Optimization 1

Lecture notes, University of Technology, Graz
based on the lecture by Bettina Klinz

Lukas Prokop

July 2, 2019

Contents

0 Course

↓ *This lecture took place on 2019/03/04.*

- Lecture
  - Monday, 12:15–14:00
  - Tuesday, 16:15–18:00
- First week, the practicals session will be used for the lecture
- Practicals will take place usually on Wednesday, 16:15–18:00
  in exceptional cases on Thursday, 16:15–18:00
- 2 websites (work in progress):
  - http://www.math.tugraz.at/~klinz/optimvo (list of literature)
  - http://www.math.tugraz.at/~klinz/optimue (practicals mode, practicals exercises, additional content)

- Two large topics in this lecture
  - Linear optimization (linear target function, linear side conditions)
  - Unconstrained, non-linear optimization
    where non-linear optimization denotes that the target function is non-linear
- Be aware, that this class might be the lecture requiring previous results of classes the most.
- Advanced lecture in masters
  - Lecture "Non-linear optimization" (includes non-linear optimization with sub conditions)
- Exam
  - written + orally, in case of negotiation and few candidates only orally
  - 1st date will be at the end of the semester, optionally in summer holidays
  - 2 written exams for the practicals

1 Linear optimization

1.0 Introduction

We have already seen optimization problems in high school or previous semesters. But, for example, handling constraints consisting of inequalities was tedious or trivial. We consider more sophisticated techiques here. In practice, linear models occur rarely. But they often provide a sufficient heuristic.

1.0.1 Examples for linear optimization models

Example 1 (Production planning model)**.** *A factory can produce n goods. The revenue per unit of good j is given by c_j units of money. Production is limited by restrictions, that result from constrained availability of staff, equipment and raw materials. Let m denote the number of these resources and b_i is the maximum availability of resource i ($i = 1, \dots, m$). Let a_{ij} with $1 \leq i \leq m$ and $1 \leq j \leq n$ denote the quantity of resource i required to produce 1 unit of good j. Our goal is to maximize revenue with respect to the given constraints.*

Decision variable: X_j *is the quantity of good j*

$$\textit{target function:} \ \max \sum_{j=1}^{n} c_j x_j$$

$$\textit{subject to (constraints)} \ \sum_{j=1}^{n} a_{ij} x_j \leq b_i \qquad i \in \{1, \dots, m\}$$

$$\text{with } x_j \geq 0 \qquad \text{sign condition}$$

Pay attention! We do not require $x_j \in \mathbb{Z}$. For $x_j \in \mathbb{Z}$ we get an integral linear program which is not *a linear programm! (this is a difficult subproblem of optimization)*

Example 2 (Mixture problem)**.** *Consider n kinds of raw materials. Our goal: We want to create a new material by mixing existing raw materials to reduce costs.*

Example (Alloys)**.** *Consider n different base alloys $L_1, \ldots, L_n$. For each material we have the lead content in percent $(a_1, \ldots, a_n)$ and the costs per unit of weight $(c_1, \ldots, c_n)$. We have to produce alloys with lead content $b\%$. The decision variable x_j is given by the ratio of L_j.*

Formally, the problem can be defined as

$$\min \sum_{j=1}^{n} c_j x_j \ \ s.t. \ \ \sum_{j=1}^{n} x_j = 1, \sum_{j=1}^{n} a_j x_j = b, x_j \geq 0$$

These are typical mixture constraints and they ensure a given lead content.

Example 3 (Linear transportation problem)**.** *Given m firms, n customers, a_i is the offer by firm i and b_j is the demand by customer j. c_{ij} are the transportation costs from firm i to customer j.*

Find an admissible transportation plan with minimal costs. The decision variable x_{ij} is the quantity of goods transported from firm i to customer j.

Formally,

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \ \ s.t. \ \ \sum_{i=1}^{m} x_{ij} = b_j, \sum_{j=1}^{n} x_{ij} = a_i, x_{ij} \geq 0 \quad j \in \{1, \ldots, n\}, i \in \{1, \ldots, m\}$$

Example (Diet problem)**.** *The following problem has a strong historical background in optimization sciences: Stigler diet problem (in the year 1939) by Georg Stigler.*

Given a list of 77 ingredients. Per ingredient we are given features such as calories and proteins. Find an optimal combination to minimize costs.

His heuristic results were confirmed as almost optimal in 1947.

In the following lectures, our goal will be:

- Theory of linear optimization (Knowledge about fundamentals and background)
- Algorithmic solutions procedures: in the lecture we will discuss two procedures:
  - Simplex method (G. Dantzig, 1947, in practice useful, no polynomial runtime)

- Inner point method (in practice useful, polynomial runtime)

Outside this lecture:

- Ellipsoid method: polynomial runtime, but in practice useless

1.1 Geometrical considerations of linear optimization

Definition. *Standard form of a linear program (canonical representation)*

$$\max z(x) = z_0 + \sum_{j=1}^{n} c_j x_j \text{ such that } \sum_{j=1}^{n} a_{ij} x_j \leq b_i, x_j \geq 0 \qquad i \in \{1, \dots, m\}$$

In compact notation:

$$\max z_0 + c^t x \text{ s.t. } Ax \leq b, x \geq 0$$

typically $\max c^t x$. z_0 *does not influence the result.*

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \qquad a_{ij} \in \mathbb{R}, b_i \in \mathbb{R}, c_j \in \mathbb{R}$$

↓ *This lecture took place on 2019/03/05.*

Revision. *The canonical/standard form of a linear program is given by*

$$\max c^t x (+ \text{ optionally } z_0) \text{ such that } Ax \leq b \qquad x \geq 0$$

Remark (Observation)**.** *Every arbitrary linear program (min/max of an affine linear function over a linear constraint) can be transformed into the canonical form above.*

1. *The minimization over* $c^t x$ *corresponds to* $-c^t x$ *as maximization problem.*
2. *Constraints of form* $\alpha^t x \geq b$ *can be written as* $-\alpha^t x \leq -\beta$.
3. *A constraint of form* $\alpha^t x = \beta$ *can be written as* $\alpha^t x \leq \beta$ *with* $\alpha^t x \geq \beta$. *Or equivalently as* $\alpha^t x \leq \beta$ *with* $-\alpha^t x \leq -\beta$.

   Remark.

   - *Disadvantage: One constraint is transformed into two.*
   - *In practice, the explicit handling of equality constraints should be preferred.*

4. *Let* x_j *not be restricted w.r.t. the sign. Write* x_j *as* $x_j = x_j^+ - x_j^-$ *with* $x_j^+ \geq 0$ *and* $x_j^- \geq 0$. x_j *will be replaced by* $x_j^+ - x_j^-$ *with* $x_j^+ \geq 0$ *and* $x_j^- \geq 0$.

Remark. *Disadvantage: Number of variables is increased.*

Remark (Terminology)**.**

- *2 points* $u, v \in \mathbb{R}^n$ *define a* line $G(u,v) = \{u + \lambda(v-u) \mid \lambda \in \mathbb{R}\}$.
- *A* halfline *is formally given by* $\{u + \lambda(v-u) \mid \lambda \geq 0\}$
- *A* closed interval *(or* segment*) is defined as* $[u,v] = \{u + \lambda(v-u) \mid \lambda \in [0,1]\}$ *and an* open interval *is defined as* $(u,v) = \{u + \lambda(v-u) \mid \lambda \in (0,1)\}$.

Lemma 1.1.1.
- *An affine linear function* $z_0 + c^t x$ *takes up its maximum/minimum in segment* $[u,v]$ *in its end points* u *or* v.
- *An affine linear function* $z_0 + c^t x$ *takes up its maximum/minimum on a half-line in the end points.*

Proof. Left as an exercise to the reader (use parameter representation and insert it into the function) □

1.1.1 Hyperplane and Halfspaces

By the linear inequality $\alpha^t x \leq \beta$ ($\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}$) with $\alpha \neq 0$ (zero vector) we define a *closed halfspace.*

$$H_{\leq} := \left\{x \in \mathbb{R}^n \mid \alpha^t x \leq \beta\right\}$$

Analogously we can define open halfspaces:

$$H_{<} := \left\{x \in \mathbb{R}^n \mid \alpha^t x < \beta\right\}$$

Hyperplane:

$$H_{=} := \left\{x \in \mathbb{R}^n \mid \alpha^t x = \beta\right\}$$

In $\mathbb{R}^2$, hyperplanes are halflines. In $\mathbb{R}^3$, hyperplanes are halfplanes.

Possible cases for the position of lines $G(u,v)$ w.r.t. to the hyperplane $H = \{x \mid \alpha^t x = \beta\}$.

Case 1 Line G is contained in H (denoted $G \subseteq H$)
$\alpha^t u = \beta$, $\alpha^t(u-v) = 0$

Case 2 G is parallel to H
$\alpha^t u \neq \beta$, $\alpha^t(u-v) = 0$

Case 3 G intersects H in one point c
$\alpha^t(u-v) \neq 0$

Lemma 1.1.2. *If the line* $G(u,v)$ *is neither contained in halfplane* $H = \{x \mid \alpha^t x = \beta\}$ *nor in some open halfspace constrained by* H, G *intersects the halfplane* H *in one point.*

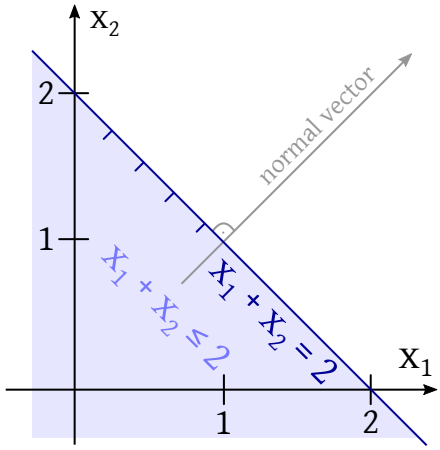

Figure 1: Constraint $x_1 + x_2 \leq 2$ visualized in $\mathbb{R}^2$. To determine the halfplane (bright blue) resulting from the constraint, it helps to express the constraint with x_2 on the left-hand side: $x_2 \leq 2 - x_1$. Then determine x_2 for two different x_1 assuming an equality operator, e.g. $x_1 = 0$ with $x_2 = 2$ and $x_1 = 1$ with $x_2 = 1$. Then choose some large value x_1 and x_2 (e.g. $x_1 = 10$ and $x_2 = 10$). Does it satisfy $x_1 + x_2 \leq 2$? No, the halfplane containing $(10, 10)$ is not the one, we are looking for. Some people prefer to consider the normal vector. Sometimes dashes are used to mark the side of the considered halfplane.

Remark (Observation)**.** *The parameter representation of a line is ambiguous. We can choose u und v wisely: $G(u,v), \overline{x} \in G$. We can always choose u and v such that $\overline{x} \in (u,v)$.*

Let $\overline{x} \in H_< = \{x \in \mathbb{R}^n \mid \alpha^t x < \beta\}$. If $G \subseteq H_<$, then $u, v \in H_<$. Otherwise, due to $\overline{x} \in G \cap H_<$, G intersects the hyperplane $H = \{x \in \mathbb{R}^n \mid \alpha^t x = \beta\}$.

Furthermore we can choose a representation (u,v) such that $\overline{x} \in (u,v) \subseteq H_<$ and $v \in H_=$. This can be generalized to multiple halfspaces.

Definition. *A* polyhedron*(dt. "Polyeder") is the intersection of finitely many halfspaces. A bounded polyhedron is also called* polytope*.*

Remark (Observation)**.** *The admissible set $P(A,b)$ of the linear program (given in the previous revision) is a polyhedron.*

Lemma 1.1.3. *Halfspaces and thus also polyhedrons are convex sets.*

Furthermore affine-linear functions are convex *and* concave. Linear optimization is a special case of convex optimization minimizing a convex function over a convex set. The neat property of such convex optimization tasks is that local and global minima/maxima collapse (just like in the linear case).

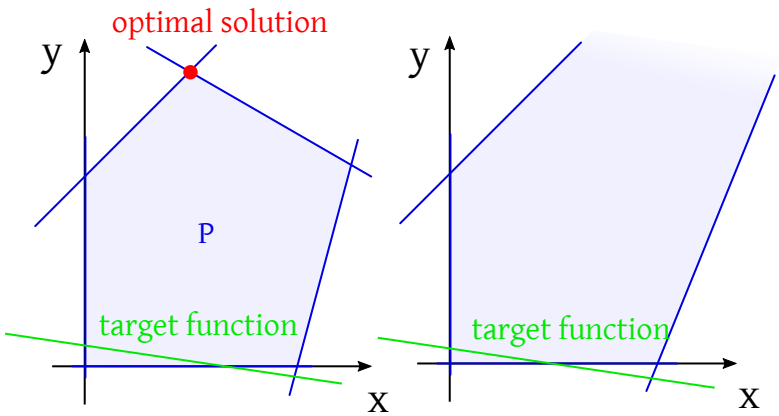

Figure 2: Bounded (left) and unbounded (right) sets. The unbounded set does not have an optimal solution.

Geometric illustration for $n = 2$

$n = 2$ means that we consider 2 variables.

$$\max c_1x_1 + c_2x_2 \qquad a_{i1}x_1 + a_{i2}x_2 \leq b_i \quad i = 1, \dots, m \qquad x_1, x_2 \geq 0$$

Remark (Observation)**.** *Optimal solutions occur at vertices of the set (intersection of two constraints). Compare with Figure 2.*

This approach provides a graphical solution method for linear programs with 2 variables.

The generic geometric consideration of a polyhedron is

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\} \qquad A \in \mathbb{R}^{m \times n}$$

Without loss of generality, we assume all row vectors a_i of A are non-zero.

$$H_i := \left\{x \in \mathbb{R}^n \mid a_i^t x = b_i\right\} \qquad i\text{-th halfplane} \quad i = 1, \dots, m$$

Let $I \subseteq \{1, \dots, m\}$. Consider

$$\hat{H}_I := \bigcap_{i \in I} H_i = \left\{x \in \mathbb{R}^n \mid a_i^t x = b_i \text{ for } i \in I\right\} \qquad \text{affine subspace of } \mathbb{R}^n$$

If $\hat{H}_I \cap P \neq P \neq \emptyset$, then this set is called *face* of P. The face of P is called minimal, if it does not contain any other face properly.

Example. *A cube in $\mathbb{R}^3$ has 27 faces:*

- *The cube itself, $I = \emptyset$*

- *6 faces,* $I = \{1\}, I = \{2\}, \ldots, I = \{6\}$
- *12 face edges,* $|I| = 2$
- *8 vertices,* $|I| = 3$

Example (Faces in the left subfigure in Figure 2)**.** *11 faces in total (5 vertices of dimension 0, 5 with dimension 1, 1 with dimension 2)*

More formally: Let P be described by a hyperplane H_i ($a_i^t x = b_i$). Let $S \subseteq \mathbb{R}^n, S \neq 0$.

Definition. $I(S) \coloneqq \{i \in \{1, \ldots, m\} \mid S \subseteq H_i\}$

For $S = \{x_0\}$, we write $I(x_0)$ instead of $I(\{x_0\})$.

Definition.

$$L(S) \coloneqq \left\{x \mid a_i^t x = b_i, i \in I(S)\right\}$$

is the smallest affine subspace containing S*. For a non-empty* S*,* S *is a face iff* $S = L(S) \cap P$*.*

If S *is a minimal face, then* $S = L(S)$ *(*$L(S)$ *is then a part of the polyhedron).*

Definition (Dimension of a face)**.**

$$\dim S \coloneqq \dim L(S)$$

so the dimension of the smallest affine subspace containing S*.*

Definition (Vertex)**.** *A vertex is a face of dimension* 0*.*

Remark. *For polyhedron, the vertex term from above is an alternative definition for the vertex term for convex sets (the terms correspond).*

Remark. *A circle has infinitely many vertices; not none.*

Let S be convex set in $\mathbb{R}^n$. x is called vertex of S if it is not possible to represent x as $x = y + (1 - \lambda)z$ with $y, z \in S, y \neq z, \lambda \in [0, 1]$.

↓ *This lecture took place on 2019/03/06.*

Not every polyhedron has vertices. For example, if the boundaries are given as two parallels, then no vertices can be identified.

Remark. *The problem with two parallels is not representable in canonical form.*

Definition. *A polyhedron* with *vertices is called* acute*.*

A vertex is a minimal face. A face results as unique solution of the corresponding equation system.

$$a_i^t x = b_i \qquad \forall i \in I(z)$$

If face $S = L(\overline{x}) \cap P$ for $\overline{x} \in \overline{P(A,b)} =: P$ has dimension ≥ 1 (hence no vertex), then you can let some line G pass through $\overline{x}$ that lies in $L(\overline{x})$.

$\overline{x}$ lies in the intersection of open halfspaces $\left\{x \mid a_i^t x < b_i, i \notin I(\overline{x})\right\}$. Thus we can provide a representation of line G passing through two points c and d with $c, d \in S, \overline{x} \in (c,d)$.

Assume the halfspace with end point $\overline{x}$ in direction d is bounded by some of the constraining hyperplanes of these open halfspaces. Then d can be chosen as the closest intersection point of the line with one of these hyperplanes H_i for $i \notin I(\overline{x})$.

$$\implies \left|I(d)\right| > \left|I(\overline{x})\right| \qquad \implies \dim L(d) < \dim L(\overline{x})$$

Analogously, the same applies to the constraint by the halfline with end point $\overline{x}$ and in direction c. $\dim L(c) < \dim L(\overline{x})$. Step by step, we can reduce the dimension to end up with a vertex.

Theorem 1.1.4 (Statement about acute-angled polyhedrons). *1. A non-empty polyhedron is acute iff it does not contain any line.*

2. Every face of an acute polyhedron contains one vertex.

3. A polyhedron has (at most) finitely many vertices.

Proof. 1a. Assume P does not contain any lines. Let $x_0 \in P$ ($P \neq 0$). If x_0 is a vertex, then P is acute. If x_0 is not a vertex, then consider the face $S := L(x_0) \cap P$. Then $\dim S \geq 1$ holds true. Now we use the idea, that was sketched above right before the theorem.

We choose a line $G(c,d) \subseteq L(x_0)$ where c and d are chosen as described before. Because P does not contain a line, S also does not contain any. Without loss of generality, we assume that the halfline with end point x_0 in direction d is bounded and thus $\dim L(d) < \dim L(x_0)$. If d is not a vertex, repeat this construction with some new $x_0 = d$.

In every step, we are losing at least one dimension. After at most n steps, we are going to have a vertex.

1b. Let P be acute, then P has a vertex. Choose a vertex $\overline{x}$ and n inequalities with maximum row rank. $I \subseteq I(\overline{x})$, submatrix A_I of A has $\operatorname{rank}(A_I) = n$.

Assume there exists a line $G(u,v)$ with $u \neq v$, that in entirely contained in P. The inequalities $A_I u + \lambda \cdot A_I(v-u) \leq b_I$ must be true for all $\lambda \in \mathbb{R}$.

Because λ is unbounded, it should be true that $A_I(v-u) = 0$ where A_I is a matrix of full rank. So $v - u = 0 \implies v = u$. This is a contradiction.

2. A face S of an acute polyhedron P cannot contain any line because $S \subseteq P$ und S is itself a polyhedron. By the first statement, S has one vertex.

3. Let P be described by some $m \times n$ matrix A. $\{1, \ldots, m\}$ has only finitely many subsets. Thus we have only finitely many faces.

$$\leq \binom{m}{n} \text{ vertices}$$

□

Remark. *By Theorem 1.1.4, it is immediate that non-empty polyhedrons, that result from linear programs in canonical form,* $P = \{x \in \mathbb{R}^n \mid Ax \leq b, x \geq 0\}$ *are always acute. So it has at least one vertex (because the set* $\{x \in \mathbb{R}^n \mid x_i \geq 0, i \in \{1, \ldots, n\}\}$ *does not contain a line).*

1.1.2 Fundamental theorem of Linear Optimization

Theorem 1.1.5 (Fundamental theorem of Linear Optimization)**.** *1. If an affine-linear function* $z(x)$ *takes up its maximum/minimum in a polyhedron* P *in* $\overline{x} \in P$*, then also in all points of face* $S = L(\overline{x}) \cap P$*.*

2. Especially the optimum is taken up in a vertex of P*, if* P *is acute.*

Proof. 1. Let $\overline{x}$ be a maximum (analogously for minima) and let $y \neq \overline{x}$ be another point at S. Then the line $G := G(\overline{x}, y)$ in $L(\overline{x})$. For G there exists a representation $G = G(c, d)$ with $c, d \in P, \overline{x} \in (c, d)$. By Lemma 1.1.1 the affine-linear function $z(x)$ takes up its maximum in line segment $[c, d]$ in c or d. Without loss of generality, we assume its maximum in c. Thus $z(c) \geq z(\overline{x})$, because $\overline{x} \in (c, d)$.

On the other hand, we have $z(\overline{x}) \geq z(c)$, because $c \in P$ and the maximum is reached in $\overline{x}$.

Thus $z(c) = z(\overline{x})$. Hence, $z(x)$ is constant in G and therefore $z(y) = z(\overline{x})$. y was chosen arbitrarily, then z is constant at face S.

2. Follows by Theorem 1.1.4 (b).

□

The polyhedron for linear programs in canonical form are empty or acute (have vertices). The Fundamental theorem of Linear Optimization followingly states that for such linear programs (and thus any linear program becausc every linear program can be represented in canonical form) it suffices to investigate all vertices.

Corollary 1.1.6. *If* $\max\{c^{\perp}x : x \in P\}$ *has a linear optimization solution* x^**, then* $c^t x^* = \max\{c^t x : x \in V(P)\}$ *where* $V(P)$ *is the set of vertices of* P*.*

Thus we retrieve a finite method for linear programs: Determine all vertices and filter the vertex optimizing the target function.

Remark (Disadvantage)**.** *Because there are exponentially (in* n *and* m*) many vertices in general, there is no practically useful method of this idea.*

1.2 The generic Simplex method

The Simplex method goes back to George Dantzig (1947). The method relies on the Fundamental theorem of Linear Optimization and tries to find an optimal vertex. It utilizes convexity to claim a local minimum as global one. Thus for a

given vertex x^*, any adjacent vertex (reachable by one edge) has a worse target function value.

The basic idea is:

1. Determine an initial vertex x (if none exists, the polyhedron is empty because we utilize the canonical form).
2. Test whether x is a local optimum. Consider the edges starting from x (they are either unbounded or lead to adjacent vertices).
3. If x is a local optimum, then stop.
4. Otherwise either an unbounded problem is given or we replace x by some adjacent vertex with a better target function value.
5. Iterate this process.

This process is necessarily finite, because there are only finitely many vertices. This process gives rise to the generic Simplex algorithm:

1. Choose an arbitrary vertex x of P as initial vertex. If none exists ($P = \emptyset$), then stop.
2. While there exists some edge k starting from x increasing along the target function value, do
   (a) Choose such an edge
   (b) If k is not a halfline of our polyhedron P then
      i. substitute x by edge $\tilde{x}$ at the other end of k

      else stop, as the problem is unbounded
3. Return vertex x

↓ *This lecture took place on 2019/03/11.*

Our next goal is to implement of this algorithmic idea algebraically.

Remark (Observation)**.** *It is difficult to transform inequality systems of form $Ax \leq b$.*

Transformation of an inequality system:
Canonical form $\max\{c^\perp x \text{ s.t. } Ax \leq b_1, x \geq 0\}$ with $x \in \mathbb{R}^n, b \in \mathbb{R}^m, c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$.

Polyhedron $P = \{x \in \mathbb{R}^n \mid Ax \leq b, x > 0\}$.

Introduction of auxiliary variables y_i (slack variables) (dt. "Schlupfvariable"). $y = b - Ax$ in vector notation. $y_i = b_i - a_{i1}x_1 - \cdots - a_{in}x_n \quad i = 1, \ldots, m$.

Every point $x \in \mathbb{R}^n$ corresponds to exactly one point $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{n+m}$.

Polyhedron $P \to$ polyhedron $\tilde{P} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{m+n} \mid Ax + y = b, x, y \geq 0 \right\}$. The polyhedron structure is retained in such a way that dimensions of faces are preserved and vertices will become vertices.

The following correspondence will become useful:

$$x_{n+1} \coloneqq y_1 \qquad x_{n+2} \coloneqq y_2 \qquad \dots \qquad x_{n+m} \coloneqq y_m$$

This provides a uniform naming of variables.

This results in the following representation, we call *normal form*

$$\max c_1 x_1 + \dots + c_n x_n + c_{n+1} x_{n+1} + \dots + c_{n+m} x_{n+m}$$

subject to

$$\begin{array}{lllllll}
a_{11}x_1 & +\dots & +a_{1n}x_n & +x_{n+1} & & = b_1 \\
a_{11}x_1 & +\dots & +a_{1n}x_n & & +x_{n+1} & = b_2 \\
a_{11}x_1 & +\dots & +a_{1n}x_n & & \vdots & \\
a_{m1}x_1 & +\dots & +a_{mn}x_n & & \vdots & = b_m \\
x_1 & ,\dots, & x_{m+n} & & & \geq 0
\end{array}$$

We agree on $c_{n+1} = \dots = c_{n+m} = 0$.

In the following, we will also denote the previous coefficient matrix with A. This A results from the canonical form and a $m \times n$ unit matrix I.

$$\left(\; A_{\text{canonical}} \mid I \; \right)$$

Example (Canonical form)**.**

$$\max x_1 + x_2$$

s.t.

$$\begin{aligned}
x_1 + 2x_2 &\leq 4 \\
2x_1 - x_2 &\leq 3 \\
x_2 &\leq 1 \\
x_1, x_2 &\geq 0
\end{aligned}$$

Example (Normal form)**.**

$$\max x_1 + x_2$$

s.t.

$$\begin{aligned}
x_1 + 2x_2 + x_3 &= 4 \\
2x_1 - x_2 + x_4 &= 3 \\
x_2 + x_5 &\leq 1 \\
x_1, x_2, \dots, x_5 &\geq 0
\end{aligned}$$

Remark. *In the following, we assume a linear program in normal form. A is a $m \times (m+n)$ matrix for which we assume that it has full row rank* $\operatorname{rank}(A) = m$.

In a similar way, P denotes the polyhedron corresponding to our system $Ax = b, x > 0$. Let $J \subseteq \{1, \ldots, m+n\} \to (J(1), J(2), \ldots, J(k))$ be a map to index vectors where J is an index set $|J| = K$.

Be aware that we implicitly switch between sets and tuples.

Our next goal is to introduce the terms basis, basis solution, non-basis.

Definition. *A submatrix A_B of A with $A_B = (A_{B(1)}, \ldots, A_{B(m)})$ and* $\operatorname{rank} A_B = m$ *(thus the m columns of A_B are linear independent) is called* basis matrix *and B is called* basis*. Here we assume that A_B is regular.*

The remaining columns of A are summed up in index vector N.

$$\text{matrix } A_N = (A_{N(1)}, \ldots, A_{N(n)}) \text{ where } N = \{1, \ldots, m+n\} \setminus B$$

N is called *non-basis* and considered as set. A_N is called *non-basis matrix.*

We call x_j with $j \in B$ *basis variable* and x_j with $j \in N$ *non-basis variable.*

The following compact notations are practical:

x_B	...	vector of basis variables
x_N	...	vector of non-basis variables
c_B	...	vector of cost-coefficients c_j for $j \in B$ (basis variable)
c_N	...	vector of cost-coefficients c_j for $j \in N$ (non-basis variable)

$$\max c^t x \qquad Ax = b, x \geq 0$$

$$\Longrightarrow \max c_B^t x_B + c_N^t x_N \qquad A_B x_B + A_N x_N = b \qquad x_B, x_N \geq 0$$

We can write it as,

$$c = (c_B, c_N) \quad x = (x_B, x_N) \quad A = (A_B, A_N)$$

Definition. *A vector $x \in \mathbb{R}^{m+n}$ is called* basis solution *of a linear optimization problem in normal form (*$\max\{c^t x \mid Ax = b, x \geq 0\}$*), if there exists some basis B with $A_B x_B = b$ and $x_N = 0$ (remark: $x_B = A_B^{-1} b$).*

A basis solution is called admissible *if $x_B \geq 0$. In this case, B is called* admissible basis*.*

A basis solution is called *degenerate*, if there exists some i with $x_{B(i)} = 0$. Otherwise x_B is called non-degenerate. Analogously we define *degenerate bases* and *non-degenerate bases.*

Remark. *A basis solution x is in polyhedron P, if it is admissible.*

For the generic Simplex method, we go from vertex to vertex and thus from admissible solution to admissible solution.

Example. $N = (1, 2)$*. So, non-basis variables are x_1, x_2*
$B = (3, 4, 5)$*. So, basis variables are x_3, x_4, x_5.*

The corresponding basis solution $(0, 0, 4, 3, 1)$*.*

$$N = (1,5) \qquad B = (2,3,4)$$

is the corresponding basis solution.

Solve the system:

$$\begin{pmatrix} 2 & 1 & 0 \\ -1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 1 \end{pmatrix}$$

$$\begin{aligned} 2x_2 + x_3 &= 4 \\ -x_2 + x_4 &= 3 \\ x_2 &= 1 \end{aligned}$$

So $x_3 = 2$ and $x_4 = 4$

$$(0,1,2,4,0)$$

is admissible and non-degenerated.

$$\begin{aligned} B &= (1,2,3) \qquad & N &= (4,5) \\ B &= (1,2,4) \qquad & N &= (3,5) \\ B &= (1,2,5) \qquad & N &= (3,4) \end{aligned}$$

all lead to $x = (2,1,0,0,0)^t$ (admissible, degenerated).

Remark (Here be dragons)**.** *To some non-degenerate basis solution, there exists exactly one basis. This does not hold true for degenerated basis solutions.*

The same vertex of the polyhedron corresponds to several bases in case of degeneration.

Theorem 1.2.1. *The admissible basis solutions correspond to the vertices of the polyhedron, vice versa. If the basis solution is non-degenerated, then the corresponding basis is uniquely determined.*

Proof.

1. The basis solution $\tilde{x}$ (for basis B and non-basis N) maps [by the definition of the basis solution] to $\tilde{x}_N = 0$ and $\tilde{x}_B$ is the unique solution of $A_B \tilde{x}_B = b$ ($Ax = b$).

   $$\{\tilde{x}\} = \{x \mid x_N = 0\} \cap \{x \mid Ax = b\}$$

   If $\tilde{x}$ is an admissible basis solution, then $\tilde{x}_B \geq 0$ and thus $\tilde{x} \geq 0$ and thus $\tilde{x} \in P$. Hence $\tilde{x}$ is a vertex of P.

2. Let $\hat{x}$ be a vertex of P. Then the sign conditions must be satisfied and $\hat{x}$ is not uniquely defined by $m + n$ equations. Hence $\left\{ \begin{matrix} n \\ x \end{matrix} \right\} = \{x \mid x_N = 0\} \cap \{x \mid Ax = b\}$ with $N \subseteq \{1, \ldots, m+n\}$. A_B must be regular. $\hat{x}$ is basis solution.

3. In some non-degenerated basis solution, there are exactly m components $\neq 0$. B is uniquely defined and thus N.

□

Theorem 1.2.1 allows us to use basis solutions to implement the Simplex method numerically/algebraically.

Remark. *Let a linear program in normal form be given. We assume it was transformed from the canonical representation. With* $N = (1, \dots, n)$ *and* $B = (n{+}1, \dots, n{+}m)$ *for* $b \geq 0$*, we always get one admissible basis solution* $x \begin{pmatrix} x_N = 0 \\ x_B = b \end{pmatrix}$*. This corresponds to the origin in the coordinate system of the canonical form. It can be used as initial guess in the Simplex method. For the other cases, we are still looking for an approach.*

Now let B be a fixed basis and N is the corresponding non-basis.

$$Ax = b \iff A_B x_B + A_N x_N = b \qquad A_B \text{ regular}$$

$$x_B = A_B^{-1}(bA_N x_N) = \underbrace{A_B^{-1} b}_{:=\tilde{b}} - \underbrace{A_B^{-1} A_N}_{\tilde{A}_N} x_N$$

$$\tilde{b} := A_B^{-1} b \qquad \tilde{A}_N = A_B^{-1} A_N$$

$$x_B = \tilde{b} - \tilde{A}_N x_N$$

Polyhedron P:

$$x_B = \tilde{b} - A_N x_N \qquad x_B, x_N \geq 0$$

where the basis variables are represented by the non-basis variables. This is the reduced representation with respect to (B, N).

Representation of form:

$$x_{B(i)} = t_{i_0} + \sum_{j=1}^{n} t_{ij} x_{N(j)} \qquad i = 1, \dots, m$$

where t_{ij} are the representation coefficients t_{ij} with $i = 1, \dots, m$ and $j = 0, \dots, n$.

Projection of the polyhedron in the space of independent variables (non-basis variables).

We retrieve the canonical form representation in this space

$$\tilde{A}_N x_N \leq \tilde{b} \qquad x_N \geq 0$$

Example (continued)**.**

$$N = (1, 5) \qquad B = (2, 3, 4)$$

$$\begin{aligned} x_3 &= 2 - x_1 + 2x_5 \\ x_4 &= 4 - 2x_1 - x_5 \\ x_2 &= 1 - x_5 \end{aligned}$$

We want to insert this new representation into the target function.

$$\begin{aligned} z(x) = z &= z_0 + c^t x = z_0 + c_B^t x_B + c_N^t x_N \\ &= \underbrace{(z_0 + \underbrace{c_B^t A_B^{-1} b}_{\text{constant}})}_{\tilde{z}_0} - \underbrace{(c_B^t A_B^{-1} A_N X_N + C_N^t x_N)}_{+(c_N^t - c_B^t A_B^{-1} A_N) x_N} \\ &= \tilde{z}_0 + \tilde{c}_N^t x_N \end{aligned}$$

with $\tilde{c}_N^t = c_N^t - c_B^t \underbrace{A_B^{-1} A_N}_{\tilde{A}_N}$. $\tilde{c}_N$ is called *reduced cost coefficients*. Later, $\tilde{z}_0$ and $\tilde{c}_N$ will become the 0-th row of the coefficient tableau.

> ↓ *This lecture took place on 2019/03/12.*

Revision.

$$z := t_{00} + \sum_{j=1}^{n} t_{0j} X_{N(j)}$$

as 0-th row of the tableau.

