathbf{y} = \mathbf{x} + \lambda_{\min}\mathbf{d}\}$ where $\lambda_{\min} \in \operatorname{argmin}\{f(\mathbf{x} + \lambda\mathbf{d}) \mid \lambda \in \mathcal{L}\}$. Then, $\mathcal{M}$ is closed for $\mathcal{L}$ closed interval and f continuous and $\mathbf{d} \neq \mathbf{0}$. Why? We had $(\mathbf{x}_k, \mathbf{d}_k) \longrightarrow (\mathbf{x}, \mathbf{d})$ and for $\mathbf{y}_k$ derived by $\mathcal{M}$ from $(\mathbf{x}_k, \mathbf{d}_k)$ and $\mathbf{y}_k \longrightarrow \mathbf{y}$ we have to prove $\mathcal{M} : (\mathbf{x}, \mathbf{d}) \longmapsto \mathbf{y}$. So, $\mathbf{y}_k = \mathbf{x}_k + \lambda\mathbf{d}_k \Rightarrow$ as $\mathbf{d} \neq \mathbf{0}$ and $\mathbf{d}_k \longrightarrow \mathbf{d}$ then $\mathbf{d}_k \neq \mathbf{0}$ for big k. So, $\mathbf{y}_k - \mathbf{x}_k = \lambda_k\mathbf{d}_k$, $\|\mathbf{y}_k - \mathbf{x}_k\| = |\lambda_k|\|\mathbf{d}_k\|$, and $\lambda_k = \frac{\|\mathbf{y}_k - \mathbf{x}_k\|}{\|\mathbf{d}_k\|}$. Therefore, for $k \longrightarrow \infty$, we have and denote $\frac{\|\mathbf{y}_k - \mathbf{x}_k\|}{\|\mathbf{d}_k\|} \longrightarrow \frac{\|\mathbf{y} - \mathbf{x}\|}{\|\mathbf{d}\|} = \lambda_{\min}$ (and $\lambda_k \longrightarrow \lambda_{\min}$). So $\mathbf{y}_k = \mathbf{x}_k + \lambda_k\mathbf{d}_k$, and because of convergence, we get $\mathbf{y} = \mathbf{x} + \lambda_{\min}\mathbf{d}$. Because of $\mathcal{L}$ closed $\lambda_{\min} \in \mathcal{L}$ so $\mathbf{y}$ has the form by $\mathcal{M}$.
But is $\lambda_{\min}$ a true minimum? $\forall\lambda \in \mathcal{L}, \forall k : f(\mathbf{y}_k) \leq f(\mathbf{x}_k + \lambda\mathbf{d}_k)$ and because of limit and continuity of f: $f(\mathbf{y}) \leq f(\mathbf{x} + \lambda\mathbf{d})$. So we have $\mathcal{D}$ continuous and $\mathcal{M}$ closed. Because of Theorem 231, $\mathcal{A} = \mathcal{M}\mathcal{D}$ is closed.
The last step to a convergence proof is to show that $\forall\mathbf{x} \notin \Omega : \mathbf{y} \in \mathcal{A}(\mathbf{x}) \Rightarrow \alpha(\mathbf{y}) < \alpha(\mathbf{x})$. But $\mathbf{y} = \mathbf{x} + \lambda_{\min}\mathbf{d}$ and $\mathbf{d} = -\nabla f(\mathbf{x}) \neq \mathbf{0}$. So, for $\nabla\mathbf{f}(\mathbf{x})^\top\mathbf{d} = \nabla f(\mathbf{x})^\top(-\nabla f(\mathbf{x})) = -\|f(\mathbf{x})\|^2 < 0$ and by unconstrained optimization theory $\Rightarrow \mathbf{d}$ is a descent direction (check your notes!), so from $\alpha(\cdot) = f(\cdot) \Rightarrow f(\mathbf{y}) < f(\mathbf{x})$.
At the end, we need to suppose that $\{x_k\}$ generated by the algorithm is contained in a compact set to complete the proof (global convergence i.e. $\forall\mathbf{x}_1$ the sequence converges to $\mathbf{x}_{\min}$). E.g., $\|\mathbf{x}\| \longrightarrow \infty$ then $f(\mathbf{x}) \longrightarrow \infty$.

The algorithm looks easily implementable, it converges. Why it is not often used? Because it works well at the beginning of iteration process and for nice functions. But near to the minimum and for 'narrow valleys' zigzaging effect may appear. Why? Intuitively, because of the fact that near to the minimum $\nabla f(\mathbf{x}_k) \longrightarrow \mathbf{0}$.

In detail: $f(\mathbf{x}_k + \lambda\mathbf{d}) = f(\mathbf{x}_k) + \lambda\nabla f(\mathbf{x}_k)^\top\mathbf{d} + \lambda\|\mathbf{d}\|\alpha(\mathbf{x}_k; \mathbf{d})$ (see Taylor expansion), where $\mathbf{d}$ is a search direction, $\|\mathbf{d}\| = 1$ and for $\lambda\mathbf{d} \longrightarrow \mathbf{0}$, we have $\alpha(\mathbf{x}_k; \lambda\mathbf{d}) \longrightarrow 0$. In the case of steepest descent method, we use $\mathbf{d} = -\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ and after substitution the second term of Taylor series is $\lambda\nabla f(\mathbf{x}_k)^\top\mathbf{d} = -\lambda\|\nabla f(\mathbf{x}_k)\|$, and hence the objective function changes slowly. (It is even worse if α — involving Hessian describing curvature — contributes significantly to the description of f.)

Exercise 269.

Create and compute own examples. Draw contour graphs (e.g., using Matlab functions). Use quadratic bivariate functions to be able to compute univariate minima ($\lambda_{\min}$) easily (solving simple linear equation derived from $\theta'(\lambda) = 0$ instead of using the line search).

Remark 270 (Convergence rate)**.**

The theoretical convergence rate is linear. In practical cases it may be destroyed because of the zigzag effect. The reason is that although $-\nabla f(\mathbf{x}_k)$ is the steepest descent direction it is true only locally (linear approximation). And if the true descent direction turns out of it, it creates problems — see, e.g., Rosenbrock function computations (see Matlab demo banana function $f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$). Therefore, the line search allows just a short step.

Remark 271 (Terminology)**.**

Remember the difference between a global convergence property i.e. that for all $\mathbf{x}_1$ satisfying certain assumptions the sequence of $\mathbf{x}_k$ converges to Ω and convergence to a global minimum!

Exercise 272.

Why $\nabla f(\mathbf{x}_{k+1})$ and $\nabla f(\mathbf{x}_k)$ are orthogonal? Show using simple examples! Explain theoretically. Hint: Use that from $\theta(\lambda) = f(\mathbf{x}_k + \lambda \mathbf{d}_k)$ we obtain that $0 = \theta'(\lambda) = \mathbf{d}_k^\top \nabla f(\mathbf{x}_k + \lambda \mathbf{d}_k)$

Exercise 273.

Discuss why the steepest descent algorithm works badly for problems with ill conditioned functions of the narrow and alongated valley type.

Our discussion leads to the conclusion that from the global convergence viewpoint it si better to continue locally in non-steepest direction and slightly modify it, e.g. turn it.

In general we have two main possibilities. Either to use historical information (see the conjugate gradient method later) to update $\nabla f(\mathbf{x}_k)$ or to use the information about the function curvature obtained from the Hessian matrix of f denoted as $\mathbf{H}(\mathbf{x})$.

It is important to remember that the steepest descent (gradient) method uses only linear approximation (see use of $f'_{\mathbf{d}}$ previously). Therefore, only incomplete information about f is used for the direction choice.

Exercise 274.

Compare numerical behaviour for two problems: $\min\{x_1^2 + x_2^2\}$ and $\min\{100x_1^2 + x_2^2\}$. Draw contour graphs and compute eigenvalues. identify conditional numbers defined as the fraction maximal eigenvalue and minimal eigenvalue. And use the following formula for the estimation of the convergence rate for the steepest descent method:

$$|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_{\min})| \leq (\frac{\operatorname{cond}\mathbf{H}(\mathbf{x}_{\min} - 1)}{\operatorname{cond}\mathbf{H}(\mathbf{x}_{\min} + 1)})^2 |f(\mathbf{x}_k) - f(\mathbf{x}_{\min})|.$$

Compare bounds with results of iterations. The formula is related to β and the condition number of Hessian equals the fraction of its largest and smallest eigenvalue.

Exercise 275.

Explain the concept of linear convergence rate on the selected geometric sequence. Hint: Use the idea $r_{k+1} = qr_k$ where $r_k = |f(\mathbf{x}_k) - f(\mathbf{x}_{\min})|$.

Convergence rate analysis. Let $\{r_k\}$ be a sequence satisfying $r_k \longrightarrow \bar{r}$. It may be generated, e.g., by $f(\mathbf{x}_k)$ or $\alpha(\mathbf{x}_k)$ values. Let:

$$p_0 = \sup\{p \mid \limsup_{k \longrightarrow \infty} \frac{|r_{k+1} - \bar{r}|}{|r_k - \bar{r}|^p} = \beta < \infty\},$$

where β is a ratio of convergence (if it is smaller for the same p then faster convergence may occur), p is an order of convergence (larger p may also cause faster convergence). Remember that the formula above describes the limiting behaviour of the algorithm. See the following example:

Example. Consider the sequence defined by $r_{k+1} = (r_k + 1)/2$. For $k \longrightarrow \infty$, we get $r_k \longrightarrow 1$, so $\bar{r} = 1$. Then:

$$\lim_{k \longrightarrow \infty} \frac{|\frac{r_k + 1}{2} - 1|}{|r_k - 1|^1} = \lim_{k \longrightarrow \infty} \frac{1}{2} \frac{|r_k - 1|}{|r_k - 1|} = \frac{1}{2} < \infty$$

We may denote $e_k = r_k - \bar{r}$ and then $|e_k| = |r_k - \bar{r}|$ or in the vector case $\|\mathbf{e}_k\| = \|\mathbf{r}_k - \bar{\mathbf{r}}\|$. Without considering lim, we may write $|e_{k+1}| = \beta |e_k|^p$ or we have $|e_{k+1}| = \beta |e_k|$ for $p = 1$. Therefore, the geometric sequence can be taken as an example of the error decrease. Compare $1, 0.1, 0.01, 0.001, \ldots$ for $\beta = 0.1$ and $e_0 = 1$ with either $\beta = 0.01$ $(1, 0.01, 0.0001, 0.000001, \ldots)$ or $\beta = 0.99$ $(1, 0.99, \ldots)$. When $\beta > 1$ then the sequence diverges. With higher p (e.g. $p = 2$) we get faster convergence and error decrease $1, 0.1, 0.001, 0.0000001, \ldots$.

In our example, p cannot be equal to 2 because:

$$\lim_{k \longrightarrow \infty} \frac{|\frac{r_k + 1}{2} - 1|}{|r_k - 1|^2} = \lim_{k \longrightarrow \infty} \frac{1}{2} \frac{1}{|r_k - 1|} = \infty$$

1.10.2 Newton's method

While the gradient method works with a linear approximation of the objective f, this method utilizes its quadratic approximation (using the higher (second) derivative):

$$q(\mathbf{x}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^\top \mathbf{H}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k).$$

The necessary convergence condition $\nabla q(\mathbf{x}) = \mathbf{0}$ may by applied:

$$\mathbf{H}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k). \tag{1.4}$$

We simplify the formula to derive $\mathbf{x}$ explicitly from it and we get a new approximation $\mathbf{x}_{k+1}$, so:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k)\nabla f(\mathbf{x}_k).$$

Note that the formula satisfies $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$ principle, however, $\lambda_k = 1$ and $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)\mathbf{H}(\mathbf{x}_k)$.

Algorithm 276 (Newton).

Choose $\varepsilon > 0$, $\mathbf{x}_1$, $k := 1$.

1. Je-li $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$, STOP. Otherwise $\mathbf{d}_k := -\mathbf{H}^{-1}(\mathbf{x}_k)\nabla f(\mathbf{x}_k)$, and so, $\mathbf{x}_{k+1} := \mathbf{x}_k + \mathbf{d}_k$, $k := k + 1$ a GOTO 1.

Example 277.

Solve $\min\left[(x_1 - 2)^4 + (x_1 - 2x_2)^2\right]$ by Algorithm 276.

k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	$\nabla f(\mathbf{x}_k)$	$\mathbf{H}(\mathbf{x}_k)$	$\mathbf{H}^{-1}(\mathbf{x}_k)$	$\mathbf{d}_k$	$\mathbf{x}_{k+1}$
1	$(0,0;3,0)$ $52,0$	$(-44,0;24,0)$	$\begin{bmatrix} 50,0 & -4,0 \\ -4,0 & 8,0 \end{bmatrix}$	$\frac{1}{384}\begin{bmatrix} 8,0 & 4,0 \\ 4,0 & 50,0 \end{bmatrix}$	$(0,67;-2,67)$	$(0,67;0,33)$
2	$(0,67;0,33)$ $3,13$	$(-9,39;-0,04)$	$\begin{bmatrix} 23,2 & -4,0 \\ -4,0 & 8,0 \end{bmatrix}$	$\frac{1}{170}\begin{bmatrix} 8,0 & 4,0 \\ 4,0 & 23,2 \end{bmatrix}$	$(0,44;0,23)$	$(1,11;0,56)$
3	$(1,11;0,56)$ $0,63$	$(-2,84;-0,04)$	$\begin{bmatrix} 11,5 & -4,0 \\ -4,0 & 8,0 \end{bmatrix}$	$\frac{1}{76,0}\begin{bmatrix} 8,0 & 4,0 \\ 4,0 & 11,5 \end{bmatrix}$	$(0,30;0,14)$	$(1,41;0,70)$
4	$(1,41;0,70)$ $0,12$	$(-0,80;-0,04)$	$\begin{bmatrix} 6,18 & -4,0 \\ -4,0 & 8,0 \end{bmatrix}$	$\frac{1}{33,4}\begin{bmatrix} 8,0 & 4,0 \\ 4,0 & 6,18 \end{bmatrix}$	$(0,20;0,10)$	$(1,61;0,80)$

When the method converges then the theoretical convergence rate is quadratic. If the starting point is 'close enough' (see literature) to $\mathbf{x}_{\min}$ (to point satisfying $\nabla f(\mathbf{x}_{\min}) = \mathbf{0}$) and $r(\mathbf{H}(\mathbf{x}_{\min})) = n$ then it converges with order two (quadratic). Even one step convergence is guaranteed for strictly convex quadratic functions. However, it is impossible to verify theoretical sufficient convergence conditions during computations, so method may accidentally diverge. In practice, the method quickly converges nearly to minimum.

So, the important problems to be considered are:

1. $\mathbf{H}(\mathbf{x}_k)$ may be singular and the inverse matrix $\mathbf{H}^{-1}(\mathbf{x}_k)$ does not necessarily exist. The later idea will to replace $\mathbf{H}(\mathbf{x}_k)$ with a PD approximating matrix (e.g., Marquardt-Levenberg method).

2. For regular $\mathbf{H}(\mathbf{x}_k)$, the direction $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ is not necessarily descent, and so, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ is not guaranteed. Again the idea will be to use the aforementioned PD approximation.

3. If $\mathbf{x}_{k+1} - \mathbf{x}_k$ is not a descent step because of $\lambda_k = 1$ then we may introduce the formula with λ_k (cf. $\mathbf{x}_{k+1} := \mathbf{x}_k - \lambda_k \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$) and to use a trust region method.

4. As no global convergence property can be expected, we may use Newton's method near to optimum or to approximate $\mathbf{H}(\mathbf{x}_k)^{-1}$.

5. Regarding the storage requirements: $H(\mathbf{x}_k)$ needs the $n \times n$ table and $\mathbf{d}_k$ may be obtained from a system of linear equations $\mathbf{H}(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ instead of computing $\mathbf{H}^{-1}(\mathbf{x}_k)$ explicitly. There are also memory-less quasi-Newton methods discussed later.

6. Computational difficulties are based on computation of $\mathbf{H}(x)$, and then $\mathbf{H}(x_k)$ and the solution of SLEq to obtain $\mathbf{d}_k$. Again the approximation of $\mathbf{H}^{-1}(\mathbf{x}_k)$ may help.

We may also discuss the relation between the gradient and Newton methods (cf. $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$ and $\mathbf{x}_{k+1} := \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$). If $\mathbf{H}(\mathbf{x}_k)$ is PD then $\mathbf{H}^{-1}(\mathbf{x}_k)$ is also PD, and so $\nabla f(\mathbf{x}_k)^\top (-\mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)) < 0$. Therefore, $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$ is a descent direction here (cf. convex function properties!). If $\mathbf{H}(\mathbf{x}_k)^{-1}$ exists then $\exists \mathbf{L}$ (a lower triangular matrix with diagonal elements greater then zero) such that $\mathbf{H}(\mathbf{x}_k)^{-1} = \mathbf{L}\mathbf{L}^\top$ (Cholesky decomposition).

We further think about $\mathbf{x} = \mathbf{L}\mathbf{y}$ transformation. Then $f(\mathbf{x}) = f(\mathbf{L}\mathbf{y}) = F(\mathbf{y})$ ($F(\cdot) = F(\mathbf{L}\cdot)$ here) and $\mathbf{x}_k = \mathbf{L}\mathbf{y}_k$, $\mathbf{y}_k = \mathbf{L}^{-1}\mathbf{x}_k$ and $\nabla F(\mathbf{y}_k) = \mathbf{L}^\top \nabla f(\mathbf{x}_k)$. So, we may use the special (gradient method-like) iteration step for $\mathbf{y}$ space (after 'scaling'): $\mathbf{y}_{k+1} = \mathbf{y}_k - 1 \cdot \nabla F(\mathbf{y}_k) =$ (here $\lambda_k = 1$). Then, by substitution, $\mathbf{L}^{-1}\mathbf{x}_{k+1} = \mathbf{L}^{-1}\mathbf{x}_k - \mathbf{L}^\top \nabla f(\mathbf{x}_k)$, and finally, $\mathbf{x}_{k+1} = \mathbf{L}^{-1}\mathbf{x}_k - \mathbf{L}\mathbf{L}^\top \nabla f(\mathbf{x}_k)$ that is the original Newton method step $\mathbf{H}(\mathbf{x}_k)^{-1}$.

The previous discussions offer the idea to start with the gradient method and to finish with Newton's method (Marquardt-Levenberg method for LSQ nonlinear regression. $\mathbf{x}_{k+1}$ is derived from

$$(\mu_k \mathbf{I} + \mathbf{H}(\mathbf{x}_k))(\mathbf{x} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k),$$

where μ_k is the Marquardt parameter. μ_k is chosen such that all eigenvalues λ_i of matrix $\mathbf{M}_k = \mu_k \mathbf{I} + \mathbf{H}(\mathbf{x}_k)$ satisfy $\exists \delta : \lambda_i \geq \delta > 0$. So, the idea is to check whether $\mathbf{M}_k$ is PD trying $\mathbf{L}\mathbf{L}^\top$ factorization. If it is not PD then increase μ_k. If it is PD, solve $\mathbf{M}_k(\mathbf{x}_{k+1} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k)$ by a lower triangular factorization $\mathbf{L}\mathbf{L}^\top(\mathbf{x}_{k+1} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k)$. Precisely, compute $f(\mathbf{x}_{k+1})$, evaluate $R_k = \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{q(\mathbf{x}_{k+1}) - q(\mathbf{x}_k)}$ where $q(\mathbf{x})$ is the quadratic approximation of f at $\mathbf{x}$. Then, the following heuristic is utilized to update μ_k and $k := k + 1$:

$R_k < 0.25 \Rightarrow \mu_{k+1} := 4\mu_k$,

$R_k > 0.75 \Rightarrow \mu_{k+1} := 0.5\mu_k$,

$0.25 \leq R_k \geq 0.75 \Rightarrow \mu_{k+1} := \mu_k$,

$R_k < 0$ (no progress) $\Rightarrow$ restart $\mathbf{x}_k$.

So, positive eigenvalues imply the PD matrix that is regular, and so invertible. Then, the iteration formula $\mathbf{x}_{k+1} := \mathbf{x}_k - \mathbf{M}_k^{-1} \nabla f(\mathbf{x}_k)$ uses descent directions (as $0 > \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k = -\nabla f(\mathbf{x}_k) \mathbf{M}_k^{-1} \nabla f(\mathbf{x}_k)$).

Trust region methods use the formula $\mathbf{x}_{k+1} := \mathbf{x}_k - \lambda_k \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$ and λ_k is chosen such that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ is not too large, so the step is restricted. This region ('region of trust') is chosen such that the $q(\mathbf{x})$ (quadratic) approximation at $\mathbf{x}_k$ is reliable, so R_k is evaluated. If R_k is small then Δ_k is decreased. If R_k is big (near to 1) then Δ_k is expanded. Then $\mathbf{x}_{k+1}$ is found:

$$\mathbf{x}_{k+1} \in \operatorname{argmin}_{\mathbf{x}} \{q(\mathbf{x}) \mid \mathbf{x} \in \Omega = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_k\| < \Delta_k\}\}.$$

For the discussion of the method relation to $\mathbf{x}_{k+1} := \mathbf{x}_k - \lambda_k \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$ (also the case of $\mathbf{M}_k^{-1}$ used instead of $\mathbf{H}(\mathbf{x}_k)^{-1}$) or another ways to compute $\mathbf{x}_{k+1}$, like the dog-leg trajectory method or PARTAN method — see literature.

1.10.3 Conjugate directions

Improvement of the gradient method. The question is how to improve the convergence of ∇f method? The idea is based on several facts: At $\mathbf{x}_{\min}$, we have $\nabla f(\mathbf{x}_{mmin}) = \mathbf{0}$ and for Taylor series we get $q(\mathbf{x}) = f(\mathbf{x}_{\min}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{\min})^\top \mathbf{H}(\mathbf{x}_{\min})(\mathbf{x} - \mathbf{x}_{\min})$ and $\mathbf{H}(\mathbf{x}_{\min})$ that is PD (positive definite). Then f behaves in the $\mathcal{N}_\varepsilon(\mathbf{x}_{\min})$ like a strictly convex quadratic function! So, a general method to be efficient should quickly converge on quadratic functions, otherwise it is slow near $\mathbf{x}_{\min}$.

