

TECHNISCHE UNIVERSITÄT BERLIN

Fakultät II – Institut für Mathematik

Nonlinear Optimization

Summer term 2022

Tobias Breiten

(Translation of the lecture notes of Dietmar Hömberg)

Inhaltsverzeichnis

1	Optimization problems	1
1.1	Literature	1
1.2	Notation, basic concepts and examples	1
1.3	Existence of solutions	4
1.4	Convex optimization problems	6
1.5	Some “classical” optimization problems	9
1.6	Numerical solution of optimization problems	12
1.7	Optimization software	12
1.7.1	Programming libraries	12
1.7.2	Interactive programming languages	12
2	Derivative free optimization methods	13
2.1	Simplex method by Nelder and Mead	13
2.1.1	Basic constructions	13
2.1.2	Steps of the algorithm	15
2.2	Mutation selection methods	18
3	Unconstrained problems – theory	19
3.1	Optimality conditions	19
3.1.1	First order necessary conditions	19
3.1.2	Second order necessary conditions	21
3.1.3	Second order sufficient conditions	22
3.2	Convex optimization problems	23
4	Unconstrained problems – numerical methods	27
4.1	Basics	27
4.2	Newton’s method	29
4.3	General descent methods – general properties	32
4.3.1	Efficient step sizes	32
4.3.2	Gradient related directions	33
4.3.3	General convergence results	35
4.4	Line search methods	37
4.4.1	Exact step size	37

4.4.2	Armijo step size	38
4.4.3	Powell step size	40
4.5	Gradient descent (steepest descent)	42
4.6	Damped Newton's method	45
4.6.1	The algorithm	45
4.6.2	Interpretation of Newton direction	45
4.6.3	Convergence of the method	46
4.7	Variable metric and quasi Newton methods	47
4.7.1	General procedure	47
4.7.2	Global convergence of variable metric methods	48
4.7.3	Quasi Newton methods	48
4.7.4	BFGS-Update	49
4.7.5	BFGS for quadratic problems	51
4.7.6	BFGS for general nonlinear problems	53
4.8	Conjugate direction methods	53
4.8.1	CG for quadratic optimization problems	53
4.8.2	Analysis of the CG method	55
4.8.3	Preconditioning	57
4.8.4	CG methods for general nonlinear optimization problems	57
4.9	Trust region methods	58
4.9.1	Motivation	58
4.9.2	Trust region Newton method	59
5	Constrained problems – theory	62
5.1	Introductory examples	62
5.2	Tangent cone and constraint qualifications	67
5.3	First order necessary optimality conditions	70
5.4	Proof of Theorem 5.3.1	72
5.5	Second order optimality conditions	75
5.6	Problems with box constraints	81
5.7	Further regularity conditions	83
5.8	Geometric interpretation of the necessary optimality conditions	85
5.9	Lagrange multipliers and sensitivity	87
5.10	Duality	88

5.11	Outlook: numerical solution of nonlinear optimization problems with constraints	93
6	Problems with linear constraints - methods	94
6.1	Quadratic optimization problems	94
6.1.1	Problems with equality constraints	94
6.1.2	Problems with inequality constraints - the active set method	98
6.2	Equality constraints, nonlinear objective	106
6.3	Inequality constraints, nonlinear objective	109
7	Problems with nonlinear constraints - methods	112
7.1	The Lagrange-Newton method	112
7.2	Sequential quadratic programming	114
8	Penalty, barrier and augmented Lagrangian methods	116
8.1	The quadratic penalty method	116
8.2	The logarithmic barrier method	119
8.3	Augmented Lagrangian methods	120

1 Optimization problems

1.1 Literature

These lecture notes closely follow the expositions from the book of Prof. Dr. Walter Alt, Universität Jena [1] (in particular Chapters 1-4, 6,7) and the book of Nocedal und Wright [2] (Chapters 5 and 8). The current version is an English translation of a draft of Fredi Tröltzsch with extensions and modifications from Dietmar Hömberg.