Example.

$$N = (1, 5) \qquad B = (2, 3, 4)$$

$$\max x_1 + x_2 = 1 + x_1 - x_5$$

Let $x_2 = 1 - x_5$. Representation of the target function in the space of non-basis variables.

1.2.1 Sufficient optimality criterion for basis solutions

Basis solutions correspond to vertices.

A sufficient basis solution x for basis B (non-basis N) is optimal for the given linear program in normal form if $\tilde{c}_n \leq 0$ where $\tilde{c}_n$ is the vector of reduced cost coefficients.

> Reduced form
> $$\max \left\{ \tilde{c}_N^t x_n \mid \tilde{A}_N x_N \leq \tilde{b}_n, x_n \geq 0 \right\}$$
> with $\tilde{c}_N, \tilde{A}_N, \tilde{b}$ as established in the last lecture.

Remark. *The criterion is not necessary.*

Example (Continued)**.** $\tilde{c}_N = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \not\leq 0$. *Criterion not satisfied.*

Remark (Research question)**.** *How can we potentially improve the target function value if the optimality criterion is not satisfied?*

Currently we are in a vertex. Basis solution und non-basis variables with $\tilde{c}_{N(j)} = t_{0j} > 0$ have potential to give a better target function value if we increase $X_{N(j)}$ from 0 to some value > 0. We can only increase $X_{N(j)}$ such that admissibility is preserved.

Example (Continued)**.**

$$\max 1 + x_1 - x_5$$

$$\begin{aligned}
x_2 &= 1 - x_5 && \textit{no constraint} \\
x_3 &= 2 - x_1 + 2x_5 && \implies x_1 \leq 2 \\
x_4 &= 4 - 2x_1 - x_5 && \implies x_1 \leq 2 \\
x_i &\geq 0 \forall i \in \{2,3,4\}
\end{aligned}$$

We want to increase x_1! By how much is it admissible? The answer in this example is 2.

Which variable leaves the basis and becomes a non-basis variable instead of x_1? We have two options here: x_3 or x_4 (both values are 0 if $x_1 = 2$).

Assume we chose x_4. The new basis is $(1,2,3)$ and the new non-basis is $(4,5)$. We get a new reduced representation. And so on and so forth.

Step of improvement, general description

Let s chosen[1] such that $\tau_{N(s)} = t_{0s} > 0$ (optimality criterion is not satisfied). Our goal is to make $X_{N(s)}$ as large as possible. All other non-basis variables are fixed to be zero.

Case distinction:

Case 1: all $t_{is} \geq 0$ for all i $X_{N(s)}$ can be arbitrary large. The linear program is unbounded.

Case 2: there exists some i with $t_{is} < 0$ Then we determine

$$\varepsilon := \min \left\{ \frac{\overbrace{t_{i0}}^{\tilde{b}_i}}{-t_{is}} \mid t_{is} < 0 \right\}$$

Let r be such that $\varepsilon = \frac{t_{r0}}{-t_{rs}}$. $X_{B(r)}$ takes up the value 0.

We substitute the variable in $N(s)$ with the variable in $B(r)$.

Remark. *r is not necessarily unique. Currently, the choice in such cases is arbitrary.*

New basis:

$$\overline{B}(i) := \begin{cases} B(i) & i \neq r \\ N(s) & i = r \end{cases}$$

[1] the selection criteria will be discussed later

New non-basis:

$$\overline{N}(j) := \begin{cases} N(j) & j \neq s \\ B(r) & j = s \end{cases}$$

In the following, r will be called *pivot row*, s will be called *pivot column* and t_{rs} will be called *pivot element*. The transition from (B, N) to $(\overline{B}, \overline{N})$ is called *pivot step*.

So the basis solution for (B, N) becomes the basis solution for $(\overline{B}, \overline{N})$. The vertex x becomes the adjacent vertex $\overline{x}$.

Implementation of the basis exchange

Exchange $B(r) \leftrightarrow N(s)$. We solve the constraint belonging to pivot row r (to $x_{B(r)} = \dots$) by $X_{N(s)}$ and insert it into the remaining constraints.

$$-t_{rs} x_{N(s)} = t_{r0} - x_{B(r)} + \sum_{j \neq s} t_{rj} x_{N(j)}$$

$$\Longrightarrow \underbrace{x_{N(s)}}_{= x_{\overline{B}(r)}} = -\frac{t_{r0}}{t_{rs}} + \frac{1}{t_{rs}} x_{B(r)} + \sum_{j \neq s} \frac{t_{rj}}{t_{rs}} x_{N(j)}$$

This constraint is finished.

For $i \neq r$ we get

$$x_{\overline{B}(i)} = x_{B(i)} = t_{i0} + t_{is} \left(\frac{t_{r0}}{-t_{rs}} + \frac{1}{t_{rs}} x_{B(r)} + \sum_{j \neq s} \frac{t_{rj}}{-t_{rs}} x_{N(j)} \right) + \sum_{j \neq s} t_{ij} x_{N(j)}$$

$$= \underbrace{(t_{i0} - \frac{t_{is}}{t_{rs}} t_{r0})}_{\bar{t}_{i0}} + \underbrace{\frac{t_{is}}{t_{rs}}}_{\bar{z}_{is}} x_{\overline{N}(s)} + \sum_{j \neq s} \underbrace{\left(t_{ij} - \frac{t_{is}}{t_{rs}} t_{rj} \right)}_{\bar{t}_{ij}} x_{\overline{N}(j)}$$

$$x_{\overline{B}(i)} = \bar{t}_{i0} + \bar{t}_{is} x_{\overline{N}(s)} + \sum_{j \neq s} \bar{t}_{ij} x_{\overline{N}(j)}$$

$i \neq r$, constraint is finished.

Analogously, for the target function row (case $i = 0$).

Summary in tableau form

Tableau T is transformed into tableau F by a simplex step. The tableau is given as a table with highlighted 0th row (target function). The value t_{rs} is given by pivot row r and pivot column s.

The transformation laws for the pivot step are given by

$$\overline{t_{rs}} := \frac{1}{t_{rs}}$$

	s	j
r	A	B
i	C	D

new x-element = old x-element minus $\frac{C \cdot B}{A}$.
Pivot element $\rightarrow$ use reciprocal.
Remaining pivot column: $\cdot -1$ divided by pivot element
Remaining pivot row: divide by pivot element

$$\overline{t_{rj}} := -\frac{t_{rj}}{t_{rs}} \qquad j = 0, \ldots, n; j \neq s$$
$$\overline{t_{is}} := -\frac{t_{is}}{t_{rs}} \qquad i = 0, \ldots, m; j \neq r$$

Remark. *Internalize these rules by heart!*

In more detail:

Example (Our standard example)**.**

$$\max x_1 + x_2$$

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 4 \\ 2x_1 - x_2 + x_4 &= 3 \\ x_2 + x_5 &= 1 \\ x_1, x_2, x_3, x_4, x_5 &\geq 0 \end{aligned}$$

Initial tableau for $B = (3, 4, 5)$ *and* $N = (1, 2)$.

	x_1	x_2	
0	1	1	
4	1	2	x_3
3	2	-1	x_4
1	0	1	x_5

belongs to solution $x_1 = x_2 = 0, x_3 = 4, x_4 = 3$ *and* $x_5 = 1$.

Non-optimal, x_1 *and* x_2 *can be considered as new basis variables. Assume we choose* $s = 1$.

$$\varepsilon = \min\left\{\frac{4}{1}, \frac{3}{2}\right\} = \frac{3}{2}$$

$r = 2$, *so* x_4 *is removed from the basis. This corresponds to* $x_1 = \frac{3}{2}$ *and* $x_2 = 0$, *so target function value is* $\frac{3}{2}$.

	x_4	x_2	
$-\frac{3}{2}$	$-\frac{1}{2}$	$\frac{3}{2}$	
$\frac{5}{2}$	$-\frac{1}{2}$	$\frac{5}{2}$	x_3
$\frac{3}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	x_1
1	0	1	x_5

New basis: $(1, 3, 5)$
New non-basis: $(2, 4)$

The current values of the basis are retrievable.

It is not yet optimal. x_2 *should be removed from the non-basis.* $s > 2$ *(second column)*

$$\varepsilon = \min\left\{\frac{\frac{5}{2}}{\frac{5}{2}}, \frac{1}{1}\right\} = 1$$

Two options. Let's choose $r = 3$.

Tableau:

$$\begin{array}{c|cc|c}
 & x_4 & x_5 & \\
-3 & -\frac{1}{2} & -\frac{3}{2} & \\
\hline
0 & -\frac{1}{2} & -\frac{5}{2} & x_3 \\
2 & \frac{1}{2} & \frac{1}{2} & x_1 \\
1 & 0 & 1 & x_2
\end{array}$$

is optimal. $x_1 = 2$ *and* $x_2 = 1$ *with target function value* 3.

Next week: How can we determine an admissible solution?

↓ *This lecture took place on 2019/03/18.*

Revision. *It remains to discuss:*

- *Which solution is necessary to begin with if* b *is not* ≥ 0*?*
- *What about finiteness of the algorithm if the basis solution is degenerated?*

To determine admissible basis solutions, we have 2 approaches: The first one is called "2-phase method by Dantzig".

2-phase method by Dantzig

Given $Ax = b$ with $x \geq 0$ and $\exists i : b_i < 0$.

We sort the rows (= constraints) of A and b such that in the first rows are those with negative right-hand side.

$$b = \begin{pmatrix} \hat{b} \\ \hat{\hat{b}} \end{pmatrix} \text{ with } \hat{b} < 0, \hat{\hat{b}} \geq 0 \text{ and } A = \begin{pmatrix} \hat{A} \\ \hat{\hat{A}} \end{pmatrix}$$

System in normal form (block remains unchanged):

$$\hat{A}x + \hat{y} = \hat{b} \qquad \hat{\hat{A}}x + \hat{\hat{y}} = \hat{\hat{b}}$$

where $\hat{\hat{y}}$ is the vector of slack variables for the remainder with $b_i < 0$ and $b_i \geq 0$. By multiplication with -1:

$$\begin{aligned} -\hat{A}x - \hat{y} + u &= -\hat{b} \\ \hat{\hat{A}}x + \hat{\hat{y}} &= \hat{\hat{b}} \end{aligned}$$

Choose u and $\hat{\hat{y}}$ as basis variable.

Resolve by the basis variables

$$\begin{aligned} u &= -\hat{b} + \hat{A}x + \hat{y} \\ \hat{\hat{y}} &= \hat{\hat{b}} - \hat{\hat{A}}x \end{aligned}$$

Gives an admissible basis solution (of the new system). The remainders are non-basis variables.

$$u = -\hat{\hat{b}} \qquad \hat{\hat{y}} = \hat{\hat{b}}$$

Solutions with $u \neq 0$ are non-admissible solutions for our original problem. Our goal is to find solutions with $u = 0$ if such a solution exists.

The implementation is done by introducing an auxiliary problem

$$\min e^t u \text{ with } u \geq 0 \qquad \text{where } e = (1, \ldots, 1)$$

$$\iff \max \underbrace{-e^t u}_{Z_H}$$

means that we find a solution with $u = 0$, if possible.

$$Z_H = -e^t u = e^t(\hat{b} - \hat{A}x - \hat{y}) = e^t\hat{b} - e^t\hat{A}x - e^t\hat{y}$$

Representation in the space of non-basis variables:

Example.

$$\max -x_1 - 2x_2$$

subject to

$$\begin{aligned} x_1 + x_2 &\geq 3 \\ x_2 &\geq 2 \\ -x_1 + x_2 &\leq 3 \\ x_1 - x_2 &\leq 3 \\ x_1, x_2 &\geq 0 \end{aligned}$$

$$\begin{aligned} \Longrightarrow \quad -x_1 - x_2 &\leq -3 \\ -x_2 &\leq -2 \\ -x_1 + x_2 &\leq 3 \\ x_1 - x_2 &\leq 3 \\ x_1, x_2 &\geq 0 \end{aligned}$$

$$
\begin{aligned}
\Longrightarrow\ x_1 + x_2 - y_1 + u_1 &= 3 \\
x_2 - y_2 + u_2 &= 2 \\
-x_1 + x_2 + y_3 &= 3 \\
x_1 - x_2 + y_4 &= 3 x_1, x_2, y_1, y_2, y_3, y_4, u_1, u_2 \qquad \geq 0
\end{aligned}
$$

Here the variables before the equality sign are the basis variables to begin with. The auxiliary problem is given by

$$\max -u_1 - u_2 = \max -5 - y_1 - y_2 + x_1 + 2x_2$$

	x_1	x_2	y_1	y_2	
5	1	2	−1	−1	
0	−1	−2	0	0	
3	1	1	−1	0	u_1
2	0	1	0	−1	u_2
3	−1	1	0	0	y_3
3	1	−1	0	0	y_4

The initial solution is given by (is non-optimal):

$$u_1 = 3 \qquad u_2 = 2 \qquad y_3 = 3 \qquad y_4 = 3$$

Choose $r = 2$ and $s = 2$ and apply the pivot step method.

	x_1	u_1	y_1	y_2	
1	1	−2	−1	1	
4	−1	2	0	−2	
1	1	−1	−1	1	u_1
2	0	1	0	−1	x_2
1	−1	−1	0	1	y_3
5	1	1	0	−1	y_4

u_2 became non-basis variables and we thus we can remove (i.e. ignore) the column corresponding to u_2. This solution is non-optimal (choose $s = 1$ and $r = 1$).

	u_1	y_1	y_2	
0	1	0	0	
5	+1	−1	−1	
1	1	−1	1	x_1
2	0	0	−1	x_2
2	1	−1	2	y_3
4	−1	1	−2	y_4

optimal for the auxiliary problem. We remove the second column from left.

$$x_1 = 1 \qquad x_2 = 2$$

is an admissible solution for the original problem (is already optimal for the original problem otherwise continue with the remaining tableau in the second phase). The target function value is 0 for the auxiliary problem. $u_1 = u_2 = 0$.

Thus the algorithm looks as follows:

1. Continue until the first auxiliary problem is solved.

   Case 1 The auxiliary problem has optimal value 0, then second phase

   Case 2 The auxiliary problem has optimal value $\neq 0$, then stop because the original problem does not have an admissible solution

2. Continue until the original problem is optimally solved.

Remark. *Once variable u_i end up in the non-basis, the corresponding column in tableau can be removed. At the end of the first phase, the auxiliary row can be removed. Attention! If some u_i is a basis variable at the end of the first phase by degeneration, this variable must not be removed for the second phase.*

1.2.2 M-method

Consider the following problem:

$$\max c^t x \text{ s.t. } Ax + y = b \text{ with } x, y \geq 0, \exists i : b_i < 0$$

Now, we consider the modified problem

$$\max c^t x - M\tilde{x} \text{ s.t. } Ax + y - \tilde{e}^t \tilde{x} = b \qquad x, y, \tilde{x} \geq 0$$

$$\tilde{e}_i := \begin{cases} 1 & b_i < 0 \\ 0 & b_i \geq 0 \end{cases}$$

thus the new variable $\tilde{x}$ is substracted with $b_i < 0$ in the constraints. M is sufficiently large such that $\tilde{x}$ takes up value 0 in an optimal solution ($\exists$ a solution with $\tilde{x} = 0 \iff$ original problem has an admissible solution).

Remark. *Or we can consider*

$$Ax + y - e^t \tilde{x} = b \qquad x, y, \tilde{x} \geq 0$$

$\tilde{x}$ occurs in all constraints.

To get an admissible solution for the auxiliary problem (the problem extended by $\tilde{x}$), wlog. $b_m = \min_{x \leq 1 \leq m} b_i$ with $b_m < 0$ assuming A has m rows. Subtract the m-th row from all the others (for the variant, where $-\tilde{x}$ occurs in all constraints)

$$\begin{aligned}
(a_{11} - a_{m1})x_1 + \dots \quad & +(a_{11} - a_{mn})x_n + y_1 & & & - y_m \quad &= b_1 - b_m \\
& & & & & \qquad (1) \\
(a_{21} - a_{m1})x_1 + \dots \quad & +(a_{21} - a_{mn})x_n & & +y_2 & - y_m \quad &= b_2 - b_m \\
\vdots \quad & & & \ddots & & \vdots \\
(a_{m-1,1} - a_{m1})x_1 + \dots \quad & +(a_{m-1,r} - a_{mn})x_n & & & -y_{m-1} - y_m \quad &= b_{m-1} - b_m \\
-a_{m,1}x_1 + \dots \quad & + - a_{mn}x_n + \tilde{x} & & & - y_m \quad &= \underbrace{-b_m}_{>0} \\
& & & & & \qquad (2)
\end{aligned}$$

m-th row multiplied with -1. With $x_{n+1}, \dots, x_{n+m-1}, \tilde{x}$ we are given an admissible solution.

Remark (Problem in practice)**.** *One problem in practice is the choice of M.*

Our workaround is not to choose M explicitly. Instead we split the target function into two parts (the auxiliary part including $\tilde{x}$ and the remainder). This represents the lexicographic ordering for vectors.

Example.

	x_1	x_2	$\tilde{x}$	
0	0	0	-1	
0	-1	-2	0	
-3	-1	-1	-1	x_3
-2	0	-1	-1	x_4
3	-1	1	-1	x_5
3	1	-1	-1	x_6

must be converted into a correct initial tableau. Either by (1) *or alternatively bring $\tilde{x}$ into the basis. Choose the variable, that leaves the basis as the one, where the constraints take up $\min b_i$ (this corresponds to* (1)*).*

Remark. *Here the constraints are written in the original order. They must be modified with* (1)*.*

	x_1	x_2	x_3	
3	1	1	-1	
0	-1	-2	0	
3	1	1	-1	$\tilde{x}$
1	1	0	-1	x_4
6	0	2	-1	x_5
6	2	0	-1	x_6

The auxiliary problem is not yet optimal. Optimize the auxiliary problem first.

	x_4	x_2	x_3	
2	-1	1	0	
1	1	-2	-1	
2	-1	1	0	$\tilde{x}$
1	1	0	-1	x_1
6	0	2	-1	x_5
6	-2	0	1	x_6

Still not yet optimal for the auxiliary problem. $s = 2, r = 1$.

	x_4	$\tilde{x}$	x_3	
0	0	-1	0	
5	-1	2	-1	
2	-1	1	0	x_2
1	1	0	-1	x_1
2	2	-2	-1	x_5
4	-2	0	1	x_6

The auxiliary problem is solved and $\tilde{x} = 0$ (optimal value 0). So, there exists an admissible solution for the original problem.

If $\tilde{x}$ is in the non-basis (as in the example), this column can now be removed. Usually we continue with the remaining problem with the original target function. Here we stop, because optimality was reached.

On finiteness

Remark (Obvious observation)**.** *If no degenerated basis solutions occur, the Simplex method is finite (in every iteration the target function value increases, because $\tilde{c}_{N(s)} > 0$, $a_{rs} > 0$ and $\tilde{b}_{B(r)} > 0$).*

Question: Can it happen that, in the case of degenerated basis solutions, we traverse a cycle?
Answer: Yes, if we do not take proper prerequisites.

Such an example goes back to Gass (see practicals, exercise 45)

$$\max \frac{3}{4}x_1 - 150x_2 + \frac{1}{50}x_3 - 6x_4$$

subject to

$$\begin{aligned} \frac{1}{4}x_1 - 60x_2 - \frac{1}{25}x_3 + 9x_4 &\leq 0 \\ \frac{1}{2}x_1 - 90x_2 - \frac{1}{50}x_3 + 3x_4 &\leq 0 \\ x_3 &\leq 1 \\ x_1, x_2, x_3, x_4 &\geq 0 \end{aligned}$$

Question: Which prerequisites can we make to avoid cycles?

In the following, there are two approaches:

- Rule by Bland
- lexicographical row selection rule

Remark (About the lexicographical approach)**.** *Recall that in the case of degenerated basis solutions the choice of the pivot row is ambiguous (with the previous approach).*

Idea: *Introduce an extended criterion to choose a pivot row.*
Remark: *On constraint on the choico of a pivot column!*

A new rule for choice of the pivot columns considers the lexicographical minimum over row vectors instead of the minimum over scalars.

↓ *This lecture took place on 2019/03/19.*

1.2.3 Rules to avoid cycles

Today, we will discuss the lexicographical row selection rule.

We need appropriate measures to ensure in every pivot step a certain kind of progress. We already know that considering the target function value itself does not suffice.

In case of degenerated basis solutions, the minimum is taken up in computation of ε^* for ≥ 2 rows.

$$\begin{array}{c|ccc} & & & \\ \hline * & \dots & \square & \dots \\ * & \dots & \square & \dots \end{array}$$

So far, we ignored the columns filled with dots.

For the compact representation of the row selection rule, we start with the system $(A|I)x = b$ that results from the inequality system by introduction of slack variables.

Definition. *Let v be a vector in $\mathbb{R}^k$. v is called* lexicographically positive *(compact notation: $v \succ 0$) if its first non-zero component is positive.*

$v \in \mathbb{R}^k$ *is lexicographically*	*smaller-equal*	*than* $u \in \mathbb{R}^k$, *if*	$v = u$ *or* $u - v \succ 0$
	smaller	*than* $u \in \mathbb{R}^k$, *if*	$u - v \succ 0$
	greater-equal	*than* $u \in \mathbb{R}^k$, *if*	$v = u$ *or* $v - u \succ 0$
	greater	*than* $u \in \mathbb{R}^k$, *if*	$v - u \succ 0$

In the following, we call lexmin *the minimum with respect to this lexicographical order of vectors.*

Example. *Thus,* $(0, 0, 0, 4, 1, -7)$ *is lexicographically positive.* $(0, -1, 2, 3)$ *is lexicographically negative.*

Example. *Compare this example with the practicals exercise 45.*

$$\max \frac{3}{4}x_1 - 150x_2 + \frac{1}{50}x_2 - 6x_4$$

subject to

$$\begin{aligned} \frac{1}{4}x_1 - 60x_2 - \frac{1}{25}x_3 + 9x_4 + x_5 &\leq 0 \\ \frac{1}{2}x_1 - 90x_2 - \frac{1}{50}x_3 + 3x_4 + x_6 &\leq 0 \\ x_3 + x_7 &\leq 1 \\ x_1, x_2, x_3, x_4 &\geq 0 \end{aligned}$$

Choose $s = 1$. The classical row selection rule provides no distinction between

row 1 and 2 ($\frac{0}{\frac{1}{4}}$ versus $\frac{0}{\frac{1}{2}}$). Here the lexicographic rule chooses $r = 1$. $u \prec v$.

$$\begin{array}{c|ccccc|ccc}
 & x_1 & x_2 & x_3 & x_4 & & x_5 & x_6 & x_7 \\
0 & \frac{3}{4} & -150 & \frac{1}{50} & 6 & & 0 & 0 & 0 \\
\hline
0 & \frac{1}{4} & -60 & -\frac{1}{25} & 9 & & 1 & 0 & 0 \\
0 & \frac{1}{2} & -90 & -\frac{1}{50} & 3 & & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & & 0 & 0 & 1
\end{array}$$

with first and second row as,

$$\begin{array}{c|cccc|ccc}
0 & 1 & -240 & -\frac{4}{25} & 36 & 4 & 0 & 0 \\
0 & 1 & -180 & -\frac{1}{25} & 6 & 0 & 2 & 0
\end{array}$$

In the following, we need the lexicographical sorting for vectors. Without loss of generality we assume that the rows of our tableau in extended form is lexicographically positive (otherwise apply column exchanges).

The remaining example is left for the practicals.

Definition (Lexicographical row seletion rule)**.** *Let s be the chosen pivot column. The pivot row r is chosen as the lexicographically smallest of the weighted row vectors of rows i with $t_{is} > 0$. t_{is} is the tableau entry in row i and column s.*

$$\operatorname{lexmin}_{i \in \{1,\dots,n\}} \left\{ \frac{t_i}{t_{is}} \mid t_{is} > 0 \right\}$$

where t_i is the vector of the i-th tableau row.

Remark. *The choice above provides a unique solution. (Assumption: the choice is ambiguous.)*

$$\frac{t_i}{t_{is}} = \frac{t_k}{t_{ks}} \text{ for } i \neq k \implies \text{ the } i\text{-th row is a multiple of the } k\text{-th row}$$
$$\implies x \neq 0 \text{ with } t_i = \lambda \cdot t_k$$

This is a contradiction with $\operatorname{rank}(A) = m$.

It remains to show that the lexicographical row selection rule satisfies its purpose.

Theorem 1.2.2. *If you choose the pivot row by the lexicographical row selection rule, then*

1. *The vector in the target function row decreases strictly lexicographically.*
2. *All row vectors stay positive.*

Proof. 1. The new target function coefficients result from

$$\bar{t}_{vj} := t_{oj} - t_{rj} \underbrace{\frac{t_{os}}{t_{rs}}}_{>0} \qquad j \in \{1, \dots, n+m\}$$

Because the first non-vanishing $t_{rj} > 0$, we get $T_0 \prec t_0$ where T_0 is the new vector (in the tableau in the target function row) and t_0 is the old vector. Thus, we get lexicographical decline.

2. For $t_{is} > 0$, due to choice of r,

$$\frac{1}{t_{is}} t_i \succ \frac{1}{t_{rs}} t_r$$

$$\Longrightarrow \overline{t_i} = t_i - \frac{t_{is}}{t_{rs}} \cdot t_r \succ 0$$

where $\overline{t_i}$ is the i-th row. For $t_{is} \leq 0$ due to $t_{rs} > 0$, we have

$$\overline{t_i} = t_i - \frac{t_{is}}{t_{rs}} \cdot t_r \succ t_i$$

Thus all vectors retain lexicographically positive.

□

Theorem 1.2.3. *The Simplex method with lexicographical row selection rule is finite.*

Proof. Immediate by Theorem 1.2.2 (1). No basis solution can occur twice and there are only finitely many basis solutions. □

Remark (Where does the lex. selection rule come from?)**.** *Perturbation of the polyhedron/equation system:*

$$b \to b + \varepsilon \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \qquad 0 < \varepsilon_m \ll \varepsilon_{m-1} \ll \cdots \ll \varepsilon_1 \ll \text{ all other data}$$

Geometrically, a small deviation (perturbation) is introduced to the polyhedron.

In practice this cannot be implemented, because proper choice of ε is unknown. The lexicographical rule is an implicit implementation of this idea.

Remark (About the Rule by Bland)**.** *We constrain the choice of the pivot column and the pivot row. We choose the pivot column s as follows:*

$$N(s) := \left\{ N(j) \mid t_{0j} > 0 \quad j \in \{0, \ldots, n\} \right\}$$

(among the non-basis variables that violate the optimization condition, choose the one with smallest index).

Choose the pivot row r as follows:

$$B(r) := \min \left\{ B(q) \mid \tilde{b}_q = \min_{\substack{i \in \{1, \ldots, m\} \\ t_{is} > 0}} \tilde{b}_i \right\}$$

with $\tilde{b}_q := \frac{t_{q0}}{t_{qs}}$ and $\tilde{b}_i := \frac{t_{i0}}{t_{is}}$. So this chooses the variable with the smallest index *among all the variables we consider for leaving the basis.*

Theorem 1.2.4. *The Simplex method enforced with the Rule by Bland is finite.*

Proof. Compare with appropriate literature. □

In the previous example, we need to choose $s = 1$ (due to the lexicographical rule, $s = 3$ would be possible as well). $r = 1$ (x_5 is the smallest index).

This conclude the generic Simplex method.

1.3 Extensions and Algorithmic Aspects of the Simplex method

In this chapter, we are going to discuss extensions as well as efficiency in the algorithmic implementation of the Simplex method.

1.3.1 Possible rule for choice of the pivot column

If we do not use the rule by Bland, we can choose every column that violates the optimality condition (thus $t_{0j} > 0$ with $t_{0j} = \tilde{c}_{N(j)}$). In literature the following rules are recommended:

Method of steepest slope in the space of non-basis variables. Also called *Rule by Dantzig.* Choose column s with $t_{0s} = \max t_{0j}$, i.e. $\max\left\{t_{0j} \mid t_{0j} > 0\right\}$.

Method of largest absolute increase of the target function. For every possible choice of j of the pivot column (hence $t_{0j} > 0$), we determine the corresponding pivot row $r(j)$. For every possible choice of j, this results in one pivot element $t_{r(j)j}$. Choose j (pivot column s) only such that we maximize

$$\frac{t_{0j}t_{r(j)0}}{t_{r(j)j}}$$

where t_{0j} corresponds to $(\tilde{c}_n)_j$ and $t_{r(j)0}$ corresponds to $\tilde{b}_{r(j)}$ and the fraction represents the increase of the target function.

Remark. *Disadvantage: determination is computationally intense.*

Method of steepest slope in the space of all variables. If we choose $x = (x_B, x_N)^t$ in some vertex, then the solution is given by $(\tilde{b}, 0)^t$.

If $x_{N(j)}$ increases from 0 to 1 (by 1 unit), then the target function value increases by $(\tilde{c}_N)_j$ (corresponds to t_{0j}), by the one hand, and x_B changes to

$$x_B = \underbrace{A_B^{-1}b}_{\tilde{b}-\tilde{a}_j} - A_B^{-1}A_N X_{N(j)}$$

on the other hand. Here j denotes the j-th column of $\tilde{A}_N$ (column corresponding to $X_{N(j)}$).

- Change in $\mathbb{R}^{m+n}$ by $(-\tilde{a}_{1j}, \ldots, -\tilde{a}_{mj}, 0, \ldots, 0, 1, 0, \ldots, 0)$ where 1 is given for $X_N(j)$

- for $N(j)$ corresponding component of the gradient of the target function in $\mathbb{R}^{m+k}$
- Row selection rule: Choose s such that $\frac{t_{0j}}{\sqrt{1+\sum_{i=1}^{m} t_{ij}^2}}$ (where t_{0j} corresponds to $(\tilde{c}_N)_j$ and t_{ij}^2 corresponds to $\tilde{a}_{ij}^2$) becomes maximal for $s = j$ and below the columns j with $(\tilde{c}_N)_j > 0$.

↓ *This lecture took place on 2019/03/25.*

1.3.2 Extensions to the Simplex method

Ideas: Direct handling of equations, no sign-restricted variables, upper bounds.

Basically this is not required, because every linear program can be transformed to canonical form. This happens to the disadvantage on the number of restrictions or the number of variables.

1st case: no sign-restricted variables Adjustment of the optimality criterion, adjustments to determine the pivot row. The details are given in the practicals.

2nd case: equations The approach is similar to the 2-phases method. Introduce one auxiliary variable for each equation and auxiliary target function. The minimum number of auxiliary variables corresponds to the maximum minus the sum of auxiliary variables. Compare this with the practicals.

3rd case: lower bound

$$l_j \leq x_j$$

Transformation to $\tilde{x}_j \geq 0$ by $\tilde{x}_j = x_j - l_j$. Compare this with the practicals.

4th case: upper bound

$$x_j \leq u_j \qquad (0, \ldots, 0, 1, 0, \overbrace{\ldots}^{x_j}, 0)$$

Goal: implicit handling of such residue classes without maintaining them in the tableau.

Idea: $x_j + \overline{x_j} = u_j$. We call $\overline{x_j}$ *complementary variables* to x_j and x_j complementary to $\overline{x_j}$. The value of x_j results from value $\overline{x_j}$; vice versa with $\overline{x_j}$ and x_j. It suffice to maintain one of the two variables; the value of the other can be easily determined.

Determine K, the index set of the variables bounded by above. We have $0 \leq x_j \leq u_j$ for $j \in K$ and $0 < u_j < \infty$. N denotes the index vector of the current non-basis and B is the index vector of the current basis. S is the index of the pivot column.

We need to distinguish three cases:

Case 1 $x_{N(s)}$ is bounded, so $x_{N(s)} \leq u_{N(s)}$

Case 2 $x_{B(i)} \leq 0 \rightarrow x_{B(i)} = \tilde{b}_i - \tilde{a}_{is} x_{N(s)} \geq 0 \implies x_{N(s)} \leq \frac{\tilde{b}_i}{\tilde{a}_{is}}$ for $\tilde{a}_{is} > 0$

Case 3 $x_{B(i)}$ is bounded by $u_{B(i)}$

$$x_{B(i)} = \tilde{b}_i - \tilde{a}_{is} x_{N(s)} \leq u_{B(i)} \implies x_{N(s)} \leq \frac{\tilde{b}_i - u_{B(i)}}{\tilde{a}_{is}} \text{ for } \tilde{a}_{is} < 0$$

By the choice of the pivot row, we need to determine the following minimum.

$$\min \left\{ \underbrace{U_{N(s)}}_{\text{Case 1}}, \underbrace{\min \left\{ \frac{\tilde{b}_i}{\tilde{a}_{is}} \mid \tilde{a}_{is} > 0 \right\}}_{\text{Case 2}}, \underbrace{\min \left\{ \frac{\tilde{b}_i - u_{B(i)}}{\tilde{a}_{is}} \mid \tilde{a}_{is} < 0 \text{ and } B(i) \in K \text{ i.e. } x_{B(i)} \text{ is bounded by above} \right\}}_{\text{Case 3}} \right\}$$

It remains to discuss what happens if the minimum results from the expression in case 1 or 3.

If it results from ...

Case 1, then
- $x_{N(s)}$ turns from 0 to $\tilde{U}_{N(s)}$
- $\overline{x_{N(s)}}$ becomes zero
- We replace $x_{N(s)}$ by $\overline{x}_{N(s)}$
- So a new non-basis variable is introduced, the basis retains unchanged

$$\overline{x_{N(s)}} = u_{N(s)} - x_{N(s)} \qquad \text{insertion into tableau representation}$$

$$\tilde{a}_{i1} x_{N(1)} + \cdots + \tilde{a}_{is} \underbrace{x_{N(s)}}_{(u_{N(s)} - \tilde{x}_{N(s)})} + \cdots + \tilde{a}_{in} x_{N(n)} + x_{B(i)} = \tilde{b}_i$$

Transformation $T(s)$ s-th column with -1 and substitute the right hand side $\tilde{b}_i$ by $\tilde{b}_i - \tilde{a}_{is} u_{N(s)}$.

Case 2, then classical pivot operation with $\tilde{a}_{rc} > 0$.