Definition 278 (D-conjugacy)**.**

Let $\mathbf{D}$ be a $n \times n$ symmetric matrix. Vectors $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are called conjugate (with respect to $\mathbf{D}$) iff they are linearly independent and $\forall i, j = 1, \ldots, n \; i \neq j \; \mathbf{d}_i^\top \mathbf{D} \mathbf{d}_j = 0$.

Remark 279 (Separability achieved by conjugate directions)**.**

Let us show the importance of conjugate directions. Assume that we minimize a quadratic function $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{D}\mathbf{x}$ ($\mathbf{D}$ symmetric). Then, because of linear independence of $\mathbf{d}_1, \ldots, \mathbf{d}_n$, we may express arbitrary $\mathbf{x} \in \mathbb{R}^n$ as a sum of the initial point $\mathbf{x}_1$ and linear combination of conjugate directions i.e. $\mathbf{x} = \mathbf{x}_1 + \sum_{j=1}^n \lambda_j \mathbf{d}_j$. We substitute it in $f(\mathbf{x})$ and we obtain a function of variables $\lambda_1, \ldots, \lambda_n$ denoted as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$, so

$$F(\boldsymbol{\lambda}) = \mathbf{c}^\top (\mathbf{x}_1 + \sum_{j=1}^n \lambda_j \mathbf{d}_j) + \frac{1}{2}(\mathbf{x}_1 + \sum_{j=1}^n \lambda_j \mathbf{d}_j)^\top \mathbf{D} (\mathbf{x}_1 + \sum_{j=1}^n \lambda_j \mathbf{d}_j).$$

Exercise 280.

Simplify and show that because of conjugate directions F is separable in $\lambda_1, \ldots, \lambda_n$. Answer why it can be helpful. Hint:

$$F(\boldsymbol{\lambda}) = \mathbf{c}^\top \mathbf{x}_1 + \tfrac{1}{2}\mathbf{x}_1^\top \mathbf{D}\mathbf{x}_1 + \sum_{j=1}^n \lambda_j (\mathbf{c}^\top + \mathbf{x}_1^\top \mathbf{D})\mathbf{d}_j + \tfrac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \mathbf{d}_i^\top \mathbf{D} \lambda_j \mathbf{d}_j =$$
$$\mathbf{c}^\top \mathbf{x}_1 + \tfrac{1}{2}\mathbf{x}_1^\top \mathbf{D}\mathbf{x}_1 + \sum_{j=1}^n (\lambda_j (\mathbf{c}^\top + \mathbf{x}_1^\top \mathbf{D})\mathbf{d}_j + \tfrac{1}{2}\lambda_j^2 \mathbf{d}_j^\top \mathbf{D}\mathbf{d}_j).$$

Think about multivariate minimization realized by repeated independent univariate minimizations for conjugate directions to $\lambda_{j,\min}$, and hence, $\mathbf{x}_{\min}$). Hint: For positive definite $\mathbf{D}$, we have (using derivatives $F'_{\lambda_j}(\lambda_j) = 0$):

$$\lambda_{j,\min} = \frac{-(\mathbf{c}^\top \mathbf{d}_j + \mathbf{x}_1^\top \mathbf{D}\mathbf{d}_j)}{\mathbf{d}_j^\top \mathbf{D}\mathbf{d}_j}$$

So, $\lambda_{j,\min}, j = 1, \ldots, n$ are explicitly computed and repeatedly used. Then, the sum is computed. It is similar when f is sequentially minimized from $\mathbf{x}_1$ along directions $\mathbf{d}_j, j = 1, \ldots, n$.

Notice that conjugate directions may be used also in algorithms without derivatives.

Example 281.

Solve $\min(-12x_2 + 4x_1^2 + 4x_2^2 - 4x_1x_2)$. Derive Hessian matrix $\mathbf{H}$ and choose $\mathbf{d}_1^\top = (1;0)$. If $\mathbf{d}_2^\top = (a;b)$ then $0 = \mathbf{d}_1^\top \mathbf{H}\mathbf{d}_2 = 8a - 4b$. Therefore, $\mathbf{d}_2$ is not unique. Select $\mathbf{d}_2^\top = (1;2)$. Minimize f from $\mathbf{x}_1^\top = (-0,5;1)$ along $\mathbf{d}_1$ and get $\mathbf{x}_2^\top = (0,5;1)$. Continue along $\mathbf{d}_2$ and get $\mathbf{x}_{\min} = \mathbf{x}_3^\top = (1;2)$.

Emphasize that minimization of quadratic function using conjugate directions require (only) n times realized univariate minimization.

Conjugate directions for quadratic functions: We may use the following algorithm: At first $j := 1$. Main step: Start at $\mathbf{x}_j$, follow $\mathbf{d}_j$ to obtain $\lambda_{j,\min}$ and $\mathbf{x}_{j+1}$. Then, $j := j + 1$. If $j = n + 1$ then $\boxed{\text{STOP}}$ and $\mathbf{x}_{\min} := \mathbf{x}_{n+1}$, otherwise repeat the main step.

Theorem: Let $f(\mathbf{x}) = \mathbf{c}^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\mathbf{D}\mathbf{x}$ (and minimum exists) and $\mathbf{D}$ is $n \times n$ symmetric matrix, $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are $\mathbf{D}$-conjugate, $\mathbf{x}_1 \in \mathbb{R}^n$ is a starting point. Let $\forall k = 1, \ldots, n : \lambda_{k,\min} \in \operatorname{argmin}\{f(\mathbf{x}_k + \lambda_k\mathbf{d}_k) \mid \lambda_k \in \mathbb{R}\}$ where $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_{k,\min}\mathbf{d}_k$. Then $\forall k = 1, \ldots, n$:

1. $\nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_j = 0, j = 1, \ldots, k.$
2. $\nabla f(\mathbf{x}_1)^\top \mathbf{d}_k = \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k.$
3. $\mathbf{x}_{k+1} \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} - \mathbf{x}_1 \in \mathcal{L}(\mathbf{d}_1, \ldots, \mathbf{d}_k)\}$ where $\mathcal{L}(\mathbf{d}_1, \ldots, \mathbf{d}_k) = \{\sum_{j=1}^k \mu_j\mathbf{d}_j \mid forall j : mu_j \in \mathbb{R}\}$ and $\mathbf{x}_{n+1} \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$.

Proof: (1) We already know it (cf. gradient method and use $F'(\boldsymbol{\lambda})$; $\nabla f(\mathbf{x}_{k+1}) = \nabla f(\mathbf{x}_k + \lambda_k\mathbf{d}_k)$). (2) Use conjugate directions — see literature. (3) The technical proof — see literature. □
Exercise: *Interpret theorem geometrically.*

Explanatory remarks: Conjugate direction method converges finitely in at most n steps in the case of minimizing convex quadratic function $f(\mathbf{x}) = \mathbf{c}^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\mathbf{D}\mathbf{x}$. We have $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k\mathbf{d}_k$ ($\lambda_k \in \operatorname{argmin}\{f(\mathbf{x}_k + \lambda\mathbf{d}_k) \mid \lambda \in \mathbb{R}\}$) and $\mathbf{x}_{n+1} := \mathbf{x}_1 + \sum_{j=1}^n \lambda_j\mathbf{d}_j$. We know that $\nabla f(\mathbf{x}_{n+1}) = \mathbf{0}$ (minimum achieved). Then, $\nabla f(\mathbf{x}_{n+1}) = \mathbf{D}\mathbf{x}_{n+1} + \mathbf{c} = \mathbf{0}$ and from the theorem $0 = \nabla f(\mathbf{x}_{k+1})^\top \mathbf{d}_k = (\mathbf{D}\mathbf{x}_{k+1} + \mathbf{c})^\top \mathbf{d}_k = ((\mathbf{D}(\mathbf{x}_k) + \lambda_k\mathbf{d}_k))^\top \mathbf{d}_k + \mathbf{c}^\top\mathbf{d}_k \Rightarrow \lambda_k = -\frac{(\mathbf{D}\mathbf{x}_k + \mathbf{c})^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k}$ and for $\mathbf{x}_k = \mathbf{x}_1 + \sum_{j=1}^{k-1} \lambda_j\mathbf{d}_j$ we obtain $\mathbf{d}_k^\top\mathbf{D}\mathbf{x}_k = \mathbf{d}_k^\top\mathbf{D}\mathbf{x}_1 + \sum_{j=1}^{k-1} \lambda_j\mathbf{d}_k^\top\mathbf{D}\mathbf{d}_j = \mathbf{d}_k^\top\mathbf{D}\mathbf{x}_1$, and hence,

$$\lambda_k = -\frac{\mathbf{d}_k^\top(\mathbf{D}\mathbf{x}_1 + \mathbf{c})}{\mathbf{d}_k^\top\mathbf{D}\mathbf{d}_k}.$$

So, only one question remains: How to generate conjugate directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$? We have certain degrees of freedom to choose them. Therefore, two classes of algorithms may be derived: (1) conjugate gradients, (2) quasi-Newton methods. Or from the practitioner's viewpoint, we may consider two basic ideas (1) cross-country skiing with slow turns (based on averaging of $\nabla f(\mathbf{x}_k)$ — conjugate gradients) and (2) downhill skiing with the experienced (risk-less) skier (based on approximation of $\mathbf{H}(\mathbf{x}_k)^{-1}$ — quasi-Newton).

1.10.4 Conjugate gradient methods

Remark 282 (Basic idea)**.**

Because of problems with the gradient method, methods of fixed metrics (conjugate gradients) using the information from previous iterations to modify the current gradient direction have been developed.

Algorithm for quadratic functions: We further discuss the conjugate gradient method (related references are Hestens and Stiffel 1952 who developed the method for systems of linear equations and later in 1964 Fletcher and Reeves applied it to optimization problems). These methods are often called fixed metric methods because of its relation to metric spaces.

We have a quadratic function $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{D}\mathbf{x}$. The idea is to construct $\mathbf{d}_1, \dots \mathbf{d}_n$ $\mathbf{D}$-conjugate directions combining (linearly) $\mathbf{x}_k$, $\nabla f(\mathbf{x}_k)$, and $\{\mathbf{d}_j\}_{j=1}^{k-1}$. We denote $\mathbf{g}_k = \nabla f(\mathbf{x}_k) = \mathbf{D}\mathbf{x}_k + \mathbf{c}$.

1. Set $\mathbf{x}_1$, compute $\mathbf{g}_1 = \nabla f(\mathbf{x}_k) = \mathbf{D}\mathbf{x}_1 + \mathbf{c}$, assign $\mathbf{d}_1 := -\mathbf{g}_1$ and $k := 1$

2. Compute $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$ where $\lambda_k = -\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k}$. Update $\mathbf{d}_{k+1} := -\mathbf{g}_{k+1} + \alpha_k \mathbf{d}_k$ where $\alpha_k = \frac{\mathbf{g}_{k+1}^\top \mathbf{D}\mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k}$. If $k = n+1$ then $\mathbf{x}_{n+1}$ is $\mathbf{x}_{\min}$ and STOP. Otherwise, $k = k+1$ and GOTO 2.

It is important to know that (for the quadratic function and existing minimum) the conjugate gradient algorithm achieves the minimum (converges) after one complete main step (n univariate minimizations — line searches). The following theorem answers why:

Theorem: Let $f(\mathbf{x})$ be a quadratic function, $\mathbf{x}_1 \in \mathbb{R}^n$ and the conjugate gradients algorithm is used. Then:

1. $\mathbf{d}_1, \dots, \mathbf{d}_n$ generated by the algorithm are $\mathbf{D}$-conjugate directions.
2. $\mathbf{d}_1, \dots, \mathbf{d}_n$ are descent directions.
3. $\alpha_k = \frac{\mathbf{d}_k^\top \mathbf{D}\nabla f(\mathbf{x}_{k+1})}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k} = \frac{\mathbf{g}_{k+1}^\top \mathbf{g}_{k+1}}{\mathbf{g}_k^\top \mathbf{g}_k} = \frac{\|\nabla f(\mathbf{x}_{k+1})\|^2}{\|\nabla f(\mathbf{x}_k)\|^2} = \frac{\mathbf{g}_{k+1}^\top (\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{g}_k^\top \mathbf{g}_k}$.

Remark: We already know that when optimum has not yet been reached then

$$\lambda_k = \frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k} = -\frac{\mathbf{g}_k^\top \mathbf{d}_k}{\mathbf{d}_k \mathbf{D}\mathbf{d}_k} = -\frac{\mathbf{g}_k^\top (-\mathbf{g}_k + \alpha_{k-1}\mathbf{d}_{k-1})}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k}$$

and because of theorem and conjugate directions, we know that $\mathbf{g}_k^\top \mathbf{d}_{k-1} = 0$.

Proof: For readers: The proof is rather technical, just check the main ideas and improve the ability to do matrix calculations! Be able to do small computational simple steps.

$$\begin{aligned}
\mathbf{g}_{k+1} - \mathbf{g}_k &= \mathbf{D}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \lambda_k \mathbf{D}\mathbf{d}_k \\
\mathbf{g}_{k+1}\mathbf{D}\mathbf{d}_k &= \frac{1}{\lambda_k}\mathbf{g}_{k+1}^\top (\mathbf{g}_{k+1} - \mathbf{g}_k) \\
\mathbf{g}_{k+1}\mathbf{D}\mathbf{d}_k &= \frac{1}{\frac{\mathbf{g}_k^\top \mathbf{g}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k}}\mathbf{g}_{k+1}^\top (\mathbf{g}_{k+1} - \mathbf{g}_k) \\
\alpha_k = \frac{\mathbf{g}_{k+1}\mathbf{D}\mathbf{d}_k}{\mathbf{d}_k^\top \mathbf{D}\mathbf{d}_k} &= \frac{\mathbf{g}_{k+1}^\top (\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{g}_k^\top \mathbf{g}_k} = \frac{\mathbf{g}_{k+1}^\top \mathbf{g}_{k+1}}{\mathbf{g}_k^\top \mathbf{g}_k}
\end{aligned}$$

because $\mathbf{g}_k = \mathbf{d}_k - \beta_{k-1}\mathbf{d}_{k-1} \in \mathcal{L}(\mathbf{d}_1, \ldots, \mathbf{d}_k)$ and $\mathbf{g}_{k+1} \perp \mathcal{L}(\mathbf{d}_1, \ldots, \mathbf{d}_k)$, so $\mathbf{g}_{k+1}^\top \mathbf{g}_k = 0$. To prove 1. and 2., we use induction to move from $\mathbf{d}_1, \ldots, \mathbf{d}_k$ to $\mathbf{d}_{k+1}$. We obtain 1. $\mathbf{d}_{k+1}^\top \mathbf{D}\mathbf{d}_k = 0$ using substitution $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \alpha_k \mathbf{d}_k$ (and again α_k is replaced). Similarly, $\mathbf{d}_{k+1}^\top \mathbf{D}\mathbf{d}_j = 0, j = 1, \ldots, k$ by induction and recursion. □

Remark 283 (Convergence).

Under quite general assumptions, conjugate gradient methods converge superlinearly.

Algorithm 284 (Fletcher-Reeves).

Choose $\varepsilon > 0$, $\mathbf{x}_1$, $\mathbf{d}_1 := -\nabla f(\mathbf{x}_1)$. Assign $\mathbf{y}_1 := \mathbf{x}_1$ ($\mathbf{d}_1 = -\mathbf{g}_1 = -\nabla f(\mathbf{x}_1) = \nabla f(\mathbf{y}_1)$), $j := 1$, $k := 1$.

1. If $\|\nabla f(\mathbf{y}_j)\| < \varepsilon$ then optimum is reached and STOP, otherwise solve $\min\{f(\mathbf{y}_j + \lambda \mathbf{d}_j) \mid \lambda \geq 0\}$. Solution λ_j is used to compute $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j = n$ then GOTO 3.. In opposite case, when $j < n$, continue GOTO 2..

2. $\mathbf{d}_{j+1} := -\nabla f(\mathbf{y}_{j+1}) + \alpha_j^{\mathrm{FR}} \mathbf{d}_j$, where

$$\alpha_j^{\mathrm{FR}} = \frac{\|\nabla f(\mathbf{y}_{j+1})\|^2}{\|\nabla f(\mathbf{y}_j)\|^2}.$$

   Then $j := j + 1$ and GOTO 1.

3. Define $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}$ (i.e. $= \mathbf{y}_{j+1}$) and $\mathbf{y}_1 := \mathbf{x}_{k+1}$. Further set up $\mathbf{d}_1 := -\nabla f(\mathbf{y}_1)$, $k := k + 1$, $j := 1$, and GOTO 1.

We see that the algorithm is based on the following update of the descent direction: $\mathbf{d}_{j+1} := -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$. The coefficient α_j is chosen in such a way that directions $\mathbf{d}_j$ are conjugate with respect to Hessian matrix $\mathbf{H}$ in the case of the quadratic objective function. Other modifications of conjugate gradient methods are obtained for other choices of α_j. Hestens and Stiefel suggested for $\mathbf{q}_j = \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)$ (also further used in Algorithm 288):

$$\alpha_j^{\mathrm{HS}} = \frac{\nabla f(\mathbf{y}_{j+1})^\top \mathbf{q}_j}{\mathbf{d}_j^\top \mathbf{q}_j}$$

and Polak and Ribiere suggested to specify

$$\alpha_j^{\mathrm{PR}} = \frac{\nabla f(\mathbf{y}_{j+1})^\top \mathbf{q}_j}{\|\nabla f(\mathbf{y}_j)\|^2}.$$

Exercise: *You may show that for quadratic functions $f(\mathbf{x})$, these choices α_j^{HS}, α_j^{PR}, and α_j^{FR} are equivalent.*

Example 285.

Solve $\min\left[(x_1-2)^4+(x_1-2x_2)^2\right]$ by Algorithm 284.

k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $\mathbf{f}(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\|\nabla f(\mathbf{y}_j)\|$	α_1	$\mathbf{d}_j$	λ_j
1	(0,00; 3,00) 52,00	1	(0,00; 3,00) 52,00	(−44,00; 24,00)	50,12		(44,00; −24,00)	0,062
		2	(2,70; 1,51) 0,34	(0,73; 1,28)	1,47	0,0009	(−0,69; −1,30)	0,23
2	(2,54; 1,21) 0,10	1	(2,54; 1,21) 0,10	(0,87; −0,48)	0,99		(−0,87; 0,48)	0,11
		2	(2,44; 1,26) 0,04	(0,18; 0,32)	0,37	0,14	(−0,30; −0,25)	0,63
3	(2,25; 1,10) 0,008	1	(2,25; 1,10) 0,008	(0,16; −0,20)	0,32		(−0,16; 0,20)	0,10
		2	(2,23; 1,12) 0,003	(0,03; 0,04)	0,05	0,04	(−0,036; −0,032)	1,02
4	(2,19; 1,09) 0,0017	1	(2,19; 1,09) 0,0017	(0,05; −0,04)	0,06		(−0,05; 0,04)	0,11
		2	(2,185; 1,094) 0,0012	(0,02; 0,01)	0,02			

Practical implementation of conjugate gradient method requires to discuss the restart possibilities (see 3. in the FR algorithm), the update of Hessian matrix, and the computations of $\nabla f(\mathbf{x})$.

Remark (Convergence): Step 3. (outer loop) ensures the global convergence property that is a consequence of the global convergence property of the ∇f method. The influence of the inner loop inserted step 2. will be discussed later together with the DFP method. See Bazaraa-Shetty book for details and the restart idea.

Remark (Computational properties): The conjugate gradient methods have useful computational properties as they have modest memory requirements — it is enough to save 3 n-dimensional vectors.

Remark (Convergence rate): Even convergence rate is superior to the rate of ∇f method. For the gradient method, we had:

$$\limsup_{k\longrightarrow\infty}\frac{f(\mathbf{x}_{k+1})-f(\mathbf{x}_{\min})}{f(\mathbf{x}_k)-f(\mathbf{x}_{\min})}=\beta\leq\left(\frac{\operatorname{cond}\mathbf{H}(\mathbf{x}_{\min})-1}{\operatorname{cond}\mathbf{H}(\mathbf{x}_{\min})+1}\right)^2=\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2.$$

So, for $f(\mathbf{x})=x_1^2+100x_2^2$, we obtain $\mathbf{H}=\begin{pmatrix}1&0\\0&100\end{pmatrix}$ and eigenvalues $\lambda_{\min}=1$ and $\lambda_{\max}=100$. It implies that $\beta\leq\left(\frac{99}{101}\right)^2$. If you think about the related sequence $e_{k+1}=\beta e_k$ where $e_0=1$ and $\beta=0.99$ we see the slow decrease of error e_{k+1}! In contrast, the conjugate gradient method has a superlinear convergence as $\frac{\|\mathbf{x}_{k+1}-\mathbf{x}_{\min}\|}{\|\mathbf{x}_k-\mathbf{x}_{\min}\|}\longrightarrow 0$. (remember that for the algorithm step from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$ we need n single univariate minimization steps generating $\mathbf{y}_2,\ldots,\mathbf{y}_{n+1}$!). Under Lipschitz continuity property (If $\exists K\ \forall\mathbf{d}\in\mathbb{R}^n,\forall\mathbf{x}\in\mathcal{N}_\varepsilon(\mathbf{x}_{\min}):\|\mathbf{H}(\mathbf{x})\mathbf{d}\|\leq K\cdot\|\mathbf{d}\|$) we even obtain a quadratic convergence: $\limsup\frac{\mathbf{x}_{k+1}-\mathbf{x}_{\min}\|}{\|\mathbf{x}_k-\mathbf{x}_{\min}\|^2}<\infty$. Theoretically, this is true even with inexact line searches. The practical considerations show that $2n$ iterations are often enough. (Bazaraa-Shetty also discuss how eigenvalues have smaller influence than in the case of gradient method.)

1.10.5 Quasi-Newton methods

Remark 286 (Ideas and properties)**.**

These methods use conjugate directions. They are also called variable metric methods (cf. fixed metric methods — conjugate gradients). The main idea is to express the descent direction in the form $\mathbf{d}_j = -\mathbf{D}_j \nabla f(\mathbf{x})$, where $\mathbf{D}_j$ is a symmetric positive definite matrix, approximating inverse matrix $\mathbf{H}^{-1}(\mathbf{x})$ (the reason for the name quasi-Newton methods). These methods converge under quite general assumptions. The convergence rate is superlinear.