1. Alt, W., *Nichtlineare Optimierung*. Vieweg, Braunschweig/Wiesbaden 2002.
2. Nocedal, J. and Wright, S.J., *Numerical Optimization*. Springer, New York 2006.
3. Gill, P.E., Murray, W., and M.H. Wright, *Practical Optimization*. Academic Press, London 1981.
4. Kelley, C.T., *Iterative Methods for Optimization*. SIAM, Philadelphia 1999.
5. Spelluci, P., *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel 1993.
6. Luenberger, D.G., *Optimization by Vector Space Methods*. Wiley, 1969.
7. Luenberger, D.G., *Linear and Nonlinear Programming*. Addison Wesley, London 1984.
8. Großmann, C. und Terno, J., *Numerik der Optimierung*. Teubner-Verlag, Stuttgart 1993.
9. Moré, J.J. and Wright, S.J., *Optimization Software Guide*. SIAM, Philadelphia 1993.

1.2 Notation, basic concepts and examples

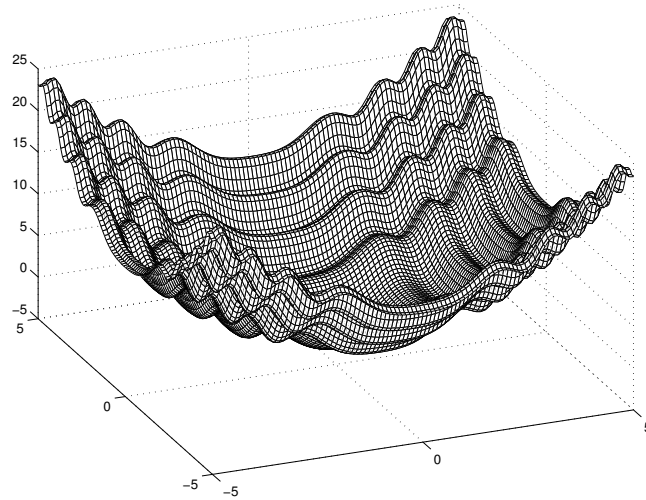
This course deals with the problem of characterizing and computing the minimum of a given function $f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ where D is open. Let us begin with some simple examples and introduce the notation used throughout these lecture notes:

Example 1.2.1

- $f(x) = x^2, f: \mathbb{R} \rightarrow \mathbb{R}$ has exactly one min in $\tilde{x} = 0$.
- $f(x) = x, f: \mathbb{R} \rightarrow \mathbb{R}$ is not bounded from below; the minimization problem does not have a solution.

Example 1.2.2

$f(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) - \cos(x_1^2) - \cos(x_2^2), f: \mathbb{R}^2 \rightarrow \mathbb{R}$ has a (strict) **global** minimum and multiple (strict) **local** minima and maxima.



Notation:

For $x \in \mathbb{R}^n$, we call

$\ x\ = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$	Euclidean norm
$B(x, r) = \{y \in \mathbb{R}^n \mid \ y - x\ < r\}$	open ball
$\overline{B}(x, r) = cl B(x, r)$	closed ball
x_i	i -th component of x
$x^{(k)}$	k -th iterate of a sequence of vectors $(x^{(k)})_{k=1}^\infty$

In what follows, let us be given a fixed, open set $D \subset \mathbb{R}^n$, a (sub)set $\Omega \subset D$ and a function $f: D \rightarrow \mathbb{R}$. We then consider the optimization problem

$$(P) \quad \boxed{\min_{x \in \Omega} f(x)}$$

We call f - **objective function**
 Ω - **admissible set** or feasible set

In case of $\Omega = D$, we call (P) an **unconstrained** oder **free optimization problem** (cf. Example 1.2.1). If Ω is characterized by additional constraints, (P) is called **constrained optimization problem**. Typically, we assume that Ω is described by equations and inequalities. The elements of Ω are called **admissible points**.

Example 1.2.3

$$\begin{array}{ll} \min_{x \in \mathbb{R}} & x^3 \\ \text{s.t.} & x \geq 1 \end{array}$$

This is an optimization problem with linear inequality constraints, $D = \mathbb{R}$, $\Omega = [1, \infty)$, the (unique) solution is $\tilde{x} = 1$.