Case 3, then combination of exchange of basis variables with non-basis variables and transition to complementary variables

Computational implementation:

- first, make a pivot step with $\tilde{a}_{rs}$ (consider $\tilde{a}_{rs} = 0$)
- second, apply transformation $T(s)$

Remark (Compact representation)**.** *Use negative indices for complementary variables $\overline{x_3}$ in basis, so use -3 instead of 3 as index in B.*

Example 4.

$$\max -x_1 + 4x_2$$

subject to

$$\begin{aligned} x_1 - x_2 &\leq 2 \\ -x_1 + x_2 &\leq 3 \\ x_2 &\leq 4 \\ x_1, x_2 &\geq 0 \end{aligned}$$

...becomes ...

$$\max -x_1 + 4x_2$$

subject to

$$\begin{aligned} x_1 - x_2 + x_3 &= 2 \\ x_2 &\leq 4 \\ x_3 &\leq 5 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

where $x_2 \leq 4$ and $x_3 \leq 5$ are two variables with upper bounds.

$$\begin{array}{c|ccc} & x_1 & x_2^* & \\ 0 & -1 & 4 & \\ \hline 2 & 1 & -1 & x_3 \end{array}$$

$$s = 2 \qquad \min \left\{ \underbrace{4}_{\text{Case 1}}, \underbrace{\frac{2-5}{-1}}_{\text{Case 3}} \right\} = 3$$

(a):

$$\begin{array}{c|ccc} & x_1 & x_3 & \\ 8 & 3 & 4 & \\ \hline -2 & -1 & -1 & x_2 \end{array}$$

(b): Transformation $T(2)$:

$$\begin{array}{c|ccc} & x_1 & \overline{x}_3 & \\ -12 & 3 & -4 & \\ \hline -3 & -1 & 1 & x_2 \end{array}$$

$-1 \implies$ admissible tableau.

Now $s = 1$.

$$\min \left\{ \underbrace{\frac{3-4}{-1}}_{\text{Case 3}} \right\} = 1$$

$$\begin{array}{c|cc}
 & x_2 & \overline{x_3} \\
-3 & 3 & -1 \\
\hline
-3 & -1 & -1 \quad x_1
\end{array}$$

Transformation $T(1)$

$$\begin{array}{c|cc}
 & \overline{x}_2 & \overline{x_3} \\
-15 & 3 & -1 \\
\hline
1 & 1 & -1 \quad x_1
\end{array}$$

$$\Longrightarrow x_1 = 1 \qquad \overline{x_2} = 0 \Longrightarrow x_2 = 4 \qquad \overline{x_3} = 0 \Longrightarrow x_3 = 5$$

1.3.3 The revised Simplex method

Recall:

- $x_B = A_B^{-1} b$
- $\tilde{C}_n^t = C_N^t - C_B^t A_B^{-1} A_N$
- $\tilde{A}_N = A_B^{-1} A_N$

One interpretation of the steps in the context of pivot operations of the Simplex method is, that 3 linear equation systems are solved.

- $A_B x_B = b$ to retrieve the current basis solution, $x_B = \tilde{b}$
- $A_B^t \Pi = c_B$, $\rightarrow \tilde{c}_N^t = c_N^t - \Pi_t A_n$
- $A_B \tilde{a}_s = a_s$ where a_s is the s-th column of A and $\tilde{a}_s$ is the s-th column of $\tilde{A}$

Regarding dimensions:

$$A_B, A_B^t \ldots m \times n \text{ matrices} \qquad \text{tableau size } O(mn)$$

If we solve the 3 equation systems in every step manually, we need $O(m^3)$ computational resources (independent of n). In the tableau $O(nm)$, so only useful for n at least $O(m^2)$.

In practice, we use the fact that A_B changes in every step only marginally (1 column!). The update formulas can be given explicitly (also for corresponding basis inverse). From a numerical perspective, it is advantageous to combine the process with known methods of linear algebra (for example, LU decomposition).

Conclusion: There exists numerically stable implementations of the Simplex method. Without proper prerequisites the rounding errors accumulate, cancellation effects occur such that results for larger ill-conditioned problems become unuseful.

1.3.4 Brief consideration of the runtime of the Simplex method

Obviously, the application of one pivot step can be done in polynomial time. The essential question is, does a polynomial boundary exist for the maximum number of pivot steps? A partial answer can be given: No such boundary is known and for all known column selection rules pathological examples have been found such that exponentially many pivot steps are required.

By current research, the Simplex method does not provide a polynomial solution for the linear program. We are going to cover one such method in the section about the inner point method. But in practice, the Simplex method works very good. Empirically, $O(n)$ pivot steps suffice.

Further analysis was done:

- Average case analysis by Borgwardt
- Smoothed analysis by Spielman, et al.

1.4 Duality for linear optimization

We will cover the definition of duality, duality theorems and alternative theorems.

1.4.1 Example for motivation

Example (Transportation example)**.** *Linear transportation problem: transportation costs c_{ij}, m firms, offer a_i, n customers and demand b_j.*

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}$$

subject to

$$\begin{aligned} \sum_{j=1}^{n} x_{ij} &= a_i \text{ for } i = 1, \dots, m \\ \sum_{i=1}^{m} x_{ij} &= b_j \text{ for } j = 1, \dots, n \\ x_{ij} &\geq 0 \text{ for } i = 1, \dots, m; j = 1, \dots, n \end{aligned}$$

↓ *This lecture took place on 2019/03/26.*

External transportation service: Purchase at firm with price u_i, sale at customer j with price v_j.

$$\max\left(\sum_{i=1}^{n} b_j v_j - \sum_{i=1}^{m} a_i u_i\right) \qquad \textit{models the profit}$$

$$v_j - u_i \le c_{ij} \qquad i = 1, \dots, m; j = 1, \dots, n$$

$$v_j + \overline{u_i} \le c_{ij} \qquad i = 1, \dots, m; j = 1, \dots, n$$

or correspondingly with $\overline{u}_i = -u_i$. *More examples in an economical context can be found.*

1.4.2 Definition of a dual linear program

Definition. *Let (P) be a linear problem* $\max\{c^t x \mid Ax \le b, x \ge 0\}$. *The dual linear problem (D) has form* $\min\{b^t y \mid A^t y \ge c, y \ge 0\}$.

Remark. *The problem (P) is also called* primal problem.

Example 5. *We compare the models side-by-side.*

$$\max x_1 + 2x_2 \qquad (P)$$

subject to

$$\begin{aligned} x_1 &\le 4 & y_1 \\ 2x_1 + x_2 &\le 10 & y_2 \\ -x_1 + x_2 &\le 5 & y_3 \\ x_1, x_2 &\ge 0 \end{aligned}$$

$$\min 4y_1 + 10y_2 + 5y_3 \qquad (D)$$

subject to

$$\begin{aligned} y_1 + y_2 - y_3 &\ge 1 \\ y_2 + y_3 &\ge 2 \\ y_1, y_2, y_3 &\ge 0 \end{aligned}$$

Example 6.

$$\min 4x_1 + 3x_2$$

subject to

$$x_1 + 2x_2 \ge 7 \tag{3}$$

$$2x_1 - x_2 \ge 5 \tag{4}$$

$$3x_1 + x_2 \ge -2 \tag{5}$$

First, we model it canonically:

$$\max -4x_1 - 3x_2$$

$$\begin{aligned} -x_1 - 2x_2 &\le -7 \\ -2x_1 + x_2 &\le -5 \\ -3x_1 - x_2 &\le 2 \end{aligned}$$

Secondly, we create sign constraints with $x_1 = x_1^+ - x_1^-$ *and* $x_2 = x_2^+ - x_2^-$.

$$\max -4x_1^+ + 4x_1^- - 3x_2^+ + 3x_1^-$$

subject to

$$\begin{aligned} -x_1^+ + x_1^- - 2x_2^+ + 2x_2^- &\leq -7 && y_1 \\ -2x_1^+ + 2x_1^- + x_2^+ - x_2^- &\leq -5 && y_2 \\ -3x_1^+ + 3x_1^- - x_2^+ + x_2^- &\leq 2 && y_3 \\ x_i^\pm &\geq 0 \end{aligned}$$

Thus we get the dual problem:

$$\min -7y_1 - 5y_2 + 2y_3$$

subject to

$$\begin{aligned} -y_1 - 2y_2 - 3y_3 &\geq -4 \\ y_1 + 2y_2 + 3y_3 &\geq 4 \\ -2y_1 + y_2 - y_3 &\geq -3 \\ 2y_1 - y_2 + y_3 &\geq 3 \\ y_1, y_2, y_3 &\geq 0 \end{aligned}$$

$$\iff \min -7y_1 - 5y_2 + 2y_3$$

subject to

$$\begin{aligned} y_1 + 2y_2 + 3y_3 &= 4 \\ 2y_1 - y_2 + y_3 &= 3 \\ y_1, y_2, y_3 &\geq 0 \end{aligned}$$

Example 7. *Take the resulting dual problem in Example 2 and determine its dual problem.*

Remark (Observation)**.** *The dual problem of the primal problem is the dual problem. The dual problem of the dual problem is the primal problem.*

In general:

1. An equation in the primal problem corresponds to one unbounded (non-sign-restricted) variable in the dual problem
2. A non-sign-restricted variable in the primal problem corresponds to one equation in the dual problem
3. The dual problem of the dual problem is the primal problem

Example 8. *Consider the transportation problem. Begin with the minimization problem (the dual maximization problem).*

$$\max \sum_{i=1}^{m} a_i \cdot \alpha_i + \sum_{j=1}^{n} b_j \beta_j$$

subject to

$$\alpha_i + \beta_j \leq c_{ij}$$

Primal problem	Dual problem
Maximization of the target function	Minimization of the target function
target function coefficients (c)	RHS vector
RHS vector (b)	target function vector
coefficients matrix (A)	transposed coefficients matrix
i-th constraint is $\leq$ constraint	i-th dual variable $y_i \geq 0$
i-th constraint is $=$ constraint	i-th dual variable y_i is not sign-restricted
$x_j \geq 0$	j-th constraint $\geq$ constraint
x_j is not sign-restricted	j-th constraint $=$ constrtaint (equation)

Remark. *α_i and β_j are not sign-restricted, because in the primal problem we have equations. Compare with the transportation problem.*

Original constraints:

$$(D) \leftarrow (P) \left(\begin{array}{ccccccccc} 1 & \dots & 1 & 0 & \dots & & & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ & & & & \vdots & & & & \\ 0 & \dots & & & & 0 & 1 & \dots & 1 \\ \hline 1 & \ddots & 0 & 1 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & & 1 & 0 & & 1 & 1 & & 1 \end{array}\right)$$

The horizontal separates the firms model (top m rows) from the customers model (bottom n rows) giving $m+n$ rows in total.

Remark (Semantics of dual variables).

- *β_j correponds to v_j in the introductory transportation problem (where v_j represents the price to be paid for transportation at customer j).*
- *α_i corresponds to u_i (where u_i represents the price to be paid at firm i)*

Remark. *Do* not *mix models! So either model a linear program*

- *as maximization problem with constraints $\leq$ or $=$*
- *as minimization problem with constraints $\geq$ or $=$*

Then apply the rules.

Remark. *Consider Example 2. Every point satisfying* (3)–(5) *also satisfies the corresponding linear combinations. e.g. 2 times* (3) + (4)*. Searching for the best boundaries exactly leads to the dual linear program.*

1.4.3 Duality- and alternative theorems

2 central duality theorems:

- strong duality theorem

  $c^t x^* = b^t y^*$ $\quad$ x^* optimal for primal problem, y^* optimal for dual problem

- weak duality theorem

$$c^t x \leq b^t y \qquad \text{for } x \in M_P = \{x \mid Ax \leq b, x \geq 0\}, y \in M_D = \left\{y \mid A^t y \geq c, y \geq 0\right\}$$

and the theorem of the complementary slackness (optimality check).

Remark. *In the following, let M_D denote the admissible set of the dual problem. Let M_P denote the admissible set of the primal problem.*

Theorem 1.4.1 (Weak duality theorem)**.** *1. $x \in M_P, y \in M_D$. Then $c^t x \leq b^t y$*

2. *Let $M_P \neq \emptyset$ (hence the primal problem has an admissible solution) ($c^t x$ over M_P). Then M_P is empty, if the primal theorem is unbounded.*
3. *Let $M_D \neq \emptyset$ (hence the dual problem has an admissible solution). Then M_P is empty iff the dual problem is unbounded.*

Proof. 1. $x \in M_P$, so $Ax \leq b$ with $x \geq 0$. $y \in M_D$, so $A^t y \geq c$, $y \geq 0$.

$$c^t x \leq (A^t y)^t x = y^t A x \leq y^t b = (b^t y)$$

2. It suffices to prove (2) or (3).

   We will prove (3) und for this purpose, we are going to prove an auxiliary result first. And we are going to show this later.

□

We are going to briefly cover the topic of alternative theorems. The following is an introductory example:

Theorem 1.4.2. *A is an $m \times n$ matrix over $\mathbb{R}$. Let $b \in \mathbb{R}^m$. Exactly one of the following alternatives holds true:*

1. $\exists x \in \mathbb{R}^n : Ax = b$
2. $\exists y \in \mathbb{R}^m : y^t A = 0, y^t b = 1$

Hence, either the linear equation system $Ax = b$ has a linear solution or the linear equation system $y^t A = 0$ and $y^t b = 1$ has a linear solution.

Proof. Assume both alternatives are simultaneously true. This immediately gives a contradiction because $0 = g^t A x = y^t b = 1$.

The first case has no solution iff b is not in the space spanned by columns of A, hence $\operatorname{rank}([A|b]) = \operatorname{rank}(A) + 1$

$$\rightarrow \operatorname{rank}\left(\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix}\right) = \operatorname{rank}(A) + 1$$

if and only if the last row of the matrix above is linear dependent on the rows of $\begin{pmatrix} A & b \end{pmatrix}$, hence

$$y^t A = 0, y^t b = 1$$

has a solution. □

Theorem 1.4.3. *Exactly one of the following alternatives holds true:*

1. $\exists x \in \mathbb{R}^n : Ax = b, x \geq 0$
2. $\exists y \in \mathbb{R}^m : y^t A \geq 0, y^t b < 0$

Proof. If both alternatives hold true, then $0 \leq y^t Ax = y^t b < 0$. This gives a contradiction

Without loss of generality, we can assume that $b \geq 0$. Otherwise, we change the signs in the i-th row of $Ax = b$ and y_i simultaneously.

If the first alternative is not solvable, then the linear optimization problem

$$\begin{aligned}\gamma &:= \max\left\{-e^t u \mid Ax + u = b, x \geq 0, u \geq 0\right\} \text{ where } e \text{ is the zero-vector} \\ &:= \max\left\{-\sum u_i \mid Hz = b, z \geq 0, u \geq 0\right\}\end{aligned}$$

has only solutions with negative target function value (because no solution with $u = 0$ exists).

$$\max\left\{-e^t u \mid Ax + u = b, x, u \geq 0\right\} \text{ with } -e^t u = \begin{pmatrix} 0 & -e \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix} \text{ and } -e^t u = -\sum u_i \tag{o}$$

has only solutions with negative target function values (but exists, call it γ).

So $\gamma < 0$. Now we apply the fundamental theorem of linear optimization. Thus there exists an optimal (finite) basis solution for (o) with target function value γ. □

↓ *This lecture took place on 2019/04/01.*

Revision 1 (Theorem 1.4.3)**.**

1. *either $Ax = b$ with $x \geq 0$ has a solution*
2. *or $y^t A \geq 0, y^t b < 0$ has a solution*

Corollary 1.4.4 (Farkas Lemma, 1894)**.** *TFAE:*

- $\exists x \in \mathbb{R}^n : Ax = b, x \geq 0$
- $\forall y \in \mathbb{R}^m : y^t A \geq 0 \implies y^t b \geq 0$

Also has a geometric interpretation: The halfspace $\{y \mid y^t b \geq 0\}$ contains a *polyhedral cone* $\{y \mid y^t A \geq 0\}$ iff $b \in \{Ax \mid x \geq 0\}$ where $\{Ax \mid x \geq 0\}$ is the cone $K(A)$.

Now we can complete the proof of Theorem 1.4.1. Thus we need to show statement (3) of Theorem 1.4.1.

Revision 2. *Let $M_D \neq \emptyset$. Then M_P is empty iff $y^t b$ is unbounded by below.*

Proof.

- Let $M_p \neq \emptyset$. By (1), we have that $c^t x \leq b^t y \forall x \in M_p$, $y \in M_D$, thus every $x \in M_p$ provides a lower bound for the target function values of the primal problem.
- Let $M_p = \emptyset$, thus $Ax = b$ with $x \geq 0$ has no solution. P has no admissible solution, so $y^t A \geq 0, y^t b < 0, y \geq 0$ has a solution (why? we will discuss it later). Let $\hat{y}$ be such a solution. Let $\overline{y} \in M_D$. Consider $\overline{y} + \lambda \hat{y} \in M_D$ for $\lambda \geq 0$ (half-line).

Target function value: $(\overline{y} + \lambda \hat{y})^t b = \overline{y}^t b + \lambda \hat{y}^t b$ is unbounded by below. □

$$(P) \qquad z_P = \max\left\{c^t x \mid Ax \leq b, x \geq 0\right\} \qquad M_p = \{x \mid Ax \leq b, x \geq 0\}$$

$$(D) \qquad p_D = \min\left\{b^t y \mid A^t y \geq c, y \geq 0\right\} \qquad M_D = \left\{y \mid A^t y \geq c, y \geq 0\right\}$$

If $M_P \neq \emptyset$ and $M_D \neq \emptyset$, then $z_P \leq z_D$ by Theorem 1.4.1.

Question: In this case, can $z_P < z_D$?
Answer: No, see Theorem 1.4.6.

Lemma 1.4.5. *$P_z = \{x \mid z \leq c^t x, Ax \leq b, x \geq 0\}$. Let $M_p \neq \emptyset$. Then $P_z = \emptyset \iff y^t b < z, A^t y \geq c, y \geq 0$ solvable.*

Theorem 1.4.6 (Strong duality theorem)**.** *If one of the linear problems, dual to each other, has a finite optimal solution, then also the other and the optimal target function values correspond (thus, $z_P = z_D$).*

Proof. There are various, different proofs, including one directly over the Simplex method. It suffices to prove one of the two statements.

Let (D) have a finite optimal solution y^*, so $b^t y^* = z_D$. By Theorem 1.4.1 (3), we have $M_p \neq \emptyset$ ($b^t y$ is bounded by below). Because there is no admissible solution for (D) with target function value $< z_D$, the system $y^t b < z_D$, $A^t y \geq c, y >\geq 0$ is not solvable. Now consider P_{z_D} with $P_z = \{x \mid x \leq c^t x, Ax \leq b, x \geq 0\}$. We can conclude $P_{z_D} \neq \emptyset$ (represented in Lemma 1.4.5), thus $\exists x^*$ with $z_D \leq c^t x^*$, $Ax^* \leq b, x^* \geq 0$ with $x^* \in M_P$. Now we apply the weak duality theorem (Theorem 1.4.1), so $c^t x^* \leq b^t y^*$ where $b^t y^* = z_D$

$$\implies c^t x^* = b^t y^* = z_D$$

hence x^* is optimal for the primal problem and y^* is optimal for the dual problem. □

Proof of Lemma 1.4.5. $M_P \neq \emptyset$ (equivalently $\exists x : Ax = b, x \geq 0$) iff $y^t b < 0, A^t y \geq 0, y \geq 0$ has no solution (remember this as "the criterion").

$$P_z = \emptyset \iff \begin{matrix} (y_0, y^t)\binom{-z}{y} < 0 \\ (y_0, y)^t\binom{-c^t}{A} \geq 0 \end{matrix} \text{ is solvable}$$

A solution with $y_0 = 0$ is not possible due to "the criterion". So $y_0 > 0$. We can substitute $y_0 \geq 0$ by $y_0 = 1$ (because the set of solutions creates a cone)

without restricting the set of solutions. Thus

$$P_z = \emptyset \iff y^t b < z, y^t A \geq c^t, y \geq 0 (A^t y \geq c) \text{ solvable}$$

□

Remark (P has no admissible solution, so $y^t A \geq 0$ has a solution)**.** *It is trivial to see that the alternative theorem 1.4.3 can be generalized to a combination of equations and inequalities.*

Theorem 1.4.7 (Theorem of complementary slackness)**.** *Let $x \in M_p$ and $y \in M_D$. Then x is optimal for the primal problem and y is optimal for the dual problem if and only if (1) $x^t(A^t y - c) = 0$ and (2) $y^t(Ax - b) = 0$. Then (x, y) is called* optimal pair.

Remark (Interpretation of $x^t(A^t y - c) = 0$ and $y^t(Ax - b) = 0$)**.**

- $x_i = 0$ *(i-th primal variable is 0) or $(A^t y - c)_i = 0$ (the i-th restriction of the dual problem is fulfilled with equality) (corresponds to the i-th slack variable in the dual problem).*
- $y_i = 0$ *(i-th primal variable is 0) or $(Ax - b)_i = 0$ (the i-th primal restriction is fulfilled with equality).*

This follows because

$$x \geq 0 \qquad (A^t y - c) \geq 0 \text{ because } A^t y \geq c$$

hence every summand must be contained in $x^t(A^t y - c) \geq 0$ and thus equals zero.

$y \geq 0$ $(A^t x - b) \leq 0$, *thus every summand is in $y^t(Ax - b) \leq 0$ und thus equals zero.*

Remark. *The interpretation above plays a vital role in optimality tests (examples will be provided in the practicals). Also they can be used to derive an optimal solution of the primal problem given the solution of the dual problem; vice versa.*

Proof of Theorem 1.4.7.

- Let (x, y) be an optimal pair, thus $Ax \leq b$, $x \geq 0$, $A^t y \geq c$, $y \geq 0$, $c^t x = b^t y$.

$$(Ax)^t y \leq b^t y = c^t x \implies \underbrace{x^t}_{\geq 0} \underbrace{(A^t y - c)}_{\geq 0} \leq 0 \implies x^t(A^t y - c) = 0$$

Analogously, $(A^t y)^t x \geq c^t x = b^t y$,

$$\underbrace{y^t}_{\geq 0} \underbrace{(Ax - b)}_{\leq 0} \geq 0 \qquad \to \qquad y^t(Ax - b) = 0$$

- If $x \in M_P$ and $y \in M_D$ satisfy $x^t(A^t y - c) = 0$ and $y^t(Ax - b) = 0$, then

$$x^t(A^t y - c) = y^t(Ax - b) = 0 \implies x^t A^t y - \underbrace{x^t c}_{c^t x} = y^t A x - \underbrace{y^t b}_{b^t y}$$

hence $c^t x = b^t y$ so (x, y) is an optimal pair.

□

Remark. *We can even show sometimes more restrictive: There always exists an optimal pair (x^*, y^*) that is strictly complementary, thus*

$$\begin{aligned} x_j^* = 0 &\iff (A^t y^* - c)_j > 0 \qquad j = 1, \dots, n \\ y_i^* = 0 &\iff (Ax - b)_i < 0 \qquad i = 1, \dots, m \end{aligned}$$

Attention! *Not all optimal pairs (x, y) are strictly complementary. (A proof is not given at this point)*

1.4.4 The dual Simplex method

In contrast to the primal problem, we now consider ...

$$(P)\ \max\left\{c^t x \mid Ax \le b, x \ge 0\right\} \qquad (D)\ \min\left\{b^t y \mid A^t y \ge c, y \ge 0\right\}$$

$$(D')\ \max\left\{-b^t y \mid -A^t y \le -c, y \ge 0\right\}$$

This gives an equivalent problem to (D) in canonical form.

Idea: Solve (D') with the primal Simplex method and we consider this as a method to solve the original problem (*dual Simplex method*, Lemke, 1954)

Again, we use the tableau form. Let $s = -c$ be the left-most column (the set of basis variable values) and $z = -b^t$ be the top row (cost coefficients) in the tableau. A Simplex tableau T is called

dually admissible	if $c \le 0$ (i.e. $-c > 0$)
primally admissible	if $b \ge 0$ (i.e. $s \ge 0$)
dually optimal	if $-b \le 0$ (i.e. $b \ge 0$)
primally optimal	if $z \le 0$

Example 9.

$$\max -x_1 - 2x_2$$

subject to

$$\begin{aligned} x_1 + x_2 &\ge 3 \\ -x_2 &\le -2 \\ x_1, x_2 &\ge 0 \end{aligned}$$

In canoncial form, we have

$$\max -x_1 - 2x_2$$

subject to

$$\begin{aligned} -x_1 - x_2 &\le -3 \\ -x_2 &\le -2 \\ x_1, x_2 &\ge 0 \end{aligned}$$

0	−1	−2	
−3	−1	−1	x_3
−2	0	−1	x_4

The initial tableau is dually admissible, but not primally.

Remark. *The tableau is optimal if it has at least one of the following states:*

- *primally admissible + primally optimal*
- *dually admissible + dually optimal*
- *primally admissible + dually admissible*

The dual Simplex method begins with a dually admissible tableaus and we consecutively apply pivot steps until a dually optimal tableau is given.

Question: What do pivot steps in the dual method look like?

Optionally, we need an approach to determine a dually admissible initial solution.

↓ *This lecture took place on 2019/04/02.*

Dual Simplex method:

- Choice of a pivot row. Choice of a row with entry < 0 in 0-th column or if none exists, the optimal solution was reached.
  Analogously to the choice of the pivot column in the primal process.
- Choice of the pivot column (ensures that the tableau stays dually admissible).
  Choose column s with $\frac{t'_{0s}}{t'_{rs}} = \min\left\{\frac{t'_{0s}}{t'_{rj}} \mid t'_{rj} < 0\right\}$.
- Application of the pivot step

$$\bar{t}'_{rs} = \frac{1}{t'_{rs}} \qquad \bar{t}'_{rj} = \frac{t'_{rj}}{t'_{rs}} \text{ for } j \neq s$$

$$\bar{t}'_{is} = -\frac{t'_{is}}{t'_{rs}} \text{ for } i \neq r \qquad \bar{t}'_{ij} = t'_{ij} - \frac{t'_{is} \cdot t'_{rj}}{t'_{rs}}$$

Choose a column s with [column selection rule]

$$\frac{t_{0s}}{t_{rs}} = \min\left\{\frac{-t_{0j}}{-t_{rj}} \mid t_{rj} < 0\right\} = \min\left\{\frac{t'_{0j}}{t'_{rj}} \mid t'_{rj} < 0\right\}$$

Remark. *We can generalize this to the lexicographical column selection rule (this way we can avoid circles)*

Example 10.

$$\max -x_1 - 2x_2$$

subject to

$$\begin{aligned} -x_1 - x_2 + x_3 &= -3 \\ -x_2 + x_4 &= -2 \\ -x_1 + x_2 + x_5 &= 3 \\ x_1 - x_2 + x_6 &= 3 \\ x_1, x_2 &\geq 0 \end{aligned}$$

This gives the following tableau:

	x_1	x_2	
	-1	-2	
-3	-1	-1	x_3
-2	0	-1	x_4
3	-1	1	x_5
3	1	-1	x_6

where the row of coefficients is dually admissible (hence not primally optimal) and the non-basis is not primally admissible and not dually optimal.

We choose pivot element $r = 1$ and $s = 1$.

	x_3	x_2	
3	-1	-1	
3	-1	1	x_1
-2	0	-1	x_4
6	-1	2	x_5
0	1	-2	x_6

The coefficient row is still dually admissible. The non-basis column is not yet dually optimal.

We choose pivot element $r = 2$ and $s = 2$.

	x_3	x_4	
5	-1	-1	
1	-1	1	x_1
2	0	-1	x_2
2	-1	2	x_5
4	1	-2	x_6

The non-basis column is dually optimal and thus dually admissible.

The optimal solution is given with $x_1 = 1$ and $x_2 = 2$. The optimal target function value is -5.

Remark (Advantages of the dual Simplex method).

- *Assume the initial tableau is dually admissible (as in the example before) and the right-hand side is not greater-equal zero. Then we can begin immediately with the dual process whereas the primal process requires an initial solution first.*
- *The main field of application is given by the following property: Assume some linear program is already optimally solved and we attach additional*

restrictions. The dual Simplex method enables us to continue directly with the given/previous solution basis solution. This solution stays dually admissible. If the solution is not yet dually optimal, we apply dual Simplex steps until we get a dually optimal solution.

Typically the processing time is reduced due to the reoptimization strategy (a so-called "warm start") compared to a restart.

Remark (Cutting planes method)**.** *The dual Simplex method has an important role in the cutting planes method. This method solves integral (or mixed) linear programs.*

Cutting planes discard undesirable (non-integral) optimal solutions of linear relaxation (problem of non-integrability) without loss of admissible integer solutions. The quality of relaxation is improved step by step.

- *Handling variables with upper bounds*

$$\max -x_1 + 4x_2$$

subject to

$$\begin{aligned} x_1 - x_2 &\leq 2 \\ -x_1 + x_2 &\leq 3 \\ x_1 + x_2 &\geq 3 \\ x_2 &\leq 4 \\ x_1, x_2 &\geq 0 \end{aligned}$$

This leads to an initial tableau, that is neither primally nor dually admissible.

The fourth restriction is transformed to $x_2 + \overline{x}_2 = 4$ *with*

$$\max 16 - x_1 - 4\overline{x}_2$$

subject to

$$\begin{aligned} x_1 + \overline{x}_2 &\leq 6 \\ -x_1 - \overline{x}_2 &\leq -1 \\ -x_1 + \overline{x}_2 &\leq 1 \\ x_1 &\geq 0 \\ \overline{x}_2 &\geq 0 \\ \overline{x}_2 &\leq 4 \end{aligned}$$

This results in the following tableau:

	x_1	$\overline{x}_2$	
-16	-1	-4	
6	1	1	x_3
-1	-1	-1	x_4
1	-1	1	x_5

This is a dually admissible tableau. The pivot element is $r = 2$ and $s = 1$.

	x_4	$\overline{x}_2$	
-15	-1	-3	
5	1	0	x_3
1	-1	1	x_1
2	-1	2	x_5

$\Longrightarrow$ *optimal (dually admissible and optimal) with $x_1 = 1, x_2 = 4$ (because $\overline{x}_2 = 0$)*

1.4.5 Final remarks on dual solutions

Question: How can we determine an optimal solution of the dual problem if the primal problem was solved by the Simplex method?

Consider the normal form $\max c^t x$ with $Ax = b$ and $x \geq 0$ (with slack variables) and the optimal basis solution (x_B, x_N where N is the non-basis and B is the basis). We can determine vector $\tilde{y}$ as $\tilde{y}^t := c_B^t A_B^{-1}$.

Claim. *$\tilde{y}$ represents an optimal basis for the primal problem.*

Proof. B is an optimal basis for our primal problem, so the reduced cost coefficients satisfy $c \leq 0$, thus $\tilde{c}^t \leq 0$.

$$c^t - \tilde{y}^t A \leq 0$$

This gives 0 for the basis variables and the reduced cost coefficients for non-basis variables. Thus $\tilde{y}$ is admissible for the primal problem.

Furthermore it is trivial to see that $(\tilde{x}, \tilde{y})$ is an optimal pair, because $\tilde{x}_i$ is admissible for the primal problem and $\tilde{y}$ is admissible for the dual problem and have the same target function value.

$$c^t \tilde{x} = c_B^t \tilde{x}_B = c_B^t A_B^{-1} b \text{ (because } \tilde{x} \text{ is a basis solution)}$$

$$b^t \tilde{y} = \tilde{y}^t b \underbrace{=}_{\text{definition of } \tilde{y}} c_B^t A_B^{-1} b$$

□

Remark.
- *Due to the property above, another proof of the strong law of duality follows intuitively (original proof by Dantzig). Above we have seen it for the case that primal and dual problems have a finite optimal solution.*
- *A second method for determining a solution of*
  - *the dual problem if the primal optimal solution is given*
  - *the primal problem if the dual optimal solution is given*

  is the application of the Theorem of complementary slackness 1.4.7.

Question: What are other examples for the economical interpretation of the dual problem?

Example 11. *Consider Stigler's diet problem introduced after Example 3.*

Let n be the number of food products and m be the total number of ingredients. Let x_j be the quantity of food product j. Let a_{ij} be the ratio of the i-th ingredient in food product j. And b_i denotes the minimal demand of ingredient i. Finally c_j are the costs per unit for ingredient j.

Our goal is to design an admissible diet plan minimizing costs.

Assume pharmaceutical company offers pills per ingredient. Now let y_i denote the price of pill type i.

The goal of the pharmaceutical company is to maximize the profit.

The primal problem is given by

$$\min\left\{c^t x \mid Ax \geq b, x \geq 0\right\}$$

$$A^t y \leq c \qquad y^t A \leq c^t$$

The dual problem is given by

$$\max b^t y$$

subject to

$$\begin{aligned} A^t y &\leq c \\ y &\geq 0 \end{aligned}$$

The duality theorem states that the minimal costs of the end consumer match the maximum profit of the pharmaceutical company.

About the optimal prices of the pharmaceutical company: Transformation of the primal problem to the normal form (for the minimization problem):

$$\min\left\{c^t x + 0^t z \mid Ax - z = b, x \geq 0, z \geq 0\right\}$$

$$\iff \min\left\{d^t u \mid Hu = b, u \geq 0\right\}$$

is the alternate problem denoted by $(\tilde{D})$ (where z are the slack variables).

$$\begin{pmatrix} x \\ z \end{pmatrix} = u \qquad d = \begin{pmatrix} c \\ 0 \end{pmatrix} \qquad H = (A \mid -I)$$

If B is the optimal basis for problem $(\tilde{P})$, then we get an optimal solution for the dual problem $(\tilde{D})$ $\max\{b^t y \mid H^t y \leq d\}$ given by $\tilde{y}^t = d_B^t H_B^{-1}$ is given by $\tilde{y}^t = d_B^t H_B^{-1}$ (compare with the first question in this section 1.4.5). The dually admissible solution $\tilde{y}^t H \leq d^t$ implies $\tilde{y} \geq 0$ (H contains the negated unit matrix).

Assume the end consumer only wants to cover a specific ratio of the demand $\triangle b$ by pills. For simplification assume $H_B^{-1}(b - \triangle b) \geq 0$. Then the basis b stays optimal. The costs of the ingredients get reduced by $\triangle z = \sum_i \triangle z_i$. The relative reduction $\frac{\triangle z_i}{\triangle b_i}$ is also called marginal price of b_i

$$\triangle z = z_p - d_B^t H_B^{-1}(b - \triangle b) = d_B^t H_B^{-1} \triangle b = \overline{y} \triangle b$$

The marginal costs are now market prices $\overline{y}_i$.