Remark (Conjugate directions by $\mathbf{H}(\mathbf{x}_k)^{-1}$ approximation): The main idea is to deflect the gradient direction $\nabla f(\mathbf{y}_j)$ (cf. conjugate gradients and the inner loop specification) by $\mathbf{D}_j$ $n \times n$ PD approximation of $\mathbf{H}(y_j)^{-1}$. As we already know, with $\mathbf{D}_j$ PD then $\mathbf{d}_j = -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ is a descent direction (for $\nabla f(\mathbf{y}_j) \neq \mathbf{0}$) because $\mathbf{d}_j^\top \nabla f(\mathbf{y}_j) = -\nabla f(\mathbf{y}_j) \mathbf{D}_j \nabla f(\mathbf{y}_j) < 0$.

Remark (The simplest $\mathbf{D}_j$ update): For the next step $\mathbf{D}_{j+1}$ is derived as $\mathbf{D}_j + \mathbf{C}_j$ (the addition is the simplest full matrix update). The form $\mathbf{C}_j$ specifies the quasi-Newton's method.

Remark (Rank 1 method): The computationally simplest form of $\mathbf{C}_j$ is $\alpha_j \mathbf{u}_j \mathbf{u}_j^\top$. Why? It needs only $n + n^2$ multiplications. The rank of $\mathbf{C}_j$ is 1 (just formed from multiplications of the same row $\mathbf{u}_j$).

Algorithm 287 (Rank 1)**.**

Choose $\varepsilon > 0$, starting point $\mathbf{x}_1$, symmetric, positive definite matrix $\mathbf{D}_1$ (often $\mathbf{D}_1 := \mathbf{I}$). Then we assign $\mathbf{y}_1 := \mathbf{x}_1$, $j := 1$, $k := j$.

1. If $\|\nabla f(\mathbf{y}_j)\| < \varepsilon$ then STOP (and minimum is found), otherwise $\mathbf{d}_j := -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ and solve $\min\{f(\mathbf{y}_j + \lambda \mathbf{d}_j) \mid \lambda \geq 0\}$. Solution λ_j is used to compute $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j = n$ then $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}$, $\mathbf{y}_1 := \mathbf{x}_{k+1}$, $k := k + 1$, $j := 1$ (D_j is given and guarantees the restart) and GOTO 1. Otherwise, when $j < n$, we continue with GOTO 2.

2. Assign $\mathbf{D}_{j+1} := \mathbf{D}_j + \mathbf{C}_j$, where

$$\begin{aligned} \mathbf{C}_j = \mathbf{C}_j^{\mathbf{R1}} &:= \frac{(\mathbf{p}_j - \mathbf{D}_j \mathbf{q}_j)(\mathbf{p}_j - \mathbf{D}_j \mathbf{q}_j)^\top}{\mathbf{q}_j^\top (\mathbf{p}_j - \mathbf{D}_j \mathbf{q}_j)} \\ \mathbf{p}_j &:= \mathbf{y}_{j+1} - \mathbf{y}_j \quad (= \lambda_j \mathbf{d}_j), \\ \mathbf{q}_j &:= \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j), \end{aligned}$$

$j := j + 1$, and GOTO 1.

Note that $\mathbf{C}_j^{\mathbf{R1}} = \alpha \mathbf{u}_j \mathbf{u}_j^\top$ because $\mathbf{u}_j = \mathbf{p}_j - \mathbf{D}_j \mathbf{q}_j$ and $\alpha_j = (\mathbf{q}_j^\top (\mathbf{p}_j - \mathbf{D}_j \mathbf{q}_j))^{-1}$.

Remark (How it was derived?): For $f(\mathbf{x}) = \mathbf{c}^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x}$ ($\mathbf{H}$ is PD, so regular), we have $\nabla f(\mathbf{x}) = \mathbf{H}\mathbf{x} + \mathbf{c}$, and for $\mathbf{y}_{j+1}$, $\mathbf{y}_j$, we obtain $\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j) = \mathbf{H}(\mathbf{x}_{j+1} - \mathbf{x}_j) \Rightarrow \mathbf{q}_j = \mathbf{H}\mathbf{p}_j$ (as $\mathbf{q}_j = \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)$ and $\mathbf{p}_j = \mathbf{x}_{j+1} - \mathbf{x}_j$). It also implies that $\mathbf{p}_j = \mathbf{H}^{-1}\mathbf{q}_j$. It must be also satisfied (in certain form) by $\mathbf{D}_{j+1}$ to guarantee the algorithm nice behaviour in the case of $f(\mathbf{x})$ quadratic (or near to quadratic, e.g., in the neighbourhood of the minimum).

At first, we use $\mathbf{C}_j = \alpha_j\mathbf{u}_j\mathbf{u}_j^\top$ (because it is easily computable and symmetric, so $\mathbf{D}_{j+1}$ is a sum of two symmetric matrices, and hence also symmetric).

We consider the idea to replace property $\mathbf{p}_j = \mathbf{H}^{-1}\mathbf{q}_j$ by $\mathbf{p}_j = \mathbf{D}_{j+1}\mathbf{q}_j$ specifying $\mathbf{D}_{j+1}$ update. So: $\mathbf{D}_{j+1}\mathbf{q}_j = (D_j + \mathbf{C}_j)\mathbf{q}_j = (\mathbf{D}_j + \alpha_j\mathbf{u}_j\mathbf{u}_j^\top)\mathbf{q}_j = \mathbf{p}_j$. So $\alpha_j\mathbf{u}_j\mathbf{u}_j^\top\mathbf{q}_j = \mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j$.

We may also multiply the equality by $\mathbf{q}_j^\top$ from the left and obtain: $\mathbf{q}_j^\top\mathbf{D}_j\mathbf{q}_j + \alpha_j(\mathbf{q}_j^\top\mathbf{u}_j)(\mathbf{u}_j^\top\mathbf{q}_j) = \mathbf{q}_j^\top\mathbf{p}_j$. We may express the term containing α_j. So: $\alpha_j(\mathbf{q}_j^\top\mathbf{u}_j)(\mathbf{u}_j^\top\mathbf{q}_j) = \mathbf{q}_j^\top\mathbf{p}_j - \mathbf{q}_j^\top\mathbf{D}_j\mathbf{q}_j$. Therefore, $\alpha_j(\mathbf{q}_j^\top\mathbf{u}_j)^2 = \mathbf{q}_j^\top(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)$.

Now we return back to $\mathbf{C}_j$ (remember its symmetry!).

$$\mathbf{C}_j = \alpha_j\mathbf{u}_j\mathbf{u}_j^\top = \frac{\alpha_j^2\mathbf{u}_j\mathbf{u}_j^\top}{\alpha_j} = \frac{\alpha_j\mathbf{u}_j\alpha_j\mathbf{u}_j^\top}{\alpha_j} = \frac{(\mathbf{u}_j^\top\mathbf{q}_j)\alpha_j\mathbf{u}_j\alpha_j\mathbf{u}_j^\top}{\mathbf{q}_j^\top\mathbf{u}_j\alpha_j} = \frac{\alpha_j\mathbf{u}_j(\mathbf{u}_j^\top\mathbf{q}_j)\alpha_j\mathbf{u}_j^\top(\mathbf{q}_j^\top\mathbf{u}_j)}{\alpha_j\mathbf{q}_j^\top\mathbf{u}_j\mathbf{u}_j^\top\mathbf{q}_j} =$$
$$\frac{\alpha_j(\mathbf{u}_j\mathbf{u}_j^\top)\mathbf{q}_j\alpha_j(\mathbf{q}_j^\top\mathbf{u}_j)\mathbf{u}_j^\top}{\alpha_j(\mathbf{u}_j^\top\mathbf{q}_j)^\top\mathbf{u}_j^\top\mathbf{q}_j} = \frac{\alpha_j(\mathbf{u}_j\mathbf{u}_j^\top)\mathbf{q}_j\alpha_j\mathbf{q}_j^\top(\mathbf{u}_j\mathbf{u}_j^\top)}{\alpha_j(\mathbf{u}_j^\top\mathbf{q}_j)^\top\mathbf{u}_j^\top\mathbf{q}_j} = \frac{\alpha_j\mathbf{u}_j\mathbf{u}_j^\top\mathbf{q}_j(\alpha_j\mathbf{u}_j\mathbf{u}_j^\top\mathbf{q}_j)^\top}{\alpha_j(\mathbf{u}_j^\top\mathbf{q}_j)^2} = \frac{(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)^\top}{\mathbf{q}_j^\top(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)}.$$

The method converges in n (inner) steps if $\mathbf{C}_j$ is PD and f is a quadratic function. Even exact line search is not required!

However, there are the following problems: $\mathbf{D}_j$ is not necessarily PD and denominator may approach 0.

Remark (Rank 2 method): Because rank 1 method is not so good, rank 2 methods are further used:

$$\mathbf{C}_j := \alpha_j\mathbf{u}_j\mathbf{u}_j^\top + \beta_j\mathbf{v}_j\mathbf{v}_j^\top.$$

By the form of α_j, $\mathbf{u}_j$, β_j, and $\mathbf{v}_j$, you get various methods. We discuss DFP (Davidon-Fletcher-Powell: good but sensitive to the precision of line search method) and BFGS (Broyden-Fletcher-Goldfarb-Shanno: today the best choice that is robust enough) algorithms.

We have $\mathbf{q}_j = \mathbf{H}\mathbf{p}_j$ and $\mathbf{p}_j = \mathbf{H}^{-1}\mathbf{q}_j$. Building new matrices $\mathbf{Q}$ (with columns $\mathbf{q}_j, j = 1, \ldots, n$) and $\mathbf{P}$ (with columns $\mathbf{p}_j, j = 1, \ldots, n$), we may write $\mathbf{Q} = \mathbf{H}\mathbf{P}$ and $\mathbf{P} = \mathbf{H}^{-1}\mathbf{Q}$ or even $\mathbf{H} = \mathbf{Q}\mathbf{P}^{-1}$ and $\mathbf{H}^{-1} = \mathbf{P}\mathbf{Q}^{-1}$. And this the key idea. We want to obtain $\mathbf{D}_{n+1}\mathbf{H} = \mathbf{I}$ (to have a precise approximation of $\mathbf{H}^{-1}$ at minimum). Then $\mathbf{D}_{n+1}\mathbf{H}\mathbf{p}_j = \mathbf{p}_j, j = 1, \ldots, n$. We also want $\mathbf{D}_{j+1}\mathbf{H}\mathbf{p}_k = \mathbf{p}_k, k = 1, \ldots, j$. So, $\mathbf{p}_k$ are linearly independent eigenvectors (they form the matrix $\mathbf{P}$) of $\mathbf{D}_{j+1}\mathbf{H}$ with eigenvalues equal to 1 (see above). This gives the idea how to derive $\mathbf{C}_j^{\mathrm{DFP}}$. We know the requirements: $Q = \mathbf{H}\mathbf{P}$, $\mathbf{P}$ columns must be $\mathbf{H}$-conjugate and linearly independent. In addition, $\mathbf{p}_k, j = 1, \ldots, k$ are eigenvectors of $\mathbf{D}_{j+1}\mathbf{H}$ with eigenvalues equal to 1 (as $\mathbf{D}_{j+1}\mathbf{q}_k = \mathbf{p}_k$ and $\mathbf{D}_j\mathbf{H}\mathbf{p}_k = \mathbf{p}_k, k = 1, \ldots, j-1$). So, $\mathbf{p}_k = \mathbf{D}_j\mathbf{q}_k + \mathbf{C}_j\mathbf{q}_k = \mathbf{D}_j\mathbf{H}\mathbf{p}_k + \mathbf{C}_j\mathbf{q}_k = \mathbf{p}_k + \mathbf{C}_j\mathbf{q}_k, k = 1, \ldots, j-1$. So, $\mathbf{C}_j\mathbf{q}_k = \mathbf{0}, k = 1, \ldots, j-1$. For $k = j$, we obtain $\mathbf{D}_{j+1}\mathbf{H}\mathbf{p}_j = \mathbf{p}_j$ (that is a secant condition $\mathbf{D}_{j+1}\mathbf{q}_j = \mathbf{p}_j$). And so, $(\mathbf{D}_j + \mathbf{C}_j)\mathbf{q}_j = \mathbf{p}_j \Rightarrow \mathbf{C}_j\mathbf{q}_j = \mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j$. If $\mathbf{C}_j$ contains the term $\frac{\mathbf{p}_j\mathbf{p}_j^\top}{\mathbf{p}_j^\top\mathbf{q}_j}$ then multiplying the numerator of the fraction by $\mathbf{q}_j$ from the left, we obtain $\mathbf{p}_j$. If $\mathbf{C}_j$ contains the term $\frac{-\mathbf{D}_j\mathbf{q}_j(\mathbf{D}_j\mathbf{q}_j)^\top}{(\mathbf{D}_j\mathbf{q}_j)^\top\mathbf{q}_j}$ then multiplying the numerator of the fraction by $\mathbf{q}_j$ from the left, we obtain $-\mathbf{D}_j\mathbf{q}_j$. So, if $\mathbf{C}_j^{\mathrm{DFP}} = \frac{\mathbf{p}_j\mathbf{p}_j^\top}{\mathbf{p}_j^\top\mathbf{q}_j} + \frac{-\mathbf{D}_j\mathbf{q}_j(\mathbf{D}_j\mathbf{q}_j)^\top}{(\mathbf{D}_j\mathbf{q}_j)^\top\mathbf{q}_j}$ then even $\mathbf{C}_j^{\mathrm{DFP}}\mathbf{q}_k = \mathbf{0}$.

So, one of the important possibilities how to choose $\mathbf{d}_j$ directions and approximating matrix $\mathbf{D}_j$ is introduced as follows.

Algorithm 288 (Davidon-Fletcher-Powell)**.**

Choose $\varepsilon > 0$, starting point $\mathbf{x}_1$, symmetric, positive definite matrix $\mathbf{D}_1$. Then we assign $\mathbf{y}_1 := \mathbf{x}_1$, $j := 1$, $k := 1$.

1. If $\|\nabla f(\mathbf{y}_j)\| < \varepsilon$ then STOP, otherwise $\mathbf{d}_j := -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ and solve $\min\{f(\mathbf{y}_j + \lambda \mathbf{d}_j) \mid \lambda \geq 0\}$. Solution λ_j is used to compute $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j = n$ then $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}$, $\mathbf{y}_1 := \mathbf{x}_{k+1}$, $k := k + 1$, $j := 1$ and GOTO 1. Otherwise, when $j < n$, we continue.

2. Assign $\mathbf{D}_{j+1} := \mathbf{D}_j + \mathbf{C}_j^{\mathrm{DFP}}$, where

$$\begin{aligned} \mathbf{C}_j^{\mathrm{DFP}} &= \frac{\mathbf{p}_j \mathbf{p}_j^\top}{\mathbf{p}_j^\top \mathbf{q}_j} - \frac{\mathbf{D}_j \mathbf{q}_j \mathbf{q}_j^\top \mathbf{D}_j}{\mathbf{q}_j^\top \mathbf{D}_j \mathbf{q}_j}, \\ \mathbf{p}_j &= \mathbf{y}_{j+1} - \mathbf{y}_j \quad (= \lambda_j \mathbf{d}_j), \\ \mathbf{q}_j &= \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j), \end{aligned}$$

$j := j + 1$, and GOTO 1.

Note that $\mathbf{C}_j^{\mathrm{DFP}} = \alpha \mathbf{u}_j \mathbf{u}_j^\top + \beta \mathbf{v}_j \mathbf{v}_j^\top$ because $\mathbf{u}_j = \mathbf{p}_j$, $\alpha_j = (\mathbf{p}_j^\top \mathbf{q}_j)^{-1}$, $\mathbf{v}_j = \mathbf{D}_j \mathbf{q}_j$, and $\beta_j = (\mathbf{q}_j^\top \mathbf{D}_j \mathbf{q}_j)^{-1}$.

Example 289.

Solve $\min[(x_1 - 2)^4 + (x_1 - 2x_2)^2]$ by Algorithm 288.

k	j	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\mathbf{D}_j$	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$
1	1	(0,0; 3,0) 52,0	(0,0; 3,0) 52,0	(−44,0; 24,0)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(44,0; −24,0)	0,062	(2,70; 1,51)
	2		(2,70; 1,51) 0,34	(0,73; 1,28)	$\begin{bmatrix} ,25 & ,38 \\ ,38 & ,81 \end{bmatrix}$	(−0,67; −1,31)	0,22	(2,55; 1,22)
2	1	(2,55; 1,22) 0,1036	(2,55; 1,22) 0,1036	(0,89; −0,44)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(−0,89; 0,44)	0,11	(2,45; 1,27)
	2		(2,45; 1,27) 0,0490	(0,18; 0,36)	$\begin{bmatrix} ,65 & ,45 \\ ,45 & ,46 \end{bmatrix}$	(−0,28; −0,25)	0,64	(2,27; 1,11)
3	1	(2,27; 1,11) 0,008	(2,27; 1,11) 0,008	(0,18; −0,20)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(−0,18; 0,20)	0,10	(2,25; 1,13)
	2		(2,25; 1,13) 0,004	(0,04; 0,04)	$\begin{bmatrix} ,80 & ,38 \\ ,38 & ,31 \end{bmatrix}$	(−0,05; −0,03)	2,64	(2,12; 1,05)
4	1	(2,12; 1,05) 0,0005	(2,12; 1,05) 0,0005	(0,05; −0,08)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(−0,05; 0,08)	0,10	(2,12; 1,06)
	2		(2,12; 1,06) 0,0002	(0,004; 0,004)				

Exercise: *Be able to realize one algorithm iteration. In this case, use calculus minimization instead of line search algorithm for univariate minimization (to obtain λ_j).*

Remark: It is known that $\mathbf{D}_j$ is PD, and hence, $\mathbf{d}_j$ is a descent direction.

Theorem (DFP properties): Let $f(\mathbf{x})$ be a quadratic function with PD Hessian $\mathbf{H}$ to be minimized ($\mathbf{x} \in \mathbb{R}^n$). We solve it by DFP and $\forall j = 1, \ldots, n : \nabla f(\mathbf{y}_j) \neq \mathbf{0}$. Then:

1. $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are $\mathbf{D}$-conjugate directions,
2. $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$,
3. $\mathbf{y}_{n+1} \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$.

Proof: We emphasize main ideas of several steps. For the complete proof — see literature. We consider the situation $\forall j \leq n$. So, we take j arbitrarily fixed. Then for 1. we have to prove that $\mathbf{d}_k$ are linearly independent $\mathbf{H}$ conjugate directions for $k \leq j$. It is true, because of the way how $\mathbf{d}_k$ are constructed.

We would like to show that $\mathbf{D}_{j+1}\mathbf{H}\mathbf{p}_k = \mathbf{p}_k, k \leq j$ and $\mathbf{D}_{j+1}\mathbf{H}\mathbf{d}_k = \mathbf{d}_k$ (we know that $\mathbf{p}_k = \lambda_k \mathbf{d}_k$). So, $\mathbf{H}\mathbf{p}_k = \mathbf{H}(\lambda_k \mathbf{d}_k) = \mathbf{H}(\mathbf{y}_{k+1} - \mathbf{y}_k) = \mathbf{q}_k = \nabla f(\mathbf{y}_{k+1}) - \nabla f(\mathbf{y}_k)$, Therefore by induction $\mathbf{H}\mathbf{p}_1 = \mathbf{q}_1 \Rightarrow \mathbf{D}_2\mathbf{H}\mathbf{p}_1 = (\mathbf{D}_1 + \frac{\mathbf{p}_1\mathbf{p}_1^\top}{\mathbf{p}_1^\top\mathbf{q}_1} + \frac{-\mathbf{D}_1\mathbf{q}_1(\mathbf{D}_1\mathbf{q}_1)^\top}{(\mathbf{D}_1\mathbf{q}_1)^\top\mathbf{q}_1})\mathbf{q}_1 = \mathbf{D}_1\mathbf{q}_1 + \mathbf{p}_1 - \mathbf{D}_1\mathbf{q}_1 = \mathbf{p}_1$ and then we continue by induction. □

Remark 290 (Why DFP choice?)**.**

Let us discuss the DFP Algorithm 288 in the case of quadratic objective function $f(\mathbf{x}) = \mathbf{c}^\top\mathbf{x} + \frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x}$ with positive definite Hessian matrix $\mathbf{H}$. The algoritmus is constructed in such a way that descent directions $\mathbf{d}_j$ are conjugate and linearly independent. In addition, $\mathbf{D}_j$ approximates $\mathbf{H}^{-1}$ in such a way that it is a positive definite and symmetric. In addition, $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$ and the algorithm stops after one complete iteration (composed of n univariate minimizations).

Remark 291 (Choice of vectors)**.**

For the j'th iteration we know vectors $\mathbf{p}_k$, $k = 1, \ldots, j-1$ that satisfy $\mathbf{D}_j\mathbf{H}\mathbf{p}_k = \mathbf{p}_k$, and hence, they are eigenvectors of matrix $\mathbf{D}_j\mathbf{H}$ with unit eigenvalues. In addition, they are linearly independent and conjugate. For known point $\mathbf{y}_j$ we find a direction $\mathbf{d}_j = -\mathbf{D}_j\nabla f(\mathbf{y}_j)$ to get a new iterate $\mathbf{y}_{j+1}$. Then $\mathbf{p}_j = \mathbf{y}_{j+1} - \mathbf{y}_j$ and $\mathbf{q}_j = \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j) = \mathbf{H}(\mathbf{y}_{j+1} - \mathbf{y}_j) = \mathbf{H}\mathbf{p}_j$.