Definition 1.2.1 A point $\tilde{x} \in \Omega$ is called

- **local minimum** of f in Ω or **local solution** of (P) , if $\exists r > 0$, s.t.

$$f(x) \geq f(\tilde{x}) \quad \forall x \in \Omega \cap B(\tilde{x}, r)$$

- *analogously* **strict local minimum**, if

$$f(x) > f(\tilde{x}) \quad \forall x \in \Omega \cap B(\tilde{x}, r), \quad x \neq \tilde{x}$$

- *analogously* **global minimum** or **global solution**, if

$$f(x) \geq f(\tilde{x}) \quad \forall x \in \Omega$$

- **strict global minimum** or **strict global solution**, if

$$f(x) > f(\tilde{x}) \quad \forall x \in \Omega, \quad x \neq \tilde{x}$$

Nonlinear optimization problems can have multiple local or global minima, e.g., $f(x) = \sin x$, $f(x) = x \sin\left(\frac{1}{x}\right)$.

Remark: We restrict ourselves to the case of minimizing f . If we search for maxima \tilde{x}

$$f(x) \leq f(\tilde{x}) \quad \forall x \in \Omega$$

we can use the equivalence $-f(x) \geq -f(\tilde{x})$ and instead minimize $\tilde{f} := -f$.

Example 1.2.4 (*Linear regression*)

We want to find an affine function

$$\eta(\xi) = x_1 \xi + x_2$$

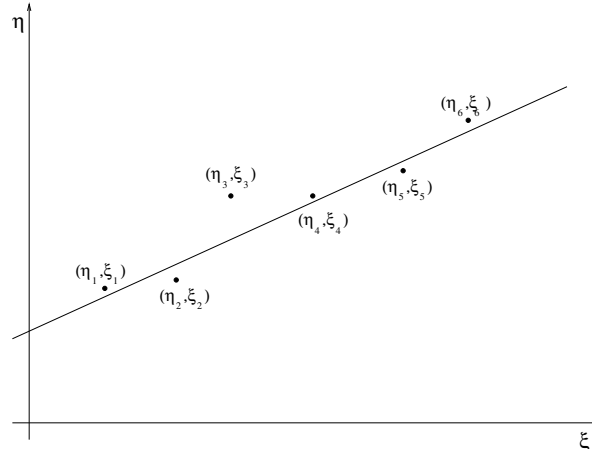
with coefficients x_1, x_2 to be determined such that it fits “optimally” with the data pairs $(\xi_i, \eta_i), i = 1, \dots, m$ (e.g., measurements). We set

$$\eta(\xi) = g(x_1, x_2, \xi) = x_1 \xi + x_2$$

and search for x_1, x_2 such that the objective function

$$\begin{aligned} f(x) = f(x_1, x_2) &= \sum_{i=1}^m (\eta_i - g(x_1, x_2, \xi_i))^2 \\ &= \sum_{i=1}^m (\eta_i - x_1 \xi_i - x_2)^2 \end{aligned}$$

is minimized. f is a polynomial of degree 2 in x_1, x_2 , i.e., a quadratic objective function.



Throughout this course, we will address the following problems:

- Existence and uniqueness of solutions
- Necessary optimality conditions
- Sufficient optimality conditions
- Numerical methods for solving optimization problems.