If $a_i^t\overline{x} > b_i$ *is true for some optimal solution* $\overline{x}$ *of the primal problem, then the demand is more than expected. By the Theorem of complementary slackness,* $\overline{y}_i = 0$ *and thus the pill* i *is unsellable.*

1.5 Inner point methods

↓ *This lecture took place on 2019/04/08.*

Goal: Method to solve linear programs with polynomial runtime (in the worst case) and is practical.

Remark. *The Simplex method does not provide these runtime guarantees. There are examples with an exponential number of pivot steps. The Ellipsoid method has a polynomial runtime, but is impractical and not covered in this lecture.*

There are 3 kinds of inner point methods:

- primal
- dual
- primal-dual

In this course, we will cover the basics of the primal-dual path following method. Reminder of the basics of the Simplex algorithm: only vertices are considered.

In case of inner point methods, the interior plays a major role. Essential basis for primal-dual methods (esp. path following method) is the theorem of complementary slackness.

$$\begin{aligned} (P)\ \max\ & c^t x \\ \text{s.t. } & Ax \leq b \\ & x \geq 0 \end{aligned} \qquad \begin{aligned} (D)\ \min\ & b^t y \\ \text{s.t. } & A^t y \geq c \\ & y \geq 0 \end{aligned}$$

$$\begin{aligned} (\overline{P})\ \max\ & c^t x \\ \text{s.t. } & Ax + w = b \\ & x, w \geq 0 \end{aligned} \qquad \begin{aligned} (\overline{D})\ \min\ & b^t y \\ \text{s.t. } & A^t y - z = c \\ & y, z \geq 0 \end{aligned}$$

Definition. (x, w, y, z) *is an optimal pair for* $(\overline{P})$ *and* $(\overline{D})$ *if*

OPT $Ax + w = b; A^t y - z = c; x, w, y, z \geq 0$

Complementary condition $y^t(Ax - b) = y^t w = w^t y = 0; x^t(A^t y - c) = x^t z = z^t x = 0$ *and accordingly* $w_i y_i = 0 \forall i, z_j \cdot x_j = 0 \forall j$

We define the interior of primal-dual problems

$$F = \left\{(x, w, y, z) \mid Ax + w = b; A^t y - z = c; w^t y = z^t x = 0; x, y, w, z \geq 0\right\}$$

$$F^0 = \{(x, w, y, z) \in F \mid x, y, w, z > 0\}$$

Remark. *In Analysis: Lagrange method is used to handle equations as constraints.*

Here, we have (P) *and* (D) *and also sign-restricted constraints in the primal-dual system. One approach to handle such constraints are so-called* barrier functions. *These are special penalty expressions or penalty functions. Those penalize negative values of sign-restricted variables. Such a function is the logarithm.*

1.5.1 Primal barrier problem

$$
\begin{aligned}
(BP\mu)\ \max\, & c^t x + \mu\left(\sum_{j=1}^{n} \log x_j + \sum_{i=1}^{b} \log w_i\right) \qquad \mu > 0 \\
\text{s. t. } & Ax + w = b
\end{aligned}
$$

$(BP\mu)$ is closely related with (P), or equivalently, $(\overline{P})$.

$$
\begin{aligned}
(BD\mu)\ \min\, & b^t y - \mu\left(\sum_{j=1}^{m} \log y_j + \sum_{i} \log z_i\right) \qquad \mu > 0 \\
\text{s. t. } & A^t y - z = c
\end{aligned}
$$

We can apply the Lagrange method to $(BP\mu)$ and $(BD\mu)$.

Remark. *y and z is newly defined. It turns out, that y and z are equal to the original y and z.*

By $BP\mu$:

$$
\begin{array}{lll}
\overline{L}(x,w,y) & = c^t x + \mu \sum_j \log x_j & \\
 & +\mu \sum_i \log w_i + y^t(b - Ax - w) & y \text{ is a Lagrange multiplier} \\
\frac{\partial \overline{L}}{\partial x_j} & = c_j + \mu \frac{1}{x_j} - \sum_i y_i a_{ij} \overset{!}{=} 0 & j = 1,\dots,n \quad A^t y - z = c \\
\frac{\partial \overline{L}}{\partial w_i} & = \mu \frac{1}{w_i} - y_i \overset{!}{=} 0 & i = 1,\dots,m \quad w_i y_i = \mu \\
\frac{\partial \overline{L}}{\partial y_i} & = b_i - \sum_j a_{ij} x_j - w_i \overset{!}{=} 0 & i = 1,\dots,m \quad Ax + w = b
\end{array}
$$

$$
\begin{aligned}
\Longrightarrow\ & w_i y_i = \mu \qquad i = 1,\dots,n \\
& Ax + w = b \qquad A^t y - z = c
\end{aligned}
$$

Let $z_j = \mu \frac{1}{x_j}$ and accordingly in matrix form $z = \mu X^{-1} e$ with the following conventions (in this section):

- e is always the vector full of ones

- x vector $\rightarrow$

$$X = \operatorname{diag}(x) \cdot \begin{pmatrix} x_1 & & 0 \\ & \ddots & \\ 0 & & x_n \end{pmatrix} \qquad X^{-1} = \begin{pmatrix} \frac{1}{x_1} & & \\ & \ddots & \\ & & \frac{1}{x_n} \end{pmatrix} \qquad x_i \neq 0$$

The same applies to the dual problem:

$$\Longrightarrow \; A^t y - z = c \qquad Ax + w = b \qquad x_j - z_j = \mu$$

We can also create the primal-dual barrier problem. In total we get (let's denote the problem as (0)):

$$\begin{aligned} A^t y - z &= c \\ Ax + w &= b \\ x_j \cdot z_j &= \mu \quad \forall j \\ w_i \cdot y_i &= \mu \quad \forall i \\ (w_i x_i y_i z &\geq 0 \quad \text{implicit}) \end{aligned}$$

for $\mu \rightarrow 0$ against constraints from (OPT) (page 58). 3rd and 4th constraints are generalized complementary constraints.

A graphical illustration is given in Figure 3. An analogous construction can be made for the dual problem and the primal-dual variant.

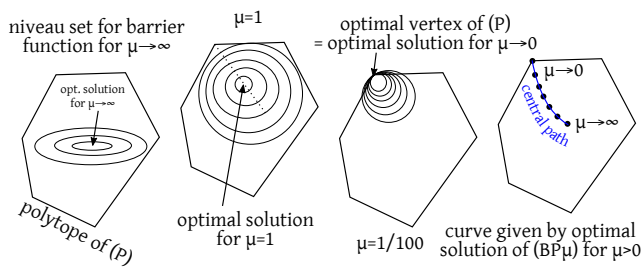

Figure 3: Graphical illustration

(0) in vector notation:

$$\begin{aligned} A^t y - z &= c \\ Ax + w &= b \\ XZe &= \mu e \\ YWe &= \mu e \end{aligned}$$

Question: Is there always a solution of the corresponding barrier problem?
Answer: No.

We consider two examples:

Example.

$$\max 0 \xrightarrow{\textit{barrier function}} \max \mu \log x \qquad \mu > 0$$

subject to $x \geq 0$ does not have a finite optimum.

Example.

$$\max -x \xrightarrow{\textit{barrier function}} \max -x + \mu \log x$$

(takes maximum with $x = \mu$) subject to $x \geq 0$ has a unique solution $x = 0$.

Theorem 1.5.1. *There exists a (finite) solution for the barrier problem if and only if the admissible set for the primal as well as the dual problem has a non-empty interior.*

Proof. $\Longrightarrow$ trivial

$\Longleftarrow$ Assume there are inner points for (P) and (D), thus

- there exists $(\overline{x}, \overline{w})$ an admissible solution for $(\overline{P})$ with $\overline{x} > 0, \overline{w} > 0$ and
- there exists $(\overline{y}, \overline{z})$ an admissible solution for $(\overline{D})$ with $\overline{y} > 0, \overline{z} > 0$

so $(\overline{x}, \overline{w}, \overline{y}, \overline{z}) \in F^0$. Now consider $\overline{z}^t x + \overline{y}^t w$ for arbitrary admissible (x, w) for $(\overline{P})$. Insert slack variables:

$$\Longleftrightarrow \ (A^t \overline{y} - c)^t x + \overline{y}^t (b - Ax) = b^t \overline{y} - c^t x$$

Hence our target function is given as $c^t x = \overline{z}^t x - \overline{y}^t w + b^t \overline{y}$.

$\rightarrow$ barrier problem of primal program

$$\begin{aligned} f(x, w) &= c^t x + \mu \sum_j \log x_j + \mu \sum_i \log w_i \\ &= \sum_j \left(-\overline{z}_j x_j + \mu \log x_j \right) + \sum_i \left(-\overline{y}_i w_i + \mu \log w_i \right) + \underbrace{b^t \overline{y}}_{\text{constant}} \end{aligned}$$

Every summand in both sums is a function in only one variable. Those functions are all of form

$$\rho(\xi) = -\alpha \xi + \mu \cdot \log \xi \qquad \xi \in (0, \infty), \alpha > 0$$

Takes up a unique maximum for $\frac{\mu}{\alpha}$. Limit for $\xi \to \infty$ and $\to -\infty$

$$\left\{ (x, w) \in \mathbb{R}^{n+m} \mid f(x, w) \geq c \right\}$$

is bounded for constant c.

↓ *This lecture took place on 2019/04/09.*

$$f(x,w) = c^t x + \mu \sum_j \log x_j + \mu \sum_i \log w_i$$

$$S := \left\{(x,w) \in \mathbb{R}^{n+m} \mid f(x,w) \geq c\right\} \text{ bounded}$$

Let $\overline{f} = f(\overline{x}, \overline{w})$. $(\overline{x}, \overline{w})$ is strictly admissible for $(\overline{P}, \overline{x} > 0, \overline{w} > 0)$. Consider

$$\overline{Q} := \left\{(x,w) \mid Ax + w = b, x \geq 0, w \geq 0, f(x,w) \geq \overline{f}\right\}$$

Then $\overline{Q} \neq \emptyset$ because $(\overline{x}, \overline{w}) \in \overline{Q}$. Because S is bounded, $\overline{Q}$ is bounded. Moreover, $\overline{Q}$ is closed. We show this the following way:

$$\overline{Q} = \{(x,w) \mid Ax + w = b\} \cap \{(x,w) \mid x \geq 0, w \geq 0\} \cap \{(x,w) \mid f(x,w) \geq F\}$$

$\{f(x,w) \geq f\}$ is the inverse image of $[\overline{f}, \infty)$ for the continuous function f, thus also closed. $\overline{Q}$ is the intersection of three closed set, thus closed itself.

Thus, $\overline{Q} \subseteq R^{n+m}$ is compact. Hence, continuous functions on non-empty compact sets $\overline{Q}$ take up a maximum.

f takes up the maximum on $\{(x,w) \mid x > 0, w > 0\}$ because by definition of $\overline{Q}$, the maximum is taken up in a region with values $\geq \overline{f}$.

Thus, the existence of a solution for the barrier problem is guaranteed. In the following, we will show uniqueness:

$$f(x,w) = c^t x + \mu \sum_j \log(x_j) + \mu \sum_i \log(w_i)$$

$$\frac{\partial f}{\partial x_j} = c_j + \frac{\mu}{x_j} \qquad \frac{\partial^2 f}{\partial x_j^2} = -\frac{\mu}{x_j^2} \qquad \frac{\partial f}{\partial x_j w_i} = 0$$

$$\frac{\partial f}{\partial w_i} = \frac{\mu}{w_i} \qquad \frac{\partial^2 f}{\partial w_i^2} = -\frac{\mu}{w_i^2}$$

$\Longrightarrow$ The Hessian matrix of f is a diagonal matrix full of negative values on the diagonal (f is strictly concave). Every critical point of f is a maximum.

$\Longrightarrow$ Furthermore here there can be at most one single critical point and if it exists, it represents a global maximum.

Thus uniqueness is given.

□

Corollary 1.5.2. *If the admissible set of $(\overline{P})$ (and accordingly $(\overline{D})$) contains a strictly admissible set (thus has a non-empty interior) and is bounded, then the system (we call it S)*

$$Ax + w = b \qquad A^t y - z = c \qquad XZe = \mu e \quad (x_j z_j = \mu \forall j) \qquad YWe = \mu e \quad (y_i w_i = \mu \forall i)$$

For all $\varepsilon > 0$ there exists a unique solution $(x_\mu, w_\mu, y_m u, z_\mu)$.

Proof. Follows from Theorem 1.5.1 and Lemma 1.5.3, which will be proven in the practicals. □

Definition (Null variable)**.** *A variable that takes up value* 0 *in all admissible solutions of the linear program.*

Lemma 1.5.3. *If a linear program has admissible solutions and the set of admissible solutions is bounded, then the corresponding dual program has strictly admissible solutions.*

More specifically,

$$M_{\overline{P}} = \{(x, w) \mid Ax + w = b, x \geq 0, w \geq 0\} \qquad M_{\overline{P}} \neq \emptyset \text{ bounded}$$

$$M_{\overline{D}} = \left\{(y, z) \mid A^t y - z = c, y, z \geq 0\right\}$$

$\exists (z, w) \in M_{\overline{D}}$ *with* $z, w > 0$.

Remark (Equivalently and easier to prove)**.** *If a linear program has admissible solution and the dual problem has null variables, then the admissible set of the linear problem is unbounded.*

Definition.

$$\left\{(x_\mu, w_\mu, y_\mu, z_\mu) \mid (x_\mu, w_\mu, y_\mu, z_\mu) \text{ is solution of the system } S \text{ from above}, \mu > 0\right\}$$

is called central path.

Remark. *Assuming that* $(\overline{P})$ *and* $(\overline{D})$ *has strictly admissible solutions (in the following, this is assumed), the central path is well-defined.*

Remark. *Fundamentals finished. Now we want to achieve algorithmic implementation. These are the basic steps of a primal-dual path traversal method.*

The method discussed here is a one-phase method. We begin with $(x, w) > 0$, *but* (x, w) *must not be admissible for* $(\overline{P})$. $(y, z) > 0$, ..., (y, z) *for* $(\overline{D})$.

Our goal is to solve $(\overline{P})$ *and* $(\overline{D})$ *optimally. The method is a non-finite iteration process requiring a termination condition.*

0. Begin with an arbitrary $(x, w, y, z) > 0$
1. Determine an appropriate value for μ
2. Determine the direction $(\triangle x, \triangle w, \triangle y, \triangle z)$ (here we use path point $(x_\mu, w_\mu, y_\mu, z_\mu)$ as guide)
3. Determine the step size ϑ such that $(\tilde{x}, \tilde{w}, \tilde{y}, \tilde{z}) > 0$ with

$$\begin{aligned} \tilde{x} &:= x + \vartheta \cdot \triangle x \\ \tilde{w} &:= w + \vartheta \cdot \triangle w \\ \tilde{y} &:= y + \vartheta \cdot \triangle y \\ \tilde{z} &:= z + \vartheta \cdot \triangle z \end{aligned}$$

4. Substitute (x, w, y, z) by $(\tilde{x}, \tilde{w}, \tilde{y}, \tilde{z})$.

As long as the termination condition is not met, repeat steps 1–4. But many details require further discussion. l

Remark (Direction determination (step 2))**.** *Our goal is to determine* $(\triangle x, \triangle w, \triangle y, \triangle z)$ *such that* $(x + \triangle x, w + \triangle w, y + \triangle y, z + \triangle z)$ *lies in the neighborhood of point* $(x_\mu, w_\mu, y_\mu, z_\mu)$ *(point at the central path at this* μ*).*

Insert $(x + \triangle x, w + \triangle w, y + \triangle y, z + \triangle z)$ *into the central path system.*

$$A(x + \triangle x) + (w + \triangle w) = b \tag{6}$$
$$A^t(y + \triangle y) - (z + \triangle z) = c \tag{7}$$
$$(X + \triangle X)(Z + \triangle Z)e = \mu e \tag{8}$$
$$(Y + \triangle Y)(W + \triangle W)e = \mu e$$

where the fourth equation belongs to 8 and unknown variables are $\triangle x, \triangle y, \triangle X, \triangle Y, \triangle w, \triangle z, \triangle W$ *and* $\triangle Z$*. Equations 6 and 7 are linear equation systems.*

Recall that,

$$\triangle X = \begin{pmatrix} \triangle x_1 & & 0 \\ & \ddots & \\ 0 & & \triangle x_n \end{pmatrix}$$
$$(x_j + \triangle x_j)(z_j + \triangle z_j) = \mu \forall j$$
$$(y_i + \triangle y_i)(w_i + \triangle w_i) = \mu \forall i$$
$$(x_j + \triangle x_j)(z_j + \triangle z_j) = x_j z_j + z_j \triangle x_j + x_j \triangle z_j + \triangle x_j \triangle z_j$$

This is a non-linear term. To handle the non-linearity of the third expression, we apply the Newton's method.

The generic problem statement is:

Given: *function* $F : \mathbb{R}^n \to \mathbb{R}^n$
Find: $\xi^* \in \mathbb{R}^n$ *with* $F(\xi^*) = 0$

$$F(\xi) = \begin{pmatrix} F_1(\xi) \\ \vdots \\ F_N(\xi) \end{pmatrix} \qquad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_N \end{pmatrix}$$

The application looks as follows: Starting with $\xi \in \mathbb{R}^N$ *with* $F(\xi) \neq 0$*, we determine* $\triangle \xi$ *and* $\tilde{\xi} = \xi + \triangle \xi$ *substitutes* ξ*.*

In the perfect setting, we have $F(\xi + \triangle \xi) = 0$ *but for non-linear functions this is is a non-realistic goal in one step. Taylor approximation of* F *up to terms of order* 1*.*

$$F(\xi + \triangle \xi) \approx F(\xi) + F'(\xi) \triangle$$

Solve $F'(\xi) \triangle \xi = -F(\xi)$ *and so on for the next point* $\xi + \triangle \xi$ *etc.*

$$F'(\xi) = \begin{bmatrix} \frac{\partial F_1}{\partial \xi_1} & \frac{\partial F_1}{\partial \xi_2} & \cdots & \frac{\partial F_1}{\partial \xi_N} \\ \frac{\partial F_2}{\partial \xi_1} & \cdots & \cdots & \vdots \\ \vdots & & \ddots & \\ \frac{\partial F_N}{\partial \xi_1} & \frac{\partial F_N}{\partial \xi_2} & \cdots & \frac{\partial F_N}{\partial \xi_N} \end{bmatrix}$$

In our case:

$$\xi = \begin{pmatrix} x \\ y \\ w \\ z \end{pmatrix} \qquad F(\xi) = \begin{pmatrix} Ax + w - b \\ A^t y - z - c \\ XZe - \mu e \\ YWe - \mu e \end{pmatrix}$$

I is the corresponding unit matrix. 0 is the corresponding zero matrix.

$$F'(\xi) = \begin{pmatrix} A & 0 & I & 0 \\ 0 & A^t & 0 & -I \\ Z & 0 & 0 & X \\ 0 & W & Y & 0 \end{pmatrix}$$

in block matrix notation.

$$\triangle\xi = \begin{pmatrix} \triangle x \\ \triangle y \\ \triangle w \\ \triangle z \end{pmatrix}$$

As solution for $\triangle\xi$, we get the solution of

$$\begin{aligned} A\triangle x + \triangle w &= \rho \\ A^t \triangle y - \triangle z &= \sigma \\ Z\triangle x + X\triangle z &= \mu e - XZe \\ W\triangle y + Y\triangle w &= \mu e - YWe \end{aligned}$$

$$A\triangle x + \triangle w = b - Ax - w =: \rho \qquad \textit{primal residue}$$

$$A^t\triangle y - \triangle z = c - A^t y + z =: \sigma \quad \textit{dual residue} Z\triangle x + X\triangle z + \triangle X\triangle Ze = \mu e - XZe$$

$$W\triangle y + Y\triangle w + \triangle Y\triangle We = \mu e - YWe$$

We want to apply linearization to the third and fourth equation. The omission of non-linear expressions and solving the linear equation system corresponds to the Newton step.

Remark (About the choice of μ)**.** *If μ is chosen too large, then the danger is convergence to the analytical center of admissible sets, which is undesirable. If μ is chosen too small, the deviation from the central path can become too large (thus we are pushed to non-optimal marginal solutions).*

At the beginning, in general (x, w, y, z) does not lie on the central path. There are various possibilities for the choice of μ. For example, we could compute x_j, z_j for some fixed j and w_i, y_i for some fixed i. Or

$$\mu = \frac{x^t z + y^t w}{n + m} \qquad \text{“Average value”}$$

which gives the exact value if (X, W, Y, Z) is on the central path. Also

$$\mu = \delta \cdot \frac{x^t z + y^t w}{n + m} \qquad \delta \in (0, 1)$$

In particular, $\delta = \frac{1}{10}$ was established as practical parameter.

↓ *This lecture took place on 2019/04/29.*

Revision.

$$\max c^t x$$
$$s.t.\ \ Ax + w = b; x, w \geq 0$$

$$\min b^t y$$
$$s.t.\ \ A^t y - z = c; y, z \geq 0$$

$$\begin{aligned} Ax + w &= b \\ A^t y - z &= c \\ XZe &= \mu e \\ YWe &= \mu e \\ X, W, Y, Z &\geq 0 \end{aligned}$$

$$X_j Z_j = \mu \forall j \qquad Y_i W_i = \mu \forall i$$

$$x, w, y, z \to (x + \theta\triangle x, w + \theta\triangle w, y + \theta\triangle y, z + \theta\triangle z)$$

all components should stay positive (hence admissible and not on the boundary).

Choice of μ:

$$\mu = \triangle \frac{x^t z + y^t w}{n + m} \qquad \triangle \in (0, 1),\ \ e.g.\ \triangle = \frac{1}{10}$$

By the choice of θ (step size),

$$\begin{aligned} x_j + \theta\triangle x_j > 0 &\to \frac{1}{\theta} > -\frac{\triangle x_j}{x_j} \forall j \\ z_j + \theta\triangle z_j &> 0 \\ w_i + \theta\triangle w_i &> 0 \\ y_i + \theta\triangle y_i &> 0 \end{aligned}$$

$$\frac{1}{\theta^*} = \max_{i,j} \left\{ -\frac{\triangle x_j}{x_j}, -\frac{\triangle z_j}{z_j}, \frac{\triangle w_i}{w_i}, \frac{\triangle y_i}{y_i} \right\}$$

To avoid contact with boundary, we choose θ such that

$$\theta = \min \left\{ r \left\{ \max_{i,j} \left\{ -\frac{\triangle x_j}{x_j}, -\frac{\triangle z_j}{z_j}, \frac{\triangle w_i}{w_i}, \frac{\triangle y_i}{y_i} \right\} \right\}^{-1}, 1 \right\}$$

Thus $r < 1$, but close to 1.

Revision (Linear Newton system). *$\triangle x, \triangle y, \triangle z, \triangle w$ as solution of*

$$\begin{aligned} A\triangle x + \triangle w &= \triangle \\ A^t \triangle y - \triangle z &= \sigma \\ Z\triangle x + X\triangle z &= \mu e - XZe \\ W\triangle y + Y\triangle w &= \mu e - ZWe \end{aligned}$$

with

$$\begin{aligned} \rho &= b - Ax - w \\ \sigma &= c - A^t y + z \end{aligned}$$

The algorithm is completely specified. The analysis remains to be done.

Measuring progress:

1. Measuring primal admissibility
2. Measuring dual admissibility
3. Performance (size of duality gap) and accordingly complement notion

Use $\rho = b - Ax - w$, $\rho = c - A^t y + z$ and $\gamma = z^t x + y^t w$.

Progress in one iteration

1. Measuring primal admissibility

$$\hat{g} = b - A\hat{x} - \hat{w} = \underbrace{b - Ax - w}_{\rho} - \theta(A\triangle x + \triangle w)$$

   where $A\triangle x + \triangle w = \rho$ according to the Linear Newton system. Hence,

$$\hat{g} = b - A\hat{x} - \hat{w} = (1-\theta)\rho$$

2. Analogously, we get,

$$\hat{\sigma} = c - A^t \hat{y} + \hat{z} = \cdots = (1-\theta)\sigma$$

3.

$$\begin{aligned} \hat{y} &= \hat{z}^t + \hat{x} + \hat{y}^t \hat{w} \\ &= (z + \theta\triangle z)^t (x + \theta\triangle x) + (y + \theta\triangle y)^t + (w + \theta\triangle w) \\ &= \underbrace{z^t x + y^t w}_{\gamma} + \theta(z^t\triangle x + \triangle z^t x + y^t \triangle w + \triangle y^t w) + \theta^2(\triangle z^t \triangle x + \triangle y^t \triangle w) \end{aligned} \tag{9}$$

where $\triangle z^t \triangle x + \triangle y^t \triangle w = (A^t \triangle y - \sigma)^t \triangle x + \triangle y^t (\rho - A\triangle x) = \triangle y^t \rho - \sigma^t \triangle x$. By the Linear Newton system, we have,

$$\begin{aligned} z^t \triangle x + \triangle z^t x &= e^t \left(Z\triangle x + X\triangle z\right) \\ &= e^t \left(\mu e - ZXe\right) = \mu n - z^t x \end{aligned}$$

Analogously,

$$y^t \triangle w + \triangle y^t w = e^t \left(Y\triangle w + W\triangle y\right) + e^t \left(\mu e - YWe\right) = \mu m - y^t w$$

$$\begin{aligned} \triangle z^t \triangle x + \triangle y^t \triangle w &= (A^t \triangle y - \sigma)^t \triangle x + \triangle y^t (\rho - A\triangle x) \\ &= \triangle y^t \rho - \sigma^t \triangle x \end{aligned}$$

Thus, by (9),

$$\hat{y} = \underbrace{z^t x + y^t w}_{\gamma} + \theta \underbrace{\Big(\mu(n+m) - (z^t x + y^t w)\Big)}_{\delta\gamma} + \theta^2 \left(\triangle y^t \rho - \sigma^t \triangle x\right)$$

So, $\hat{\gamma} = (1 - (1-\delta)\theta)\gamma + \theta^2 \left(\triangle y^t \rho - \sigma^t \triangle x\right)$. We continue using estimates. The following special case of Hölder's inequality is helpful:

$$\left|v^t u\right| = \left|\sum_j v_j u_j\right| \leq \sum_j \left|v_j\right| \left|u_j\right| \leq \left(\max_j \left|v_j\right|\right) \cdot \sum_j \left|u_j\right| = \|v\|_\infty \cdot \|u\|_1$$

In our case, we get:

$$\begin{aligned} \left|\triangle y^t \rho\right| &\leq \left\|\rho\right\|_1 \left\|\triangle y\right\|_\infty \\ \left|\sigma^t \triangle x\right| &\leq \|\sigma\|_1 \cdot \|\triangle x\|_\infty \end{aligned}$$

Hence,

$$\hat{y} \leq (1 - (1-\delta)\theta)\gamma + \theta \Big(\left\|\rho\right\|_1 \cdot \left\|\theta\triangle y\right\|_\infty + \|\sigma\|_1 \cdot \|\theta\triangle x\|_\infty\Big)$$

By choice of θ,

$$\theta \leq \frac{x_j}{\left|\triangle x_j\right|} \text{ for all } j$$

$$\rightarrow \|\theta\triangle x\|_\infty \leq \|x\|_\infty$$

Analogously, $\left\|\theta\triangle g\right\|_\infty \leq \left\|y\right\|_\infty$. Assume $\|x\|_\infty$ and $\left\|y\right\|_\infty$ are bounded by above, then $\exists M \in \mathbb{R} : \|x\|_\infty \leq M$ and $\left\|y\right\|_\infty \leq M$ (M can be very large). Thus,

$$\hat{y} \leq (1 - (1-\delta)\theta)\gamma + \theta(M\left\|\rho\right\|_1 + M \cdot \|\sigma\|_1)$$

About the termination condition

Let $\varepsilon > 0$ be a small tolerance parameter and $0 < M < \infty$ a large tolerance parameter.

- If $\|x\|_\infty > M$ is true at some point, then STOP with error "(P) is unbounded" (M must be chosen sufficiently large!)
- If $\|y\|_\infty > M$ is true at some point, then STOP with error "(D) is unbounded"
- If $\|\rho\|_1 < \varepsilon$ (sufficiently primaly admissible), $\|\sigma\|_1 < \varepsilon$ (sufficiently dually admissible) and $\gamma < \varepsilon$ (suffienciently optimal), then STOP and return current x (i.e. (x,z)) and current y (i.e. (y,w)) as optimal solution of (P) and accordingly (D).

Relation of γ and the performance of the duality gap

$$\begin{aligned}
\gamma &= z^t x + y^t w = (\sigma + A^t y - c)^t x + y^t (b - Ax - \rho) \\
&= \underbrace{b^t y - c^t x}_{\text{duality gap}} - \sigma^t x - \rho^t y \\
\left|b^t y - c^t x\right| &\leq \gamma + \left|\sigma^t x\right| + \left|y^t \rho\right| \\
&\leq \gamma + \|\rho\|_1 \cdot \|x\|_\infty + \|\rho\|_1 \|y\|_\infty
\end{aligned}$$

So, if $\gamma, \|\sigma\|_1, \|\rho\|_1$ are sufficiently small (and $\|x\|_\infty$ and $\|y\|_\infty$ are sufficiently large), then the duality gap is sufficiently small.

You should not expect that the duality gap becomes sufficiently small before x and y are almost admissible ($\|\sigma_1\|$ and $\|\rho\|_1$ sufficiently small) (this is also confirmed in numerical experiments).

Analysis of progress over multiple iterations

Denote $\rho^{(k)}, \sigma^{(k)}, x^{(k)}, y^{(k)}, z^{(k)}, w^{(k)}, \gamma^{(k)}, \theta^{(k)}$, et cetera denote the corresponding value in iteration k (iteration 0 defines the initial values).

Theorem 1.5.4. *Assume $\exists t \in \mathbb{R}, t > 0, M \in \mathbb{R}, M < \infty$ and $K \in \mathbb{N}$ such that*

$$\theta^{(k)} \geq t \qquad \left\|x^{(k)}\right\| \leq M \qquad \left\|y^{(k)}\right\| \leq M$$

is true for all $k \leq K$. Then $\exists \hat{M} < \infty$ such that

$$\begin{aligned}
\left\|\rho^{(k)}\right\|_1 &\leq (1-t)^k \left\|\rho^{(0)}\right\|_1 \\
\left\|\sigma^{(k)}\right\|_1 &\leq (1-t)^k \left\|\sigma^{(0)}\right\|_1 \\
\gamma^{(k)} &\leq (1-\tilde{t})^k \cdot \hat{M}
\end{aligned}$$

for all $k \in K$ with $\tilde{t} := (1-\delta)$.

Proof. By analysis of the progress in one iteration, we get:

$$\left\|g^{(k)}\right\|_1 \leq (1-t)\cdot\left\|g^{(k-1)}\right\|_1 \leq \cdots \leq (1-t)^k\cdot\left\|\rho^{(0)}\right\|_1$$
$$\left\|\sigma^{(k)}\right\|_1 \leq (1-t)\left\|\sigma^{(k-1)}\right\|_1 \leq \cdots \leq (1-t)^k\left\|\sigma^{(0)}\right\|_1$$

More difficult for $\gamma(k)$, we get

$$\begin{aligned}
\gamma(k) &\leq (1-t(1-\delta))\gamma^{(k-1)} + M(1-t)^{k-1}\left(\left\|\rho^{(0)}\right\|_1 + \left\|\rho^{(0)}\right\|_1\right) \\
&= (1-\tilde{t})\gamma^{(k-1)} + \tilde{M}(1-t)^{(k-1)} \text{ with } \tilde{M} = M(\left\|\rho^{(0)}\right\|_1 + \left\|\sigma^{(0)}\right\|_1) \\
&\leq (1-\tilde{t})\left((1-\tilde{t})\gamma^{(k-2)} + \tilde{M}(1-t)^{k-2}\right) \\
&= (1-\tilde{t})^2\gamma^{k-2} + \tilde{M}(1-t)^{k-1}\cdot\left(\frac{1-\tilde{t}}{1-t}+1\right) \\
&\vdots \\
&\leq (1-\tilde{t})^k\gamma^{(0)} + \tilde{M}(1-t)^{k-1}\left[\left(\frac{1-\tilde{t}}{1-t}\right)^{k-1} + \cdots + \frac{1-\tilde{t}}{1-t} + 1\right]
\end{aligned}$$

gives a geometrical sum. Hence,

$$\begin{aligned}
&= (1-\tilde{t})^k\gamma^{(0)} + \tilde{M}\cdot(1-t)^{k-1}\cdot\left(1-\left(\frac{1-\tilde{t}}{1-t}\right)^k\right)\left(1-\frac{1-\tilde{t}}{1-t}\right)^{-1} \\
&= (1-\tilde{t})^k\cdot\gamma^{(0)} + \tilde{M}\left((1-\tilde{t})^k - (1-t)^k\right)(t-\tilde{t})^{-1} \\
&\leq (1-\tilde{t})^k\cdot\gamma^{(0)} + \tilde{M}\left(1-\tilde{t}\right)^k\cdot(\delta t)^{-1} \\
&= (1-\tilde{t})^k\cdot(\gamma^{(0)} + \tilde{M}\cdot(\delta t)^{-1})
\end{aligned}$$

□

Remark (Remark about Theorem 1.5.4)**.** *Theorem 1.5.4 is a constrained convergence result, because the assumption $\theta^{(n)} \geq t \forall k \leq K$ was made such that the step size of θ is path bounded. The latter can be achieved by a modification of the algorithm and a proper choice of an initial solution. Details can be looked up in literature.*

Remark. *Consider that factor $(1-t)$ per iteration for ρ and σ and factor $(1-\tilde{t})$ for γ. Recognize that in practice, it can be shown that the dual gap converges slower than primal and dual inadmissibility.*

For practice, a lot of implementation details need to be discussed, but we won't do it in class.

By convergence analysis we get that the primal-dual path traversal method leads to a method that can determine an almost-optimal pair (x, y) in polynomial time. The termination condition is left to be discussed.

↓ *This lecture took place on 2019/04/30.*

2 Unconstrained, non-linear optimization

2.1 Basic terminology

We consider functions of $\mathbb{R}^n \to \mathbb{R}$.