Remark 292 (Choice of matrix)**.**

We search for symmetric matrix $\mathbf{C}_j$, which simply updates $\mathbf{D}_j$ i.e. $\mathbf{D}_{j+1} := \mathbf{D}_j + \mathbf{C}_j$. We also want $\mathbf{D}_{j+1}$ satisfying conditions similar to conditions satisfied by $\mathbf{D}_j$, and so, $\mathbf{C}_j\mathbf{q}_k = \mathbf{0}$ for $k = 1, \ldots, j-1$ and $\mathbf{D}_{j+1}\mathbf{q}_j = \mathbf{p}_j$ must be satisfied. These conditions are satisfied for the choice $\mathbf{C}_j = \mathbf{C}_j^{\mathrm{DFP}}$ used above.

Remark (Convergence rate): Assuming f continuously differentiable and that $\mathbf{H}(\mathbf{x}_{\min})$ is PD then:

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_{\min}\|}{\|\mathbf{x}_{k+1} - \mathbf{x}_{\min}\|} \longrightarrow 0$$

that is the superlinear convergence rate to $\mathbf{x}_{\min}$. With $\|\mathbf{H}(\mathbf{x})\mathbf{y}\| \leq K\|\mathbf{y}\|$ $\forall \mathbf{y}$ and $\forall \mathbf{x} \in \mathcal{N}_\varepsilon(\mathbf{x}_{\min})$ then

$$\limsup_{k \longrightarrow \infty} \frac{\|\mathbf{y}_{k+1} - \mathbf{x}_{\min}\|}{\|\mathbf{y}_k - \mathbf{x}_{\min}\|^2} < \infty$$

that is quadratic convergence rate. Compare with conjugate gradients, where we obtain

$$\limsup_{k \longrightarrow \infty} \frac{\|\mathbf{y}_{k+n} - \mathbf{x}_{\min}\|}{\|\mathbf{y}_k - \mathbf{x}_{\min}\|^2} < \infty.$$

So, n-times more steps might be necessary for the same behaviour. However, for DFP storage requirements are $\sim n^2$ (cf. $3n$) and intermediate matrix products also spend time.

Till now, the algorithm derivation is based on several ideas:

1. The secant property is required.
2. The transformed eigenvectors condition is needed.
3. The symmetry and positive definiteness of approximating matrices cannot be excluded.

Remark 293 (Numerical behaviour and BFGS).

The DFP method sometimes generates matrices near to singular. Because there are some degrees of freedom in the choice of $\mathbf{C}_j$ (see e.g. BAZARAA 1993), Broyden has suggested another update (for $\phi > 0$: $\mathbf{C}_j$ is PD and also $\mathbf{C}_j\mathbf{q}_k = \mathbf{0}$):

$$\mathbf{C}_j^{\mathrm{B}} = \mathbf{C}_j^{\mathrm{DFP}} + \frac{\phi\tau_j \mathbf{v}_j \mathbf{v}_j^\top}{\mathbf{p}_j^\top \mathbf{q}_j},$$

where $\mathbf{v}_j = \mathbf{p}_j - (1/\tau_j)\mathbf{D}_j\mathbf{q}_j$ (to satisfy $\mathbf{C}_j\mathbf{q}_j = \mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j$) and

$$\tau_j = \frac{\mathbf{q}_j^\top \mathbf{D}_j \mathbf{q}_j}{\mathbf{p}_j^\top \mathbf{q}_j} > 0.$$

For the choice of $\phi = 1$, we get $\mathbf{C}_j$ in the form extensively studied by Broyden, Fletcher, Goldfarb, and Shanno. We denote it as $\mathbf{C}_j^{\mathrm{BFGS}}$:

$$\mathbf{C}_j^{\mathrm{BFGS}} := \frac{\mathbf{p}_j\mathbf{p}_j^\top}{\mathbf{p}_j^\top\mathbf{q}_j}\left(1 + \frac{\mathbf{q}_j^\top\mathbf{D}_j\mathbf{q}_j}{\mathbf{p}_j^\top\mathbf{q}_j}\right) - \frac{\mathbf{D}_j\mathbf{q}_j\mathbf{p}_j^\top + \mathbf{p}_j\mathbf{q}_j^t op\mathbf{D}_j}{\mathbf{p}_j^\top\mathbf{q}_j}.$$

The idea is to approximate $\mathbf{H}$ by $\mathbf{B}_j$ (instead of approximating H^{-1} by $\mathbf{D}_j$). So, $\mathbf{H}\mathbf{p}_k = \mathbf{q}_k$ and $\mathbf{B}_{j+1}\mathbf{p}_k = \mathbf{q}_k$ for $k = 1, \ldots, j$ (with DFP, we have had $\mathbf{H}^{-1}\mathbf{q}_k = \mathbf{p}_k$ and $\mathbf{D}_{j+1}\mathbf{q}_k = \mathbf{p}_k$ for $k = 1, \ldots, j$). As $\mathbf{B}_{n+1} = \mathbf{H}$ then we obtain $\mathbf{H}^{-1}\mathbf{B}_{n+1} = \mathbf{I}$. So, we get from $\mathbf{B}_{j+1}\mathbf{p}_k = \mathbf{q}_k$ that $\mathbf{H}^{-1}\mathbf{B}_{j+1}\mathbf{p}_k = \mathbf{H}^{-1}\mathbf{q}_k$ and $\mathbf{H}^{-1}\mathbf{B}_{j+1}\mathbf{p}_k = \mathbf{p}_k, k = 1, \ldots, j$. So, p_k are eigenvectors of $\mathbf{H}^{-1}\mathbf{B}_{j+1}$ with eigenvalues equal to 1. Then $\mathbf{B}_{j+1} := \mathbf{B}_j + \bar{\mathbf{C}}_j$ and $\mathbf{H}^{-1}(\mathbf{B}_j + \bar{\mathbf{C}}_j)\mathbf{p}_k = \mathbf{p}_k$. As $\mathbf{q}_k = \mathbf{H}\mathbf{p}_k$ then $\mathbf{B}_j\mathbf{p}_k = \mathbf{q}_k$ ($\bar{\mathbf{C}}_j\mathbf{p}_k = \mathbf{0}, k \leq j-1$, $\bar{\mathbf{C}}_j\mathbf{p}_j = \mathbf{q}_j - \mathbf{B}_j\mathbf{p}_j$). Then cf. the situation for $\mathbf{D}_j$ and understand the formula based on the replacements:

$$\bar{\mathbf{C}}_j = \frac{\mathbf{q}_j\mathbf{q}_j^\top}{\mathbf{q}_j^\top\mathbf{p}_j} - \frac{\mathbf{B}_j\mathbf{p}_j\mathbf{p}_j^\top\mathbf{B}_j}{\mathbf{p}_j^\top\mathbf{B}_j\mathbf{p}_j},$$

Then we need new $\mathbf{D}_{j+1}$ (we do not want to compute inverse of $\mathbf{B}_j$). It should satisfy $\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j^{\mathrm{BFGS}} = \mathbf{B}_{j+1}^{-1} = (B_{j+1} + \bar{\mathbf{C}}_j)^{-1}$. In this case the Sherman-Morrison formula from linear algebra may help to obtain $\mathbf{C}_j^{\mathrm{BFGS}}$ as introduced above. As we need $\mathbf{D}_{j+1}$, we may also solve a system of linear equations $\mathbf{B}_j\mathbf{d}_j = -\nabla f(\mathbf{y}_j)$ by decomposition to obtain $\mathbf{d}_j$.
It is true that $\mathbf{C}_j^{\mathrm{B}} = (1-\phi)\mathbf{C}_j^{\mathrm{DFP}} + \phi\mathbf{C}_j^{\mathrm{BFGS}}$. If we replace DFP matrix $\mathbf{C}_j^{\mathrm{DFP}}$ by $\mathbf{C}_j^{\mathrm{BFGS}}$ in Algorithm 288, we get the BFGS algorithm, which is currently considered as the most efficient method for unconstrained minimization.

Remark (Computational comments): There are partial quasi-Newton methods when the restart (reset of $\mathbf{D}_j$) is used earlier than after n inner iterations. It is used to save memory.

Scaling is often realized by multiplication of $\mathbf{D}_j$ by $s_j > 0$ before update to avoid situation when $\mathbf{D}_1\mathbf{H}$ eigenvalues are $>> 1$! Then $\mathbf{D}_j$ update may move in the wrong way. So, conditional number of matrix is big and computational problems may occur.

There are also memory-less quasi-Newton methods based on the idea to forget $\mathbf{D}_j$ — often setting $\mathbf{D}_j := \mathbf{I}$ in BFGS formula. In this case you do not need so large memory and it will work

even with weak conditions (regarding the descent directions) even for inexact line search. With exact line search, the equivalence to conjugate gradients is obtained. See literature for more information.

Remark (Convergence of conjugate directions methods): For f quadratic convex was discussed. For non-quadratic case, we have composed $\mathcal{A} = \mathcal{C}\mathcal{B}$ composed map and we need to show that: (1) $\forall \mathbf{x} \in X \setminus \Omega : \mathcal{B}$ is closed map, (2) $\forall \mathbf{x} \in X \setminus \Omega : \mathbf{y} \in \mathcal{B}(\mathbf{x}) \Rightarrow f(\mathbf{y}) \leq f(\mathbf{x})$, (3) $\mathbf{z} \in \mathcal{C}(\mathbf{y}) \Rightarrow f(\mathbf{z}) \leq f(\mathbf{y})$, (4) $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_1)\}$ is a compact set ($\mathbf{x}_1$ is a starting point). We get $\mathbf{y} \in \mathcal{B}(\mathbf{x})$ by minimization along $\mathbf{d}$ from $\mathbf{x}$ where $\mathbf{d} = -\mathbf{D}\nabla f(\mathbf{x})$ ($\mathbf{D}$ is PD) For $\mathbf{D} = \mathbf{I}$, we obtain conjugate gradient method, otherwise quasi-Newton's method. $\mathcal{C}$ gives the minimum along directions by methods (decrease, so (3) is guaranteed). With $\Omega = \{\mathbf{x} \mid \nabla f(\mathbf{x}) = \mathbf{0}\}$ we may show that: $\mathcal{B}$ is closed (cf. discussion with the gradient method) and f is decreasing by $\mathbf{d}$ descent (as $\mathbf{D}$ is PD). As usually, the assumption (4) should be included. The key idea is the inserted gradient-like step.

Remark 294 (Comparison)**.**

Quasi-Newton methods utilize the modification of descent direction by multiplication of the minus gradient by approximating matrix $\mathbf{H}(\mathbf{x})$. It may be interpreted as taking into account the information about the curvature of the surface defined by $z = f(\mathbf{x})$. In the case of large scale problems, there are troubles with memory requirements (huge matrices must be stored). In this case, conjugate gradient methods are preferable.

1.11 Theory of constrained optimization

1.11.1 Introduction

In this section, we study properties of constrained optimization problems. Remember from convex analysis that in the case of differentiable convex function f and convex set S the condition $\nabla f(\bar{\mathbf{x}})^\top(\mathbf{x}-\bar{\mathbf{x}}) \geq \mathbf{0}$, $\forall \mathbf{x} \in S$ is necessary and sufficient condition for attaining the global minimum of f on S at $\bar{\mathbf{x}}$.

Remark 295 (Notation)**.**

By a symbol $\nabla \mathbf{g}(\mathbf{x})^\top$, we denote a transposed Jacobi matrix having n rows and m columns with elements $\frac{\partial g_i(\mathbf{x})}{\partial x_j}$ (see j'th row a i'th column). Therefore, matrix $\nabla \mathbf{g}(\mathbf{x})^\top$ has gradients of functions $g_i(\mathbf{x})$ as its columns. Similarly, we denote $\nabla \mathbf{h}(\mathbf{x})^\top$.

The following sequence of examples show how to solve constrained optimization (NLP — nonlinear programming — problems) analytically. Theoretical notes follow, so the reader has to use forward references in the example.

Example 296 (Cookbook)**.**

Completely solve $\min\{(x_1 - 1)^2 + (x_2 - 2)^2 \mid x_1 - x_2 \leq 1, x_2 \geq 0\}$. Hint: At first, we split the example to identify main (educational) steps in the solution procedure.

Example 297 (Graphical solution).

Solve the problem graphically (can be utilized only if $\mathbf{x} \in \mathbb{R}^2$ or partially $\mathbf{x} \in \mathbb{R}^3$).

Use your experience from LP. Notice that the contour graph is composed from circles.

Example 298 (Reformulation).

Express the given NLP in the following equivalent form: $\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\}$.

$\min\{(x_1 - 1)^2 + (x_2 - 2)^2 \mid x_1 - x_2 - 1 \leq 0, -x_2 \leq 0\}$

Example 299 (Lagrangian).

Build Lagrangian in the form $L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{g}(\mathbf{x})^\top \mathbf{u} + \mathbf{h}(\mathbf{x})^\top \mathbf{v}$.

$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = L(x_1, x_2, u_1, u_2) = (x_1 - 1)^2 + (x_2 - 2)^2 + u_1(x_1 - x_2 - 1) + u_2(-x_2)$

Example 300 (Karush-Kuhn-Tucker conditions).

For the given example, derive the KKT conditions from Lagrangian: $\nabla_x L = \mathbf{0}$ (or $\nabla_x L \geq \mathbf{0}, \mathbf{x} \geq \mathbf{0}, \mathbf{x}^\top \nabla_x L = 0$) $\mathbf{u} \geq \mathbf{0}$, $\nabla_u L \leq \mathbf{0}$, $\mathbf{u}^\top \nabla_u L = 0$, and $\nabla_v L = \mathbf{0}$.

$\nabla_x L = \mathbf{0}$:
$L'_{x_1} = 2(x_1 - 1) + u_1 = 0,$
$L'_{x_2} = 2(x_2 - 2) + u_1 - u_2 = 0,$
$\mathbf{u} \geq \mathbf{0}$:
$u_1 \geq 0, u_2 \geq 0,$
$\nabla_u L \leq \mathbf{0}$:
$L'_{u_1} = x_1 + x_2 - 1 \leq 0,$
$L'_{u_2} = -x_2 \leq 0,$
$\mathbf{u}^\top \nabla_u L = 0$:
$u_1 L'_{u_1} = u_1(x_1 - x_2 - 1) = 0,$
$u_2 L'_{u_2} = u_2(-x_2) = 0.$

Example 301 (Check KKT for the point).

Check validity of the KKT conditions for the given point $(x_1 = 1, x_2 = 0)$. Remember also the regularity condition! (columns of $\nabla \mathbf{g}(\mathbf{x})$ have to be linearly independent vectors).

For the given point, we get: $u_1 + 2(1 - 1) = 0$ and $u_1 - u_2 - 4 = 0$. Therefore, $u_2 = -4 < 0$ and the KKT conditions are not satisfied.

Example 302 (Geometrical interpretation)**.**

Interpret the KKT conditions geometrically. Hint: Draw the feasible region, contour graph as before, add vectors $-\nabla f(\mathbf{x}) = ()$, and columns of $\nabla \mathbf{g}(\mathbf{x})$.

For the point specified by $x_1 = 1$, and $x_2 = 0$ we have $u_1 = 0$ and $u_2 = -4$. We also have $-\nabla f(\mathbf{x}) = (-2(x_1 - 1), -2(x_2 - 2))^\top = (0, 4)^\top$, $\nabla g_1(\mathbf{x}) = (1, 1)^\top$, and $\nabla g_2(\mathbf{x}) = (0, -1)^\top$. So, we may show that for this point $-\nabla f(\mathbf{x})$ is not a nonnegative linear combination of gradients of active constraints i.e. $\nabla g_1(\mathbf{x})$and $\nabla g_2(\mathbf{x})$.

Example 303 (One solution satisfying KKT)**.**

Find at least one solution satisfying the KKT conditions.

For finding solutions, a good policy is to subsequently setup u_i coefficients and analyze simplified systems of equalities and inequalities:

1. $u_1 = 0, u_2 = 0$: $\Rightarrow x_1 = 1$, $x_2 = 2$, $1 + 2 - 1 \not\leq 0$. Infeasibility appeared!
2. $u_1 \neq 0, u_2 = 0$: $\Rightarrow x_1 = 1 - x_2$, $-2(x_2 - 2) = u_1 = -2(x_1 - 1)$, so $x_1 = x_2 - 1$. Then, $x_1 = 0$, $x_2 = 1$ and this point satisfies the KKT conditions. We finish the search for all optimal solutions now.
3. $u_1 = 0, u_2 \neq 0$: $\Rightarrow x_1 = 1$, $x_2 = 0$, so $u_1 = 0, u_2 - 4$ and contradiction.
4. $u_1 \neq 0, u_2 \neq 0$: $\Rightarrow x_1 = 1$, $x_2 = 0$, so $u_1 = 0, u_2 - 4$ and the same contradiction.

Example 304 (All solutions)**.**

Find all solutions satisfying the KKT conditions. Discuss also points that do not satisfy the regularity condition mentioned above.

Solution procedure is completed above. Because ∇g_1 and ∇g_2 are constant and linearly independent, no irregular points may occur.

Example 305 (Classify found solutions)**.**

Discuss which solutions are optimal using geometry or the second order conditions.

Because the feasible set S is a convex set and f is a convex function that is minimized then by the previous theory any local minimum is also a global one.

1.11.2 Review of calculus

Remark 306 (Equality constraint)**.**

Let us consider a simple program $\min\{f(x,y) \mid h(x,y)=0\}$, where $f:\mathbb{R}^2 \longrightarrow \mathbf{R}$ and $h:\mathbb{R}^2 \longrightarrow \mathbf{R}$ are differentiable functions. For minimum point $(x_{\min}, y_{\min})$, the following equality must be satisfied $\nabla f(x_{\min}, y_{\min}) = v \cdot \nabla h(x_{\min}, y_{\min})$. The requirement guarantees that there is no feasible descent direction at the minimizing point.

Remark 307 (Implicit function use)**.**

Another explanation is given by the idea that y depends functionally on x. Then, the objective function derivative f by x can be computed using the formula for computation of the implicit function derivative.

$$f'_x + f'_y y'_x = f'_x - (h'_x/h'_y) f'_y = f'_x - v \cdot h'_x = 0,$$

and so $f'_y - v \cdot h'_y = 0$, which follows from the notation of v. Generalization is given by the following theorem:

Theorem 308 (Lagrange)**.**

Let $\bar{\mathbf{x}}$ be a feasible solution of the following minimizing program Eq with constraints in the form of equalities $\min\{f(\mathbf{x}) \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. Let functions f and $\mathbf{h}$ have the continuous first order partial derivatives in the neighbourhood of $\bar{\mathbf{x}}$. Let columns of matrix $\nabla \mathbf{h}(\bar{\mathbf{x}})^\top$ are linearly independent (the regularity conditions). If $\bar{\mathbf{x}}$ is a point of local minimum of the solved program Eq then exists a vector $\mathbf{v}$ such that

$$\nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{h}(\bar{\mathbf{x}})^\top \mathbf{v} = \mathbf{0}.$$

Theorem 308 specifies the necessary first order optimality conditions for the program Eq. Lagrange multipliers $\mathbf{v}$ present the information about the sensitivity of the objective function values at minimum point $\bar{\mathbf{x}}$ with respect to the right-hand-side (RHS) changes for constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ (cf. with shadow prices in linear programming).

Remark 309 (Lagrange function (1760))**.**

For program Eq, we define Lagrange function (Lagrangian):

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^\top \mathbf{v} = f(\mathbf{x}) + \sum_{i=1}^{m} v_i h_i(\mathbf{x}).$$

Remark 310 (Stationary points)**.**

We search for stationary points of Lagrangian, satisfying the condition $\nabla L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$, i.e. $\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$ and $\nabla_{\mathbf{v}} L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$. Namely, theorem 308 says how to search the minimum of Eq using stationary points of Lagrangian. We obtain a system of $n + m$ (usually) nonlinear equations of $n + m$ unknown variables $\mathbf{x}$ a $\mathbf{v}$, including conditions of Theorem 308, and constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$.

$$\begin{aligned} \frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial x_j} &= \frac{\partial f(\mathbf{x})}{\partial x_j} + \sum_{i=1}^{m} v_i \frac{\partial h_i(\mathbf{x})}{\partial x_j} = 0 \quad j = 1, \ldots, n \\ \frac{\partial L(\mathbf{x}, \mathbf{v})}{\partial v_k} &= h_k(\mathbf{x}) = 0 \quad k = 1, \ldots, m, \end{aligned}$$

The obtained solution $\bar{\mathbf{x}}$ is a point of the possible solution of Eq. The system of equations gives the necessary conditions, however not sufficient conditions (see the example below).

Example 311.

Solve $\min\{-x_1^2 - 4x_2^2 \mid x_1 + 2x_2 = 6\}$. Then $L(x_1, x_2, u) = -x_1^2 - 4x_2^2 + u \cdot (x_1 + 2x_2 - 6)$ and the solution of $\nabla L = \mathbf{0}$ is obtained $x_1 = 3$, $x_2 = 3/2$ a $u = 6$. However, for $x_1 \longrightarrow \infty$ and $x_2 \longrightarrow \infty$ the following conclusion is valid $f(x_1, x_2) \longrightarrow -\infty$.

Remark 312 (Possible simplifications?)**.**

The discussed program and its objective function Eq can be modified by substitutions of variables using equality constraints. Then, the unconstrained optimization problem is obtained. The use of this approach may reduce the dimension of the solved problem. However, it is important to consider whether such substitution is also simplifying for the numerical algorithm use.
The additional argument to study NLPs is that it may be impossible to get completely describing expressions for separate variables from $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. Therefore, NLPs have to be studied also without such simplifications.
Note that analytical computations based on the Lagrangian function are exceptionally used to get optimal solutions. Instead of this, Lagrangian properties are very important for the development of constrained optimization numerical algorithms.

1.11.3 Constrained optimization — inequalities

Remark 313 (Transforming inequalities to equalities)**.**

Let us consider the program Ineq of the form $\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}$. We introduce slack variables $\mathbf{y} = (y_1, \ldots, y_m)^\top$ for separate constraints and we denote $\mathbf{y}^* = (y_1^2, \ldots, y_m^2)$. The program Ineq can be rewritten in the form Eq, so we obtain $\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) + \mathbf{y}^* = \mathbf{0}\}$. We have the Lagrange function in the form $L(\mathbf{x}, \mathbf{u}, \mathbf{y}) = f(\mathbf{x}) + (\mathbf{g}(\mathbf{x}) + \mathbf{y}^*)^\top \mathbf{u}$ and we may use Theorem 308. From $\nabla L = \mathbf{0}$, we get necessary conditions in the form $\nabla_{\mathbf{x}} L = \nabla f(\mathbf{x}) + \nabla \mathbf{g}(\mathbf{x})^\top \mathbf{u} = \mathbf{0}$, $\nabla_{\mathbf{u}} L = \mathbf{g}(\mathbf{x}) + \mathbf{y}^* = \mathbf{0}$ (and hence $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$), and finally $\nabla_{\mathbf{y}} L = (2u_1 y_1, \ldots, 2u_m y_m)^\top = \mathbf{0}$. We suggest to compare these conditions with further developed KKT conditions.