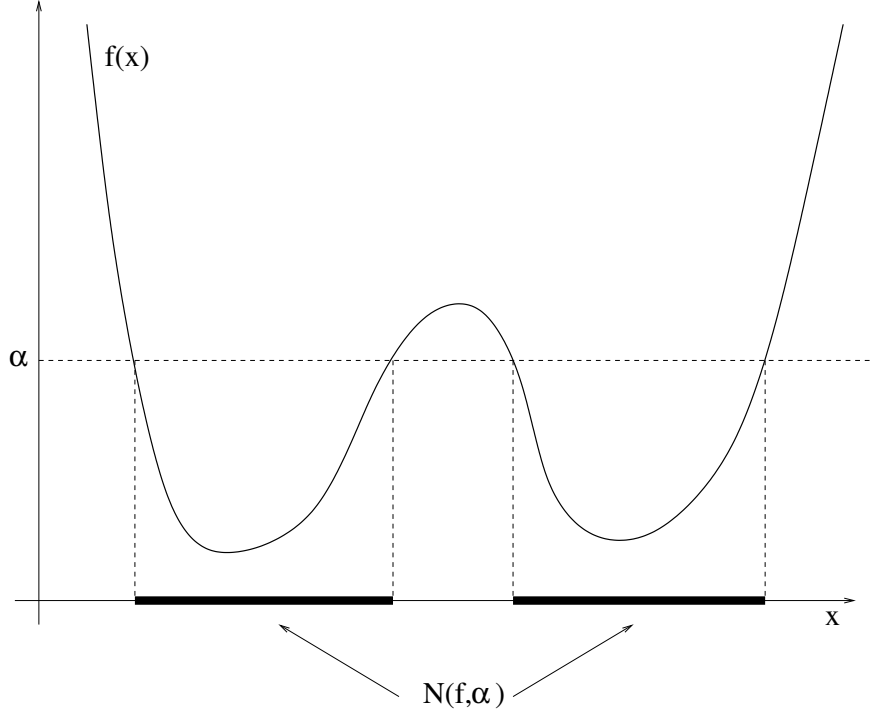
1.3 Existence of solutions

Many existence results are based on the well-known

Theorem 1.3.1 (*Weierstraß*)

If $f: \mathbb{R}^n \supset D \rightarrow \mathbb{R}$ is continuous and $K \subset D$ compact, then f attains its infimum (and its supremum) in K , i.e., there exists a global minimum (and maximum) of f in K .

Definition 1.3.1 $f: D \rightarrow \mathbb{R}, D \subset \mathbb{R}^n, \alpha \in \mathbb{R}$. The sets $N(f, \alpha) = \{x \in D \mid f(x) \leq \alpha\}$ are called **sublevel sets** of f .



Theorem 1.3.2 $D \subset \mathbb{R}^n$, $f : D \rightarrow \mathbb{R}$ continuous and $\Omega \subset D$ closed. For at least one $w \in \Omega$ the sublevel set

$$N(f, f(w)) = \{x \in D \mid f(x) \leq f(w)\}$$

is assumed to be compact. Then there exists at least one global minimum of f in Ω .

Proof: Let $\alpha = \inf_{x \in \Omega} f(x)$. It obviously holds that $\alpha \leq f(w)$. The set $\Omega \cap N(f, f(w))$ is compact and only within this set, we can find elements of Ω which are smaller or equal to $f(w)$. Hence

$$\alpha = \inf_{x \in \Omega \cap N(f, f(w))} f(x) = f(\tilde{x}),$$

where $\tilde{x} \in \Omega$ exists due to Theorem 1.3.1. □

As a (direct) consequence, we obtain the following result:

Corollary 1.3.1 D, Ω as in Theorem 1.3.2, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuous. Additionally, assume that f is **coercive**, i.e., it holds that

$$\lim_{\|x\| \rightarrow \infty} f(x) = \infty.$$

Then the minimization problem

$$\min_{x \in \Omega} f(x)$$

has at least one global solution.

Proof: Due to $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$ all sublevel sets $N(f, \alpha)$ are compact (exercise). The result then follows with Theorem 1.3.2. \square

A matrix $H \in \mathbb{R}^{n \times n}$ is called **positive semidefinite**, if $x^\top H x \geq 0$ for all $x \in \mathbb{R}^n$. It is called **positive definite**, if

$$x^\top H x > 0 \quad \forall x \in \mathbb{R}^n, x \neq 0$$

One can show that this is equivalent to the existence of $\alpha > 0$, s.t.

$$x^\top H x \geq \alpha \|x\|^2 \quad \forall x \in \mathbb{R}^n$$

(exercise). This obviously implies $x^\top H x \rightarrow \infty, \|x\| \rightarrow \infty$.

Example 1.3.1 (Unconstrained quadratic optimization)

Let us consider

$$(QU) \quad \min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^\top H x + b^\top x$$

with $b \in \mathbb{R}^n$ and positive definite $H \in \mathbb{R}^{n \times n}$. We obviously have that $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$. Due to Corollary 1.3.1, problem (QU) thus has at least one global solution.