Definition (minimum, maximum)**.** *Function $f : \mathbb{R}^n \to \mathbb{R}$ has a*

global minimum *in $x^* \in \mathbb{R}$ if $f(x^*) \leq f(x) \forall x \in \mathbb{R}^n$*

strict global minimum *in $x^* \in \mathbb{R}$ if $f(x^*) < f(x) \forall x \in \mathbb{R}^n \setminus \{x^*\}$*

local minimum *in $x^* \in \mathbb{R}$ if $\exists \varepsilon > 0$ such that $f(x^*) \leq f(x) \forall x \in \mathcal{U}_\varepsilon(x^*)$*

strict local minimum *if $\exists \varepsilon > 0$ such that $f(x^*) < f(x) \forall x \in U_\varepsilon(x^*) \setminus \{x^*\}$*

Analogously for maximum/maxima.

The minimum/maximum x^* satisfying the (strict/weak local/global) criteria, is also called (local/global) minimizer*.*

Remark. *Every global minimum is also a local minimum, but not vice versa.*
For convex functions, the two definitions collapse.
For simplicity, we always consider minimization problems.

Remark. *The typical goal in the field of non-linear optimization (and thus also what software on this field typically does), is determination of local optima (min or max) or potential candidates. Determination of global optima is significantly more difficult. Thus a separate field of global optimization was established.*

Example. *A typical problem in non-linear unconstrained optimzation is: given f, find local minima of f.*

$$\min_{x \in \mathbb{R}^n} f(x)$$

This is different from non-linear optimzation with *constraints: Given f and $X \subset \mathbb{R}^n$, find* $\min f(x)$ *such that $\forall x \in X$.*

One important aspect are optimality criteria (criteria for local optima). In the following, we assume that f is differentiable.

Definition. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then we call $x^* \in \mathbb{R}^n$* stationary point *of f if $\nabla f(x^*) = 0$.*

Theorem 2.1.1. *Let $f : X \to \mathbb{R}$ be a continously differentiable function and let $X \subseteq \mathbb{R}^k$ be an open set. If $x^* \in X$ is a local minimum of f (on X), then $\nabla f(x^*) = 0$, thus x^* is a stationary point on f.*

Proof. Suppose x^* is a local minimum of f with $\nabla f(x^*) \neq 0$. Then there exists $d \in \mathbb{R}^n$ with $(\nabla f(x^*))^t d < 0$ (e.g. choose $d = -\nabla f(x^*) \neq 0$). Because f is continuously differentiable by assumption, the directional derivative of f in x^* in direction d satisfies:

$$f'(x^*, d) = \lim_{t \to 0} \frac{f(x^* + td) - f(x^*)}{t} = (\nabla f(x^*))^t d < 0$$

$$\exists \bar{t} > 0 : x^* + td \in X \text{ and } \frac{f(x^* + td) - f(x^*)}{t} < 0 \forall t \in (0, \bar{t}]$$

$$f(x^* + td) < f(x^*) \forall t \in (0, \bar{t}]$$

This is a contradiction to x^* as a local minimum. □

Remark. *1. The approach within the proof and the relevance of the directional derivative will reoccur in our algorithmic considerations. Keyword "line search algorithms".*

2. The condition $\nabla f(x^*) = 0$ is not sufficient for existence of a local minimum (e.g. such a x^* can also be local maximum of f or neither nor).

Theorem 2.1.2. *Let $X \subseteq \mathbb{R}^n$ be an open set and let $f : X \to \mathbb{R}$ be two times differentiable. If x^* is a local minimum of f (on X), then the Hessian matrix $\nabla^2 f(x^*)$ of f in x^* is positive definite.*

This is a necessary condition of second order.

Proof. Assume x^* is a local minimum, but $\nabla^2 f(x^*)$ is not positive semi-definite, thus $\exists d \in \mathbb{R}^n$ with $d^t \nabla^2 f(x^*) d < 0$.

By Theorem 2.1.1 ($\to \nabla f(x^*) = 0$) together with Taylor's theorem, we get that for all sufficiently small $t > 0$,

$$f(x^* + td) = f(x^*) + \frac{1}{2} t^2 d^t \nabla^2 f(\xi_t) d \text{ where } \xi_t = x^* + \vartheta_t t d \text{ for some } \vartheta_t \in (0, 1)$$

Using continuity, we can conclude that $\exists \bar{t} > 0$ such that $f(x^* + td) < f(x^*) \forall t \in (0, \bar{t}]$. This is a contradiction to x^* being a local minimum. □

Remark. *Also these two criteria are insufficient for the existence of a local minimum.*

Example. $n = 2$. $f(x) = x_1^2 - x_2^4$, $x^* = (0, 0)$.

Theorem 2.1.3. *Let $X \subseteq \mathbb{R}^n$ be open and let $f : X \to \mathbb{R}$ be two times differentiable. If*

1. *$\nabla f(x^*) = 0$ and*
2. *$\nabla^2 f(x^*)$ positive definite*

then x^* is a strict local minimum of f (on X).

This is a sufficient optimality criterion of second order.

Proof sketch. It can be shown that the second condition yields,

$$\exists \mu > 0 : d^t \nabla^2 f(x^*) d \geq \mu d^t d \forall d \in \mathbb{R}^n$$

By Taylor's theorem, $\forall d \in \mathbb{R}^n$, that are sufficiently close to the zero-vector, that

$$f(x^* + td) = f(x^*) + \nabla f(x^*)^t d + \frac{1}{2} d^t \nabla^2 f(\xi_d) d$$

with $\xi_d = x^* + \vartheta_d$ and $\vartheta_d \in (0,1)$. By the first condition and the Cauchy-Schwarz inequality, we get

$$\begin{aligned} f(x^* + d) &= f(x^*) + \frac{1}{2} d^t \nabla^2 f(x^*) d + \frac{1}{2} d^t \left(\nabla^2 f(\xi_d) - \nabla^2 f(x^*) \right) d \\ &\geq f(x^*) + \frac{1}{2} \left(\mu - \left\| \nabla^2 f(\xi_d) - \nabla^2 f(x^*) \right\| \right) \|d\|^2 \\ &\implies f(x^* + d) > f(x^*) \forall d \neq 0 \text{ sufficiently close to } 0 \\ &\implies x^* \text{ is a strict local minimum} \end{aligned}$$

□

Remark. *The criterion above is not sufficient.*

Example. $n = 2$, $f(x) = x_1^2 + x_2^4$, $x^* = (0,0)$

2.2 Convex functions

This is a brief, introductory section.

Definition. $X \subseteq \mathbb{R}^n$ *is called* convex *if* $\forall x, y \in X \forall \lambda \in [0,1] : \lambda x + (1-\lambda) y \in X$

Definition. *Let* $X \subseteq \mathbb{R}^n$ *be convex. A function* $f : X \to \mathbb{R}$ *is called*

convex on X *if* $\forall x, y \in X \forall \lambda \in [0,1] : f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$

strictly convex on X *if* $\forall x, y \in X \forall \lambda \in (0,1) : x \neq y \implies f(\lambda x + (1-\lambda) y) < \lambda f(x) + (1-\lambda) f(y)$

uniformly convex on X *if* $\exists \mu > 0 : f(\lambda x + (1-\lambda) y) + \mu \lambda (1-\lambda) \left\| x - y \right\|^2 \leq \lambda f(x) + (1-\lambda) f(y) \forall x, y \in X \forall \lambda \in (0,1)$. μ *is also called* modulo.

In case of a convex function, no point of a segment between two points $(x, f(x)), (y, f(y)) \in \mathbb{R}^{n+1}$ lies below the graph of f.

Remark. *Every strict convex function is convex. Every uniform convex function is strict convex. The other directions do not hold in general.*

Example.

- $f(x) = x$ *is convex*
- $f(x) = e^x$ *is strictly convex, but not uniform*
- $f(x) = x^2$ *is uniformly convex*

On the opposite side, x^4 *is strictly convex, but not uniformly convex. For the special case of quadratic functions, the following result can be shown easily:*

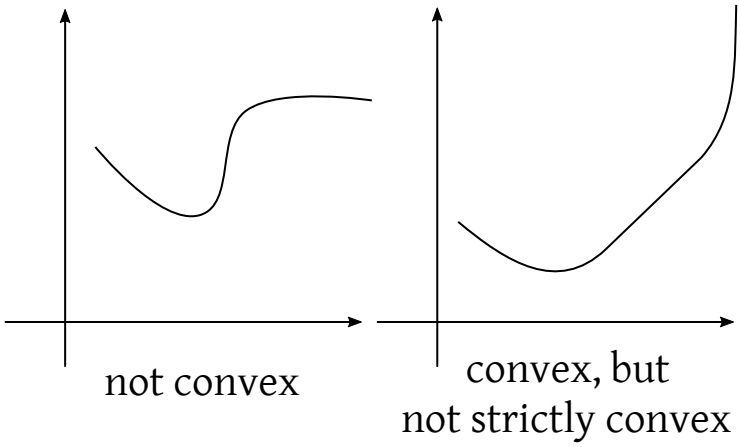

Figure 4: Convexity versus strict convexity

Remark. *Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) = \frac{1}{2}x^tQx + c^tx + \gamma$ with Q as symmetric $n \times n$ matrix over $\mathbb{R}$, $c \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$. Then,*

1. *f is convex iff Q is positive semidefinite*
2. *f is strictly convex iff f is uniformly convex iff Q is positive definite*

$\downarrow$ *This lecture took place on 2019/05/06.*

Our next goal are characterizations (of first or second order) for convexity, strict convexity and uniform convexity.

Theorem 2.2.1. *Let $X \subseteq \mathbb{R}^n$ open and convex and $f : X \to \mathbb{R}$ is continuously differentiable. Then f is ... (on X)*

convex *iff* $f(x) - f(y) \geq (\nabla f(y))^t(x-y) \forall x, y \in X$

strictly convex *iff* $f(x) - f(y) > (\nabla f(y))^t(x-y) \forall x, y \in X$

uniformly convex *iff* $f(x) - f(y) \geq (\nabla f(y))^t(x-y) + \mu \left\|x-y\right\|^2 \forall x, y \in X$

Proof. **(1) and (2)** are left as an exercise for the reader

(3) Let f be uniformly convex. $\forall x, y \in X \forall \lambda \in (0,1)$ for some $\mu > 0$,

$$f(y + \lambda(x-y)) \leq \lambda f(x) + (1-\lambda)f(y) - \mu\lambda(1-\lambda)\left\|x-y\right\|^2$$

$$\frac{f(y+\lambda(x-y))-f(y)}{\lambda} \leq f(x)-f(y)-\mu(1-\lambda)\left\|x-y\right\|^2$$

For $\lambda \to 0^+$, we get (by continuous differentiability of f)

$$(\nabla f(y))^t(x-y) = \lim_{\lambda\to 0^+} \frac{f(y+\lambda(x-y))-f(y)}{\lambda} \leq f(x)-f(y)-\mu\left\|x-y\right\|^2$$

Thus, the third property follows.

The rest of the proof is left as an exercise to the reader. □

Definition. *Let $X \subseteq \mathbb{R}^n$. Consider $F : X \to \mathbb{R}^n$. Then F is called*

1. *monotonous (on X) if $(x-y)^t(F(x)-F(y)) \geq 0 \forall x \neq y \in X$*
2. *strictly monotonous (on X) if $(x-y)^t(F(x)-F(y)) > 0 \forall x \neq y \in X$*
3. *uniformly mon. (on X) if $(x-y)^t(F(x)-F(y)) \geq \mu\left\|x-y\right\|^2 \forall x \neq y \in X$*

Remark. *For $n = 1$, this corresponds to scalar functions and the usual notion of monotonicity.*

Theorem 2.2.2. *Let $X \subseteq \mathbb{R}^n$ be open and convex and let $f : X \to \mathbb{R}$ be continuously differentiable. Then*

1. *f is convex (on X) iff ∇f is monotonous*
2. *f is strictly convex (on X) iff ∇f is strictly monotonous*
3. *f is uniformly convex (on X) iff ∇f is uniformly monotonous*

Theorem 2.2.3. *Let $X \subseteq \mathbb{R}^n$ be open and convex and let $f : X \to \mathbb{R}$ be two times differentiable ($\to \nabla f$ is continuously differentiable). Then*

1. *f is convex (on X) iff $\nabla^2 f(x)$ is positive semidefinite $\forall x \in X$*
2. *f is strictly convex (on X)* if *$\nabla^2 f(x)$ is positive definite $\forall x \in X$*
3. *f is uniformly convex (on X) iff $\nabla^2 f(x)$ is uniformly positive definite on X, thus*
$$\exists \mu > 0 : d^t \nabla^2 f(x) d \geq \mu \left\|d\right\|^2 \forall x \in X \forall d \in \mathbb{R}^n$$

Proof. Properties (1) and (2) are left as an exercise to the reader. We prove (3):

Direction $\Longrightarrow$ Only property 3 is new to us. Let f be uniformly convex. Use Theorem 2.2.2 (3). So ∇f is uniformly monotonous. By appropriately chosen $\mu > 0$, we get

$$\begin{aligned} d^t \nabla^2 f(x) d &= d^t \lim_{t\to 0} \frac{\nabla f(x+td)-\nabla f(x)}{t} = \lim_{t\to 0} \frac{td^t(\nabla f(x+td)-\nabla f(x))}{t^2} \\ &\geq \lim_{t\to 0} \frac{1}{t^2}\mu \left\|td\right\|^2 = \mu \left\|d\right\|^2 \qquad \forall x \in X, d \in \mathbb{R}^n \end{aligned}$$

So, $\nabla^2 f(x)$ is uniformly positive definite.

Direction $\Longleftarrow$ Let the third property be true. By the mean value theorem of differential calculus (in integral form), we get

$$\begin{aligned}(x-y)^t(\nabla f(x)-\nabla f(y)) &= \int_0^1 (x-y)^t \nabla^2 f(y+\alpha(x-y))(x-y)\,d\alpha \\ &\geq \mu \int_0^1 \left\|x-y\right\|^2 \,d\alpha = \mu \left\|x-y\right\|^2 \\ &\Longrightarrow \nabla f \text{ is uniformly monotonous on } X \\ &\xRightarrow{\text{Theorem 2.2.2 (3)}} f \text{ is uniformly convex on } X\end{aligned}$$

□

Remark. *To verify that* $\nabla^2 f(x)$ *is*

positive semidefinite *use* $\forall \lambda \in \operatorname{spec}(\nabla^2 f(x)) : \lambda \geq 0$

positive definite *use* $\forall \lambda \in \operatorname{spec}(\nabla^2 f(x)) : \lambda > 0$

uniformly positive definite *use* $\forall \lambda \in \operatorname{spec}(\nabla^2 f(x) - \mu I) : \lambda \geq 0$

Thus all eigenvalues of $\nabla^2 f(x)$ *are greater-equal to* μ*. Recognize that constant* μ *is independent of* x*. Thus the eigenvalue of* 0 *is negatively bounded.*

Remark (Remark about (2))**.** $f(x) = x^4$ *is strictly convex but* $\nabla^2 f(x) = f''(x) = 12x^2$*.* $f''(0) = 0$*, so* f *is only positive semidefinite. There is only one direction in (2).*

Our goal is to establish a result about the existence of a minimum.

Definition. *Let* $f : \mathbb{R}^n \to \mathbb{R}$*. A* level set *(dt. "Niveaumenge") is defined by*

$$\mathcal{L}(\tilde{x}) = \{x \in \mathbb{R}^n \mid f(x) \leq f(\tilde{x})\}$$

Lemma 2.2.4. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable. Let* $\tilde{x} \in \mathbb{R}^n$ *with* $\mathcal{L}(\tilde{x})$ *be convex and let* f *be uniformly convex on* $\mathcal{L}(\tilde{x})$*. Then* $\mathcal{L}(\tilde{x})$ *is compact.*

Proof. $\mathcal{L}(\tilde{x}) \neq \emptyset \implies \exists x \in \mathcal{L}(\tilde{x})$. Because f is uniformly convex on $\mathcal{L}(\tilde{x})$, we get (by $X = \frac{1}{2}$ and appropriate $\mu > 0$)

$$\begin{aligned}\frac{1}{4}\mu \left\|x-\tilde{x}\right\|^2 &\leq \frac{1}{2}\left(\frac{1}{2}f(x) - f(\tilde{x})\right) - \left(f(\frac{1}{2}(x+\tilde{x})) - f(\tilde{x})\right) \\ &\leq -\left(f(\frac{1}{2}(x+\tilde{x})) - f(\tilde{x})\right) \\ &\leq -\frac{1}{2}\nabla f(\tilde{x})^t (x-\tilde{x}) \\ &\leq \frac{1}{2}\left\|\nabla f(\tilde{x})\right\| \left\|x-\tilde{x}\right\|\end{aligned}$$

Thus $\left\|x-\tilde{x}\right\| \leq c$ with $c = \frac{2\left\|\nabla f(\tilde{x})\right\|}{\mu}$ (a constant independent of x) for all $x \in \mathcal{L}(\tilde{x})$.

Thus $\mathcal{L}(\tilde{x})$ is bounded. Because f is continuous, $\mathcal{L}(\tilde{x})$ is closed. So $\mathcal{L}(\tilde{x})$ is compact. □

Remark. *Assume f to be uniformly convex on entire $\mathbb{R}^n$ or only on a convex set X, that contains $\mathcal{L}(\tilde{x})$. Then the convexity of $\mathcal{L}(\tilde{x})$ follows immediately and we can omit the explicit precondition $\mathcal{L}(\tilde{x})$.*

Theorem 2.2.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $X \subseteq \mathbb{R}^n$ be convex. We consider the optimization problem (P) $\min_{x \in X} f(x)$. Then*

1. *If f is convex on X, then the solution set of (P) is convex (potentially empty)*
2. *If f is strictly convex on X, then (P) has at most one solution.*
3. *If f is uniformly convex on X and $X \neq \emptyset$ and X is closed (e.g. for $X = \mathbb{R}^n$), then (P) has exactly* one *solution*

Remark. *The third property illustrates the particular significance of the class of convex, strictly convex and uniformly convex functions in non-linear optimization.*

Proof.

1. Let $\bar{x}$ and $\bar{\bar{x}}$ be solutions of (P) with $\bar{x}, \bar{\bar{x}} \in X$.

$$f(\bar{x}) = f(\bar{\bar{x}}) = \min_{x \in X} f(x)$$

For $\lambda \in (0, 1)$, we have $\lambda \bar{x} + (1 - \lambda)\bar{\bar{x}} \in X$, because X is convex.

Because f is convex,

$$\begin{aligned} f(\lambda \bar{x} + (1 - \lambda)\bar{\bar{x}}) &\leq \lambda f(\bar{x}) + (1 - \lambda) f(\bar{\bar{x}}) \\ &= f(\bar{x}) = \min_{x \in X} f(x) \\ &\Longrightarrow \quad \text{also } \lambda \bar{x} + (1 - \lambda)\bar{\bar{x}} \text{ is solution of } (P) \end{aligned}$$

2. Assume there are 2 different solutions $\bar{x}$ and $\bar{\bar{x}}$ with $\bar{x} \neq \bar{\bar{x}}$. For $\lambda \in (0, 1)$,

$$f(\lambda \bar{x} + (1 - \lambda)\bar{\bar{x}}) < \lambda f(\bar{x}) + (1 - \lambda) f(\bar{\bar{x}}) = f(\bar{x}) = \min_{x \in X} f(x)$$

as f is strictly convex. This is a contradiction to $\bar{x}$ and $\bar{\bar{x}}$ as solutions of (P).

3. Let $\tilde{x} \in X$ be arbitrary. By Lemma 2.2.4, $\mathcal{L}(\tilde{x})$ is compact. Thus $X \cap \mathcal{L}(\tilde{x})$ with $X \neq \emptyset$ and $\mathcal{L}(\tilde{x})$ is compact and non-empty.

   Hence, the continuous function f has a global minimum over $X \cap \mathcal{L}(\tilde{x})$. It has to be a solution of (P).

□

Remark. *Consider $f(x) = e^x$ over $X = \mathbb{R}$. f in strictly convex, but the solution set of (P) is empty.*

Strict convexity does not generally suffice to ensure the existence of (P). Remember that e^x is not uniformly convex.

- *If f is uniformly convex and $X = \emptyset$, then there obviously does not exist a solution of (P)*

- *If f is uniformly convex and $X \neq \emptyset$, but X is not closed, (P) might not have a solution.*

$$f(x) = x^2 \qquad X = (0,1] \qquad \nexists \text{ solution}$$

Lemma 2.2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\tilde{x} \in \mathbb{R}^k$ with $\mathcal{L}(\tilde{x})$ convex. f is uniformly convex on $\mathcal{L}(\tilde{x})$ and let $x^* \in \mathbb{R}^n$ be the unique solution of $\min_{x \in \mathbb{R}^n} f(x)$. Then there exists $\mu > 0$ such that*

$$\mu \|x - x^*\|^2 \subseteq f(x) - f(x^*) \qquad \forall x \in \mathcal{L}(\tilde{x})$$

Proof. Use Theorem 2.2.2 (3). Observation x^* as global minimum of f is a stationary point of f. □

↓ *This lecture took place on 2019/05/07.*

Theorem 2.2.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and convex. Let x^* be a stationary point of f. Then x^* is a global minimum of f (on $\mathbb{R}^n$).*

Proof. Follows by Theorem 2.2.2.

$$f(x) - f(x^*) \geq \underbrace{(\nabla f(x^*))^t}_{=0}(x - x^*) = 0$$

$$\implies f(x) \geq f(x^*) \qquad \forall x \in \mathbb{R}^n \implies x^* \text{ is global minimum}$$

□

Remark. *The result above can be generalized.*

$$f : X \to \mathbb{R} \text{ such that } (\nabla f(y))^t (x - y) \geq 0 \implies f(x) \geq f(y) \forall x, y \in X$$

defines pseudoconvex functions *and thus the class of pseudoconvex functions which is a more generic class than the class of convex functions.*

2.3 Generic descent methods

The generic setup gives a function $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable.

$$\min_{x \in \mathbb{R}^n} f(x)$$

We look for local minima (or candidates for such). Consider f as mountain range and local minima as valleys.

The abstract idea of the descent method: Begin with $x^{(0)}$. Let $k := 0$. Verify whether x^* is a local minimum and test whether there is any direction $d^{(k)}$ such that beginning from $x^{(k)}$ a smaller target function value can be achieved. If not, a local minimum has been reached. If yes, take a small step in direction $d^{(k)}$ such that $x^{(k+1)} = x^{(k)} + t_k \cdot d^{(k)}$ with $t_k > 0$ as step size.

This establishes the notion of direction of descent.

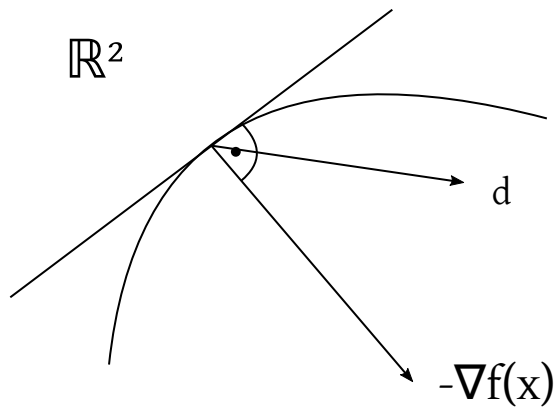

Figure 5: Geometrical interpretation of the descent

Definition 2.3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}, x \in \mathbb{R}^n$. A vector $d \in \mathbb{R}^n$ is called* direction of descent *for f in x if $\exists \hat{t} > 0$ such that $f(x + td) < f(x) \forall t \in (0, \hat{t}]$.*

However, this notion is difficult to verify algorithmically. We need an alternative approach.

Lemma 2.3.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Let $(\nabla f(x))^t d < 0$ (thus a negative direction of descent is given). Then d is the* direction of descent *of f in x.*

Proof. Because f is continuously differentiable, we consider the directional derivative of f in x:

$$f'(x, d) = \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t} = (\nabla f(x))^t d < 0$$

$\implies f(x + td) - f(x) < 0$ for all $t > 0$ sufficiently small, thus we get value $\hat{t}$. □

Remark. *Does Lemma 2.3.2 provide a necessary criterion for the existence of a direction of descent? Yes. But it is not sufficient. x can be a strict local maximum. All directions $d \neq 0$ are directions of descent, but $\nexists d$ such that $(\nabla f(x))^t d < 0$ (because x is a stationary point).*

Remark (Geometrical interpretation of $(\nabla f(x))^t d < 0$)**.**

$$\left\langle (-\nabla f(x))^t, d \right\rangle > 0 \implies \textit{the angle between } -\nabla f(x) \textit{ and } d \textit{ is } < \frac{\pi}{2}$$

Remark. *Colloquially the condition $(\nabla f(x))^t d < 0$ is used as definition of direction of descent.*

We are going to use this condition in our algorithms.

Remark (Observation)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $x \in \mathbb{R}^n$. x is not a stationary point of f. Let B be a symmetric, positive definite $n \times n$ matrix. Then $d = -B\nabla f(x)$ is a direction of descent.*

Special case: *$d = -\nabla f(x)$ is a direction of descent. In particular, the direction of* steepest descent.

Algorithm 2.3.3 (Generic descent method)**.**

1. *Choose $x^{(0)} \in \mathbb{R}^n$ and let $k := 0$*
2. *If $x^{(k)}$ suffices the chosen termination condition, then STOP.*
3. *Determine a direction of descent $d^{(k)}$ of f in $x^{(k)}$*
4. *Determine a step size $t_k > 0$ with $f(x^{(k)} + t_k d^{(k)}) < f(x^{(k)})$*
5. *Let $x^{(k+1)} := x^{(k)} + t_k d^{(k)}$. Let $k := k + 1$. Go to (1).*

Step size, direction of descent and termination condition are left to be discussed.

For analysis of convergence, we consider the infinite sequence $\left\{x^{(k)}\right\}, \left\{d^{(k)}\right\}$ and $\{t_k\}$ without satisfied termination condition. The central question for us is under which conditions of $d^{(k)}$ and t_k, convergence is achieved. With arbitrary chosen $d^{(k)}$ and t_k we don't get a useful generic approach. We need to establish some conditions on t_k and $d^{(k)}$.

Our goal: Every cluster point of sequence $\left\{x^{(k)}\right\}$ should be at least one stationary point of f.

Remark (Find appropriate step size)**.** *Given (x, d) with $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$, find step size $t > 0$.*

Consider map T, that associates a subset $T(x, d)$ of $R^{++} = \{t \in R \mid t > 0\}$ to every $(x, d) \in \mathbb{R}^n \times \mathbb{R}^n$. We call T a step size strategy *(or* Step size rule*).*

T is called well-defined *if—under some particular conditions—the set $T(x, d)$ for every pair $(x, d) \in \mathbb{R}^n \times \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$ is non-empty.*

Definition 2.3.4 (Efficient step size strategy)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ be a direction of descent of f in x. A step size strategy T is called* efficient*, if $\exists \theta > 0$ (of x and d) such that*

$$\forall t \in T(x, d) : f(x + td) \leq f(x) - \theta \left(\frac{(\nabla f(x))^t d}{\|d\|} \right)^2$$

Remark. *One motivating case for this definition is the special case of quadratic functions (compare with practicals).*

In the next subchapter, we are going to discuss specific (efficient) step size strategies.

Definition. *$t \in T(x, d)$ with efficient T is called* efficient step size.

Theorem 2.3.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let sequence $\left\{x^{(k)}\right\}$ be generated by the generic descent method. If $\left\{x^{(k)}\right\}$ is satisfies the two conditions (E1) and (E2), then every cluster point of $\left\{x^{(k)}\right\}$ is a stationary point of f.*

Condition (E1) is called "angle condition":

$$\exists c > 0 \forall k \in \mathbb{N} : \frac{-\left(\nabla f(x^{(k)})\right)^t d^{(k)}}{\left\|\nabla f(x^{(k)})\right\| \left\|d^{(k)}\right\|} \geq c$$

Condition (E2) requires step sizes $t_k > 0$ are efficient for all $k \in \mathbb{N}$.

Remark. *The direction of the steepest descent $d^{(k)} = -\nabla f(x^{(k)})$ satisfies the angle condition. More examples will be provided in the upcoming classes.*

Remark (About the angle condition E1)**.** *Let φ_k be an angle between $d^{(k)}$ and $-\nabla f(x^{(k)})$. Then*

$$\cos \varphi_k = -\frac{\left(\nabla f(x^k)\right)^t d^{(k)}}{\left\|\nabla f(x^k)\right\| \left\|d^{(k)}\right\|}$$

If this angle is smaller than $\frac{\pi}{2}$, then the direction of descent is $d^{(k)}$ by Lemma 2.3.2. The condition (E1) states that the angle between $-\nabla f(x^{(k)})$ and $d^{(k)}$ is uniformly bounded (sufficiently far away from $\frac{\pi}{2}$).

Proof of Theorem 2.3.5. For every efficient $t_k > 0$,

$$\exists \theta > 0 : f(x^{(k+1)}) > f(x^{(k)} + t_k \cdot d^{(k)}) \leq f(x^{(k)}) - \theta \left(\frac{(\nabla f(x^{(k)}))^t d^{(k)}}{\left\|d^{(k)}\right\|} \right)^2 \forall k \in \mathbb{N}$$

By E1, we get

$$(*) : \quad f(x^{(k+1)}) \leq f(x^{(k)}) - \kappa \cdot \left\|\nabla f(x^{(k)})\right\|^2 \text{ with } \kappa := \theta c^2$$

Now let x^* be a cluster point of $\left\{x^{(k)}\right\}$.

The sequence $\left\{f(x^{(k)})\right\}$ is monotonically decreasing and converges on a subsequence towards $f(x^*)$. Thus $\left\{f(x^{(k)})\right\}$ converges towards $f(x^*)$.

$$\Longrightarrow f(x^{(k+1)}) - f(x^{(k)}) \to 0 \qquad \overset{(*)}{\Longrightarrow} \left\|\nabla f(x^{(k)})\right\| \to 0$$

Thus every accumulation point of $\left\{x^{(k)}\right\}$ is a stationary point of f. □

A second convergence result requires a weakening of the angle condition. (E1) forbids sequence $\frac{\pi}{2}$ to approach $\{\varphi_k\}$. In this case, we discard this requirement, but it must not happen too fast.

Theorem 2.3.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. The level set $\mathcal{L}(x^{(0)}) = \left\{x \in \mathbb{R}^n \mid f(x) \leq f(x^{(k)})\right\}$ is convex and f is uniformly convex on $\mathcal{L}(x^{(0)})$. Let $\left\{x^{(k)}\right\}$ be the iteration sequence generated by the descent method satisfying*

$\overline{E1}$ **("Zouten dijk condition")** $\sum_{k=0}^{\infty} \delta_k = \infty$ *with*

$$\delta_k = \left(\frac{\left(\nabla f(x^{(k)})\right)^t d^{(k)}}{\left\|\nabla f(x^{(k)})\right\| \left\|d^{(k)}\right\|} \right)$$

$\overline{E2}$ *as before*

In this case, $\left\{x^{(k)}\right\}$ *converges towards the uniquely determined global minimum of* f

$\downarrow$ *This lecture took place on 2019/05/13.*

Revision 3 (Generic descent method)**.** $\left\{x^{(k)}\right\}$ *sequence of iteration points*

$\left\{d^{(k)}\right\}$ *sequence of search directions*

$\{t_k\}$ *sequence of step sizes*

Proof of Theorem 2.3.6. The global minimum must lie in $\mathcal{L}(x^{(0)})$, hence f has exactly one global minimum (Lemma 2.2.5 (3), the preconditions are satisfied here). Let x^* be this particular minimum.

Let $\mu > 0$ be the module of f (f uniformly convex). Trivially,

$$\left\| \sqrt{\frac{\mu}{2}} \left(x^* - x^{(k)}\right) + \sqrt{\frac{1}{2\mu}} \nabla f(x^{(k)}) \right\|^2 \geq 0$$

By simple transformations, we get,

$$-\frac{1}{2\mu} \left\|\nabla f(x^{(k)})\right\| \leq \frac{\mu}{2} \left\|x^* - x^{(k)}\right\|^2 + \left(\nabla f(x^{(k)})\right)^t (x^* - x^{(k)})$$

By Theorem 2.2.1 (3), we have that

$$-\frac{1}{2\mu} \left\|\nabla f(x^{(k)})\right\|^2 \leq f(x^*) - f(x^{(k)})$$

Because the step size sequence is efficient, we get,

$$\begin{aligned}
f(x^{(k+1)}) = f(x^{(k)} + t_k d^{(k)}) &\leq f(x^{(k)}) - \theta \left(\frac{\left(\nabla f(x^{(k)})\right)^t d^{(k)}}{\left\|d^{(k)}\right\|} \right)^2 \\
&= f(x^{(k)}) - \theta \left\|\nabla f(x^{(k)})\right\|^2 \delta_k \leq f(x^{(k)}) - 2\mu\theta\delta_k (f(x^{(k)}) - f(x^*)) \\
\implies 0 \leq f(x^{(k+1)}) - f(x^*) &= f(x^{(k+1)}) - f(x^{(k)}) + f(x^{(k)}) - f(x^*) \\
&\leq (1 - 2\mu\theta\delta_k) \left(f(x^{(k)}) - f(x^*)\right)
\end{aligned}$$

Iterative application gives,

$$0 \leq f(x^{(k+1)}) - f(x^*) \leq \prod_{l=1}^{k} (1 - 2\mu\theta\delta_l) \left(f(x^{(0)}) - f(x^*)\right)$$

Because $e^z \geq 1 + z \forall z \in \mathbb{R}$ (with $e^z := \exp(z)$),

$$\begin{aligned}
&\leq \prod_{l=0}^{k} \exp\left(-2\mu\theta\delta_l\right) \left(f(x^{(0)}) - f(x^*)\right) \\
&= \exp\left(-2\mu\theta \cdot \sum_{l=0}^{k} \delta_l\right) \left(f(x^{(0)}) - f(x^*)\right) \\
&\xrightarrow{k\to\infty} \infty
\end{aligned}$$

So, $\left\{f(x^{(k)})\right\} \to f(x^*)$ for $k \to \infty$. By Lemma 2.2.6 we have

$$\begin{aligned}
f(x^{(k)}) - f(x^*) \geq \mu \left\|x^{(k)} - x^*\right\|^2 \forall k \in \mathbb{N} \\
\implies \left\{x^{(k)}\right\} \to x^*
\end{aligned}$$

There are bunch of convergence results, but they always have requirements for $\{t_k\}$ and $\left\{d^{(k)}\right\}$. □

In order to implement the generic descent method, we need to specify a lot of details. Especially, we need to determine step sizes $\{t_k\}$ and also $\left\{d^{(k)}\right\}$.