Definition 314 (Cone of feasibility directions)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, $\bar{\mathbf{x}} \in \operatorname{cl} S$. The cone of feasibility directions of S at $\bar{\mathbf{x}}$ is:

$$D = \{\mathbf{d} \mid \mathbf{d} \neq \mathbf{0}, \exists \delta > 0, \forall \lambda \in (0, \delta) : \bar{\mathbf{x}} + \lambda \mathbf{d} \in S\}.$$

Exercise 315.

Consider f differentiable at $\bar{\mathbf{x}}$. Draw a figure illustrating the fact that for minimum at $\bar{\mathbf{x}}$, $\nabla f(\bar{\mathbf{x}})^\top \mathbf{d} \not< 0$. Hint: Otherwise $\mathbf{d}$ is a feasible descent direction of f at $\bar{\mathbf{x}}$.

Definition 316 (Cone of improving directions)**.**

We define a cone of improving (descent) directions of f at $\bar{\mathbf{x}}$ as

$$F = \{\mathbf{d} \mid \exists \delta > 0, \forall \lambda \in (0, \delta) : f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})\}.$$

Definition 317.

We have f differentiable at $\bar{\mathbf{x}}$. We define a halfspace F_0 by

$$F_0 = \{\mathbf{d} \mid \nabla f(\mathbf{x}_0)^\top \mathbf{d} < 0\}$$

Theorem 318 (Geometric optimality conditions)**.**

Let $S \subset \mathbb{R}^n$, $S \neq \emptyset$, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ differentiable at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ $\Rightarrow F_0 \cap D = \emptyset$.
Conversely, $F_0 \cap D = \emptyset$, f is pseudoconvex at $\bar{\mathbf{x}}$ and $\exists \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$, $\varepsilon > 0$ $\forall \mathbf{x} \in S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$: $\mathbf{d} = (\mathbf{x} - \bar{\mathbf{x}}) \in D \Rightarrow \bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$

Proof: By contradiction: $\exists \mathbf{d} \in F_0 \cap D \Rightarrow \exists \delta_1, \forall \lambda \in (0, \delta_1) f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})$ and $\exists \delta_2, \forall \lambda \in (0, \delta_2) \bar{\mathbf{x}} + \lambda \mathbf{d} \in S$. Because of intersection, we get a contradiction.
Conversely: $\forall \mathbf{x} \in S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) : f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$. Otherwise, $\exists \hat{\mathbf{x}} \in S \cap \mathcal{N}_\varepsilon : f(\hat{\mathbf{x}}) < f(\bar{\mathbf{x}})$ but $\mathbf{d} = (\hat{\mathbf{x}} - \bar{\mathbf{x}}) \in D$ and by pseudoconvexity $\nabla f(\bar{\mathbf{x}})^\top \mathbf{d} < 0$ or else $\nabla f(\bar{\mathbf{x}})^\top \mathbf{d} \geq 0 \Rightarrow f(\hat{\mathbf{x}}) = f(\mathbf{x} + \lambda \mathbf{d}) \geq f(\bar{\mathbf{x}})$. $\bar{\mathbf{x}}$ is not a local minimizer, $\exists \mathbf{d} \in F_0 \cap D$ and contradiction. □

Remark 319.

Notice that $F_0 \subset F \subset F_0'$, where $F_0' = \{\mathbf{d} \neq \mathbf{0} \mid \nabla f(\bar{\mathbf{x}})^\top \mathbf{d} \leq 0\}$. If f is pseudoconvex at $\bar{\mathbf{x}}$ then $F = F_0$. If f is strictly pseudoconcave then $F = F_0'$.

To be able to use geometric interpretation of optimality conditions for computations, we specify $S = \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$. We define $G_0 \subset D$ using ∇g_i, and so $F_0 \cap D = \emptyset$ at $\bar{\mathbf{x}} \Rightarrow F_0 \cap G_0 = \emptyset$ at $\bar{\mathbf{x}}$. Then, we may use gradients as follows.

Lemma 320.

Let $S = \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$, X be a nonempty open set in $\mathbb{R}^n$, $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, $\bar{\mathbf{x}} \in S$, $I = \{i \mid g_i(\bar{\mathbf{x}}) = 0\}$ be an index set of binding (active) constraints. Assume that g_i are differentiable for $i \in I$ at $\bar{\mathbf{x}}$ g_i are continuous at $\bar{\mathbf{x}}$ for $i \notin I$. Define $G_0 = \{\mathbf{d} \mid \nabla g_i(\bar{\mathbf{x}})^\top \mathbf{d} < 0, \forall i \in I\}$, $G_0' = \{\mathbf{d} \neq \mathbf{0} \mid \nabla g_i(\bar{\mathbf{x}})^\top \mathbf{d} \leq 0, \forall i \in I\}$. Then $G_0 \subset D \subset G_0'$.
If $g_i, i \in I$ are strictly pseudoconvex $\Rightarrow D = G_0$. If $g_i, i \in I$ are strictly pseudoconvex $\Rightarrow D = G_0'$.

Proof: Not included. □

Theorem 321 (Geometric optimality conditions II)**.**

Let the feasible set S be defined by $S = \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$, X be a nonempty open set in $\mathbb{R}^n$, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i = 1, \ldots, m$, and Ineq: $? \in \operatorname{arglocmin}_{\mathbf{x}} \{f(\mathbf{x}) \mid \mathbf{x} \in S\}$
Let $\bar{\mathbf{x}} \in S$ and I be an index set of indices of active constraints. In addition, $f, g_i, i \in I$ are differentiable at $\bar{\mathbf{x}}$ and $g_i, i \notin I$ are continuous at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}}$ solves Ineq $\Rightarrow F_0 \cap G_0 = \emptyset$. Conversely, $F_0 \cap G_0 = \emptyset$, f is pseudoconvex at $\bar{\mathbf{x}}$ and $g_i, i \in I$ are strictly pseudoconvex (over some existing $\mathcal{N}_\varepsilon(\bar{\mathbf{x}})$, $\varepsilon > 0$) $\Rightarrow \bar{\mathbf{x}} \in \operatorname{arglocmin}_{\mathbf{x}} \{f(\mathbf{x}) \mid \mathbf{x} \in S\}$.

Proof: $\bar{\mathbf{x}}$ is a local minimum. Then $F_0 \cap D = \emptyset \Rightarrow F_0 \cap G_0 = \emptyset$.
Conversely: $F_0 \cap G_0 = \emptyset$ S redefined by active constraints, pseudoconvexity considered then $G_0 = D$ by Lemma and $F_0 \cap D = \emptyset$. Then $g_i(\bar{\mathbf{x}}) \leq 0$ related level sets are convex on $\mathcal{N}_\varepsilon(\bar{\mathbf{x}}) \Rightarrow S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ is a convex set. $F_0 \cap D = \emptyset$ and f is pseudoconvex at $\bar{\mathbf{x}} \Rightarrow \bar{\mathbf{x}}$ is a local minimum (valid when binding included). □

Exercise 322.

Choose different points, compute gradients and draw figures for programs: $\min\{(x_1 - 3)^2 + (x_2 - 2)^2 \mid x_1^2 + x_2^2 \leq 5, x_1 + x_2 \leq 3, x_1, x_2 \geq 0\}$ and $\min\{(x_1 - 1)^2 + (x_2 - 1)^2 \mid (x_1 + x_2 - 1)^3 \leq 0, x_1, x_2 \geq 0\}$. Compare results.

Remark 323.

In addition if $g_i, i \notin I$ continuous then $\bar{\mathbf{x}} \in \text{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Leftrightarrow F_0 \cap D = \emptyset \Leftrightarrow F_0 \cap G_0 = \emptyset$ (f pseudoconvex) $F_0 \cap D = \emptyset \Leftrightarrow F \cap D = \emptyset$. But it is not still useful for computations. The next step is to introduce computationally useful Fritz John (FJ) conditions based on Farkas' (Gordon's) Theorem (see section about convex cones).

Theorem 324 (Fritz John).

$X \subset \mathbb{R}^n$, $X \neq \emptyset$ open set, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i = 1, \ldots, m$,
Let $S = \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$, X be a nonempty open set in $\mathbb{R}^n$, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i = 1, \ldots, m$, Ineq: $? \in \text{arglocmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$
Let $\bar{\mathbf{x}} \in S$ and I be an index set of indices of active constraints ($I = \{i \mid g_i(\bar{\mathbf{x}}) = 0\}$), f, $g_i, i \in I$ are differentiable at $\bar{\mathbf{x}}$ and $g_i, i \notin I$ are continuous at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}}$ solves Ineq $\Rightarrow$ $\exists u_0, u_i, i \in I, \mathbf{u} = (u_i)_{i \in I}$:

$$u_0 \nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} u_i \nabla \mathbf{g}_i(\bar{\mathbf{x}}) = \mathbf{0}$$
$$\forall i \in I : u_0, u_i \geq 0, \ (u_0, \mathbf{u}^\top) \neq (0, \mathbf{0}^\top).$$

Proof: Let $\bar{\mathbf{x}} \in\in \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\} \Rightarrow$ by Theorem 321 $\nexists \mathbf{d} : \nabla f(\bar{\mathbf{x}})\mathbf{d} < 0$ and $\nabla g_i(\bar{\mathbf{x}})\mathbf{d} < 0, i \in I$. We denote $\mathbf{A} = (\nabla f(\bar{\mathbf{x}}), (\nabla g_(\bar{\mathbf{x}}))_{i \in I}))^\top$. So, we assume that $\mathbf{Ad} < \mathbf{0}$ is inconsistent. By Farkas's Theorem 57 either $\{\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{0} \wedge \mathbf{c}^\top \mathbf{x} > 0\} \neq \emptyset$ or $\{\mathbf{y} \mid \mathbf{A}^\top \mathbf{y} = \mathbf{c} \wedge \mathbf{y} \geq \mathbf{0}\} \neq \emptyset$. By its Gordon's reformulation in Theorem 59 either $\{\mathbf{x} \mid \mathbf{Ax} < \mathbf{0}\} \neq \emptyset$ or $\{\mathbf{y} \mid \mathbf{A}^\top \mathbf{y} = \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \mathbf{y} \neq \mathbf{0}\} \neq \emptyset$. (Gordon's Theorem follows from Farkas' Theorem for the choice $\begin{pmatrix} \mathbf{A} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ s \end{pmatrix} \leq \mathbf{0}$ and $\begin{pmatrix} \mathbf{0}^\top & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ s \end{pmatrix} > \mathbf{0}$ $\Rightarrow \begin{pmatrix} \mathbf{A}^\top \\ \mathbf{1}^\top \end{pmatrix} \mathbf{y} = \begin{pmatrix} \mathbf{0}^\top \\ s \end{pmatrix}$ and $\mathbf{y} \geq \mathbf{0}$.) Then, for $\mathbf{Ad} < \mathbf{0}$ unsolvable $\exists \mathbf{y} \neq \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \mathbf{A}^\top \mathbf{y} = \mathbf{0}$. Denote $\mathbf{y} = (u_0, (u_1)_{i \in I})^\top$ and proof is complete. □

Corollary 325.

If, in addition, $g_i, i \notin I$ are differentiable at $\bar{\mathbf{x}}$ then we may write that exists $u_0, \mathbf{u}$ such that

$$u_0 \nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{g}(\bar{\mathbf{x}})^\top \mathbf{u} = \mathbf{0}, \qquad \mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0,$$
$$(u_0, \mathbf{u}^\top) \geq (0, \mathbf{0}^\top), \qquad (u_0, \mathbf{u}^\top) \neq (0, \mathbf{0}^\top).$$

Proof: The proof is trivial just zero terms are added for inactive constraints. □

These conditions include dual feasibility, complementary slackness, and Lagrange multipliers. Remember! When these conditions are used to discover $\bar{\mathbf{x}}$ possible minimum, the solution still must be checked for its primal feasibility ($\bar{\mathbf{x}} \in S$).

Example 326.

Solve $\min\{(x_1-3)^2+(x_2-2)^2 \mid x_1^2+x_2^2-5\leq 0,\ x_1+2x_2-4\leq 0,\ -x_1\leq 0; -x_2\leq 0\}$. At first graphically. Check points $(2;1)$ and $(0;0)$. For the first point: $u_3=u_4=0$ and $u_1=u_0/3$ and $u_2=2u_0/3$. For example $u_0=3$, $u_1=1$ and $u_2=2$ and $(2;1)$ may be minimum. Contradiction occurs for the second point as $u_1=u_2=0$ and $u_3=-6u_0$ and $u_4=-4u_0$. So, some u_i must be negative.

Example 327 (Kuhn and Tucker (1951))**.**

Solve $\min\{-x_1 \mid x_2-(1-x_1)^3\leq 0,\ -x_2\leq 0\}$. Then, the minimum is $\mathbf{x}_0=(1;0)^\top$, so $\nabla f(\mathbf{x}_0)=(-1;0)^\top$, $\nabla g_1(\mathbf{x}_0)=(0;1)^\top$ and $\nabla g_2(\mathbf{x}_0)=(0;-1)^\top$. We get $u_0=0$ and $u_1=u_2$ and Fritz John conditions are not useful as for $u_0=0$ no information about $-\nabla f(\mathbf{x}_0)$ is used.

Remark 328 (Problems with FJ conditions)**.**

These conditions may also be satisfied trivially (see also example above). When any gradient $\nabla g_i(\bar{\mathbf{x}})=\mathbf{0}$ then we may set the related $u_i\neq 0$ and other $u_i=0$ and conditions are satisfied. When $g_i(\mathbf{x})=0$ constraint is replaced by the couple $g_i(\mathbf{x})\leq 0$ and $-g_i(\mathbf{x})\geq 0$ then Fritz John (FJ) conditions are satisfied by any feasible solutions, as we may choose multipliers u_i related to these constraints nonzero and equal. In general, we may fulfill Fritz John conditions for any feasible solution $\mathbf{x}_0$ by adding redundant constraint of the form $\|\mathbf{x}-\mathbf{x}_0\|^2\geq 0$. It is satisfied for any feasible $\mathbf{x}$. The constraint is active in $\mathbf{x}_0$. Its gradient equals zero in $\mathbf{x}_0$. Therefore, $G_0=\emptyset$, and so $F_0\cap G_0=\emptyset$.

These observations led to the development in two directions: (1) the formulation of Fritz John sufficient conditions and (2) its improvement avoiding such cases (KKT conditions).

Theorem 329 (Fritz John sufficient conditions)**.**

We have Ineq problem and $\bar{\mathbf{x}}$ satisfying FJ conditions. Binding constraints are defined by index set I. We assume that $S=\{\mathbf{x} \mid g_i(\mathbf{x})\leq 0, i\in I\}$ (a relaxed feasible set).
Let $\mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ be a neighbourhood of $\bar{\mathbf{x}}$, f be pseudoconvex and $g_i, i\in I$ strictly pseudoconvex over $\mathcal{N}_\varepsilon(\bar{\mathbf{x}})\cap S\Rightarrow \bar{\mathbf{x}}\in$ arglocmin of Ineq.
Let f be pseudoconvex at $\bar{\mathbf{x}}$ and $g_i, i\in I$ strictly pseudoconvex at $\bar{\mathbf{x}}$ and $g_i, i\in I$ quasiconvex at $\bar{\mathbf{x}}\Rightarrow \bar{\mathbf{x}}\in$ argglobmin of Ineq.

Proof: Use various definitions generalizing convex functions. □

Regarding improvement of FJ, the most important idea is to avoid the situation when $F_0\cap G_0=\emptyset$ because of $G_0=\emptyset$ (for any f). Therefore, the condition $G_0\neq\emptyset$ (constraint qualification) will be sufficient to solve our problem. The Karush-Kuhn-Tucker (KKT) conditions follows.

Theorem 330 (Karush-Kuhn-Tucker conditions — inequalities)**.**

Let $X \subset \mathbb{R}^n$, $X \neq \emptyset$, X open, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i = 1, \ldots, m$ be given. Ineq: $? \in \operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m, \mathbf{x} \in X\}$. Let $\bar{\mathbf{x}}$ be feasible, I denotes set of indices of binding constraints at $\bar{\mathbf{x}}$, $f, g_i, i \in I$ are differentiable at $\bar{\mathbf{x}}$ and $g_i, i \notin I$ are continuous at $\bar{\mathbf{x}}$. Furthermore, suppose that $\nabla g_i(\bar{\mathbf{x}}), i \in I$ are linearly independent (constraint qualification).
If $\bar{\mathbf{x}}$ solves Ineq (locally) then $\exists u_i, i \in I$:

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\mathbf{x}}) = \mathbf{0}, \qquad u_i \geq 0.$$

Proof: By FJ exist u_0 and $\hat{u}_i, i \in I$ not all equal zero such that

$$u_0 \nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \hat{u}_i \nabla \mathbf{g}_i(\bar{\mathbf{x}}) = \mathbf{0}, \; u_0, \hat{u}_i \geq 0, i \in I.$$

If $u_0 = 0$ then $\nabla g_i(\bar{\mathbf{x}})$ are linearly dependent $\Rightarrow$ contradiction. So, $u_0 \neq 0$ and $u_0 > 0$. We define $u_i = \frac{\hat{u}_i}{u_0}$ and divide FJ equality with u_0 by u_0. □

Remark 331 (Geometric interpretation of KKT)**.**

KKT theorem 332 introduces necessary conditions for existence of minimum at $\bar{\mathbf{x}}$. The theorem is illustrated by Figure 5. It says that for $\bar{\mathbf{x}}$ the steepest descent vector $-\nabla f(\mathbf{x}_0)$ can be expressed as nonnegative linear combination of gradients $\nabla g_i(\mathbf{x}_0)$ defined only by binding (active) constraints.

Theorem 332 (Karush-Kuhn-Tucker - differentiable version)**.**

Let f a $\mathbf{g}$ be differentiable at $\bar{\mathbf{x}}$ minimum of $\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}$. Let columns of $\nabla \mathbf{g}(\bar{\mathbf{x}})^\top$ related to gradients of active constraints are linearly independent (regularity — constraint qualification). Then exist coefficients $\mathbf{u}$ such that:

$$\nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{g}(\bar{\mathbf{x}})^\top \mathbf{u} = \mathbf{0}, \qquad \mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0, \qquad \mathbf{u} \geq \mathbf{0}.$$

Proof: For the proof, see the previous theorem. Compare with FJ and linear programming complementary slackness. Note that $g_i(\bar{\mathbf{x}}) \leq 0, i = 1, \ldots, m$ and $u_i \geq 0$ by KKT. So, $u_i g_i(\bar{\mathbf{x}}) \leq 0, i = 1, \ldots, m$. It is equal to zero iff $u_i = 0$ or $g_i(\bar{\mathbf{x}}) = 0$. Then, $\mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = \sum_{i=1}^m u_i g_i(\bar{\mathbf{x}}) = 0$ iff $\wedge_{i=1}^m (u_i g_i(\bar{\mathbf{x}}) = 0)$. So, the condition $\mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0$ guarantees that for non-binding constraints $g_i(\bar{\mathbf{x}}) < 0, i \notin I$ u_i must be equal to zero and related gradients are not further considered. □

Example 333.

Find all points satisfying KKT conditions for $\min\{-3x_1-x_2 \mid x_1^2+x_2^2 \leq 5,\ x_1-x_2 \leq 1\}$. Using Lagrange function, we get:

$$-3+u_1\cdot 2x_1+u_2=0 \qquad -1+u_1\cdot 2x_2-u_2=0 \tag{1.5}$$
$$u_1\cdot(x_1^2+x_2^2-5)=0 \qquad u_2\cdot(x_1-x_2-1)=0 \tag{1.6}$$
$$u_1,u_2\geq 0 \tag{1.7}$$

We consider $u_1=0$, then (5) is not solvable. Therefore, $u_1\neq 0$, and (6) implies $x_1^2+x_2^2-5=0$. If $u_2=0$, we express x_1 and x_2 from (5) using u_1. However, the substitution in (6) leads to contradiction. Therefore, $u_2\neq 0$, and from (6) we derive $x_1-x_2-1=0$. We express $\mathbf{x}_2$ and we get a quadratic equation by replacement. For the first root, we have $x_1=-1$, $x_2=-2$, that do not satisfy KKT conditions as $u_1=-2/3\not\geq 0$. For the second root $x_1=2$, $x_2=1$, $u_1=2/3$ and $u_2=1/3$, so KKT conditions are satisfied.

Remark 334 (Sufficient KKT for inequalities)**.**

They are very similar to FJ (see later for the general case combining inequalities and equalities). For FJ sufficient conditions strict pseudoconvexity and quasiconvexity have been required. With KKT sufficient conditions $g_i, i\in I$ must differentiable and quasiconvex.

Exercise 335.

Think about the fact, that convexity assumptions are not necessary even for convex programs. Hint: Check $\operatorname{argmin}\{-x_2 \mid (x_1-1)^2+x_2^2\leq 1,(x_1+1)^2+x_2^2\leq 1\}$ where only feasible point is $(0,0)$ and it is also optimal. However, although it is a trivial convex program, KKT are not satisfied as constraint qualification is not valid.

1.11.4 Constrained optimization — inequalities and equalities

Remark 336 (Generalization)**.**

Theoretical results of previous sections discussed separately for equalities and inequalities is possible to generalize and unify. For computational purposes, we have usually considered feasible solutions $\mathbf{x}\in\mathbb{R}^n$ satisfying certain constraints. The most of theory also remains valid for the general case when $\mathbf{x}\in X$ where X is an open nonempty set. So, we further consider $\underline{\mathrm{P}}$: $?\in\operatorname{argmin}_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x})\leq\mathbf{0},\mathbf{h}(\mathbf{x})=\mathbf{0},\mathbf{x}\in X\}$. Notice that $\mathbf{g}(\mathbf{x})=(g_1(\mathbf{x}),\ldots,g_m(\mathbf{x}))^\top$ and $\mathbf{h}(\mathbf{x})=(h_1(\mathbf{x}),\ldots,h_l(\mathbf{x}))^\top$.