2.4 Step size determination

Two fundamental approaches:

1. Exact method (exact line search)
2. Approximation methods (inexact line search)

In case of the exact method, we consider

$$\min_{\alpha>0} f(x^{(k)} + \alpha d^{(k)}) \qquad t_k := \operatorname{argmin}_{\alpha>0} f(x^{(k)} + \alpha d^{(k)})$$

The disadvantage of this method is that is computationally intense to determine a solution accurately. An explicit solution is available only in a few cases (e.g. for quadratic functions, see practicals).

The application of numerics to solve one-dimensional ($\mathbb{R}^1$) optimization problems is—in general—inefficient. This is because of the increase in additional iteration steps compared to the increased resource requirements per iteration.

We are going to handle approximation methods.

In the following, we are going to omit $^{(k)}$ and we will use x, d and t.

The most famous representatives are the *Armijo rule* and *Wolfe-Powell rule*.

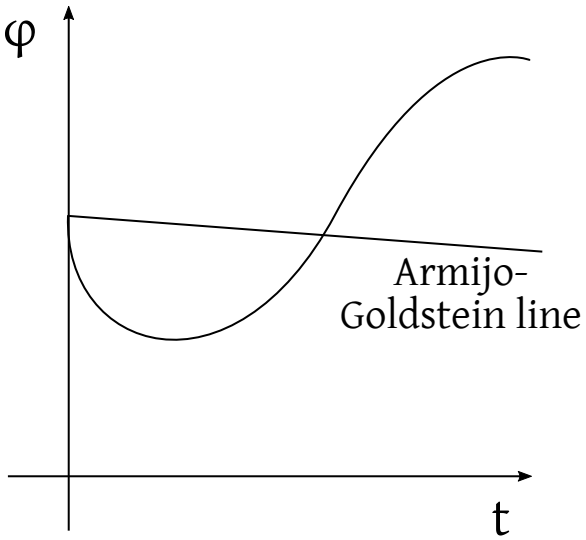

Figure 6: Armijo rule

2.4.1 Armijo rule

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\sigma, \beta \in (0,1)$ be fixed. For $x \in \mathbb{R}^n, d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$. Armijo rule determines the step size t as $\max\left\{\beta^l \mid l = 0, 1, \dots\right\}$ such that

$$f(x + td) \leq f(x) + \sigma \cdot t \left(\nabla f(x)\right)^t d \qquad (\triangle)$$

Algorithmically, we begin with $t = 1 = \beta^0$. We check whether condition $(\triangle)$ is satisfied. If so, use Armijo step size $t = 1$. Otherwise check $t = \beta \dots \beta^2 \dots$. Stop if $(\triangle)$ is satisfied for the first time.

$$\varphi(t) \coloneqq f(x + td) \qquad \varphi(0) = f(x)$$

Consider $\varphi'(t)$

$$\varphi'(0) = \left(\nabla f(x)\right)^t d$$

Condition $(\triangle)$ can be written as,

$$\varphi(t) \leq \varphi(0) + \sigma t \varphi'(0)$$

This actually gives a line, the *Armijo-Goldstein line.*

Remark. *The candidate set β^l for the Armijo step size is discrete.*

To determine the Armijo step size, the relevant region is wherever the graph of φ is below the Armijo-Goldstein line.

Theorem 2.4.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\beta, \sigma \in (0,1)$ be fixed. For $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$, there exists a finite $\tilde{l} \in \mathbb{N}$ with*

$$f(x + \beta^{\tilde{l}} d) \leq f(x) + \sigma \beta^{\tilde{l}} (\nabla f(x))^t d$$

Hence the Armijo rule is well-defined.

Proof. Assume $\forall l \in \mathbb{N}$,

$$f(x + \beta^l d) > f(x) + \sigma \beta^l \left(\nabla f(x)\right)^t d$$

$$\Longrightarrow \frac{f(x + \beta^l d) - f(x)}{\beta^l} \geq \sigma \left(\nabla f(x)\right)^t d$$

For $l \to \infty$ and f differentiable,

$$(\nabla f(x))^t d \geq \sigma (\nabla f(x))^t d$$

Because $\sigma \in (0,1)$, $(\nabla f(x))^t d \geq 0$ and we get a contradiction to the Remar $(\nabla f(x))^t d < 0$. □

Remark. *Often a more generic version of the Armijo rule is used: Let an additional scaling parameter $s > 0$ be given. Determine t as $\max\left\{s, \beta^l \mid l = 0, 1, \ldots\right\}$ such that $f(x + td) \leq f(x) + \sigma t \left(\nabla f(x)\right)^t d$. Theorem 2.4.1 can be mapped as well.*

2.4.2 Wolfe-Powell step size

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\sigma \in (0, \frac{1}{2})$ and $\rho \in [\sigma, 1)$ be fixed. For $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ with $\left(\nabla f(x)\right)^t d < 0$, determine some $t > 0$ with

$$f(x + td) \leq f(x) + \sigma t \left(\nabla f(x)\right)^t d \qquad \text{(WP1)}$$

$$\nabla f(x + td)^t d \geq \rho \left(\nabla f(x)\right)^t d \qquad \text{(WP2)}$$

For interpretation $\varphi(t) = f(x + td)$,

$$\varphi(t) \leq \varphi(0) + \sigma t \varphi'(0) \qquad \text{(WP1)}$$

$$\varphi'(t) \geq \rho \varphi'(0) \qquad \text{(WP2)}$$

The WP step size is chosen from the area where the graph of φ lies below the Armijo line and also the graph of φ is not so steeply decreasing as in point 0 (or already increases).

Remark. *Compare with Figure 7.*

- *There can also be Wolfe-Powell step sizes right of T_{WP}. This makes the proof difficult.*
- *The choice $\sigma < \frac{1}{2}$ results from the requirement that for quadratic functions as special case the exact minimum is accepted as WP step size.*

Our next goal will be to show that WP step sizes are well-defined and efficient.

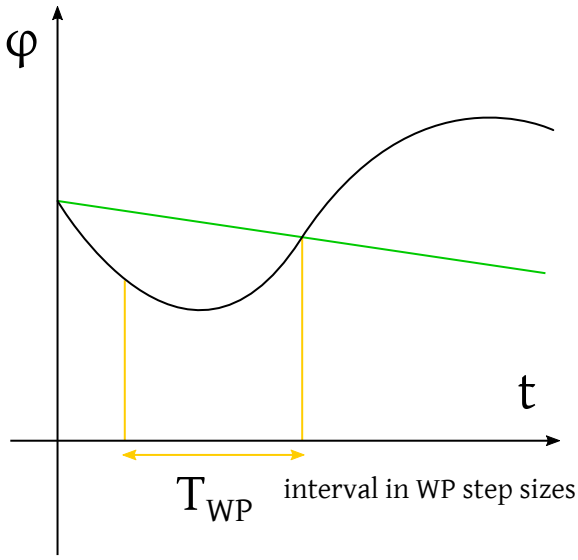

Figure 7: Wolfe-Powell step sizes

↓ *This lecture took place on 2019/05/14.*

Theorem 2.4.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\sigma \in \left(0, \frac{1}{2}\right)$ and $\rho \in (\sigma, 1)$. Let $x^{(0)} \in \mathbb{R}^n$ be fixed. For $x \in \mathcal{L}(x^{(0)})$ and $d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$, let*

$$T_{WP}(x,d) := \left\{t > 0 \mid f(x+td) \leq f(x) + \sigma t(\nabla f(x))^t d, (\nabla f(x+td))^t d \geq \rho(\nabla f(x))^t d\right\}$$

be the set of Wolfe-Powell step sizes in x in direction d. Then

- *If f is bounded by below, then $T_{WP}(x,d) \neq \emptyset$, hence the WP rule is well-defined.*
- *If ∇f is Lipschitz continuous on $\mathcal{L}(x^{(0)})$, then there exists $\theta > 0$ such that*

$$f(x+td) \leq f(x) - \theta\left(\frac{(\nabla f(x))^t d}{\|d\|}\right)^2$$

  For all $t \in T_{WP}(x,d)$, hence the WP rule is efficient, which is an important result for convergence.

Proof. 1. Compare with proof of strict WP rule.

2. Let $t \in T_{WP}(x,d)$. We know,

$$f(x+td) \leq f(x) \implies x + td \in \mathcal{L}(x^{(0)})$$

By the WP rule, we get,

$$(\rho - 1)(\nabla f(x))^t d \leq (\nabla f(x + td) - \nabla f(x))^t d$$

Now we use the Cauchy-Schwarz inequality and use Lipschitz continuity of ∇f on $\mathcal{L}(x^{(0)})$. Let L be the corresponding Lipschitz constant.

$$(\rho - 1)(\nabla f(x))^t d \leq \left\|\nabla f(x + td) - \nabla f(x)\right\| \|d\| \leq Lt \|d\|^2$$

$$\implies t \geq \frac{(\rho - 1)(\nabla f(x))^t d}{L \|d\|^2}$$

Thus,

$$\underbrace{f(x + td) \leq f(x) + \sigma t(\nabla f(x))^t d}_{\text{first WP condition}} \leq f(x) - \theta \left(\frac{(\nabla f(x))^t d}{\|d\|} \right)^2$$

with $\theta = \frac{(1-\rho)\sigma}{L}$.

□

Remark (Simple sufficient condition for Lipschitz continuity of ∇f on $\mathcal{L}(x^{(0)})$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be 2 times continuously differentiable and $x^{(0)} \in \mathbb{R}^n$. Assume one of the two following conditions is satisfied:*

(B1) $\left\|\nabla^2 f(x)\right\|$ *is bounded on a convex superset of* $\mathcal{L}(x^{(0)})$

(B2) *The level set* $\mathcal{L}(x^{(0)})$ *is compact.*

Then $\nabla f(x)$ *is Lipschitz continuous on* $\mathcal{L}(x^{(0)})$.

Proof. Left as an exercise to the reader. □

One variation of the WP rule is the *strict WP rule.*

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ be fixed. For $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$, determine some $t > 0$ with

$$f(x + td) \leq f(x) + \sigma t(\nabla f(x))^t d$$

$$\left|(\nabla f(x + td))^t d\right| \leq -\rho \cdot (\nabla f(x))^t d$$

Geometrical illustration:

$$\begin{aligned} \varphi(t) &:= f(x + td) \\ \varphi(t) &\leq \varphi(0) + \sigma t \varphi'(0) \\ \left|\varphi'(t)\right| &\leq -\rho \varphi'(0) \end{aligned}$$

Refinement of WP rule: The slope must not be too steep.

Theorem 2.4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let $\sigma \in \left(0, \frac{1}{2}\right)$, $\rho \in [\sigma, 1)$. Let $x^{(0)} \in \mathbb{R}^n$ be fixed. For $x \in \mathcal{L}(x^{(0)})$ and $d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$, let*

$$T_{SWP} := \{t > 0 \mid \textit{(SWP1) and (SWP2) are satisfied}\}$$

be the set of strict WP step sizes in x in direction of d. Then

- *If f is bounded by below, then $T_{SWP} \neq \emptyset$, hence the strict WP rule is well-defined.*
- *If ∇f is Lipschitz continuous on $\mathcal{L}(x^{(0)})$, then the strict WP rule is efficient. (This is an analogous formulation to Theorem 2.4.2 (2))*

Proof.

- This also proves Theorem 2.4.2 (1). Let

$$\varphi(t) := f(x + td)$$
$$\psi(t) := f(x) + \sigma t \left(\nabla f(x)\right)^t d$$

Show that there exists $t > 0$ such that

(SWP1) $\varphi(t) \leq \psi(t)$

(SWP2) $\left|\varphi'(t)\right| \leq -\rho \cdot \varphi'(0)$

Because $\varphi'(0) < \psi'(0)$, the graph of φ lies underneath the graph of ψ for sufficiently small $t > 0$. Let $\hat{t}$ be the smallest value $t > 0$ such that $\varphi(t) = \psi(t)$ (hence $\varphi(\hat{t}) = \psi(\hat{t})$). Recognize that $\hat{t}$ exists because $\psi(t) \to -\infty$ with $t \to \infty$ and f is bounded by below. Thus, $\varphi'(\hat{t}) \geq \psi'(\hat{t})$.

Now we distinguish between two cases:

$\varphi'(\hat{t}) < 0$**:** For $t = \hat{t}$, $\varphi(t) = \psi(t)$ and

$$\left|\varphi'(\hat{t})\right| = -\varphi'(\hat{t}) \leq -\psi'(\hat{t}) = -\sigma(\nabla f(x))^t d = -\sigma\varphi'(0) \leq -\rho\varphi'(0)$$

because $\sigma \leq \rho$. So $\hat{t} \in T_{SWP}(x, d)$.

$\varphi'(\hat{t}) \geq 0$**:** Because $\varphi'(0) < 0$, $\exists \hat{\hat{t}}$ also $\left|\varphi'(t)\right| \leq -\rho\varphi'(0)$. Thus $\hat{\hat{t}} \in T_{SWP}(x, d)$.

- We have $T_{SWP}(x, d) \subseteq T_{WP}(x, d)$. Thus the result follows by Theorem 2.4.2 (2), Theorem 2.4.2 (1) and 2.4.3 (1).

□

So far, we have discussed Armijo in a scaled variation. With certain requirements Wolfe-Powell and the strict Wolfe-Powell method are efficient. All of them are well-defined.

Armijo is not efficient; not even under requirements of Wolfe-Powell. But there exists an corresponding result for the proper choice of the scaling factor.

Furthermore there exists a bunch of further results, for example the Curry conditions or Goldstein conditions.

Remark (In terms of algorithmic implementation)**.** *For WP and SWP, this is non-trivial and it must be handled (in contrast to Armijo).*

The following 2-phase approach provides a WP step size.

$$\varphi(t) := f(x + td) \qquad \tilde{\psi}(t) := \varphi(t) - \varphi(0) - \sigma t \varphi'(0)$$

Phase 1: Determine an interval $[a, b]$ that contains an interval of points which satisfy $\tilde{\psi}(t) \leq 0$ (WP1) and $\varphi'(t) \geq \rho\varphi'(0)$ (WP2).

Phase 2: The interval $[a, b]$ is successively reduced until we can find some $t > 0$ with $t \in T_{WP}(x, d)$.

Lemma 2.4.4 (Fundamental auxiliary result)**.** *Let $\sigma < \rho$ and $\varphi'(0) < 0$. If $[a, b]$ is an interval with $0 \leq a < b$ with $\tilde{\psi}(a) \leq 0$, $\tilde{\psi}(b) \geq 0$ and $\tilde{\psi}(a) < 0$, then $[a, b]$ contains a point $\bar{t}$ with $\tilde{\psi}(\bar{t}) < 0$ and $\tilde{\psi}'(\bar{t}) = 0$. $\bar{t}$ is an inner point of the interval I such that for all $z \in I$:*

$$\tilde{\psi}(t) \leq 0 \qquad \varphi'(t) \geq \rho \cdot \varphi'(0)$$

Hence $I \subseteq T_{WP}(x, d)$.

Proof. Let $\bar{t}$ be a global minimum of $\tilde{\psi}$ on interval $[a, b]$. Because $\tilde{\psi}(a) \leq 0$, $\tilde{\psi}(b) \geq 0$ and $\tilde{\psi}(a) < 0$, $\bar{t}$ is an inner point of $[a, b]$. Thus $\tilde{\varphi}'(\bar{t}) = 0$.

Furthermore $\tilde{\psi}(\bar{t}) < \tilde{\psi}(a) \leq 0$. By $\tilde{\psi}(\bar{t}) < 0$, $\tilde{\psi}'(\bar{t}) = 0$ and $\sigma < \rho$, the existence of interval I with inner point $\bar{t}$ follows such that $\forall t \in I : \tilde{\psi}(t) \leq 0$, $\tilde{\varphi}'(t) \geq (\rho - \sigma) \cdot \varphi'(0) \iff \tilde{\psi}(t) \leq 0, \varphi'(t) \geq \rho\varphi'(0) \forall t \in I$. □

Remark. *In the lemma, we required $\sigma < \rho$. So far, we allowed $\sigma = \rho$. In practice, this is not a relevant constraint, because σ is commonly chosen much smaller than ρ.*

Algorithm (Algorithm to determine WP step size)**.** *Given $x, d \in \mathbb{R}^n$ with $(\nabla f(x))^t d < 0$.*

Phase 1*:*

1.0. Choose $t_0 > 0$, $\gamma > 1$ and let $i := 0$

1.1. If $\tilde{\psi}(t_i) \geq 0$, then $a := 0$, $b := t_i$. Go to phase 2.
If $\tilde{\psi}(t_i) < 0$ and $\varphi'(t_i) \geq \rho\varphi'(0)$, then let $t := t_i$ and abort (STOP 1)
If $\tilde{\psi}(t_i) < 0$ and $\varphi'(t_i) < \rho\varphi'(0)$, then let $t_{i+1} := \gamma t_i$, $i := i + 1$ and go to (1.1)

Phase 2*:*

2.0. Choose τ_1, $\tau_2 \in \left(0, \frac{1}{2}\right]$. Let $j := 0$ and let $a_0 := a$, $b_0 := b$ (of phase 1)

2.1. Choose $\tilde{t}_j \in [a_j + \tau_1(b_j - a_j), b_j - \tau_2(b_j - a_j)]$

2.2. If $\tilde{\psi}(\tilde{t}_j) \geq 0$, then let $a_{j+1} := a_j$, $b_{j+1} := \tilde{t}_j$, $j := j + 1$ and go to (2.1).
If $\tilde{\psi}(\tilde{t}_j) < 0$ and $\varphi'(\tilde{t}_j) \geq \rho \cdot \varphi'(0)$. Then let $t := \tilde{t}_j$ and abort (STOP 2)
If $\tilde{\psi}(\tilde{t}_j) < 0$ and $\varphi'(\tilde{t}_j) < \rho \cdot \varphi'(0)$. Then let $a_{j+1} := \tilde{t}_j$, $b_{j+1} := b_j$, $j := j + 1$ and go to step (2.1)

Remark. *In case of abortion, provided t is the WP step size (remains to be shown).*

Remark (About the choice of t_0 in (1.0))**.** *With step 2 and following in any descent method, t_0 can be chosen as the step size in the step before. If we know some lower bound $\underline{f}$ of f in $\mathbb{R}^n$, then*

$$t \leq \frac{\underline{f} - \varphi(0)}{\sigma\varphi'(0)} := t^*$$

follows from $\tilde{\psi}(t) \leq 0$. Then it makes sense to choose $t_0 \in (0, t^*]$.

Theorem 2.4.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and bounded by below. Let $\sigma \in (0, \frac{1}{2})$, $\rho \in (\sigma, 1)$. Then the algorithm above aborts after finitely many steps with STOP 1 or STOP 2 with some WP step size.*

> ↓ *This lecture took place on 2019/05/16.*

Proof. **About phase 1**

If the abortion occurs in phase 1, then with STOP 1.

$$\tilde{\psi}(t_i) < 0 \qquad \varphi'(t_i) \geq \rho\varphi'(0) \implies t := t_i \text{ satisfies (WP1) and (WP2)}$$

If no abortion occurs in phase 1, then phase 2 is entered with an interval $[a, b]$. By construction of the algorithm, we have that

$$\varphi'(a) < \rho\varphi'(0)$$

Assume phase 1 neither aborts with STOP 1 nor a transition to phase 2 happens (step 2.0), then $t_i = \gamma^i \cdot t_0$ for $i \in \mathbb{N}$ with $\tilde{\psi}(t_i) < 0$.

$$\implies \varphi(t_i) < \varphi(0) + \sigma t_i \varphi'(0)$$

Because $\gamma > 1$, $\varphi'(0) < 0$ and f bounded by below, we get a contradiction such that $\varphi(t_i) < \varphi(0) + \sigma t_i \varphi'(0)$ cannot be satisfied.

About phase 2

You can immediately verify that $t := t_j$ is an admissible WP step size, when abortion STOP 2 is triggered.

First, we are going to show that for all j of the interval $[a_j, b_j]$ the following properties are satisfied:

$$\tilde{\psi}(a_j) \leq 0, \tilde{\psi}(b_j) \geq 0, \tilde{\psi}(a_j) < 0 \qquad \text{(compare with Lemma 2.4.4)}$$

and $\varphi'(a_j) < \rho\varphi'(0)$.

We are going to prove this statement by induction over j.

Induction base $j = 0$: Results from considerations of phase 1.

Induction step $j \to j+1$ Assume $[a_j, b_j]$ has desired properties. Show that $[a_{j+1}, b_{j+1}]$ has desired properties.

Case $\tilde{\psi}(t_j) < 0$ $\varphi'(t_j) < \rho\varphi'(0)$. $\to a_{j+1} = t_j$ and $b_{j+1} = b_j$.
We have

$$\begin{aligned}
\tilde{\psi}(a_{j+1}) &= \tilde{\psi}(t_j) < 0 \\
\tilde{\psi}(b_{j+1}) &= \tilde{\psi}(b_j) \geq 0 \\
\tilde{\psi}'(a_{j+1}) &= \tilde{\psi}'(t_j) \overset{\substack{\text{induction}\\ \text{hypothesis}}}{<} \rho\varphi'(0) < 0
\end{aligned}$$

Case $\tilde{\psi}(t_j) \geq 0$

$$\begin{aligned}
a_{j+1} &= a_j \\
b_{j+1} &= t_j \\
\tilde{\psi}(a_{j+1}) &= \tilde{\psi}(a_j) \leq 0 \\
\tilde{\psi}(b_{j+1}) &= \tilde{\psi}(t_j) \overset{\substack{\text{induction}\\ \text{hypothesis}}}{\geq} 0 \\
\tilde{\psi}'(a_{j+1}) &= \tilde{\psi}'(a_j) \overset{\substack{\text{induction}\\ \text{hypothesis}}}{\leq} 0
\end{aligned}$$

Now, we want to show that phase 2 aborts after a finite number of steps. Assume there is no abortion.

The interval length $\left|b_j - a_j\right|$ decreases step by step at least by factor $\max\{1-\tau_1, 1-\tau_2\} < 1$. This results in a contraction in one point $\hat{t}$.

Now apply Lemma 2.4.4 on $[a_j, b_j] \forall j \in \mathbb{N}$. For all $j \in \mathbb{N} \exists t_j \in (a_j, b_j)$ with $\tilde{\psi}(t_j) < 0$ and $\tilde{\psi}'(t_j) = 0$. For $j \to \infty$, we have $\lim_{j\to\infty} t_j = \hat{t}$ and thus $\tilde{\psi}'(\hat{t}) = 0$.

$$\Longrightarrow \varphi'(\hat{t}) = \sigma\varphi'(0)$$

Because $\sigma < \rho$ and $\varphi'(0) < 0$, we get a contradiction to $\varphi'(t) \leq \rho\varphi'(0)$ (this results from $\varphi'(a_j) < \rho\varphi'(0)$). □

Remark (About the choice of t_j in phase 2)**.** *One possibility is the use of interpolation polynomials. For example, we can uniquely determine the cubic (Hermitian) interpolation polynomial p_j such that*

$$p_j(a_j) = \varphi(a_j) \qquad p_j'(a_j) = \varphi'(a_j) \qquad p_j(b_j) = \varphi(b_j) \qquad p_j'(b_j) = \varphi'(b_j)$$

The local minimum t_j^* of p_j (if it exists), can be determined explicitly and if t_j^* lies within the desired interval, then choose $t_j = t_j^*$. Otherwise e.g. as midpoint of the interval.

If you now use only a quadratic interpolation polynomial q_j, the number of gradient evaluations is smaller.

$$q_j(a_j) = \varphi(a_j) \qquad q_j'(a_j) = \varphi'(a_j) \qquad q_j(b_j) = \varphi(b_j)$$

In the same manner, an algorithm for determining strict WP step sizes can be devised.

2.5 Convergence speed

Conventions in this chapter:

$$\text{sequence } \left\{x^{(k)}\right\} \to \underbrace{x^*}_{\in \mathbb{R}^n} \text{ with } x^{(k)} \in \mathbb{R}^n$$

Definition (Q-convergence). *1. Let $p \in [1, \infty)$.*

$$Q_p\left(\left\{x^{(k)}\right\}\right) = \begin{cases} \limsup_{k\to\infty} \frac{\left|x^{(k+1)} - x^*\right|}{\left|x^{(k)} - x^*\right|^p} & \text{if } x^{(k)} \neq x^* \text{ for } k \geq k_0 \\ 0 & \text{if } x^{(k)} = x^* \text{ for } k \geq k_0 \\ \infty & \end{cases}$$

is called quotient factor *or* Q-factor *of* $\left\{x^{(k)}\right\}$.

2. The value $\inf\left\{p \in [1, \infty) \mid Q_p(\left\{x^{(k)}\right\}) = \infty\right\}$ *is called* Q-convergence order *of* $\left\{x^{(k)}\right\}$.

Remark.

- *The Q-factor depends on the choice of the norm. Convergence order does not. (The proof is left as an exercise to the reader)*
- *There exists some value $p_0 \in [1, \infty)$ with*

$$Q_p\left(\left\{x^{(k)}\right\}\right) = \begin{cases} 0 & p \in [1, p_0) \\ \infty & p \in (p_0, \infty) \end{cases}$$

- *We consider the following special cases:*

$$\begin{array}{ll} Q_1\left(\left\{x^{(k)}\right\}\right) = 0 & \textit{Q-superlinear convergence} \\ Q_1\left(\left\{x^{(k)}\right\}\right) < 1 & \textit{Q-linear convergence} \\ Q_2\left(\left\{x^{(k)}\right\}\right) = 0 & \textit{Q-superquadratic convergence} \\ 0 < Q_2\left(\left\{x^{(k)}\right\}\right) < \infty & \textit{Q-quadratic convergence} \end{array}$$

Example 12.

- *$x^{(k)} = a^k$ with $a \in (0, 1)$. $\lim_{k\to\infty} x^{(k)} = 0 = x^*$.*

$$\frac{\left|x^{(k+1)} - x^*\right|}{\left|x^{(k)} - x^*\right|} = \frac{a^{k+1}}{a^k} = a$$

$$\limsup_{k\to\infty} \cdots = a \qquad \textit{linear Q-convergence}$$

Consider $p > 1$.

$$\limsup_{k\to\infty} \frac{a^{k+1}}{a^{kp}} = \limsup_{k\to\infty} a^{k(1-p)+1} = \infty \textit{ for } k \to \infty$$

Q-order of $\left\{x^{(k)}\right\}$ is 1.

- *$x^{(k)} = a^{2^k}$ and $a \in (0,1)$. $x^* = 0$. Consider $p = 2$.*

$$Q_2(\{x^{(k)}\}) = \limsup_{k\to\infty} \frac{\left|a^{2^{k+1}} - 0\right|}{\left|a^{2^k} - 0\right|} = \limsup_{k\to\infty} \frac{a^{2^{k+1}}}{\left(a^{2^k}\right)^2} = 1$$

Left as an exercise to the reader: Consider $p > 2$, then the Q-convergence of $\{x^{(k)}\}$ is 2.

Definition 2.5.1 (R-convergence). • *Let $p \in [1,\infty)$. Then*

$$R_p\left(\{x^{(k)}\}\right) := \begin{cases} \limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|^{\frac{1}{k}} & p = 1 \\ \limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|^{\frac{1}{p^k}} & p > 1 \end{cases}$$

is called root factor *or* R-factor *of $\{x^{(k)}\}$.*

- *The value $\inf\left\{p \in [1,\infty) \mid R_p\left(\{x^{(k)}\}\right) = 1\right\}$ is called R-convergence order of $\{x^{(k)}\}$..*

Example 13. • *Consider $x^{(k)} = ca^k$ with $a \in (0,1), c \in \mathbb{R} \setminus \{0\}$. $x^* = 0$. Let $p = 1$.*

$$R_1\left(\{x^{(k)}\}\right) = \limsup_{k\to\infty} \left|x^{(k)} - x^*\right|^{\frac{1}{k}} = \limsup_{k\to\infty} (\left|ca^k\right|)^{\frac{1}{k}} = \limsup_{k\to\infty} |c|^{\frac{1}{k}}\, a = a$$

Let $p > 1$.

$$\left|x^{(k)} - x^*\right|^{\frac{1}{p^k}} = |c|^{\frac{1}{p^k}}\, a^{k\frac{1}{p^k}} \to 1$$

The R-convergence order of $\{x^{(k)}\}$ is thus 1.

- *$x^{(k)} = a^{2^k}$ with $a \in (0,1)$. Consider $p = 2$.*

$$\limsup_{k\to\infty} \left(a^{2^k}\right)^{\frac{1}{2^k}} = a \qquad R_2\left(\{x^{(k)}\}\right) = a < 1$$

Consider $p > 2$.

$$\left(a^{2^k}\right)^{\frac{1}{p^k}} = a^{\left(\frac{2}{p}\right)^k} \to 1$$

Its R-convergence order is 2.

Remark. *In general, the notions of Q and R convergence do not match. But we have the generic result that Q-convergence order of $\{x^{(k)}\} \le$ R-convergence order of $\{x^{(k)}\}$ and $R_1\left(\{x^{(k)}\}\right) \le Q_1\left(\{x^{(k)}\}\right)$.*

↓ *This lecture took place on 2019/05/20.*

In the context of termination criteria, we look for estimates of $\left\|x^{(k)} - x^*\right\|$. Termination if $\left\|x^{(k)} - x^*\right\| \le \varepsilon$. Direct question is unrealistic.

Theorem 2.5.2. *1. We have*

$$1 - \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} \leq \frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)} - x^*\right\|} \forall x^{(k)} \neq x^*$$

2. If $\left\{x^{(k)}\right\} \to x^*$ is Q-superlinear and $x^{(k)} \neq x^* \forall k \geq k_0$, then $\lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)} - x^*\right\|} = 1$

Proof.

$$\begin{aligned} \left\|\left\|x^* - x^{(k)}\right\| - \left\|x^* - x^{(k+1)}\right\|\right\| &\leq \left\|x^{(k+1)} - x^{(k)}\right\| \\ &\leq \left\|x^{(k)} - x^*\right\| + \left\|x^* - x^{(k+1)}\right\| \end{aligned}$$

If $x^{(k)} \neq x^*$, then

$$\left|1 - \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|}\right| \leq \frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)} - x^*\right\|} \leq 1 + \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|}$$

By Q-superlinearity of $\left\{x^{(k)}\right\} \to x^*$, we have

$$\lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} = 0 \implies \lim_{k\to\infty} \frac{\left\|x^{(k+1)} - x^*\right\|}{\left\|x^{(k)} - x^*\right\|} = 1$$

□

Remark. *The condition of Theorem 2.5.2 (2) is equivalent to*

$$\forall \delta > 0 \exists k_0 \in \mathbb{N} \forall k \geq k_0 : (1-\delta) \leq \frac{\left\|x^{(k+1)} - x^{(k)}\right\|}{\left\|x^{(k)} - x^*\right\|} \leq 1 + \delta$$

$$\iff (1-\delta)\left\|x^{(k)} - x^*\right\| \leq \left\|x^{(k+1)} - x^{(k)}\right\| \leq (1+\delta)\left\|x^{(k)} - x^*\right\|$$

Abortion criterion $\left\|x^{(k+1)} - x^{(k)}\right\| \leq \tilde{\varepsilon}$ is meaningful if $\left\{x^{(k)}\right\}$ is Q-superlinear convergent and K sufficiently large.

Revision (Reminder)**.** *Q-factor depends on the norm. Q-order does not. (Compare with practicals)*

- Contrast to R-factor:

  R-factor is independent of the norm (thus also R-convergence order).

  Let $\|\cdot\|_a$ and $\|\cdot\|_b$ two norms on $\mathbb{R}^n$. $\exists 0 < c_1 < c_2$ such that $c_1 \|x\|_b \leq \|x\|_a \leq c_2 \|x\|_b \, \forall x \in \mathbb{R}^n$. Let $\{\gamma_k\}$ be a zero sequence, $\gamma_k > 0$.

$$\limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|_a^{\gamma_k} \leq \limsup_{k\to\infty} c_2^{\gamma_k} \left\|x^{(k)} - x^*\right\|_b = \underbrace{\lim_{k\to\infty} c_2^{\gamma_k}}_{=1} \limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|_b^{\gamma_k}$$

Analogously, we get

$$\limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|_b^{\gamma_k} \leq \limsup_{k\to\infty} \left\|x^{(k)} - x^*\right\|_a^{\gamma_k}$$

Thus the R-factor wrt. $\|\cdot\|_a$ and $\|\cdot\|_b$ is identical.

- $\exists p_0 \in [1, \infty)$ with

$$R_p\left(\left\{x^{(k)}\right\}\right) = \begin{cases} 0 & \text{for } p \in [1, p_0) \\ 1 & \text{for } p \in (p_0, \infty) \end{cases}$$

- Q-order of $\left\{x^{(k)}\right\} \leq R$ order of $\left\{x^{(k)}\right\}$
  and $R_1\left(\left\{x^{(k)}\right\}\right) \leq Q_1\left(\left\{x^{(k)}\right\}\right)$

In the following, we are going to focus on sequences $\left\{x^{(k)}\right\}$ that minimize $f : \mathbb{R}^n \to \mathbb{R}$.

Remark (Observation)**.** *If* $\left\|x^{(k+1)} - x^*\right\| = o(\left\|x^{(k)} - x^*\right\|) \implies \left\{x^{(k)}\right\}$ *converges Q-superlinear. If* $\left\|x^{(k+1)} - x^*\right\| = O(\left\|x^{(k)} - x^*\right\|^2) \implies \left\{x^{(k)}\right\}$ *converges Q-quadratic or Q-superquadratic.*

Lemma 2.5.3. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be two times differentiable. Then*

1. $\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})(x^{(k)} - x^*)\right\| = o\left(\left\|x^{(k)} - x^*\right\|\right)$
2. *If additionally* $\nabla^2 f$ *locally Lipschitz continuous with constant* L *is given, then*

$$\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})(x^{(k)} - x^*)\right\| = O\left(\left\|x^{(k)} - x^*\right\|^2\right)$$

Proof. 1.

$$\begin{aligned}\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})(x^{(k)} - x^*)\right\| &= \left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^*)(x^{(k)} - x^*)\right\| \\ &\quad + \left\|\nabla^2 f(x^{(k)}) - \nabla^2 f(x^*)\right\| \left\|x^{(k)} - x^*\right\|\end{aligned}$$

By the Taylor expansion, we get

$$\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^*)(x^{(k)} - x^*)\right\| = o\left(\left\|x^{(k)} - x^*\right\|\right)$$

$$\lim_{k\to\infty} \left\|\nabla^2 f(x^{(k)}) - \nabla^2 f(x^*)\right\| = 0$$

$$\implies \left\|\nabla^2 f(x^{(k)}) - \nabla^2 f(x^*)\right\| \left\|x^{(k)} - x^*\right\| = o\left(\left\|x^{(k)} - x^*\right\|\right)$$

2.

$$\nabla f(x^{(k)}) - \nabla f(x^*) = \int_0^1 \nabla^2 f\left(x^* + \tau(x^{(k)} - x^*)\right)(x^{(k)} - x^*) d\tau$$

$$
\begin{aligned}
&\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})\right\| \\
&= \left\| \int_0^1 \left[\nabla^2 f(x^* + \tau(x^{(k)} - x^*)) - \nabla^2 f(x^{(k)})\right](x^{(k)} - x^*)\, d\tau \right\| \\
&\leq \int_0^1 \left\|\nabla^2 f(x^* + \tau(x^{(k)} - x^*)) - \nabla^2 f(x^{(k)})\right\| d\tau \left\|x^{(k)} - x^*\right\| \\
&\leq \int_0^1 (1-\tau)\, d\tau L \left\|x^{(k)} - x^*\right\|^2 = \frac{L}{2}\left\|x^{(k)} - x^*\right\|^2 = O\left(\left\|x^{(k)} - x^*\right\|\right)
\end{aligned}
$$

□

Lemma 2.5.4. *Let A and B be $n \times n$ matrices over $\mathbb{R}$. Let I be the unit matrix.*

$$\|I - BA\| < 1 \implies A \text{ and } B \text{ regular and } \left\|B^{-1}\right\| \leq \frac{\|A\|}{1 - \|I - BA\|}$$

Proof. Result from Linear Algebra. Here without proof. □

Lemma 2.5.5. *Let f be two times differentiable and let $x^* \in \mathbb{R}^n$ with $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ regular. Then $\exists k_0 \in \mathbb{N}, \exists \beta > 0$ such that $\forall k \geq k_0$*

$$\left\|\nabla f(x^{(k)})\right\| \geq \beta \left\|x^{(k)} - x^*\right\|$$

Proof. Use Taylor expansion of second order.

$$\forall \varepsilon > 0 \exists k_0 \in \mathbb{N} : \left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^*)(x^{(k)} - x^*)\right\| \leq \varepsilon \left\|x^{(k)} - x^*\right\| \forall k \geq k_0$$

Choose $\varepsilon > 0$ such that $\varepsilon < \left\|(\nabla^2 f(x^*))^{-1}\right\|^{-1}$.

Recognize that

$$\left\|x^{(k)} - x^*\right\| = \left\|\left(\nabla^2 f(x^*)\right)^{-1} \cdot \nabla^2 f(x^*)(x^{(k)} - x^*)\right\| \leq \left\|\left(\nabla^2 f(x^*)\right)^{-1}\right\| \left\|\nabla^2 f(x^*)(x^{(k)} - x^*)\right\|$$

For $k \geq k_0$,

$$\left\|\nabla f(x^{(k)})\right\| = \left\|\nabla f(x^{(k)}) - \nabla f(x^*) + \nabla^2 f(x^*)(x^{(k)} - x^*) - \nabla^2 f(x^*)(x^{(k)} - x^*)\right\|$$

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

$$
\begin{aligned}
&\geq \left\|\nabla^2 f(x^*)\left(x^{(k)} - x^*\right)\right\| - \left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^*)(x^{(k)} - x^*)\right\| \\
&\geq \left\|(\nabla^f(x^*))^{-1}\right\|^{-1} \left\|x^{(k)} - x^*\right\| - \varepsilon \left\|x^{(k)} - x^*\right\| = \beta \left\|x^{(k)} - x^*\right\|
\end{aligned}
$$

with $\beta = \left\|(\nabla^2 f(x^{(k)}))^{-1}\right\|^{-1} - \varepsilon > 0$. □

Remark. $\left\|\nabla f(x^{(k)})\right\| \leq \hat{\varepsilon}$ *is a meaningful abortion criterion in this scenario, because*

$$\left\|\nabla f(x^{(k)})\right\| \leq \hat{\varepsilon} \implies \left\|x^{(k)} - x^*\right\| \leq \frac{\hat{\varepsilon}}{\beta}$$

for sufficiently large k.

Theorem 2.5.6. *Let f be two times continuously differentiable, $x^{(k)} \neq x^*$ for $k \geq k_0$ and $\nabla^2 f(x^*)$ regular. Then the following 3 statements are equivalent:*

- $x^{(k)} \to x^*$ *is Q-superlinear and* $\nabla f(x^*) = 0$
- $\left\|\nabla f(x^{(k)}) + \nabla^2 f(x^*)(x^{(k+1)} - x^{(k)})\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right)$
- $\left\|\nabla f(x^{(k)} + \nabla^2 f(x^{(k)})(x^{(k+1)} - x^{(k)}))\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right)$

Proof. Left as an exercise to the reader. □

Remark (About the relevance of Theorem 2.5.6)**.** *"Gradient-similar methods"*

Let $\left\{H^{(k)}\right\}$ be a sequence of invertible matrices over $\mathbb{R}$. For $H^{(k)} = I$, this gives the steepest descent method.

$$x^{(k+1)} = x^{(k)} - \left(H^{(k)}\right)^{-1} \nabla f(x^{(k)})$$

If $\left\{x^{(k)}\right\} \to x^*$ and $\nabla^2 f(x^*)$ invertible, then the following statements are equivalent:

$\overline{A}$: $x^{(k)} \to x^*$ Q-superlinear and $\nabla f(x^*) = 0$

$\overline{B}$: $\left\|\left(\nabla^2 f(x^*) - H^{(k)}\right)\left(x^{(k+1)} - x^{(k)}\right)\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right)$

$\overline{C}$: $\left\|(\nabla^2 f(x^{(k)}) - H^{(k)})(x^{(k+1)} - x^{(k)})\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right)$

This results directly from Theorem 2.5.6 and $\nabla f(x^{(k)}) = -H^{(k)}\left(x^{(k+1)} - x^{(k)}\right)$.

Also $x^{(k)} \to x^*$ Q-superlinear converges, if

$$\lim_{k\to\infty} \left\|H^{(k)} - \nabla^2 f(x^{(k)})\right\| = 0 \text{ and accordingly } \lim_{k\to\infty} H^{(k)} = \nabla^2 f(x^*)$$

Analogous results hold for quadratic and superquadratic converges if $o\left(\left\|x^{(k)} - x^*\right\|\right)$ is substituted by $O\left(\left\|x^{(k)} - x^*\right\|^2\right)$.

2.6 Gradient methods, Method of steepest Descent

In the generic descent method, we have large freedom in the choice of $d^{(k)}$. One possible choice is $d^{(k)} = -\nabla f(x^{(k)})$ pointing in direction of the steepest descent. This is a result from Analysis, but can also be proven directly with the following idea:

$$\min \left(\nabla f(x)\right)^t d \qquad \text{s. t. } \|d\| = 1$$

$x \in \mathbb{R}^n$ is fixed. Find the direction fo the steepest descent in x.

Utilizing the Cauchy-Schwarz inequality, we get

$$d = -\frac{\nabla f(x)}{\left\|\nabla f(x)\right\|}$$

as the optimal solution of the problem above.

In the following, we choose $d^{(k)} = -\nabla f(x^{(k)})$ and look at the resulting variants of the descent method.

Algorithm 2.6.1 (Version 1)**.**

1. *Choose* $x^{(0)} \in \mathbb{R}^k$, $\sigma \in (0,1)$, $\beta \in (0,1)$, $\varepsilon \geq 0$. *Let* $K := 0$.
2. *If* $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, *then stop.*
3. *Determine* $t_k := \max\left\{\beta^l \mid l = 0, 1, \dots\right\}$ *with (via Armijo)*

$$f\left(x^{(k)} + t_k d^{(k)}\right) \leq f\left(x^{(k)}\right) + \sigma t_k \left(\nabla f(x^{(k)})\right)^t d^{(k)}$$

4. *Let* $x^{(k+1)} := x^{(k)} + t_k d^{(k)}$. *Let* $k := k+1$ *and go to (1).*

Remark. *If we replace (3) by some efficient step size strategy (e.g. strict WP, scaled Armijo rule), then the convergence result can be derived directly from chapter 4. Specifically, every cluster point of* $\left\{x^{(k)}\right\}$ *is stationary point of* f *under appropriate conditions of* f.

For the special case $d^{(k)} = -\nabla f(x^{(k)})$ *we can also derive a convergence result for the (non-efficient) Armijo rule.*

↓ *This lecture took place on 2019/06/03.*

Revision 4. *Steepest descent method.*

$$d^{(k)} = -\nabla f(x^{(k)})$$

Step size with Armijo.

For this approach, our generic convergence result (efficient step size required) is not applicable.

Lemma 2.6.2. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable. If*

$$\left\{x^{(k)}\right\} \to x, x^{(k)}, x \in \mathbb{R}^n \qquad \left\{d^{(k)}\right\} \to d, d^{(k)}, d \in \mathbb{R}^n$$

$$\{t_k\} \to 0, t_k > 0 \quad t_k \in \mathbb{R}$$

Then

$$\lim_{k\to\infty} \frac{f(x^{(k)} + t_k d^{(k)} - f(x^{(k)}))}{t_k} = (\nabla f(x))^t d$$

Proof. Left as an exercise to the reader. □

Theorem 2.6.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Every cluster point resulting from a sequence $\left\{x^{(k)}\right\}$ generated by our algorithm is a stationary point of f.*

Remark. *This is an advantageous result because it has less requirements than the generic convergence result.*

Proof. Let $\tilde{x}$ be a cluster point of $\left\{x^{(k)}\right\}$. Let $\left\{x^{(k)}\right\}$ be subsequence of $\left\{x^{(k)}\right\}$ converging to $\tilde{x}$. Assume $\nabla f(\tilde{x}) \neq 0$. Consider "$\left\{f(x^{(k)})\right\}$ is monotonically decreasing" and "$\left\{f(x^{(k)})\right\}_k \to f(\tilde{x})$" implies that $\left\{f(x^{(k)})\right\} \to f(\tilde{x})$

$$f(x^{(k+1)}) - f(x^{(k)}) \to 0$$

By construction of the algorithm, we get

$$t_k \cdot \left(\nabla f(x^{(k)}))^t d^{(k)}\right) = -t_k \cdot \left\|\nabla f(x^{(k)})\right\|^2 \to 0$$

Because $\left\{\nabla f(x^{(k)})\right\}_k \to \nabla f(\tilde{x})$ with $\nabla f(\tilde{x}) \neq 0$ by assumption. Thus, $\{t_k\}_k \to 0$

$$\implies f(x^{(k)} + \beta^{l_{k-1}} d^{(k)}) > f(x^{(\alpha)}) + \sigma \beta^{l_k - 1} (\nabla f(x^{(k)}))^t d^{(k)}$$

for all $k \in K$ with sufficiently large k where $t_k = \beta^{l_k}$ (l_k is uniquely determined) is the Armijo step size in step k.

$$\implies \frac{f(x^{(k)} + \beta^{l_k - 1} d^{(k)} - f(x^{(k)}))}{\beta^{l_k - 1}} > \sigma \left(\nabla f(x^{(k)})\right)^t d^{(k)}$$

With $k \to \infty$, $k \in K$ we get $\beta^{l_k - 1} \to 0$ and with Theorem 2.6.2.

$$-\left\|\nabla f(\tilde{x})\right\|^2 \geq -\sigma \cdot \left\|\nabla f(\tilde{x})\right\|^2$$

This is a contradiction because $\nabla f(\tilde{x}) \neq 0, \sigma \in (0, 1)$. □

Remark. *The proof above shows that a stronger result holds true:*

It suffices if $\left\{x^{(k)}\right\}$, $\left\{d^{(k)}\right\}$ and $t_k > 0$ can be chosen such that

$$x^{(k+1)} = x^{(k)} + t_k \cdot d^{(k)} \qquad \forall k \in \mathbb{N}$$

$$f(x^{(k+1)}) \leq f(x^{(k)}) \qquad \forall k \in \mathbb{N}$$

Let $\tilde{x}$ be a limit of subsequence $\left\{x^{(k)}\right\}_k$ such that $d^{(k)} = -\nabla f(x^{(k)}) \forall k \in K$ and $t_k > 0 \forall k \in \mathbb{N}$ where t_k for $k \in K$ suffices the Armijo condition.

$$f(x^{(k)} + t_k \cdot d^{(k)}) \leq f(x^{(k)}) + \sigma \cdot t_k \cdot (\nabla f(x^{(k)})^t d^{(k)})$$

Then $\tilde{x}$ is a stationary point of f.

Convergence follows from Theorem 2.6.3. What about convergence speed? In the following, we analyze the special case of a quadratic function. Let $f(x) = \frac{1}{2}x^tQx + c^tx + \gamma$ with $\gamma \in \mathbb{R}, c \in \mathbb{R}^n$ and Q is a $n \times n$ matrix over $\mathbb{R}$. Let Q be symmetric (assumption wlog.) and positive definite.

In the following analysis, we are going to use the exact step size

$$t_k = \operatorname{argmin}_{t \to 0} f(x^{(k)} + td^{(k)})$$

Notation: $g^{(k)} := \nabla f(x^{(k)}) = Qx^{(k)} + c$

An explicit calculation (assert 0 for derivative) gives $t_k = -\frac{(g^{(k)})^l d^{(k)}}{(d^{(k)})^t Q d^{(k)}}$.

With $d^{(k)} = -g^{(k)}$ (steepest descent) results in

$$t_k = \frac{\left(g^{(k)}\right)^t g^{(k)}}{\left(g^{(k)}\right)^t Q g^{(k)}}$$

Lemma 2.6.4 (Kantorovich inequality)**.** *Let Q be a $n \times n$ symmetric, positive-definite matrix over $\mathbb{R}$. Let λ_{min} be the smallest eigenvalue and λ_{max} be the largest eigenvalue of Q. Then*

$$\frac{(x^tx)^2}{(x^tQx)(x^tQ^{-1}x)} \geq \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min} + \lambda_{max})^2} \qquad \forall x \in \mathbb{R}^n, x \neq 0$$

Theorem 2.6.5. *Let $f(x) = \frac{1}{2}x^tQx + c^tx + \gamma$. Let Q be symmetric, positive-definite. The steepest descent method with exact step size choice (gradient descent) converges to the (uniquely determined) global minimum x^* for every initial point $x^{(0)} \in \mathbb{R}^n$. Furthermore,*

$$f(x^{(k+1)}) - f(x^*) \leq \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}\right)^2 (f(x^{(k)}) - f(x^*))$$

where λ_{max} is the largest eigenvalue of Q and λ_{min} is the smallest eigenvalue of Q.

Proof. Because Q is positive-definite, there exists a unique global minimum x^* and it must satisfy

$$\nabla f(x^*) = Qx^* + c = 0 \implies x^* = -Q^{-1}c$$

$$f(x^*) = -\frac{1}{2}c^tQ^{-1}c + \gamma \qquad g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} + c$$

$$x^{(k+1)} = x^{(k)} - t_k \cdot g^{(k)} = x^{(k)} - \frac{(g^{(k)})^t g^{(k)}}{(g^{(k)})^t Q g^{(k)}} g^{(k)}$$

A simple calculation yields:

$$f(x^{(k+1)} - f(x^*)) = \left(1 - \frac{\left((g^{(k)})^t g^{(k)}\right)^2}{\left((g^{(k)})^t Q g^{(k)}\right)\left((g^{(k)})^t Q^{-1} g^{(k)}\right)}\right)(f(x^{(k)}) - f(x^*))$$

Filling the details is left as an exercise.

Now use the Kantorovic inequality from Lemma 2.6.4. This gives

$$f(x^{(k+1)}) - f(x^*) \leq \underbrace{\left(1 - \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}\right)}_{=\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\min}+\lambda_{\max}}\right)^2}(f(x^{(k)}) - f(x^*))$$

For proper choice of $x^{(0)}$ the reduction factor is chosen as $\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\min}+\lambda_{\max}}\right)^2$.

Relation to condition number:

$$\kappa := \text{condition}(Q) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

wrt. to Euclidean norm. Thus,

$$f(x^{(k+1)}) - f(x^*) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 (f(x^{(k)}) - f(x^*))$$

An additional result of linear algebra yields,

$$\left\|x^{(k)} - x^*\right\| \leq \sqrt{\kappa}\left(\frac{\kappa - 1}{\kappa + 1}\right)^k \left\|x^{(0)} - x^*\right\|$$

□

Lemma 2.6.6. *Let Q be a matrix over $\mathbb{R}$ and symmetric, positive-definite. Let λ_{min} be the smallest eigenvalue of Q and λ_{max} be the largest eigenvalue of Q. Then*

$$\lambda_{min} \cdot x^t x \leq x^t Q x \leq \lambda_{max} \cdot x^t x \qquad \forall x \in \mathbb{R}^n$$

$$\iff \lambda_{min} \leq \frac{x^t Q x}{x^t x} \leq \lambda_{max} \qquad \forall x \in \mathbb{R}^k, x \neq 0$$

Remark (Conclusion)**.** *If κ is large (Q ill-conditioned), then the convergence of steepest descent is very slow (Zickzack effect)*

In the generic case (f non-quadratic). Let f be two-times differentiable. Let x^k be the minimum of f. f can be approximated in the neighborhood of x^* by a quadratic function q.

$$q(x) = f(x^*) + (\nabla f(x^*))^t (x - x^*) + \frac{1}{2}(x - x^*)^t \nabla^2 f(x^*)(x - x^*)$$

Here, it depends on the condition of $\nabla^2 f(x^*)$.

Example (Rosenbrock function)**.** *Consider $n = 2$.*

$$\nabla^2 f(x^*) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \qquad \kappa \approx 2500 \qquad x^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The gradient method is very slow.

Hence, we conclude that the gradient method is often non-satisfactory.

One approach to an improvement is choosing $d^{(k)} = -H\nabla f(x^{(k)})$ with symmetric, positive-definite, properly chosen H. How do we chose H?

In the quadratic special case, we can show that H can be chosen such that

$$\frac{\lambda_{\max}(H^{-1}Q)}{\lambda_{\min}(H^{-1}Q)} \text{ is smaller than } \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$$

and $Hd^{(k)} = -\nabla f(x^{(k)})$ shall be easier to solve than $Qx + c = 0$

2.6.1 Gradient-like methods/directions: Generalization of the gradient method

Algorithm 2.6.7.

1. *Choose* $x^{(0)} \in \mathbb{R}^n, \sigma \in (0,1), \beta \in (0,1), \varepsilon \geq 0$ *and let* $k := 0$
2. *If* $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, *stop*
3. *Determine* $d^{(k)} \in \mathbb{R}^n$ *with* $(\nabla f(x^{(k)}))^t d^{(k)} < 0$
4. *Determine* $t_k := \max\left\{\beta^l \mid l = 0, 1, 2, \dots\right\}$ *such that*

$$f(x^{(k)} + t_k d^{(k)}) \leq f(x^{(k)}) + \sigma \cdot t_k \left(\nabla f(x^{(k)})\right)^t d^{(k)}$$

5. *Let* $x^{(k+1)} := x^{(k)} + t_k \cdot d^{(k)}$ *and* $k := k + 1$. *Go to 1.*

Remark (Question)**.** *Which requirements must* $d^{(k)}$ *satisfy?*

Definition (Gradient-like directions)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable and let* $\left\{x^{(k)}\right\}$ *be a sequence in* $\mathbb{R}^n$. *Let* $\left\{x^{(k)}\right\}_k$ *be a subsequence converging* not *to some stationary point of* f. *The sequence of directions* $\left\{d^{(k)}\right\}$ *with* $d^{(k)} \in \mathbb{R}^n$ *is called* gradient-like *wrt.* f *and* $\left\{x^{(k)}\right\}$ *if there exist constants* $c > 0$ *and* $\varepsilon > 0$ *for every* $\left\{x^{(k)}\right\}_k$ *such that*

$$\left\|d^{(k)}\right\| \leq c \qquad \forall k \in K \tag{E1}$$

$$\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -\varepsilon \qquad \forall k \in K \text{ sufficiently large} \tag{E2}$$

Remark. *The sequence* $\left\{d^{(k)}\right\}$ *with* $d^{(k)} = -\nabla f(x^{(k)})$ *is gradient-like.*

Remark (Question 1)**.** *Which other approaches exist?*

Remark (Question 2)**.** *Convergence result?*

Theorem 2.6.8. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable and let the sequence* $\left\{d^{(k)}\right\}$ *in the algorithm above be gradient-like wrt.* f *and* $\left\{x^{(k)}\right\}$. *Then every cluster point of* $\left\{x^{(k)}\right\}$ *is a stationary point of* f.

↓ *This lecture took place on 2019/06/04.*

Theorem 2.6.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let sequences $\left\{x^{(k)}\right\}, \left\{d^{(k)}\right\}$ and $\{t_k\}$ be generated from a gradient-like method. Let $p_1 \geq 0$, $p_2 \geq 0$, $c_1 > 0$ and $c_2 > 0$ satisfy $\forall k \in \mathbb{N}$*

- $\left\|d^{(k)}\right\| \leq c_1 \left\|\nabla f(x^{(k)})\right\|^{p_1}$
- $\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -c_2 \left\|\nabla f(x^{(k)})\right\|^{p_2}$

Every cluster point of $\left\{x^{(k)}\right\}$ is a stationary point of f.

Proof. Left as an exercise for the reader (optionally for practicals). □

Remark. *Special case $p_1 = 1$ and $p_2 = 2$. By Theorem 2.6.9 (1) and (2),*

$$-\frac{\left(\nabla f(x^{(k)})\right)^t d^{(k)}}{\left\|\nabla f(x^{(k)})\right\| \left\|d^{(k)}\right\|} \geq \frac{c_2}{c_1} \forall k \in \mathbb{N}$$

thus the angle condition is satisfied.

Corollary 2.6.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $\left\{H^{(k)}\right\}$. Let $H^{(k)}$ be a sequence of symmetric, positive-definite $n \times n$ matrices over R such that there exists $\mu_1 \geq 0$ and $\mu_2 \geq 0$*

$$\mu_1 \|x\|^2 \leq x^t H^{(k)} x \leq \mu_2 \|x\|^2 \qquad \forall x \in \mathbb{R}^n \forall k \in \mathbb{N}$$

Let $\left\{x^{(k)}\right\}$ be a sequence generated by a gradient-like method with choice $H^{(k)} d^{(k)} = -\nabla f(x^{(k)})$. Then every cluster point of $\left\{x^{(k)}\right\}$ is a stationary point of f.

$$d^{(k)} = -(H^{(k)})^{-1} \nabla f(x^{(k)})$$

Proof. Left as an exercise. □

2.7 Newton's method

Goal: Process with better convergence properties than the gradient method
Costs: Stronger requirements, 2nd derivative

In this section, we assume f is two times differentiable.

2.7.1 Local Newton's method

Idea: We approximate f by a quadratic function and refine this approximation step-by-step. We minimize the approximation function ("Modelling function").

$$q^{(k)}(x) := f(x^{(k)}) + \nabla f(x^{(k)})(x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^t \nabla^2 f(x^{(k)})(x - x^{(k)})$$

Taylor polynomial of second order. $x^{(k)}$ is the current iteration point.

$$\min q^{(k)}(x) \quad \rightarrow \quad \nabla q^{(k)}(x) = 0 \qquad \nabla^2 q^{(k)}(x) = \nabla^2 f(x^{(k)})$$

If $\nabla^2 f(x^{(k)})$ is positive definite, then $x^{(k+1)}$ is solution of $\min q^{(k)}(x)$ if $x^{(k+1)}$ is a solution of $\nabla q^{(k)}(x) = 0$.

$$\min q^{(k)}(x) \implies \nabla q^{(k)}(x) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}) \overset{!}{=} 0$$

$$\implies x^{(k+1)} = x^{(k)} - \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)}) \qquad \text{"Newton iteration"}$$

Important: we require $\exists (\nabla^2 f(x^{(k)}))^{-1}$

Special case $n = 1$:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

Remark. *Instead of computing the inverse $\nabla^2 f(x^{(k)})$ explicitly, we can use: Determine $d^{(k)}$ as solution of*

$$\nabla^2 f(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$

Let $x^{(k+1)} = x^{(k)} + d^{(k)}$.

This can be interpreted in our generic setting as $x^{(k+1)} = x^{(k)} + t_k d^{(k)}$ with step size $t_k = 1$.

Algorithm.

1. *Choose $x^{(0)} \in \mathbb{R}^n$, $\varepsilon \geq 0$. Let $k := 0$*
2. *If $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, then stop*
3. *Determine $d^{(k)}$ as solution of*
$$\nabla^2 f(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$
4. *Let $x^{(k+1)} := x^{(k)} + d^{(k)}$. Let $k := k + 1$ and go to (2)*

This algorithm only works if $d^{(k)}$ exists. But what about the convergence property?

Theorem 2.7.1 (Convergence behavior of the local Newton's method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be two-times continuously differentiable. Let $x^* \in \mathbb{R}^n$ be a stationary point of f and $\nabla^2 f(x^*)$ be regular.*

Then there exists $\varepsilon > 0$ such that for all $x^{(0)} \in U_\varepsilon(x^*)$, the following properties are true:

1. *The local Newton's method is well-defined and* $\left\{x^{(k)}\right\} \to x^*$.
2. *The convergence rate is superlinear (wrt. Q-convergence).*
3. *If* $\nabla^2 f$ *is locally Lipschitz-continuous, then the convergence rate is quadratic (wrt. Q-convergence).*

Then neighborhood of x^* *can be very small, thus the name* local *Newton's method was chosen.*

Revision (via chapter 5)**.**

Lemma (A)**.** $f : \mathbb{R}^n \to \mathbb{R}$. $\left\{x^{(k)}\right\} \to x^*$.

- f *is two-times differentiable. Then*
$$\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})(x^{(k)} - x^*)\right\| = o(\left\|x^{(k)} - x^*\right\|)$$
- f *is two times differentiable and* $\nabla^2 f$ *is locally Lipschitz continuous,*
$$\left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^{(k)})(x^{(k)} - x^*)\right\| = O(\left\|x^{(k)} - x^*\right\|^2)$$

Lemma (B)**.** *Let* f *be two-times differentiable.* $x^* \in \mathbb{R}^n$. $\nabla^2 f(x^*)$ *regular. Then there exists* $\varepsilon > 0 : \nabla^2 f(x)$ *regular in* $U_\varepsilon(x^*)$ *and* $\exists c > 0$.
$$\left\|\left(\nabla^2 f(x)\right)^{-1}\right\| \leq c \qquad \forall x \in U_\varepsilon(x^*)$$

Proof of theorem 2.7.1. Lemma B $\implies \exists \varepsilon_i > 0$ and $\exists c > 0$ such that $\nabla^2 f(x)$ is regular $\forall x \in U_{\varepsilon_1}(x^*)$ and $\left\|(\nabla^2 f(x))^{-1}\right\| \leq c$.

Lemma A
$$\implies \exists \varepsilon_2 > 0 : \left\|\nabla f(x) - \nabla f(x^*) - \nabla^2 f(x)(x - x^*)\right\| \leq \frac{1}{2}\|x - x^*\| \quad \forall x \in X_{\varepsilon_2}(x^*) \qquad (**)$$

Choose $\varepsilon := \{\varepsilon_1, \varepsilon_2\}$ and choose $x^{(0)} \leq U_\varepsilon(x^*)$. Hence $x^{(1)}$ is well-defined.
$$\begin{aligned}
\left\|x^{(1)} - x^*\right\| &= \left\|x^{(0)} - \left(\nabla^2 f(x^{(0)})\right)^{-1} \nabla f(x^{(0)}) - x^*\right\| \\
&\leq \left\|\underbrace{\left(\nabla^2 f(x^{(0)})\right)^{-1}}_{\text{by } (**)}\right\| \underbrace{\left\|\nabla f(x^{(0)}) - \nabla f(x^*) - \nabla^2 f(x^{(0)})(x^{(0)} - x^*)\right\|}_{\leq \frac{1}{2c}\left\|x^{(0)} - x^*\right\|} \\
&\leq c\frac{1}{2c}\left\|x^{(0)} - x^*\right\| = \frac{1}{2}\left\|x^{(0)} - x^*\right\| \\
&\implies x^{(1)} \in U_\varepsilon(x^*)
\end{aligned}$$

By complete induction we can show that $x^{(k)} \in U_\varepsilon(x^*) \forall k \in \mathbb{N}$
$$\left\|x^{(k)} - x^*\right\| \leq \left(\frac{1}{2}\right)^k \left\|x^{(0)} - x^*\right\|$$

Hence $\left\{x^{(k)}\right\}$ is well-defined and $\left\{x^{(k)}\right\} \to x^*$. By $\nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0$, Theorem 2.7.1 (2) and (3) follow from results in chapter 5. □

Remark. *For the local convergence result the assumption $\nabla^2 f(x^*)$ regular (not positive-definite) sufficed..*

Thus a case might occur such that x^* represents a local maximum. In the Global Newton's method, this (undesired) effect must not occur.

2.7.2 Global Newton's method

Well-definedness of the local Newton's method is just a local property.

Algorithm (Global Newton's method)**.**

1. *Choose $x^{(0)} \in \mathbb{R}^n, \rho > 0, p > 2, \beta \in (0,1), \sigma \in \left(0, \frac{1}{2}\right), \varepsilon \geq 0$. Let $k := 0$.*
2. *If $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, then stop.*
3. *Determine (if possible) some solution $d^{(k)}$ of*
$$\nabla^2 f(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$
If this system is not solvable or condition $\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -\rho \left\|d^{(k)}\right\|^p$ ("descent of the Newton's method is insufficient for us") is not satisfied, then
$$d^{(k)} := -\nabla f(x^{(k)})$$
4. *Determine $t_k = \max\left\{\beta^l \mid l = 0, 1, 2, \ldots\right\}$ with (Armijo step size strategy)*
$$f(x^{(k)} + t_k d^{(k)}) \leq f(x^{(k)}) + \sigma \cdot t_k \cdot \left(\nabla f(x^{(k)})\right)^t d^{(k)}$$
5. *Let $x^{(k+1)} := x^{(k)} + t_k \cdot d^{(k)}$ and $k := k + 1$. Go to (2)*

Remark. *It is trivial to see that this algorithm is well-defined considering our results in previous sections.*

What about convergence behavior?

Theorem 2.7.2. *If f is two-times continuously differentiable. Every cluster point, resulting from sequence $\left\{x^{(k)}\right\}$ of the Global Newton's method, is a stationary point of f.*

> ↓ *This lecture took place on 2019/06/06.*

Revision. *Local Newton's method:*

- *Useful only in theory*
- *Always uses Newton direction with step size 1*

Global Newton's method:

- *Uses the Newton direction if it exists and is satisfactory,*
- *otherwise direction of steepest descent.*
- *Step size choice by Armijo's rule (in every case)*

Proof of Theorem 2.7.2. Let $\tilde{x} \in \mathbb{R}^n$ be a cluster point of $\left\{x^{(k)}\right\}$. Let $\left\{x^{(k)}\right\}_k$ be a converging subsequence of $\left\{x^{(k)}\right\}$ towards $\tilde{x}$. If $d^{(k)} = -\nabla f(x^{(k)})$ for infinitely many $k \in K$, then the claim immediately follows from previous conclusions.

Assume that wlog. $d^{(k)}$ is the solution of the Newton equation $\nabla^2 f(x^{(k)})d^{(k)} = -\nabla f(x^{(k)}) \forall k \in K$.

Assume that $\tilde{x}$ is not a stationary point. Thus $\nabla f(\tilde{x}) \neq 0$.

We have $\left\|\nabla f(x^{(k)})\right\| = \left\|\nabla^2 f(x^{(k)})d^{(k)}\right\| \leq \left\|\nabla^2 f(x^{(k)})\right\| \left\|d^{(k)}\right\| \forall k \in K$.

$$\Longrightarrow \left\|d^{(k)}\right\| \geq \frac{\left\|\nabla f(x^{(k)})\right\|}{\left\|\nabla^2 f(x^{(k)})\right\|}$$

Remark. $\left\|\nabla^2 f(x^{(k)})\right\| \neq 0$. *Otherwise it does not fit our requirements.*

Now we can consider that constants $d_1, d_2 > 0$ exist such that

$$0 < d_1 \leq \left\|d^{(k)}\right\| \leq d_2 \qquad \forall k \in K \tag{10}$$

But do these constants exist? Regarding d_1, assume there exists a subsequence of K with $\left\|d^{(k)}\right\| \to 0$ for this subsequence. Thus as a consequence $\left\|\nabla f(x^{(k)})\right\| \to 0$ for this subsequence (consider that $\left\|\nabla^2 f(x^{(k)})\right\|$ is bounded). Then $\tilde{x}$ would be a stationary point of f opposing our assumption. Regarding d_2, we consider the Cauchy-Schwarz inequality. Thus $\left\|d^{(k)}\right\|$ cannot be unbounded by above. Hence, the constants exist.

Furthermore, we have that:

- $\left\{f(x^{(k)})\right\}$ is monotonically decreasing
- $\left\{f(x^{(k)})\right\}_K \to f(\tilde{x})$

By these two properties, $\left\{f(x^{(k)})\right\} \to f(\tilde{x})$. Thus $\left\{f(x^{(k+1)}) - f(x^{(k)})\right\} \to 0$. We consider the requirement of the global Newton's method:

$$\nabla f(x^{(k)})d^{(k)} \leq -\rho \left\|d^{(k)}\right\|^p .$$

Using this and the step size choice (Armijo), we get that

$$\left\{t_k \left(\nabla f(x^{(k)})\right)^t d^{(k)}\right\} \to 0$$

Now we show that $\{t_k\}_K$ is bounded from 0 (i.e. $\exists c : t_k \leq c < 0$). Assume there exists some subsequence $\tilde{k}$ of K with $\{t_k\}_{\tilde{K}} \to 0$. By Armijo's rule $t_k = \beta^{l_k}$ with

$l_k \in \mathbb{N}$ uniquely determined. For β^{l_k-1}, the Armijo condition is not yet satisfied. Let k be sufficiently large,

$$\frac{f(x^{(k)} - \beta^{l_k-1} d^{(k)}) - f(x^{(k)})}{\beta^{l_k-1}} > \sigma \left(\nabla f(x^{(k)})\right)^t d^{(k)} \tag{11}$$

By (10), we can assume that $\left\{d^{(k)}\right\}_K \to \tilde{d}$ with $\tilde{d} \neq 0$. By (11) and the Lemma **??** from section 2.6, we can conclude that $(\nabla f(\tilde{x}))^t \tilde{d} \geq \sigma \left(\nabla f(\tilde{x})\right)^t \tilde{d}$ for $k \to \infty$, $k \in K$. By $\sigma \in \left(0, \frac{1}{2}\right)$, we can conclude

$$\left(\nabla f(\tilde{x})\right)^t \tilde{d} \geq 0$$

Direction as in Newton's method is chosen only if

$$\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq \underbrace{-\rho \left\|d^{(k)}\right\|^p}_{<0}$$

$$\implies \left(\nabla f(\tilde{x})\right)^t \tilde{d} < 0$$

This gives a contradiction with our assumption regarding $\{t_k\}$.