Remark 337 (Overview – see literature for details)**.**

We have $\underline{\mathrm{P}}$, I set, $\bar{\mathbf{x}} \in$ arglocmin of $\underline{\mathrm{P}}$, f differentiable, $g_i, i \in I$ differentiable, $g_i, i \notin I$ continuous, and $h_i, \forall i$ continuously differentiable.
Geometrical condition: F_0, G_0 are defined as before, $H_0 = \{\mathbf{d} \mid \nabla h_i(\bar{\mathbf{x}})^\top \mathbf{d} = 0, i = 1, \ldots, l\}$ $\nabla h_i(\bar{\mathbf{x}})$ linearly independent $\Rightarrow F_0 \cap G_0 \cap H_0 = \emptyset$ (for the proof using ordinary differential equations, see literature).
Fritz John necessary conditions: We have $\underline{\mathrm{P}}$, I, f, g_i, h_i as before (and all functions are differentiable). If $\bar{\mathbf{x}} \in$ arglocmin of $\underline{\mathrm{P}} \Rightarrow$

$$u_0 \nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{g}(\bar{\mathbf{x}})^\top \mathbf{u} + \nabla \mathbf{h}(\bar{\mathbf{x}})^\top \mathbf{v} = \mathbf{0},$$
$$\mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0, \quad (u_0, \mathbf{u}^\top)^\top \geq \mathbf{0}, \quad (u_0, \mathbf{u}^\top, \mathbf{v}^\top)^\top \neq (0, \mathbf{0}^\top, \mathbf{0}^\top)^\top.$$

Fritz John sufficient conditions: We have $\underline{\mathrm{P}}$, $\bar{\mathbf{x}}$ FJ solution, I, and S is a relaxation of the feasible set at $\bar{\mathbf{x}}$. Then if h_i is affine, ∇h_i are linearly independent, and $\exists \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ such that f is pseudoconvex on $S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$ and $g_i, i \in I$ are strictly pseudoconvex on $S \cap \mathcal{N}_\varepsilon(\bar{\mathbf{x}}) \Rightarrow \bar{\mathbf{x}} \in$ arglocmin of $\underline{\mathrm{P}}$.

Theorem 338 (KKT 1st order necessary conditions)**.**

1. Let $X \subset \mathbb{R}^n$ be a nonempty open set, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and $\mathbf{h} : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ be functions. The components of $\mathbf{g}$ are denoted as g_i, components of $\mathbf{h}$ are denoted h_i.
2. We solve $\underline{\mathrm{P}}$: $? \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0},\ \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in X\}$.
3. We have $\bar{\mathbf{x}}$ feasible solution $\underline{\mathrm{P}}$. We denote $I = \{i \mid g_i(\mathbf{x}_0) = 0\}$ the index set of binding (active) inequality constraints.
4. Let f and g_i are differentiable at $\bar{\mathbf{x}}$ for $i \in I$, g_i are continuous for $i \notin I$ at $\bar{\mathbf{x}}$, and h_i are continuously differentiable at $\bar{\mathbf{x}}$ for $i = 1, \ldots, l$.
5. We also assume that gradients $\nabla g_i(\bar{\mathbf{x}}), i \in I$ and $\nabla h_i(\bar{\mathbf{x}}), i = 1, \ldots, l$ are linearly independent (constraint qualification).

If $\bar{\mathbf{x}}$ is a local minimum of $\underline{\mathrm{P}}$ then $\exists u_i, i \in I$ and $\exists \mathbf{v}_i, i = 1, \ldots, l$ such that the following KKT conditions are valid:

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{l} v_i \nabla h_i(\bar{\mathbf{x}}) = \mathbf{0}, \qquad u_i \geq 0 \qquad \text{for} \quad i \in I$$

Proof: Simple, based on FJ, similar as for inequalities — see literature. To get pure inequalities, you may set $v_i = v_i^+ - v_i^-, v_i^+, v_i^- \geq 0$. □

Remark 339.

Consider that previous Lagrange and KKT conditions may be obtained as corollaries of Theorem 338.

Remark 340 (Another formulations).

If we consider differentiability at $\bar{\mathbf{x}}$ also for $g_i, i \notin I$, then we may write KKT conditions equivalently as follows:

$$\nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{g}(\bar{\mathbf{x}})^\top \mathbf{u} + \nabla \mathbf{h}(\bar{\mathbf{x}})^\top \mathbf{v} = \mathbf{0}, \qquad \mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0, \qquad \mathbf{u} \geq \mathbf{0}.$$

Previously introduced concept of Lagrangian may be generalized as follows:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \quad = \quad f(\mathbf{x}) + \mathbf{g}(\mathbf{x})^\top \mathbf{u} + \mathbf{h}(\mathbf{x})^\top \mathbf{v}.$$

We may formulate KKT conditions using conditions for the existence of general Lagrangian stationary points and we obtain

$$\nabla_{\mathbf{x}} L = \mathbf{0}, \quad \nabla_{\mathbf{u}} L \leq \mathbf{0}, \quad \mathbf{u}^\top \nabla_{\mathbf{u}} L = \mathbf{0}, \quad \mathbf{u} \geq \mathbf{0}, \quad \nabla_{\mathbf{v}} L = \mathbf{0}. \tag{1.8}$$

If we have to introduce nonnegativity conditions $\mathbf{x} \geq \mathbf{0}$, we replace the condition $\nabla_{\mathbf{x}} L = \mathbf{0}$ in 8 by three conditions $\nabla_{\mathbf{x}} L \geq \mathbf{0}$, $\mathbf{x}^\top \nabla_{\mathbf{x}} L = \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$.

Theorem 341 (Sufficient KKT 1st order conditions).

We have P, $\bar{\mathbf{x}}$ feasible, I index set of indices of binding constraints. We suppose tha KKT conditions hold at $\bar{\mathbf{x}}$, so: $\exists \bar{u}_i \geq 0, i \in I$ $\exists \bar{v}_i \in \mathbb{R}, i = 1, \ldots, l$ such that:

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{l} \bar{v}_i \nabla h_i(\bar{\mathbf{x}}) = \mathbf{0}.$$

Let $J = \{i \mid \bar{v}_i > 0\}$, $K = \{i \mid \bar{v}_i < 0\}$ and f is pseudoconvex at $\bar{\mathbf{x}}$, g_i quasiconvex at $\bar{\mathbf{x}}$ $\forall i \in I$, and h_i is quasiconvex at $\bar{\mathbf{x}}$ $\forall i \in J$ and h_i is quasiconcave at $\bar{\mathbf{x}}$ $\forall i \in K$. Then $\bar{\mathbf{x}}$ is a global minimum of P.

Proof: Let $\bar{\mathbf{x}}$ be feasible for P. Because of quasiconvexity: $\forall i \in I$ $g_i(\mathbf{x}) \leq g_i(\bar{\mathbf{x}}) = 0$. Then, $g_i(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) = g_i(\lambda \mathbf{x} + (1 - \lambda)\bar{\mathbf{x}}) \leq \max\{g_i(\mathbf{x}), g_i(\bar{\mathbf{x}})\} = g_i(\bar{\mathbf{x}})$. So, $g_i(\mathbf{x})$ does not increase moving from $\mathbf{x}$ along $\mathbf{x} - \bar{\mathbf{x}}$: So, for $i \in I : \nabla g_(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0$. We multiply the inequality by $\bar{u}_i \geq 0$. Similarly, for h_i quasiconvex and quasiconcave: $\nabla h_i(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0, i \in J$ multiplied by $\bar{v}_i > 0$ and $\nabla h_i(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \geq 0, i \in K$ multiplied by $\bar{v}_i < 0$. So, together we have

$$(\sum_{i \in I} \bar{u}_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{l} \bar{v}_i \nabla h_i(\bar{\mathbf{x}}))^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0.$$

We use it for KKT formula and obtain $\nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \geq 0$. By pseudoconvexity of f at $\bar{\mathbf{x}}$ it implies that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$. □

Theorem 342 (Sufficient 2nd order KKT conditions).

We consider functions f, $\mathbf{g}$ a $\mathbf{h}$ from P 2nd order differentiable and vectors $\bar{\mathbf{x}}$, $\bar{\mathbf{u}}$ a $\bar{\mathbf{v}}$, satisfying the KKT conditions. We choose fixed $\bar{\mathbf{u}}$ and $\bar{\mathbf{v}}$ in Lagrangian, so we have $L(\mathbf{x}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ the function of the variable $\mathbf{x}$. We denote $\nabla_x^2 L(\mathbf{x})$ Hessian of $L(\mathbf{x}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ at $\mathbf{x}$. If $\nabla_x^2 L(\mathbf{x})$ is positive semidefinite at feasible points $\mathbf{x}$ of some neighbourhood of $\bar{\mathbf{x}}$ then $\bar{\mathbf{x}}$ is a point of a local minimum.

Remark 343 (Constraint qualification).

The KKT conditions require $\nabla g_i, \nabla h_i$ linearly independent. It is understandable and possible to check it in examples. This simple constraint qualification can can be significantly generalized (using properties of various cones at $\bar{\mathbf{x}}$)– see literature.
It is important to know that the KKT conditions require differentiability of considered functions and regularity. However, for the nondifferentiable case there are still conditions based on saddle point of L and duality — see later.

1.12 Constrained optimization — algorithms

Remark 344 (Main idea).

A lot of constrained optimization numerical algorithms uses the idea to repeatedly transform a multivariate nonlinear (constrained) program to a multivariate unconstrained optimization problem. For unconstrained minimization, several useful algorithms have been introduced. (They again repeatedly use an unconstrained problem transformation on the univariate line search problem). This 'encapsulating' principle will be often used for particular algorithms. The list of methods will completed with examples.

1.12.1 Penalty function-based algorithms

Remark 345 (Principle).

A nonlinear program is solved through the solution of a sequence of unconstrained problems. They are derived from the original NLP in two steps: (1) constraints are relaxed and (2) they are incorporated into the penalty term $\alpha(\mathbf{x})$ that is included in the objective function. The penalty term is chosen in such a way that feasible solutions (of the original NLP) are not penalized but for the infeasible solutions the objective function value is significantly increased.

Example 346.

Instead of $\min_x\{f(\mathbf{x}) \mid h(\mathbf{x}) = 0\}$ (the original constrained NLP), we solve $\min_x\{f(\mathbf{x}) + \mu h(\mathbf{x})^2 \mid \mathbf{x} \in \mathbb{R}^n\}$ (unconstrained approximating penalty problem), where μ is a penalty coefficient, and here, aforementioned $\alpha(\mathbf{x}) = \mu h(\mathbf{x})^2$. We see that for $\mathbf{x} \longrightarrow \mathbf{x}_{\min} \Rightarrow h(\mathbf{x})^2 \longrightarrow 0$.
Another possibility might be to solve $\min_x\{f(\mathbf{x}) + \mu \max\{0, g(\mathbf{x})\} \mid \mathbf{x} \in \mathbb{R}^n\}$ instead of the original $\min_x\{f(\mathbf{x}) \mid g(\mathbf{x}) \leq 0\}$. In this case, we have $\alpha(\mathbf{x}) = \mu \max\{0, g(\mathbf{x})\}$ and we may notice that if $g(\mathbf{x})$ is differentiable then $\alpha(\mathbf{x})$ need not be differentiable.

Exercise 347.

Illustrate penalty principle by own examples and figures. Start with $\min\{x \mid x \geq 2\}$ (Hint: e.g. use $\min\{x + \alpha(x) \mid x \in \mathbb{R}\}$ where $\alpha(x) = 0$ for $x \geq 0$ and $\alpha(x) = \mu(2-x)^2$ for $x < 2$). Continue with $\min\{x^2 \mid -1 \leq x \leq 2\}$.

Example 348 (The relation between solutions)**.**

Solve $\min\{x_1^2 + x_2^2 \mid x_1 + x_2 - 1 = 0\}$ graphically (the solution is $\mathbf{x}_{\min} = (0.5, 0.5)$). Reformulate the NLP given above using the penalty principle: $\min\{x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1 = 0)^2 \mid (x_1, x_2)^\top \in \mathbb{R}^2\}$. The solution $\mathbf{x}_{\mu,\min}$ of the reformulated NLP can be obtained using $\nabla f(\mathbf{x})) = \mathbf{0}$. So, we obtain $x_1 = x_2 = \frac{\mu}{2\mu+1}$. Thus, for $\mu \longrightarrow \infty$ $\Rightarrow \frac{\mu}{2\mu+1} \longrightarrow \frac{1}{2}$, and so $\mathbf{x}_{\mu,\min} \longrightarrow \mathbf{x}_{\min}$.

Remark 349 (Geometric interpretation of perturbed problem)**.**

We may consider the perturbed program: $\min\{x_1^2+x_2^2 \mid x_1+x_2-1=\varepsilon\} \Rightarrow$ We express x_2 from the equality constraint and we replace it in the objective function: The penalty reformulation is $\min\{x_1^2+(1+\varepsilon-x_1)^2 \mid x_1 \in \mathbb{R}\} \Rightarrow x_{1,\min} = x_{2,\min} = \frac{1+\varepsilon}{2}$ and the minimum value of the objective function is $v(\varepsilon) = \min\{x_1^2+x_2^2 \mid x_1+x_2-1=\varepsilon\} = \min\{x_1^2+(1+\varepsilon-x_1)^2 \mid x_1 \in \mathbb{R}\} = \frac{(1+\varepsilon)^2}{2}$.
So we may consider the mapping $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \begin{pmatrix} h(x_1,x_2) \\ f(x_1,x_2) \end{pmatrix}$. It means that the points are mapped from the space with coordinates x_1, x_2 to the space with coordinates h, f. It is also important to note that $h(\mathbf{x}) = \varepsilon$. Therefore, we may draw a graph of the function $v(\varepsilon) = \frac{(1+\varepsilon)^2}{2}$ using axes h, f – draw axes and v function.
It gives $f(\mathbf{x}_{\min})$ for different values of ε. For fixed value of ε points on the straight line specified by the equality $h = \varepsilon$ with f coordinate greater than $v(\varepsilon)$ represent the non-optimal objective function values $f(\mathbf{x})$. Therefore, all feasible points of the original perturbed problem (allowing $h(\mathbf{x}) \neq 0$) map to the epigraph of $v(\varepsilon)$. So, the solution of the original problem is derived geometrically (including the previous case $\varepsilon = 0$). Draw straight lines filling the epigraph, colour the case $\varepsilon = 0$.
Then, the penalty objective is specified by term $f + \mu h^2$. We may think about two variables h and f, and we may draw a contour graph of the function specified by this term — draw it.
Solving the penalty problem, we are interested to find a feasible point of the perturbed problem with the lowest $f + \mu h^2$ penalty objective value. Draw the point.
For different values of μ, you will get different solutions (points $(\varepsilon, v(\varepsilon)) = (h(\mathbf{x}_{\mu,\min}), f(\mathbf{x}_{\mu,\min}) + \mu h(\mathbf{x}_{\mu,\min})^2)$ different from the original problem solution $(0, v(0)) = (h(\mathbf{x}_{\min}), f(\mathbf{x}_{\min}) + \mu h(\mathbf{x}_{\min})^2) = (0, f(\mathbf{x}_{\min})))$. Check whether you draw both points.
With increasing μ curvature of parabola $f + \mu h^2$ changes and points $(h(\mathbf{x}_{\mu,\min}), f(\mathbf{x}_{\mu,\min}) + \mu h(\mathbf{x}_{\mu,\min})^2)$ are approaching to point $(0, f(\mathbf{x}_{\min}))$. Draw a sequence of points using several figures.
From figures, we see that the solution of penalty problem is infeasible (slightly with μ increasing).
We also see that this simple geometric interpretation can be generalized to $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^l$ (also to $\mathbf{g}(\mathbf{x}) \leq \mathbf{0} \in \mathbb{R}^m$, e.g., using slack variables). The whole visualizing approach is very suitable also for NLP duality.

Remark 350 (Nonconvex case)**.**

Even in the general case when $v(\varepsilon)$ might be nonconvex the approaching process works. The reason is that the parabola $f + \mu h^2$ may move along the nonconvex v (or for f and h nonconvex at $\mathbf{x}$, $f + \mu h^2$ tends to convexity at $\mathbf{x}$ with increasing μ).

Remark 351 (General penalty function).

The penalty function $\alpha(\mathbf{x})$ is defined such that it has the following property: Its value is positive for infeasible $\mathbf{x}$ and 0 for feasible $\mathbf{x}$. For optimization problem $\underline{P}$ (see previous section), we choose:

$$\alpha(\mathbf{x}) \quad = \quad \sum_{i=1}^{m} \phi(g_i(\mathbf{x})) + \sum_{i=1}^{l} \psi(h_i(\mathbf{x})),$$

where ϕ and ψ are continuous such that: $\phi(y) = 0$ for $y \leq 0$ and $\phi(y) > 0$ for $y > 0$ (cf. $g_i(\mathbf{x}) \leq 0$); $\psi(y) = 0$ for $y = 0$ and $\psi(y) > 0$ for $y \neq 0$ (cf. $h_i(\mathbf{x}) = 0$).
Typically $\phi(y) = (\max\{0; y\})^p$ a $\psi(y) = |y|^p$ where p is positive integer. In this case, the penalty function is

$$\alpha(\mathbf{x}) \quad = \quad \sum_{i=1}^{m} (\max\{0; g_i(\mathbf{x})\})^p + \sum_{i=1}^{l} |h_i(\mathbf{x})|^p.$$

We refer to the function $f(\mathbf{x}) + \mu\alpha(\mathbf{x})$ (new objective) as the auxiliary function.

Remark 352 (General formulation).

We have $X \subset \mathbb{R}^n$, $X \neq \emptyset$, $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^m$, $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^l$ and $? \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\}$ and all $f, g_i (i = 1, \ldots, m), h_i (i = 1, \ldots, l)$ are continuous on $\mathbb{R}^n$. Let α be a continuous and satisfying penalty function properties (as it was already discussed). Then (basic) penalty problem is:

$$\max\{\theta(\mu) \mid \mu \geq 0, \theta(\mu) = \inf\{f(\mathbf{x}) + \mu\alpha(\mathbf{x}) \mid \mathbf{x} \in X\}\}$$

The following lemma and theorem say:

$$\inf\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} = \sup\{\theta(\mu) \mid \mu \geq 0\} = \lim_{\mu \longrightarrow \infty} \theta(\mu)$$

So, we can get arbitrarily close to $\mathbf{x}_{\min}$ solving $\theta(\mu)$ for large μ.

Lemma 353.

Let $X \subset \mathbb{R}^n$, $X \neq \emptyset$, f, $g_1, \ldots, g_m$, $h_1, \ldots, l$ continuous functions (as above), α penalty function (as above), and $\forall \mu \; \exists \mathbf{x}_\mu \in X : \theta(\mu) = f(\mathbf{x}_\mu) + \mu\alpha(\mathbf{x}_\mu) \Rightarrow$

1. $\inf\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} \geq \sup\{\theta(\mu) \mid \mu \geq 0\}$ where $\theta(\mu) = \inf\{f(\mathbf{x}) + \mu\alpha(\mathbf{x}) \mid \mathbf{x} \in X\}$
2. $f(\mathbf{x}_\mu)$ is a nondecreasing function of $\mu \geq 0$, $\theta(\mu)$ is a nondecreasing function of μ and $\alpha(\mathbf{x}_\mu)$ is a nonincreasing function of μ.

Proof: See literature. □

Theorem 354.

Let $X \subset \mathbb{R}^n$, $X \neq \emptyset$, $f, g_1, \ldots, g_m, h_1, \ldots, l$ continuous functions (as above), α penalty function (as above), and $\forall \mu \; \exists \mathbf{x}_\mu \in X : \mathbf{x}_\mu \in \operatorname{argmin}\{f(\mathbf{x}) + \mu\alpha(\mathbf{x}) \mid \mathbf{x} \in X\}$ and the sequence $\{\mathbf{x}_\mu\}$ is contained in a compact subset of X. Then:

$$\inf\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} = \sup\{\theta(\mu) \mid \mu \geq 0\} = \lim_{\mu \longrightarrow \infty} \theta(\mu)$$

where $\theta(\mu) = \inf\{f(\mathbf{x}) + \mu\alpha(\mathbf{x}) \mid \mathbf{x} \in X\}\} = f(\mathbf{x}_\mu) + \mu\alpha(\mathbf{x}_\mu)$.
Furthermore, limit $\bar{\mathbf{x}}$ of any convergent subsequence of $\{\mathbf{x}_\mu\}$ is an optimal solution to the original problem and $\mu\alpha(\mathbf{x}_\mu) \longrightarrow 0$ as $\mu \longrightarrow \infty$.

Proof: Main ideas: We want to apply general convergence theorems. By lemma $\theta(\mu)$ monotone $\Rightarrow$ $\sup \ldots = \lim \ldots$. $\alpha(\mathbf{x}_\mu) \longrightarrow 0$ is proven using $\alpha(\mathbf{x}_\mu) \leq \varepsilon$ and $\varepsilon > 0$ and $\varepsilon \longrightarrow 0$. So, $\alpha(\bar{\mathbf{x}}) = 0$. Ten, $\bar{\mathbf{x}}$ is optimal because of convergence and continuity $\mathbf{x}_\mu \longrightarrow \bar{\mathbf{x}}$, so $\sup \ldots = f(\bar{\mathbf{x}})$. At the end $\mu\alpha(\mathbf{x}_\mu) = \theta(\mu) - f(\mathbf{x}_\mu)$, and as $\mu \longrightarrow \infty$, it approaches 0. □

Corollary 355.

$\alpha(\mathbf{x}_\mu) = 0 \Rightarrow \mathbf{x}_\mu$ solves the problem.

Remark 356.

In applications we often have X compact, so $\{\mathbf{x}_\mu\}$ belongs to compact and the assumption is satisfied.
Notice that $\mathbf{x}_\mu$ optimal for penalty problem can be made arbitrarily close to the feasible region.
$f(\mathbf{x}_\mu) + \mu\alpha(\mathbf{x}_\mu)$ can be made arbitrarily close to the optimal objective function value of the original primal problem.
$\{\mathbf{x}_\mu\}$ points with μ increasing gives the following algorithm. $\mathbf{x}_\mu$ are infeasible approaching the feasible region, so the frequent names are 'outer minimization' or 'exterior penalty function method'.

Remark 357 (KKT multipliers)**.**

With α, ϕ, ψ differentiable (Remember that function $\psi(g_i(\mathbf{x})) = |\max\{0, g_i(\mathbf{x})\}|^p$ is not necessarily differentiable with respect to inner function – ϕ' does not exist – or components of $\mathbf{x}$.) KKT multipliers can be computed in the case of the unique solution $\bar{\mathbf{x}}$. Then, the sequences may be used: $u_{\mu,i} \longrightarrow u_i$ and $v_{\mu,i} \longrightarrow v_i$ where $u_{\mu,i} = \mu\phi'(g_i(\mathbf{x}_\mu))$ $v_{\mu,i} = \mu\psi'(h_i(\mathbf{x}_\mu))$.

Algorithm 358 (SUMT — Sequential Unconstrained Minimization Technique).

Set $\varepsilon > 0$, starting point $\mathbf{x}_1$, penalty coefficient $\mu_1 > 0$ and $\beta > 1$, $k := 1$.

1. For starting solution $\mathbf{x}_k$, we solve the following penalty problem $\mathbf{x}_{\mu_k} \in \operatorname{argmin}\{f(\mathbf{x}) + \mu_k \alpha(\mathbf{x}) \mid \mathbf{x} \in X\}$. For $\mathbf{x}_{\mu_k}$ optimal solution with the objective function value $\theta(\mu_k) = f(\mathbf{x}_{\mu_k}) + \mu_k \alpha(\mathbf{x}_{\mu_k})$, we assign $\mathbf{x}_{k+1} := \mathbf{x}_{\mu_k}$ (a new starting point for the next iteration).

2. If $\mu_k \alpha(\mathbf{x}_{k+1}) < \varepsilon$ then STOP, otherwise $\mu_{k+1} := \beta \mu_k$, $k := k+1$ and GOTO 1.

Figure 1.4: Illustration of Hooke-Jeeves algorithm.

Example 359.

Solve $\min\{(x_1 - 2)^4 + (x_1 - 2x_2)^2 \mid x_1^2 - x_2 = 0, \mathbf{x} \in \mathbb{R}^2\}$ using a quadratic penalty function $\alpha(x_1, x_2) = (x_1^2 - x_2)^2$, $\mu_1 = 0,1$, $\beta = 10$ and Algorithm 358. The increasing value μ_k forces the infeasibility as small as it is possible.

k	μ_k	$\mathbf{x}_{k+1} = \mathbf{x}_{\mu_k}$	$f(\mathbf{x}_{k+1})$	$\alpha(\mathbf{x}_{k+1})$	$\mu_k \alpha(\mathbf{x}_{k+1})$	$\theta(\mu_k)$
1	0,1	(1,4539; 0,7608)	0,0935	1,8307	0,1831	0,2766
2	1,0	(1,1687; 0,7407)	0,5753	0,3908	0,3908	0,9661
3	10,0	(0,9906; 0,8425)	1,5203	0,01926	0,1926	1,7129
4	100,0	(0,9507; 0,8875)	1,8917	0,000267	0,0267	1,9184

Remark 360 (Computational difficulties).

For very large μ sometimes ill conditioning (cf. eigenvalues of Hessian of auxiliary function) may cause premature termination (near to feasible region but for from the optimum). So improvement may be required.

Remark 361 (Improvements).

The idea is to avoid computational problems for cases where we do not need $\mu \longrightarrow \infty$. Only μ large enough and still approaching $\mathbf{x}_{\min}$. It works for exact penalty functions. We may choose l_1 (absolute value) penalty function ($p = 1$). Theory says that for the KKT point and multipliers, for the convex case, to reach optimum is enough to solve the problem with l_1 penalty function and $\mu \geq \max\{u_i (i = 1, \ldots, m), v_i (i = 1, \ldots, l)\}$.

Exercise 362.

Draw figure for your own example and the case of $f + \mu|h|$ instead of $f + \mu h^2$.

Remark 363 (Augmented Lagrangian.)**.**

Therefore, recently, Lagrangian-based penalty functions are utilized (see software packages, e.g., MINOS accompanying GAMS). The idea is to shift penalty to constant θ to obtain $f(\mathbf{x}) + \mu \sum_{i=1}^{l} (h_i(\mathbf{x}) - \theta_i)^2$ (θ_i perturbed from 0). After simple computations and changed notation, we have $f(\mathbf{x}) + \sum_{i=1}^{l} v_i h_i(\mathbf{x}) + \mu \sum_{i=1}^{l} h_i^2(\mathbf{x})$. It does not need $\mu \longrightarrow \infty$ and it is differentiable (cf. l_1). For software implementations, the general objective of the form $f(\mathbf{x}) + \sum_{i=1}^{m} u_i(g_i(\mathbf{x}) + s_i^2) + \sum_{i=1}^{l} v_i h_i(\mathbf{x}) + \mu(\sum_{i=1}^{m} (g_i(\mathbf{x}) + s_i^2)^2 + \sum_{i=1}^{l} h_i(\mathbf{x})^2)$ is used. Frequently different penalty coefficients μ_i for different constraints are used. In addition, it is necessary to specify the way how Lagrange multipliers will be updated.

1.12.2 Barrier function-based algorithms

Remark 364.

We will further consider the following NLP:

$$\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in X\},$$

where $X \subset \mathbb{R}^n$, $X \neq \emptyset$ (it may involve equality constraints), f and components of $\mathbf{g}$ are continuous functions. The equality constraints are involved in X because the barrier function requires the nonempty interior of the set described by explicit constraints $\{\mathbf{x} \mid \mathbf{g}(\mathbf{x}) < \mathbf{0}\} \neq \emptyset$. (It is not enough to replace $h(\mathbf{x}) = 0$ with $h(\mathbf{x}) \geq 0$ and $h(\mathbf{x}) \leq 0$.)

Exercise 365 (Examples of barrier functions)**.**

Explain why the following barrier functions are as follows: $B(\mathbf{x}) = \sum_{i=1}^{m} (-\frac{1}{g_i(\mathbf{x})})$, $B(\mathbf{x}) = -\sum_{i=1}^{m} \ln(-g_i(\mathbf{x}))$, and $B(\mathbf{x}) = -\sum_{i=1}^{m} \ln(min\{1, -g_i(\mathbf{x})\})$.

Remark 366 (General requirements)**.**

We have $\min\{\theta(\mu) \mid \mu \geq 0\}$ and $\theta(\mu) = \inf\{f(\mathbf{x}) + \mu B(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) < \mathbf{0}, \mathbf{x} \in X\}$. There is $B(\mathbf{x}) = \sum_{i=1}^{m} \phi(g_i(\mathbf{x}))$, where $\phi(y)$ is continuous over $\{y \mid y < 0\}$, if $y < 0$ then $\phi(y) \geq 0$, and $\phi(y) \longrightarrow \infty$ with $y \longrightarrow 0^-$.

Remark 367 (Convergence).

There are $f, g_1, \ldots, g_m$ continuous functions on $\mathbb{R}^n$, $X \subset \mathbb{R}^n$, $X \neq \emptyset$ closed and $\{\mathbf{x} \in X \mid \mathbf{g}(\mathbf{x}) < \mathbf{0}\} \neq \emptyset$. There is B a barrier function, as above, continuous on set $\{\mathbf{x} \in X \mid \mathbf{g}(\mathbf{x}) < \mathbf{0}\}$. We assume that $\bar{\mathbf{x}}$ is optimal and $\forall \mathcal{N}_\varepsilon(\bar{\mathbf{x}})\ \exists \mathbf{x} \in \mathcal{N}_\varepsilon(\bar{\mathbf{x}})$: $\mathbf{g}(\mathbf{x}) < \mathbf{0}$. Then: $\min\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in X\} = \lim_{\mu \longrightarrow 0^+} \theta(\mu) = \inf\{\theta(\mu) \mid \mu > 0\}$ where $\theta(\mu) = \inf\{f(\mathbf{x}) + \mu B(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) < \mathbf{0}, \mathbf{x} \in X\}$; and so $\theta(\mu) = f(\mathbf{x}_\mu) + \mu B(\mathbf{x}_\mu)$ where $\mathbf{x}_\mu$ is feasible, i.e., $\mathbf{x}_\mu \in X \wedge \mathbf{g}(\mathbf{x}_\mu) < \mathbf{0}$; and a convergent subsequence of $\{\mathbf{x}_\mu\}$ converges to $\bar{\mathbf{x}}$ and $\mu B(\mathbf{x}_\mu) \longrightarrow 0$ as $\mu \longrightarrow 0^+$.

Remark 368 (Principles).

While the penalty method generates a sequence infeasible points that approaches the boundary of the feasible set from outside (cf. Figure **??**), the barrier method creates a sequence of feasible points that approaches the boundary from inside. Emphasize that the barrier does not allow to reach the boundary. We may discuss just to NLPs having inequalities Ineq with the nonempty interior. the reason is that under the equalities the method cannot work with interior points. certain possibility is given by the replacement of $h_i(\mathbf{x}) = 0$ by relaxed constraint $h_i(\mathbf{x})^2 \leq \varepsilon$. The barrier is involved in Ineq in such a way that constraints $\mathbf{g}(\mathbf{x}) < \mathbf{0}$ are considered and the objective function is enriched with the barrier function $B(\mathbf{x})$:

$$B(\mathbf{x}) \quad = \quad \sum_{i=1}^{m} \phi(g_i(\mathbf{x})),$$

where, usually, $\phi(y) = -1/y$, $\phi(y) = -\ln(-y)$ or $\phi(y) = -\ln(\min\{1, -y\})$.

Algorithm 369 (Barrier).

We choose $\varepsilon > 0$, $\mathbf{x}_1$, satisfying $\mathbf{g}(\mathbf{x}_1) < \mathbf{0}$, and barrier coefficients $\mu_1 > 0$ and $\beta \in (0, 1)$, $k := 1$.

1. For the initial value $\mathbf{x}_k$, we solve $\min\{f(\mathbf{x}) + \mu_k B(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) < \mathbf{0},\ \mathbf{x} \in X\}$. If $\mathbf{x}_{\mu_k}$ then the optimum solution value $\theta(\mu_k) = f(\mathbf{x}_{\mu_k}) + \mu_k B(\mathbf{x}_{\mu_k})$, is used to obtain $\mathbf{x}_{k+1} := \mathbf{x}_{\mu_k}$.

2. If $\mu_k B(\mathbf{x}_{k+1}) < \varepsilon$ then STOP, otherwise $\mu_{k+1} := \beta\mu_k$, $k := k+1$ and GOTO 1.

Example 370.

Solve $\min\{(x_1-2)^4+(x_1-2x_2)^2 \mid x_1^2-x_2\le 0, \mathbf{x}\in \mathbb{R}^2\}$ using a barrier function $B(x_1,x_2)=-1/(x_1^2-x_2)$, $\mu_1=10,0$, $\beta=0,1$ and by Algorithm 369.

k	μ_k	$\mathbf{x}_{k+1}=\mathbf{x}_{\mu_k}$	$f(\mathbf{x}_{k+1})$	$B(\mathbf{x}_{k+1})$	$\mu_k B(\mathbf{x}_{k+1})$	$\theta(\mu_k)$
1	10.0	$(0.7079; 1.5315)$	8.3338	0.9705	9.705	18.0388
2	1.0	$(0.8282; 1.1098)$	3.8214	2.3591	2.3591	6.1805
3	0.1	$(0.8989; 0.9638)$	2.5282	6.4194	0.6419	3.1701
4	0.01	$(0.9294; 0.9162)$	2.1291	19.0783	0.1908	2.3199

Figure 1.5: Illustration of KKT conditions.

Remark 371.

Fiacco and McCormick developed the SUMT algorithm combining the penalty approach for equalities with a barrier approach for inequalities.
There are also very efficient modifications (interior point methods) for linear programming.

Exercise 372.

For penalty and barrier methods, be able to formulate the approximating problem (see $\theta(\mu)$). With this problem, be able to realize one iteration (e.g., by the gradient method). Understand differences between penalty (exterior penalty function) and barrier (interior penalty function) methods. Hint: In contrast to penalty, $B(\mathbf{x})\neq 0$ (barrier) for $\mathbf{x}$ feasible (cf. $\alpha(\mathbf{x})=0$ for $\mathbf{x}$ feasible with penalty).

Remark 373 (Computational difficulty)**.**

Another computational difficulty is that we have to start from $\mathbf{x}$ relative interior point with respect to $\{\mathbf{x}\in X \mid \mathbf{g}(\mathbf{x})<\mathbf{0}\}$ and also line search, e.g., with discrete steps, nay lead out of the feasible set and, e.g., $B(\mathbf{x})<0$.

1.12.3 Methods of feasible directions

Remark 374 (The choice of feasible direction)**.**

Another possibility to solve a constrained optimization problem is based on the idea to search for feasible directions within the feasible set and use them during iterations. We start with the easiest classical and understandable Zoutendijk's method for Ineq.

Algorithm 375 (Zoutendijk).

Choose $\mathbf{x}_1$ feasible, i.e. $\mathbf{g}(\mathbf{x}_1) \leq \mathbf{0}$ and $k := 1$.

1. Denote $I = \{i \mid g_i(\mathbf{x}_k) = 0\}$ and solve

$$\min\{z \mid \nabla f(\mathbf{x}_k)^\top \mathbf{d} - z \leq 0,\ \nabla g_i(\mathbf{x}_k)^\top \mathbf{d} - z \leq 0,\ i \in I,\ -1 \leq d_j \leq 1,\ j = 1, \ldots, n\}$$

We denote $(z_k, \mathbf{d}_k)$ as the optimal solution. If $z = 0$ then STOP and $\mathbf{x}_k$ is a point satisfying conditions of Fritz John. If $z_k < 0$ then we further continue.

2. We obtain the solution λ_k of $\min\{f(\mathbf{x}_k + \lambda \mathbf{d}_k) \mid 0 \leq \lambda \leq \lambda_{\max}\}$,
where $\lambda_{\max} = \sup\{\lambda \mid g_i(\mathbf{x}_k + \lambda \mathbf{d}_k) \leq 0,\ i = 1, \ldots, m\}$.
We assign $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$, $k := k + 1$ and GOTO 1.

Remark 376 (Topkis-Veinott).

The method has several weak points (mainly the algorithmic map is not closed, and hence, the convergence is not guaranteed), however, there are several improving modifications (e.g., Topkis and Veinott achieved the convergence by replacing the constraint $\nabla g_i(\mathbf{x}_k)^\top \mathbf{d} - z \leq 0, i \in I$ by $\nabla g_i(\mathbf{x}_k)^\top \mathbf{d} - z \leq -g_i(\mathbf{x}_k), i = 1, \ldots, m$).

Remark 377 (Rosen projected gradient).

Another idea (Rosen) is to use a projected gradient instead of reduced one. With constrained problems, we often cannot use the steepest direction because it may lead (immediately) to infeasible points. The idea is to project $-\nabla f(\mathbf{x})$ in such a way that improves the objective function and maintains feasibility.

Remark 378 (Wolfe).

The most promising idea (implementation viewpoint) is Wolfe's reduced gradient method that works with the feasible directions. It depends upon reducing the dimensionality of the problem by representing all variables in terms of an independent subset of the variables. Originally has been developed for linear constraints.
We have $\min\{f(\mathbf{x}) \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, where $r(\mathbf{A}) = m$, f continuously differentiable on $\mathbb{R}^n$. Nondegeneracy assumption is satisfied: Any columns of $\mathbf{A}$ are linearly independent and each EP of the feasible set has m strictly positive variables. So, every feasible solution has at least m positive components and at most $n - m$ zero components. Let $\mathbf{x}$ be feasible $\mathbf{A} = [\mathbf{B}, \mathbf{N}]$, $\mathbf{x} = \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix}$. We denote components of $\nabla f(\mathbf{x}) = \begin{pmatrix} \nabla_B f(\mathbf{x}) \\ \nabla_N f(\mathbf{x}) \end{pmatrix}$ and $\mathbf{d}$ is an improving feasible direction (cf. LP) of f at $\mathbf{x}$ if $\nabla f(\mathbf{x})^\top \mathbf{d} < 0, \mathbf{Ad} = \mathbf{0} (x_j = 0 \Rightarrow d_j \geq 0)$. We find $\mathbf{d} = \begin{pmatrix} \mathbf{d}_B \\ \mathbf{d}_N \end{pmatrix}$ as follows: $\mathbf{0} = \mathbf{Ad} = \mathbf{B}\mathbf{d}_B + \mathbf{N}\mathbf{d}_N \Rightarrow \mathbf{d}_B = -\mathbf{B}^{-1}\mathbf{N}\mathbf{d}_N$. Therefore, (cf. with LP) $\nabla f(\mathbf{x})^\top \mathbf{d} = (\nabla_B f(\mathbf{x})^\top, \nabla_N f(\mathbf{x})^\top) \begin{pmatrix} \mathbf{d}_B \\ \mathbf{d}_N \end{pmatrix} = \nabla_B f(\mathbf{x})^\top (-\mathbf{B}^{-1}\mathbf{N}\mathbf{d}_N) + \nabla_N f(\mathbf{x})^\top \mathbf{d}_N = (\nabla_N f(\mathbf{x})^\top - \nabla_B f(\mathbf{x})^\top \mathbf{B}^{-1}\mathbf{N})\mathbf{d}_N = \mathbf{r}_N^\top$, where $\mathbf{r}_N^\top$ is a reduced gradient. Then, $\mathbf{r}^\top = \begin{pmatrix} \mathbf{r}_B \\ \mathbf{r}_N \end{pmatrix} = \nabla f(\mathbf{x})^\top - \nabla_B f(\mathbf{x})^\top \begin{pmatrix} \mathbf{B}^{-1}\mathbf{B} \\ \mathbf{B}^{-1}\mathbf{N} \end{pmatrix} = (\mathbf{0}^\top, \mathbf{r}_N^\top)$. So, we used $\mathbf{Ax} = \mathbf{b}$ to obtain a reduced description of feasible $\mathbf{d}$, we need to assign $\mathbf{d}_N$ to obtain descent direction (With LP — edge descent direction was $\mathbf{d} = \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix} \geq \mathbf{0}$.) We must choose $\mathbf{d}_N$ to obtain $\mathbf{r}_N^\top \mathbf{d}_N < 0$ (and $x_j = 0 \Rightarrow d_j \geq 0$). $\forall j$ nonbasic: $r_j \leq 0 \Rightarrow d_j = -r_j$ and $r_j > 0 \Rightarrow d_j = -x_j r_j$.

Theorem 379.

We have $\min\{f(\mathbf{x}) \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, $r(\mathbf{A}) = m$, $\mathbf{x} = \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{pmatrix}$ feasible, $\mathbf{x}_B > \mathbf{0}$, $\mathbf{A} = [\mathbf{B}, \mathbf{N}]$, and $\exists \mathbf{B}^{-1}$. f is differentiable at $\bar{\mathbf{x}}$, $\mathbf{r}^\top = \nabla f(\mathbf{x})^\top - \nabla_B f(\mathbf{x})^\top \mathbf{B}^{-1}\mathbf{A}$. $\mathbf{d} = \begin{pmatrix} \mathbf{d}_B \\ \mathbf{d}_N \end{pmatrix}$ is defined $\forall j \in \mathbb{N} : d_j = -r_j$ if $r_j \leq 0$ and $d_j = -x_j d_j$ if $r_j > 0$ and $\mathbf{d}_B = -\mathbf{B}^{-1}\mathbf{N}\mathbf{d}_N$. If $\mathbf{d} \neq \mathbf{0}$ then $\mathbf{d}$ is an improving feasible direction. $\mathbf{d} = \mathbf{0}$ iff $\mathbf{x}$ is a KKT point.

Proof: See literature. □

Algorithm 380 (Wolfe)**.**

Choose $\mathbf{x}_1$ feasible, and $k := 1$.

1. $\mathbf{d} = \begin{pmatrix} \mathbf{d}_B \\ \mathbf{d}_N \end{pmatrix}$ is obtained from (*). If $\mathbf{d} = \mathbf{0}$ then STOP and $\mathbf{x}_k$ is a KKT point (Even Lagrange multipliers are obtained for constraints $\mathbf{Ax} = \mathbf{b} \ldots \nabla_B f(\mathbf{x})^\top \mathbf{B}^{-1}, \mathbf{x} \geq \mathbf{0}, \ldots, \mathbf{r})$. $d_j = -r_j$ if $i \notin I_k$ and $r_j \leq 0$ and $d_j = -x_j r_j$ if $i \notin I_k$ and $r_j > 0$ and $\mathbf{d}_B = -\mathbf{B}^{-1}\mathbf{N}\mathbf{d}_N$. where I_k is an index set of the m largest components of $\mathbf{x}_k$, and so $\mathbf{B} = \{\mathbf{a}_j \mid j \notin I_k\}$. and $\mathbf{B} = \{\mathbf{a}_j \mid j \notin I_k\}$ and $\mathbf{r}^\top = \nabla_N f(\mathbf{x}_k)^\top - \nabla_B f(\mathbf{x}_k)^\top \mathbf{B}^{-1}\mathbf{A}$.

2. Solve the line search problem $\lambda_k \in \operatorname{argmin}\{f(\mathbf{x}_k + \lambda \mathbf{d}_k) \mid 0 \leq \lambda \leq \lambda_{\max}\}$, where $\lambda_{\max} = \infty$ for $\mathbf{d}_k \geq \mathbf{0}$ and $\min_{1 \leq j \leq n}\{-\frac{x_{jk}}{d_{jk}}\}$ otherwise $\lambda_{\max} = \infty$ for $\mathbf{d}_k \geq \mathbf{0}$. Then, $\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_k$, $k := k + 1$ and GOTO 1. .