Hence,

$$\exists \tau > 0 : t_k \geq \tau \qquad \text{for } k \in K$$

Because

$$\left\{t_k \left(\nabla f(x^{(k)})\right)^t d^{(k)}\right\}_K \to 0 \implies \left\{\left(\nabla f(x^{(k)})\right)^t d^{(k)}\right\}_K \to 0$$

By our assumption, we get $\left\{d^{(k)}\right\} \to 0$ which contradicts with above. □

Remark (Question 1)**.** *Does sequence $\left\{x^{(k)}\right\}$ converge or only its subsequence?*

Remark (Question 2)**.** *What about convergence speed?*

Lemma 2.7.3. *Let $\tilde{x} \in \mathbb{R}^n$ be an isolated cluster point of sequence $\left\{x^{(k)}\right\}$, $x^{(k)} \in \mathbb{R}^n$ ($x^{(k)}$ must not originate from Newton's method) with $\left\{\left\|x^{(k+1)} - x^{(k)}\right\|\right\}_K \to 0$ for every subsequence $\left\{x^{(k)}\right\}_K$ converging to $\tilde{x}$. Then $\left\{x^k\right\}$ converges to $\tilde{x}$.*

This lemma goes back to Moré and Sóren.

Proof. Left as an exercise. □

Theorem 2.7.4. *Let $\left\{x^{(k)}\right\}$ be generated from Global Newton's method and let $\tilde{x}$ be an isolated cluster point of $\left\{x^{(k)}\right\}$. Then $\left\{x^{(k)}\right\}$ converges to $\tilde{x}$.*

Proof. Let $\left\{x^{(k)}\right\}_K$ be a subsequence converging to an isolated cluster point $\tilde{x}$. By Theorem 2.7.2, $\tilde{x}$ is a stationary point. By continuity of f,

$$\left\{\nabla f(x^{(k)})\right\}_K \to \nabla f(\tilde{x}) \underbrace{=}_{\tilde{x} \text{ is stationary point}} 0 \tag{12}$$

Because $t_k \in (0,1]$, we have

$$\left\|x^{(k+1)} - x^{(k)}\right\| = t_k \left\|d^{(k)}\right\| \leq \left\|d^{(k)}\right\| \forall k \in \mathbb{N} \tag{13}$$

By the algorithm, we have

$$\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -\rho \left\|d^{(k)}\right\|^p \tag{14}$$

By the application of the Cauchy-Schwarz inequality,

$$\rho \left\|d^{(k)}\right\|^p \leq -\nabla f(x^{(k)}) d^{(k)} \leq \left\|\nabla f(x^{(k)})\right\| \left\|d^{(k)}\right\|$$

If all search directions $d^{(k)}$ with $k \in K$ satisfy condition (14), then immediately

$$\left\{\left\|d^{(k)}\right\|\right\}_K \to 0$$

This remains true even if $d^{(k)} = -\nabla f(x^{(k)})$ for individual $k \in K$ (or all of them) by (12). By (13),

$$\Longrightarrow \left\{\left\|x^{(k+1)} - x^{(k)}\right\|\right\}_K \to 0$$

If we apply Lemma 2.7.3, the claim is proven. □

Lemma 2.7.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be two times differentiable and $\tilde{x} \in \mathbb{R}^n$ with $\nabla^2 f(\tilde{x})$ positive-definite. Then there exists some constant $\delta_1 > 0$ and $\delta_2 > 0$ with $\delta_2 \|d\|^2 \leq d^t \nabla^2 f(x) d \; \forall x \in \mathbb{R}^n$ with $\|x - \tilde{x}\| \leq \delta_1$ and $\forall d \in \mathbb{R}^n$.*

Some kind of "uniform positive definiteness of $\nabla^2 f$"

Remark. *Especially, Lemma 2.7.5 ensures that $\nabla^2 f$ in an appropriate neighborhood of $\tilde{x}$ is positive definite (even uniform!). The proof of this property is left as an exercise to the reader.*

Lemma 2.7.6. *Let f be two times continuously differentiable. Let $\tilde{x} \in \mathbb{R}^n$ with $\nabla f(\tilde{x}) = 0$ and $\nabla^2 f(\tilde{x})$ positive-definite. Let $\left\{x^{(k)}\right\}$ with $x^{(k)} \in \mathbb{R}^n$, $\left\{x^{(k)}\right\} \to \tilde{x}$ and $\left\{d^{(k)}\right\}$ be the sequence of the Newton directions, thus satisfying*

$$\nabla^2 f(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)}) \qquad d^{(k)} = -\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$$

Then, there exists $k_0 \in \mathbb{N}$ with

$$f(x^{(k)} + d^{(k)}) \leq f(x^{(k)}) + \sigma \left(\nabla f(x^{(k)})\right)^t d^{(k)}$$

for all $k \geq k_0$ for fixed $\sigma \in \left(0, \frac{1}{2}\right)$.

Thus, beginning with index k_0, the complete step size 1 is chosen if Armijo's rule is applied. The restriction $\sigma \in \left(0, \frac{1}{2}\right)$ is necessary for this kind of result.

Theorem 2.7.7 (Central convergence theorem for Global Newton's method)**.** *Let f be two-times differentiable. Let $\left\{x^{(k)}\right\}$ be a sequence generated by the Global Newton's method. Let $\tilde{x}$ be a cluster point of $\left\{x^{(k)}\right\}$ and let $\nabla^2 f(\tilde{x})$ be positive definite. Then*

1. *The sequence* $\left\{x^{(k)}\right\}$ *converges to* $\tilde{x}$ *and* $\tilde{x}$ *is a strict local minimum of* f.
2. *For all sufficiently large* $k \in \mathbb{N}$, $d^{(k)}$ *is the Newton direction (above a certain* k_0*, the direction of steepest descent will never be chosen)*
3. *For all sufficiently large* $k \in \mathbb{N}$, $t_k = 1$ *(complete step size) is chosen.*
4. $\left\{x^{(k)}\right\}$ *converges superlinear to* $\tilde{x}$.
5. *If* $\nabla^2 f$ *is locally Lipschitz cont., then* $\left\{x^{(k)}\right\}$ *converges quadratically to* $\tilde{x}$.

↓ *This lecture took place on 2019/06/13.*

Proof. **Ad (1)** By previous results, $\tilde{x}$ is a stationary point of f. Because $\left\{f(x^{(k)})\right\}$ is monotonically decreasing and $\left\{f(x^{(k)})\right\}_k \to f(\tilde{x}) \implies \left\{f(x^{(k)})\right\} \to f(\tilde{x})$. Thus every cluster point of $\left\{x^{(k)}\right\}$ provides the same function value $\tilde{x}$. Because $\nabla^2 f(\tilde{x})$ is positive definite, $\tilde{x}$ is a strict local minimum of f (compare with the introductory section). As a consequence, $\tilde{x}$ is an isolated cluster point of $\left\{x^{(k)}\right\}$. By Theorem 2.7.4, $\left\{x^{(k)}\right\} \to \tilde{x}$.

Ad (2) Because $\left\{x^{(k)}\right\} \to \tilde{x}$, $\nabla^2 f(x^{(k)})$ $\forall k$ sufficiently large. So $\exists \kappa \in \mathbb{N}$ with $\nabla^2 f(x^{(k)})$ positive definite $\forall k \geq \kappa$, hence Newton's direction is well-defined for all $k \geq \kappa$.

It remains to show that there exists a constant $\tilde{\rho} > 0$ such that $\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -\tilde{\rho} \left\|d^{(k)}\right\|^2$ for all sufficiently large k. Because $\nabla^2 f(\tilde{x})$ is positive definite and thus regular, a previous result gives $\exists \tilde{c} > 0$ constant such that

$$\left\|\left(\nabla^2 f(x^{(k)})^{-1}\right)\right\| \leq \tilde{c} \quad \forall k \text{ sufficiently large k}$$

$$\left\|d^{(k)}\right\| = \left\|\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})\right\| \leq \tilde{c} \left\|\nabla f(x^{(k)})\right\|$$

By Theorem 2.7.4, for some constant $\alpha \geq 0$

$$\begin{aligned} -\left(\nabla f(x^{(k)})\right)^t d^{(k)} &= +\left(\nabla f(x^{(k)})\right)^t \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)}) \\ &\geq \alpha \left\|\nabla f(x^{(k)})\right\|^2 \geq \frac{\alpha}{\tilde{c}^2} \left\|d^{(k)}\right\|^2 \end{aligned}$$

for all k sufficiently large. Let $\tilde{\rho} = \frac{\alpha}{\tilde{c}^2}$. Thus, the existence of $\tilde{\rho}$ has been shown.

Furthermore, $\left\|d^{(k)}\right\| \to 0$ and from the existence of $\tilde{\rho}$, we get $p > 2$.

$$\left(\nabla f(x^{(k)})\right)^t d^{(k)} \leq -\tilde{\rho} \left\|d^{(k)}\right\|^p \qquad \forall k \text{ sufficiently large}$$

Thus for all k sufficiently large, the Newton direction is chosen.

Ad (3) Proven by (1) and (2) and Lemma 2.7.6

Ad (4) and (5) We have shown that for sufficiently large k (i.e. in a corresponding neighborhood of $\tilde{x}$), the global method behaves just like the local one. Thus the local results apply and show these properties.

□

Remark (Concluding remarks). *1. For implementation of Newton's method (in particular computation of Newton's direction), Cholesky decomposition and some adapted variant play an important role.*

2. *In numerical experiments, it can be shown that (from $\tilde{x}$) farther distant initial points $x^{(0)}$ can even give faster convergence for local settings than global ones (with Armijo). There are variants of the global method trying to allow larger step sizes t_k (non-monotonic Armijo rule)*

3. *As generalization of the Newton's method, so-called* inexact *Newton's methods were proposed and researched. Instead of satisfying the Newton equation $\nabla^2 f(x^{(k)})d^{(k)} = -\nabla f(x^{(k)})$ accurately in every Newton step [difficult anyways, because of precision and computational requirements], approximation solutions are considered acceptable. To measure inaccuracy, the* relative error

$$\frac{\left\|\nabla^2 f(x^{(k)})d + \nabla f(x^{(k)})\right\|}{\left\|\nabla f(x^{(k)})\right\|}$$

Goal*: For given $\varepsilon_k \geq 0$, we want to determine $d^{(k)}$ such that*

$$\frac{\left\|\nabla^2 f(x^{(k)})d^{(k)} + \nabla f(x^{(k)})\right\|}{\left\|\nabla f(x^{(k)})\right\|} \leq \varepsilon_k$$

This is a computationally easier task than the exact solution.

2.8 Quasi-Newton methods

Instead of the Hessian matrix (might be difficult to determine or function is not given explicitly), an appropriate [to be discussed] approximation of the Hessian (or its inverse) is used. The sequence of approximations is generated by simple update rules (e.g. rank-1-updates, rank-2-updates). Every iteration of the Quasi-Newton process is computationally easier than an iteration of Newton's method.

The central question is: Can we transfer the nice properties of Newton's method? If so, under which assumptions.

$$x^{(k+1)} := x^{(k)} - \left(H^{(k)}\right)^{-1} \nabla f(x^{(k)}) \tag{15}$$

Idea: $H^{(k)}$ should approximate $\nabla^2 f(x^{(k)})$ sufficiently good [to be quantified].
Question: Which requirements shall be satisfied by $H^{(k)}$?

$$\left\|\left(\nabla^2 f(x^{(k)}) - H^{(k)}\right)\left(x^{(k+1)} - x^{(k)}\right)\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right) \tag{16}$$

The condition (16) is necessary and sufficient for superlinear convergence of sequence $\left\{x^{(k)}\right\}$ resulting from (15) under requirements

- $f : \mathbb{R}^n \to \mathbb{R}$ two-times differentiable
- $\left\{H^{(k)}\right\}$ is a sequence of regular $n \times n$ matrices
- $\left\{x^{(k)}\right\} \to x^*$ with $x^{(k)} \neq x^*$ for all $k \in \mathbb{N}$
- $\nabla^2 f(x^*)$ regular

Because superlinear convergence is our minimal goal, (16) is our goal for $(H^{(k)})$

Goal: Rewrite (16) such that $\nabla^2 f$ does not occur anymore.

$$\begin{aligned}
&\left\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)})\left(x^{(k+1)} - x^{(k)}\right)\right\| \\
&\leq \left\|\nabla f(x^{(k+1)}) - \nabla f(x^*) - \nabla^2 f(x^*)(x^{(k+1)} - x^*)\right\| \\
&+ \left\|\nabla f(x^{(k)}) - \nabla f(x^*) - \nabla^2 f(x^*)\left(x^{(k)} - x^*\right)\right\| \\
&+ \left\|\left(\nabla^2 f(x^*) - \nabla^2 f(x^{(k)})\left(x^{(k+1)} - x^{(k)}\right)\right)\right\|
\end{aligned}$$

Thus $\left\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - \nabla^2 f(x^{(k)})\left(x^{(k+1)} - x^{(k)}\right)\right\| = o\left(\left\|x^{(k+1)} - x^{(k)}\right\|\right)$

Equation (16) is equivalent to

$$\left\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - H^{(k)}\left(x^{(k+1)} - x^{(k)}\right)\right\| = o\left(\left(\nabla f(x^{(k)})\right)^t d^{(k)}\right) \tag{17}$$

The details will be discussed in the practicals. This is motivating requiring the following condition:

$$\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = H^{(k+1)}\left(x^{(k+1)} - x^{(k)}\right)$$

Remark. *Because $x^{(k+1)}$ occurs, $H^{(k+1)}$ must have influence in the condition and not $H^{(k)}$.*

This defines the Quasi-Newton condition (sometimes called *secant method*).

↓ *This lecture took place on 2019/06/17.*

$$\underbrace{\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})}_{y} = \underbrace{H^{(k+1)}}_{H^+}\underbrace{\left(x^{(k+1)} - x^k\right)}_{s}$$

In the following, we use the following abbreviations:

$$H := H^{(k)} \qquad H^{+}s = y \qquad \text{(Quasi Newton Condition) (QNC)}$$

For given $y, s \in \mathbb{R}^n$ there are many possibilities to determine $H^{(k)}$ satisfying QNC.

Now we want to discuss the question how to choose $\left\{H^{(k)}\right\}$.

Which assumptions on $\left\{H^{(k)}\right\}$ or H^* in the QNC are made (or in general, on update rule from $H^{(k)}$ to $H^{(k+1)}$).

Common requirements:

- $H^{(k)}$ symmetric most of the time (except for few exceptions)
- QNC is satisfied
- computation requirements of update $H^{(k)}$ to $H^{(k+1)}$ is low. $H^{(k+1)}$ should not diverge too strong from $H^{(k)}$.
- the resulting Quasi-Newton Method (interpretable as Newton-like Method) shall provide useful convergence results (linear or superlinear)

In literature, a lot of different update rules (i.e. ways to design $\left\{H^{(k)}\right\}$) were investigated.

For example,

- symmetric rank-1 update formulas (update correction)
- non-symmetric rank-1 update formulas (here $H^{(k)}$ is exceptionally non-symmetric)
- versatile update formulas interpretable as rank-2 update formulas
- BFGS-formulas (with variables) (Broyden, Fletcher, Goldfarb, Shenno formula)
- DFP-formula (Davidon, Fletcher, Powell formula)
- PSB-formula (Powell-symmetric Broyden formula)

Remark (Motivation of QNC). *1. With Moré condition (compare with last lecture)*

2. *By* $\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = \int_0^1 \left(\nabla^2 f(x^{(k)} + t\left(x^{(k+1)} - x^{(k)}\right)\right) dt \left(x^{(k+1)} - x^{(k)}\right)$ *where* $M(x^{(k)}, x^{(k+1)}) := \left(\nabla^2 f(x^{(k)} + t\left(x^{(k+1)} - x^{(k)}\right)\right)$ *is called* mean value matrix *we get that* $M(x^{(k)}, x^{(k+1)})$ *satisfies the QNC. In the special case of a quadratic function* $f(x) = \frac{1}{2}x^t Q x + c^t x + \gamma$*, the Hessian matrix satisfies the QNC. In general, the Hessian matrix does* not *satisfy the QNC.*

Remark (General layout for the remaining parts of this chapter). *(mostly no proof will be provided)*

update formulas *The Frobenius norm will play an important role. Let A be a $n \times n$ matrix over $\mathbb{R}$*

$$\|A\|_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2}$$

In some way, this is analogous to then Euclidean vector norm. No induced norm.

resulting methods *(focus on local ones, short remarks on global methods, step size choice with Wolfe-Powell)*

convergence behavior

2 auxiliary results

Lemma 2.8.1.

$$\forall v \in \mathbb{R}^n : \|v\| = \max_{\|x\|=1} \left| v^t x \right|$$

Lemma 2.8.2.

$$\forall v, w \in \mathbb{R}^n : \left\| v w^t \right\| = \|v\| \, \|w\|$$

Theorem 2.8.3 (PSB formula)**.** *Let $s, y \in \mathbb{R}^n$ with $s \neq 0$ given and let H be a symmetric $n \times n$ matrix over $\mathbb{R}$. The unique solution on $\min \left\| H^+ - H \right\|_F^2$ such that $H^t s = y$ (H^+ satisfies QNC) and $(H^+)^t = H^+$ (H^+ is symmetric) is given by*

$$H_{PSB}^+ = H + \left(\frac{(y - Hs)s^t + s(y - Hs)^t}{s^t s} \right) + \left(\frac{(y - Hs)^t s}{(s^t s)^2} s s^t \right) \qquad \text{“PSB formula”}$$

Proof sketch. 1. Our target function is strictly convex (actually uniformly convex). If a solution exists, the solution is unique.

2. In the remaining parts of the proof, we need to show that H_{PSB}^+ is admissible for our optimization problem. So, we show that H_{PSB}^+ is symmetric and satisfies QNC (quite easy to see, compare with practicals).

   About the optimal solution: Let A be any $n \times n$ matrix with $As = y$ (QNC) and $At = A$ (symmetric). Because H_{PSB}^+. Because H_{PSB}^+ satisfies QNB, we have that

$$(H_{\text{PSB}}^+ - H)s = y - Hs = (A - H)s$$

3. Let $u \in \mathbb{R}^n$ with $s^t u = 0$.

$$\begin{aligned}
\left\|(H_+^{\mathrm{PSB}} - H)u\right\| &= \left\|\left(\frac{(y-Hs)s^t + s(y-Hs)^t}{s^t s}\underline{u} - \frac{(y-Hs)^t s}{(s^t s)^2} s s^t u\right)\right\| \\
&= \left\|\frac{s(y-Hs)^t}{s^t s} u\right\| = \left\|\frac{ss^t}{s^t s}(A-H)u\right\| \\
&\leq \underbrace{\left\|\frac{ss^t}{s^t s}\right\|}_{=T} \left\|(A-H)u\right\| = \left\|(A-H)u\right\| \\
&\text{by Lemma 2.8.2: } \left\|\frac{ss^t}{s^t s}\right\| = 1
\end{aligned}$$

4. It remains to show that H_{PSB}^+ is optimal solution. Because $s \neq 0$, $v^{(1)} := \frac{s}{\|s\|}$ (normed). Extend $v^{(1)}$ by Gram-Schmidt process into a orthonormal basis $v^{(1)}, v^{(2)}, \ldots, v^{(k)}$ of $\mathbb{R}^n$. By basic linear algebra and (3), it follows that $\forall A : As = y, A^t = A$.

$$\begin{aligned}
\left\|H_{\mathrm{PSB}-H}^+\right\|_F^2 &= \sum_{i=1}^n \left\|(H_{\mathrm{PSB}}^+ - H)v^{(i)}\right\|^2 \\
&= \left\|(H_{\mathrm{PSB}}^+ - H)v^{(1)}\right\|^2 + \sum_{i=2}^n \left\|\left(H_{\mathrm{PSB}}^+ - H\right)v^{(i)}\right\|^2 \\
&\leq \left\|(A-H)v^{(1)}\right\|^2 + \left\|\sum_{i=2}^n (A-H)v^{(i)}\right\|^2 \\
&= \sum_{i=1}^n \left\|(A-H)v^{(i)}\right\|^2
\end{aligned}$$

□

Lemma 2.8.4. *Let A be a $n \times n$ matrix over $\mathbb{R}^n$ and let $v^{(1)}, \ldots, v^{(n)}$ be a ONB of $\mathbb{R}^n$. Then*

$$\|A\|_F^2 = \sum_{i=1}^n \left\|Av^{(i)}\right\|^2$$

Proof. Left as an exercise to the reader. □

A possible approach to derive further update formulas is given by a generalization of Theorem 2.8.3 to the weighted case.

Lemma 2.8.5. *Let $y, s \in \mathbb{R}^n$ with $s \neq 0$ and H is symmetric $n \times n$ matrix be given. Let W be a symmetric, positive-definite $n \times n$ matrix (*weight matrix*). Then uniquely determined solution of*

$$\min \left\|W(H^+ - H)W\right\|_F^2 \text{ s.t. } H^+ s = y \text{ and } (H^+)^t = H^+$$

is given by

$$H_W^+ := H + \frac{(y - Hs)\left(W^{-2}s\right)^t + W^{-2}s(y - Hs)^t}{(W^{-2}s)^t s} - \frac{s^t(y - Hs)W^{-2}s(W^{-2}s)^t}{\left((W^{-2}s)^t s\right)^2}$$

Remark: Denominator is non-zero.

Proof. Proof by reformulation to Theorem 2.8.3. Let $D := WHW$ and $D^+ := WH^+W$

$$\implies \left\|W(H^+ - H)W\right\| = \left\|D^+ - D\right\|_F$$

It holds true that $H^+s = y \iff D^+W^{-1}s = Ws$ and $(H^+)^t = H^+ \iff (D^+)^t = D^+$.

Thus, the following problem is equivalent to our original optimization problem:

$$\min \left\|D^+ - D\right\|_F^2 \text{ s.t. } D^+S^W = y^W \qquad (D^+)^t = D$$

with $s' = W^{-1}s$ and $y^W = Wy$. $s^W \neq 0$ because $s \neq 0$. Theorem 2.8.3 provides

$$D^+ = D + \frac{(y^W - Ds^W)(s^W)^t + s^W(y^W - Ds^W)^t}{(s^W)^t s^W} - \frac{(y^W - Ds^W)^t s^W}{\left((s^W)^t s^W\right)^2} s^W (s^W)^t$$

as the unique solution the our equivalent optimization problem.

Substitution gives

$$\begin{aligned} WH^+W = WHW &- \frac{(WY - WHs)(W^{-1}s)^t + W^{-1}s(Wy - WHs)^t}{\left(W^{-1}s\right)^t \left(W^{-1}s\right)} \\ &- \frac{(Wy - WHs)^t(W^{-1}s)}{\left((W{-}1s)^t \left(W^{-1}s\right)\right)^2}(W^{-1}s)(W^{-1}s)^t \end{aligned}$$

from left and right we multiply with W^{-1}. This gives us the claimed result. □

Lemma 2.8.6. *Let $y, s \in \mathbb{R}^n$ with $s \neq 0$. A symmetric positive-definite $n \times n$ matrix Q with $Qs = y$ if $s^t y > 0$.*

Proof. Left as an exercise to the reader. □

Theorem 2.8.7 (DFP formula)**.** *Let H be a $n{\times}n$ symmetric and positive definite matrix. Let $s, y \in \mathbb{R}^n$ with $s^t y > 0$. Let Q be a symmetric positive definite matrix satisfying $Qs = y$ (QNC) (exists due to Lemma 2.8.6) and let $W := Q^{-\frac{1}{2}}$ (square root exists due to positive definiteness).*

Then the unique solution of the weighted problem of Lemma 2.8.5 is given by

$$H_{DFP}^+ = H + \frac{(y - Hs)y^t + y(y - Hs)^t}{y^t s} - \frac{(y - Hs)^t s}{(y^t s)^2} y y^t$$

Proof. Results directly from the results before. □

↓ *This lecture took place on 2019/06/24.*

Today, we want to close the topic of Quasi-Newton methods. Last time, we derived the DFP-formula with a generalized result (is based on ideas by Davidson ~1950, journal version published in 1991).

$$H^+_{\text{DFP}} = H + \frac{(y - Hs)y^t + y(y - Hs)^t}{y^t s} - \frac{(y - Hs)^t s}{(y^t s)^2} yy^t$$

with H^+ for $H^{(k+1)}$, H for $H^{(k)}$, s for $s^{(k)}$ and y for $y^{(k)}$.

Now we consider the BFGS formula (Broyden, Fletcher, Goldfarb, Shanno):

Also here, we have different approaches for derivation (one approach is the interpretation as rank-2 updates, compare with the practicals).

$$H^+ = H + \alpha uu^t + \beta vv^t$$

Here we take an approach similar to the DFP-formula.

Actually, two variants of the BFGS formula are in use. Sometimes the second variant is called *inverse BFGS formula*. One approach uses the approximation of the Hessian matrix. In the other, the approximation of the inverse of the Hessian matrix (also possible for the other Quasi-Newton formulas). In the following, we use B instead of H (and accordingly, $\left\{B^{(k)}\right\}$ instead of $\left\{H^{(k)}\right\}$ and accordingly B^+ instead of H^+) in the context of approximation of the inverse Hessian matrix.

Quasi-Newton equation (in the H-variant):

$$H^+ s = y$$

B-variant:

$$B^+ s = y$$

The following lemma results from Lemma 2.8.5 and exchanging s and y:

Lemma 2.8.8. *Let $s, y \in \mathbb{R}^n$ with $y \neq 0$ and let B be a symmetric $n \times n$ matrix over $\mathbb{R}$. Let W be a symmetric positive-definite matrix over $\mathbb{R}$. Then the uniquely determined solution*

$$\min \left\| W(B^+ - B)W \right\|_F \text{ such that}$$

- $B^+ y = s$ *(inverse form of QNB)*
- $(B^+)^t = B^+$ *(symmetric)*

be given by

$$B^+_W = B + \frac{(s - By)(W^{-2}y)^t + W^{-2}y(s - By)^t}{(W^{-2}y)^t y} - \frac{y^t(s - By)W^{-2}y(W^{-2}y)^t}{((W^{-2}y)^t y)^2}$$

Proof. Analogous to Lemma 2.8.5. □

By proper choice of W, we get

Theorem 2.8.9. *Let B be a symmetric, positive-definite $n \times n$ matrix over $\mathbb{R}$ and $s, y \in \mathbb{R}^n$ with $s^t y > 0$. There exists a symmetric, positive-definite matrix Q with $Qs = y$ and $W := Q^{\frac{1}{2}}$. The unique solution of the minimization problem of Lemma 2.8.8 is given by*

$$B^+_{BFGS} = B + \frac{(s - By)s^t + s(s - By)^t}{y^t s} - \frac{(s - By)^t y s s^t}{(y^t s)^2}$$

so-called inverse form of the BFGS-formula.

Remark.

1. *The condition $s^t y > 0$ ensures that if B is positive definite then B^+ is positive definite (thus direction $d^{(k)}$ is the direction of descent).*
2. *Direct form of the BFGS-update formula is given by*

$$H^+_{BFGS} = H + \frac{yy^t}{s^t y} - \frac{Hss^t H}{s^t H s}$$

   with H as $n \times n$ symmetric, positive-definite $s, y \in \mathbb{R}^n$ with $s^t y > 0$. H^+_{BFGS} is the inverse of the matrix B^+_{BFGS} if H is the inverse of B.
3. *As inverse form of the DFP update formula, we get*

$$B^+_{DFP} = B + \frac{ss^t}{s^t y} - \frac{Byy^t B}{y^t B y}$$

   with B as $n \times n$ symmetric, positive-definite $s, y \in \mathbb{R}^n$ with $s^t y > 0$.
4. *Sometimes BFGS is used as the name for the inverse variant.*

Other update formulas

- symmetric rank-1 update (SR1) (compare with the practicals)
- asymmetric rank-1 update formula (here H is not a symmetric matrix)

The approach for SR1 is $H^+ = H + \alpha \cdot uu^t$ with $\alpha \in \mathbb{R}, u \in \mathbb{R}^n$ leads to (to some extent) unique solution for given α and u (compare with the practicals). Disadvantage: H^+ is, in general, not positive definite (but is not necessarily a direction of descent). This gives a problem if $y - Hs = 0$ or close to 0 (as a numerical problem).

The asymmetric variant considers

$$H^+ = H + \alpha u v^t \qquad \alpha \in \mathbb{R}^n, u, v \in \mathbb{R}$$

With $v = s$ this results in a solution for u, α from the remaining parts. Similar disadvantages like SR1.

For each update formula, we get one corresponding variant of the local Quasi-Newton method (for SR1-updates compare with the practicals).

0. Choose $x^{(0)} \in \mathbb{R}^n$. $H^{(0)}$ (or $B^{(0)}$ in the inverse variant) $m \times n$ (in general symmetric) over $\mathbb{R}$. $\varepsilon \geq 0$. Let $k := 0$.
1. If $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, then stop.
2. Determine $d^{(k)} := -\left(H^{(k)}\right)^{-1} \nabla f(x^{(k)})$ (analogously for $B^{(k)}$) (solve with equation system)
3. Let $x^{(k+1)} := x^{(k)} + d^{(k)}$ and $s^{(k)} := x^{(k+1)} - x^{(k)}$ and $y^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$ and $H^{(k+1)}$ or $B^{(k+1)}$ dependending on the formula.

Remark. *Always choose step size* 1*, no step size method. In this sense, we discuss local Quasi-Newton methods.*

Remark (Major questions)**.** *1. Depending on the update formula, what can be said about convergence behavior in theory?*

2. What is the behavior in practice?

Results for the first question are more difficult to show than results for the gradient method and Newton's method. We won't discuss details here. Only for illustration, we are going to mention results without proof. For example, for PSB.

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be two times differentiable. Let $\nabla^2 f$ be locally Lipschitz continuous and $x^* \in \mathbb{R}^n$ with $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ regular. Then $\exists \hat{\varepsilon} > 0$, $\hat{\delta} > 0$ such that the local PSB algorithm (local QN-algorithm with PSB update formula) for every initial vector $x^{(0)} \in \mathbb{R}^n$ with $\left\|x^{(0)} - x^*\right\| < \hat{\varepsilon}$ and every symmetric $n \times n$ initial matrix $H^{(0)}$ over $\mathbb{R}$ with $\left\|H^{(0)} - \nabla^2 f(x^*)\right\|_F < \hat{\delta}$ is well-defined and some sequence $\left\{x^{(k)}\right\}$ is generated, that converges superlinearly to x^*.*

The first two sentences specify the same prerequisites like local Newton's method.

Remark (Analogous result for inverse BFGS variant)**.** *Let f be two-times differentiable and $\nabla^2 f$ be locally Lipschitz continuous. Let x^* with $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ be positive-definite. There exist $\hat{\varepsilon}, \hat{\delta} > 0$ with $\forall x^{(0)}$ with $\left\|x^{(0)} - x^*\right\| < \hat{\varepsilon}$ and $\forall B^{(0)}$ with $\left\|B^{(0)} - (\nabla^2 f(x^*))^{-1}\right\|_F < \hat{\delta}$. Thus we get superlinear convergence of $\left\{x^{(k)}\right\}$.*

BFGS is a well-defined method. An analogous result for direct BFGS exists.

Remark (Remark about the global Quasi-Newton method)**.** *Here the topic of step size strategies arises again.*

Goal: *Less dependence on the choice of $x^{(0)}, H^{(0)}$ and $B^{(0)}$.*

Remark (Illustrative global BFGS method)**.** *0. Choose $x^{(0)} \in \mathbb{R}^n$, $H^{(0)}$ as $n \times n$ symmetric positive-definite matrix over $\mathbb{R}$. $\sigma \in \left(0, \frac{1}{2}\right), \rho \in (\sigma, 1)$ (for Wolfe-Powell). $\varepsilon \geq 0$. Let $k := 0$*

1. If $\left\|\nabla f(x^{(k)})\right\| \leq \varepsilon$, then stop

2. *Determine* $d^{(k)} := -\left(H^{(k)}\right)^{-1} \nabla f(x^{(k)})$

3. *Determine* $t_k > 0$ *such that*

$$f(x^{(k)} + t_k \cdot d^{(k)}) \leq f(x^{(k)}) + \sigma t_k \left(\nabla f(x^{(k)})\right)^t d^{(k)} \qquad \textit{(WP1)}$$

$$\nabla f(x^{(k)} + t_k d^{(k)})^t d^{(k)} \geq \rho \left(\nabla f(x^{(k)})\right)^t d^{(k)} \qquad \textit{(WP2)}$$

4.

$$\begin{aligned}
x^{(k+1)} &:= x^{(k)} + t_k d^{(k)} \\
s^{(k)} &:= x^{(k+1)} - x^{(k)} \\
y^{(k)} &:= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\
H^{(k+1)} &:= H^{(k)} + \frac{y^{(k)} \left(y^{(k)}\right)^t}{\left(y^{(k)}\right)^t s^{(k)}} - \frac{H^{(k)} s^{(k)} \left(s^{(k)}\right)^t H^{(k)}}{(s^{(k)})^t H^{(k)} s^{(k)}} \\
k &:= k + 1
\end{aligned}$$

Go to (1).

For the implementation, detailed considerations are required to get an efficient, stable implementation.

Remark. *In practice, global variants are recommended (sometimes with different step size strategies). Experiences show that BFGS shall be preferred over DFP (because dependence on step size strategies is higher). BFGS with Wolfe-Powell is usually a proper choice for practice.*

Index