Remark 381 (Zangwill)**.**

There is also Zangwill's convex-simplex method. It is similar to Wolfe's method. But the only nonbasic variable is modified while the other nonbasic variables are fixed at their current levels (cf. LP simplex).

Remark 382.

The following Wolfe's method generalization is used for Eq. It is implemented in the CONOPT solver (cf. GAMS and AIMMS).

Remark 383 (Solved program)**.**

We solve a program Eq including bounds on $\mathbf{x}$:

$$\min\{f(\mathbf{x}) \mid \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}.$$

We assume differentiability of f and $\mathbf{h}$ functions. Therefore, we may use linearization $\mathbf{h}(\mathbf{x}_k) + \nabla \mathbf{h}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)$ and use previous remarks. We also suppose nondegeneracy that, i.e. any feasible solution $\mathbf{x}$ may be decomposed in $\mathbf{x} = (\mathbf{x}_B^\top, \mathbf{x}_N^\top)^\top$, where $\mathbf{x}_B \in \mathbb{R}^l$ a $\mathbf{x}_N \in \mathbb{R}^{n-l}$. Similarly, it is possible to split $\mathbf{a}$ and $\mathbf{b}$ in such a way that $\mathbf{a}_B < \mathbf{x}_B < \mathbf{b}_B$ is valid. In addition, Jacobi matrix $\nabla \mathbf{h}(\mathbf{x})$ may be split in two matrices: a regular square matrix $\nabla_B \mathbf{h}(\mathbf{x})$ and the remaining matrix $\nabla_N \mathbf{h}(\mathbf{x})$.

Algorithm 384 (Abadie-Carpentier)**.**

Choose the feasible solution $\mathbf{x}$ and specify $\mathbf{x}_B$ a $\mathbf{x}_N$.

1. We assign $\mathbf{r}^\top = \nabla_N f(\mathbf{x})^\top - \nabla_B f(\mathbf{x})^\top \nabla_B \mathbf{h}(\mathbf{x})^{-1} \nabla_N \mathbf{h}(\mathbf{x})$. We specify vector $\mathbf{d}_N$ of dimension $n-l$ that j-th component of d_j is equal 0 for $x_j = a_j$ and $r_j > 0$ or $x_j = b_j$ and $r_j < 0$. Otherwise $d_j := -r_j$. If $\mathbf{d}_N = \mathbf{0}$ then STOP and a point $\mathbf{x}$ satisfies the KKT conditions. Otherwise, we continue.

2. We find a solution $\mathbf{y}$ of the system of nonlinear equations $\mathbf{h}(\mathbf{y}, \bar{\mathbf{x}}_N) = \mathbf{0}$ by Newton's method ($\bar{\mathbf{x}}_N$ is specified later).

   We choose $\varepsilon > 0$ and positive integer K. Therefore, $\theta > 0$ satisfies $\mathbf{a}_N \leq \bar{\mathbf{x}}_N \leq \mathbf{b}_N$, where $\bar{\mathbf{x}}_N = \mathbf{x}_N + \theta \mathbf{d}_N$. We define $\mathbf{y}_1 := \mathbf{x}_B$, $k := 1$.

   A. We select $\mathbf{y}_{k+1} := \mathbf{y}_k - \nabla_B \mathbf{h}(\mathbf{y}_k, \bar{\mathbf{x}}_N)^{-1} \mathbf{h}(\mathbf{y}_k, \bar{\mathbf{x}}_N)$. Je-li $\mathbf{a}_B \leq \mathbf{y}_{k+1} \leq \mathbf{b}_B$, $f(\mathbf{y}_{k+1}, \bar{\mathbf{x}}_N) < f(\mathbf{x}_B, \mathbf{x}_N)$ and $\|\mathbf{h}(\mathbf{y}_{k+1}, \bar{\mathbf{x}}_N)\| < \varepsilon$ then GOTO C. Otherwise GOTO B.

   B. If $k = K$ then $\theta := \frac{1}{2}\theta$, $\bar{\mathbf{x}}_N := \mathbf{x}_N + \theta \mathbf{d}_N$, and so, $\mathbf{y}_1 := \mathbf{x}_B$, $k := 1$ and GOTO A. In contrast, $k := k + 1$ and GOTO A.

   C. We define $\mathbf{x} := (\mathbf{y}_{k+1}^\top, \bar{\mathbf{x}}_N^\top)$, identify new basis B and GOTO 1.

Exercise 385.

Explain principles of feasible direction methods. Try to reformulate own NLP problem for the selected algorithm and realize one iteration using linear programming.

1.12.4 Successive linear programming approach

Remark 386 (Approximation of programs)**.**

Engineering problems of oil industry in sixties involved less non-linear terms. Therefore, techniques based on a repeated local approximation by linear (or quadratic) programs have been developed.

Algorithm 387 (Griffith-Stewart).

We choose $\varepsilon > 0$ (terminating scalar), $\delta > 0$ (limiting iteration movement), $\mathbf{x}_1$, $k := 1$.

1. Solve a linear program

$$\min\{\nabla f(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k) \mid \nabla \mathbf{g}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) \leq -\mathbf{g}(\mathbf{x}_k),\ \nabla \mathbf{h}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) = -\mathbf{h}(\mathbf{x}_k),$$
$$\mathbf{a} \leq \mathbf{x} \leq \mathbf{b},\ -\delta \leq x_i - x_{ik} \leq \delta,\ i = 1,\ldots,n\},$$

where x_{ik} is the i-th component $\mathbf{x}_k$. The optimal solution is denoted as $\mathbf{x}_{k+1}$.

2. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \varepsilon$ and $\mathbf{x}_{k+1}$ is near feasible then STOP. Otherwise, in the case $k := k+1$ and GOTO 1.

Remark 388 (Comments).

It is useful for a few (hundreds) nonlinear terms. It is generalized in two directions: (1) l_1 penalty is used as a merit function deciding whether to reject a new iterate; (2) with the augmented Lagrangian the successive quadratic approximation is used.

1.13 Lagrangian duality

Remark 389 (Primal and dual problems).

A primal problem of NLP is P: $\min_x\{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\}$. Lagrangian dual problem D: is defined as $\max_{u,v}\{\theta(\mathbf{u},\mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\}$, where $\theta(\mathbf{u},\mathbf{v}) = \inf_x\{L(\mathbf{x},\mathbf{u},\mathbf{v}) \mid \mathbf{x} \in X\}$ i.e. $\theta(\mathbf{u},\mathbf{v}) = \inf\{f(\mathbf{x}) + \mathbf{u}^\top\mathbf{g}(\mathbf{x}) + \mathbf{v}^\top\mathbf{h}(\mathbf{x}) \mid \mathbf{x} \in X\}$.

Example 390.

Find a dual program explicit formulation (if it is possible — useful — suitable) for: $\begin{pmatrix} 2 \\ 2 \end{pmatrix} \in \operatorname{argmin}_{x_1,x_2}\{x_1^2 + x_2^2 \mid -x_1 - x_2 + 4 \leq 0, x_1, x_2 \geq 0\}$. We denote $g(\mathbf{x}) = -x_1 + x_2 + 4$ and $X = \{\mathbf{x} \mid x_1, x_2 \geq 0\}$. Then $\theta(u) = \inf_x\{x_1^2 + x_2^2 + u(-x_1 - x_2 + 4) \mid x_1, x_2 \geq 0\} = \inf_{x_1}\{x_1^2 - ux_1 \mid x_1, x_2 \geq 0\} + \inf_{x_2}\{\mathbf{x}_2^2 - ux_2 \mid x_2 \geq 0\} + 4u$. We use calculus and obtain: $x_{1,\min} = x_{2,\min} = \frac{u}{2}$ for $u \geq 0$ and $x_{1,\min} = x_{2,\min} = 0$ for $u < 0$. Hence $\theta(u) = -\frac{1}{2}u^2 + 4u$ for $u \geq 0$ and $\theta(u) = 4u$ for $u < 0$ As $\theta(u)$ is a concave function then $u_{\max} = 4$.

Example 391.

Find a dual program explicit formulation for $\begin{pmatrix} 2 \\ 1 \end{pmatrix} \in \operatorname{argmin}_{x_1,x_2}\{-2x_1+x_2 \mid x_1+x_2-3=0 \wedge (x_1,x_2)^\top \in X = \{(0,0),(0,4),(4,4),(1,2)\}\}$. Then $\theta(v) = \min_{x_1,x_2}\{-2x_1 + x_2 + v(x_1+x_2-3) \mid (x_1,x_2) \in X\}$ and we replace x_1 and x_2 with numbers and obtain $\theta(v) = -4+5v$ for $v \leq -1$, $\theta(v) = -8+v$ for $-1 \leq v \leq 2$, $-3v$ for $v > 2$ (draw a figure of graph of $\theta(v)$). Therefore, $v_{\max} = 2$.

Compare results of last two examples: $f(\mathbf{x}_{\min}) = \theta(u_{\max})$ and $f(\mathbf{x}_{\min}) > \theta(v_{\max})$.

Example 392 (Linear programming)**.**

It is possible to show that linear programming duality may be derived from Lagrangian duality of nonlinear programs. So, for primal $\min\{\mathbf{c}^\top\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \in X = \{\mathbf{x} \mid \mathbf{x} \geq \mathbf{0}\}\}$ we obtain: $\max\theta(\mathbf{v})$ where $\theta(\mathbf{v}) = \inf_x\{\mathbf{c}^\top\mathbf{x} + \mathbf{v}^\top(\mathbf{b}-\mathbf{Ax}) \mid \mathbf{x} \geq \mathbf{0}\} = \inf_x\{(\mathbf{c}^\top - \mathbf{v}^\top\mathbf{A})\mathbf{x} \mid \mathbf{x} \geq \mathbf{0}\} + \mathbf{v}^\top\mathbf{b}$. So, if $\mathbf{c}^\top - \mathbf{v}^\top\mathbf{A} \geq \mathbf{0}$ then we need to obtain infimum $\mathbf{x} = \mathbf{0}$ and we have $\theta(\mathbf{v}) = \mathbf{v}^\top\mathbf{b}$. If $\mathbf{c}^\top - \mathbf{v}^\top\mathbf{A} \not\geq \mathbf{0}$ then $\exists x_j \longrightarrow \infty$: $\theta(\mathbf{v}) = -\infty$. Hence a dual program is $\max_v\{\mathbf{v}^\top\mathbf{b} \mid \mathbf{v}^\top\mathbf{A} \geq \mathbf{c}^\top\}$.

Example 393 (Quadratic programming)**.**

We have primal $\min\{\frac{1}{2}\mathbf{x}^\top\mathbf{Hx} + \mathbf{d}^\top\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}\}$ where $\mathbf{H}$ is symmetric and PSD. Then dual is $\max\{\theta(\mathbf{u})\mathbf{u} \geq \mathbf{0}\}$ where $\theta(\mathbf{u}) = \inf\{\frac{1}{2}\mathbf{x}^\top\mathbf{Hx} + \mathbf{d}^\top\mathbf{x} + \mathbf{u}^\top(\mathbf{Ax}-\mathbf{b})\mathbf{x} \in \mathbb{R}^n\}$. So gradient must be equal to $\mathbf{0}$: $\mathbf{Hx} + \mathbf{A}^\top\mathbf{u} + \mathbf{d} = \mathbf{0}$ and dual looks as: $\max\{\frac{1}{2}\mathbf{x}^\top\mathbf{Hx} + \mathbf{d}^\top\mathbf{x} + \mathbf{u}^\top(\mathbf{Ax}-\mathbf{b}) \mid \mathbf{u} \geq \mathbf{0}, \mathbf{Hx} + \mathbf{A}^\top\mathbf{u} = -\mathbf{d}\}$. Transposing $\mathbf{Hx} + \mathbf{A}^\top\mathbf{u} + \mathbf{d} = \mathbf{0}$ and multiplying it by $\mathbf{x}$ from right, we obtain $\mathbf{u}^\top\mathbf{Ax} + \mathbf{d}^\top\mathbf{x} = -\mathbf{x}^\top\mathbf{Hx}$. We may substitute it in dual and we obtain $\max\{-\frac{1}{2}\mathbf{x}^\top\mathbf{Hx} + \mathbf{b}^\top\mathbf{u} \mid \mathbf{u} \geq \mathbf{0}, \mathbf{Hx} + \mathbf{A}^\top\mathbf{u} = -\mathbf{d}\}$. If $\mathbf{H}$ is even PD then $\exists\mathbf{H}^{-1}$ and $\mathbf{x} = -\mathbf{H}^{-1}(\mathbf{d} + \mathbf{A}^\top\mathbf{u})$, so dual has the form: $\max\{-\frac{1}{2}\mathbf{u}^\top(-A\mathbf{H}^{-1}\mathbf{A}^\top)\mathbf{u} + (-\mathbf{b}^\top - A\mathbf{H}^{-1}\mathbf{d})\mathbf{u} - \frac{1}{2}\mathbf{d}^\top\mathbf{H}^{-1}\mathbf{d} \mid \mathbf{u} \geq \mathbf{0}\}$.

Remark 394 (Geometric interpretation of Lagrange duality)**.**

Compare it with penalty algorithm geometric interpretation. So we have a simple primal program $\min\{f(\mathbf{x}) \mid g(\mathbf{x}) \leq 0, \mathbf{x} \in X\}$. We denote $y = g(\mathbf{x})$, $z = f(\mathbf{x})$ and $G = \{(y,z) \mid y = g(\mathbf{x}), z = f(\mathbf{x}), \mathbf{x} \in X\}$. So we have to find $\min\{z \mid y \leq 0, (y,z) \in G\}$. Then θ for dual is $\theta(u) = \min\{f(\mathbf{x}) + ug(x) \mid \mathbf{x} \in X\}$ or $\theta(u) = \min\{z + uy \mid (y,z) \in G\}$. If $\alpha = z + uy$ then $z = -uy + \alpha$ is a straight line($-u$ is an slope, α is an intercept). So we may draw a figure containing G, the contour graph of $z + uy$ for fixed u (slope). We may identify graphically $(y_{\min}, z_{\min})$ (like in LP), we find $\theta(u)$ value as the intersection with z axis. We may change u and obtain another $(y_{\min}, z_{\min})$ and $\theta(u)$ (like in penalty). We change u to find $u_{\max}$ such that $\theta(u)$ is maximized (as $\theta(u_{\max})$). Illustrate by figures. In contrast to LP in NLP case, the duality gap may occur (illustrate by a figure) when G is nonconvex.

1.13.1 Duality theorems

For NLP there are analogous weak and strong duality theorems like in LP.

Theorem 395 (weak duality)**.**

Let $\mathbf{x}$ be feasible of $\underline{P}$ and $(\mathbf{u}, \mathbf{v})$ be a feasible to $\underline{D}$. then $f(\mathbf{x}) \geq \theta(\mathbf{u}, \mathbf{v})$.

Proof: Based on the idea that $\inf A \leq a \in A$. □

Corollary 396.

1. "inf of $\underline{P} \geq$ sup of $\underline{D}$".
2. If $f(\bar{\mathbf{x}}) = \theta(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ and $\bar{\mathbf{x}}$, $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ are feasible to $\underline{P}$ and $\underline{D}$ respectively then they are optimal.
3. If inf of $\underline{P}$ is equal to $-\infty$ then $\forall \mathbf{u} > \mathbf{0} : \theta(\mathbf{u}, \mathbf{v}) = -\infty$.
4. If sup of $\underline{D} = - = \infty$ then $\underline{P}$ is infeasible.

Lemma 397.

$X \subset \mathbb{R}^n$, $X \neq \emptyset$ convex, $\alpha : \mathbb{R}^n \longrightarrow \mathbb{R}$, components of $\mathbf{g}; \mathbb{R}^n \longrightarrow \mathbb{R}^m$ convex, $\mathbf{h} : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ affine (defined by a linear term and constant vector). If System 1 below has no solution $\mathbf{x}$ then System 2 below has $(u_0, \mathbf{u}^\top, \mathbf{v}^\top)^\top$ solution. The converse holds if $u_0 > 0$.
System 1: $\alpha(\mathbf{x}) < 0$, $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$, $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{x} \in X$.
System 2: $\forall \mathbf{x} \in X : u_0\alpha(\mathbf{x}) + \mathbf{u}^\top \mathbf{g}(\mathbf{x}) + \mathbf{v}^\top \mathbf{h}(\mathbf{x}) \geq \mathbf{0}$, $(u_0, \mathbf{u}^\top)^\top \geq \mathbf{0}$, $(u_0, \mathbf{u}^\top, \mathbf{v}^\top)^\top =\neq \mathbf{0}$

Proof: See literature (cf. Farkas' Theorem). □

Theorem 398 (strong duality)**.**

$X \subset \mathbb{R}^n$, $X \neq \emptyset$ convex, $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, and components of $\mathbf{g} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be convex functions and $\mathbf{h} : \mathbb{R}^n \longrightarrow \mathbb{R}^l$ be affine ($\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$). Suppose that the following constraint qualification holds true $\exists \hat{\mathbf{x}} \in X : \mathbf{g}(\hat{\mathbf{x}}) < \mathbf{0}$, $\mathbf{h}(\hat{\mathbf{x}}) = \mathbf{0}$ and $\mathbf{0} \in \operatorname{int} \mathbf{h}(X)$ where $\mathbf{h}(X) = \{h(X) \mid \mathbf{x} \in X\}$.
Then $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} = \sup\{\theta(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\}$. Furthermore, if the infimum is finite then $\sup\{\theta(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\}$ is achieved in $(\bar{\mathbf{u}}\bar{\mathbf{v}})$ with $\bar{\mathbf{u}} \geq \mathbf{0}$. If the infimum is achieved at $\bar{\mathbf{x}}$ then $\mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0$.

Proof: See literature. Uses previous corollary then Lemma above. Then $u_0 > 0$ and so $\frac{1}{u_0}$ is used as a multiplier. At the end complementarity condition is utilized. □

Definition 399 (Saddle point of Lagrangian function)**.**

Let $\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{u}^\top \mathbf{g}(\mathbf{x}) + \mathbf{v}^\top \mathbf{h}(\mathbf{x})$ be Lagrangian function. Then $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ is its saddle point if $\forall \mathbf{x} \in X$, $\forall \mathbf{v}, \mathbf{u} \geq \mathbf{0}$ and $\Phi(\bar{\mathbf{x}}, \mathbf{u}, \mathbf{v}) \leq \Phi(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}}) \leq \Phi(\mathbf{x}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$.

Theorem 400 (Saddle point optimality and absence of duality gap)**.**

Let $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ with $\bar{\mathbf{x}} \in X$, $\bar{\mathbf{u}} \geq \mathbf{0}$ be a saddle point for the Lagrangian function $\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v})$ iff

a) $\Phi(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}}) = \min\{\Phi(\mathbf{x}, \bar{\mathbf{u}}, \bar{\mathbf{v}}) \mid \mathbf{x} \in X\}$,

b) $\mathbf{g}(\bar{\mathbf{x}}) \leq \mathbf{0}, \mathbf{h}(\bar{\mathbf{x}}) = \mathbf{0}$, and

c) $\mathbf{u}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0$.

Moreover $\bar{\mathbf{x}}$ solves $\underline{\mathrm{P}}$ and $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ solves $\underline{\mathrm{D}}$ and $f(\bar{\mathbf{x}}) = \theta(\bar{\mathbf{u}}, \bar{\mathbf{v}})$.

Proof: See literature. □

Corollary 401.

If X, f, $\mathbf{g}$ convex, $\mathbf{h}$ affine, $\mathbf{0} \in \operatorname{int} \mathbf{h}(X)$, and $\exists \hat{\mathbf{x}} \in X : \mathbf{g}(\hat{\mathbf{x}}) < \mathbf{0}, \mathbf{h}(\hat{\mathbf{x}}) = \mathbf{0}$. If $\bar{\mathbf{x}}$ is optimal to $\underline{\mathrm{P}} \Rightarrow \exists (\bar{\mathbf{u}}, \bar{\mathbf{v}}), \bar{\mathbf{u}} \geq \mathbf{0}$ such that $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ is a saddle point.

Theorem 402 (Relation between saddle point criterion and KKT)**.**

$S = \{\mathbf{x} \in X \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$ and $\min_{\mathbf{x}}\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ are defined. Suppose that $\bar{\mathbf{x}} \in S$ satisfies the KKT conditions, so $\exists \bar{\mathbf{u}} \geq \mathbf{0} : \nabla f(\bar{\mathbf{x}}) + \nabla \mathbf{g}(\bar{\mathbf{x}})^\top \bar{\mathbf{u}} + \nabla \mathbf{h}(\bar{\mathbf{x}})^\top \bar{\mathbf{v}} = \mathbf{0}$, $\bar{\mathbf{u}}^\top \mathbf{g}(\bar{\mathbf{x}}) = 0$. Suppose that $f, g_i (i \in I = \{i \mid g_i(\bar{\mathbf{x}}) = 0\})$ are convex at $\bar{\mathbf{x}}$ and for $\bar{v}_i \neq 0$ h_i is affine. Then $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ is a saddle point for $\Phi(\mathbf{x}, \mathbf{u}, \mathbf{v})$.
Conversely: Suppose that $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ with $\mathbf{x} \in \operatorname{int} X$ and $\bar{\mathbf{u}} \geq \mathbf{0}$ is a saddle point solution. Then $\bar{\mathbf{x}}$ is feasible to $\underline{\mathrm{P}}$ and furthermore $(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}})$ satisfies KKT.

Proof: See literature. □

All theorems together say: With certain assumptions the optimal dual variables for the Lagrangian dual problem are equal to Lagrangian multipliers for the KKT conditions and are equal to multipliers for the saddle point conditions.

Remark 403 (Further readings)**.**

The following ideas should be further studied:

- $\theta(\mathbf{u}, \mathbf{v})$ is concave under weak assumptions (useful for maximization).
- Differentiability of θ relates to assumption of uniqueness.
- Subgradient of θ is useful for ascent directions.
- Ascent directions of θ allow to search steepest ascent direction.
- Formulating dual problem and using it in its implicit form.
- Solving dual problem by a cutting plane methods (outer linearization).
- Getting primal solution from dual problem.