Example 1.3.2 (Linear regression revisited)

Expanding the objective function yields

$$\begin{aligned} f(x) &= \sum_{i=1}^m (\eta_i - (x_1 \xi_i + x_2))^2 \\ &= \underbrace{\sum_{i=1}^m \eta_i^2}_c - 2 \underbrace{\sum_{i=1}^m \eta_i (\xi_i x_1 + x_2)}_{b^\top x} + \underbrace{\sum_{i=1}^m (x_1 \xi_i + x_2)^2}_{\frac{1}{2} x^\top H x} \\ &= \frac{1}{2} x^\top H x + b^\top x + c \\ \text{mit } H &= 2 \begin{pmatrix} \sum_{i=1}^m \xi_i^2 & \sum_{i=1}^m \xi_i \\ \sum_{i=1}^m \xi_i & m \end{pmatrix} \quad b = -2 \begin{pmatrix} \sum_{i=1}^m \xi_i \eta_i \\ \sum_{i=1}^m \eta_i \end{pmatrix}. \end{aligned}$$

If at least two ξ_i are different, H is positive definite (exercise). In this case, the linear regression problem is solvable. If all ξ_i are identical, the problem is not well-posed!

1.4 Convex optimization problems

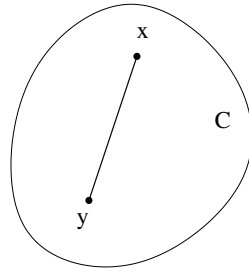
Keep in mind: convex optimization problems are nice !

For a detailed treatise of **convex analysis**, we refer to, e.g., Webster, R., Convexity. Oxford University Press 1994, or Rockafellar, R.T., Convex Analysis. Princeton University Press 1970).

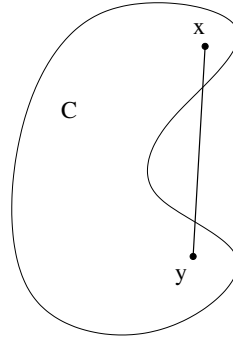
Definition 1.4.1 A set $C \subset \mathbb{R}^n$ is called convex if for all $x, y \in C$ the line segment

$$[x, y] = \{z = (1 - t)x + ty \mid 0 \leq t \leq 1\}$$

is also contained in C : $x, y \in C \Rightarrow [x, y] \subset C$.



convex set



non convex set

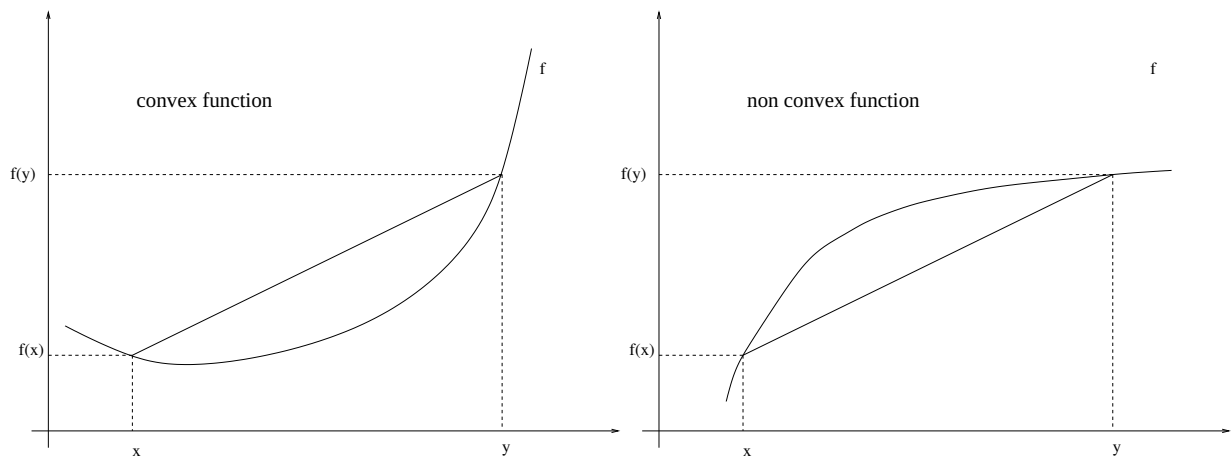
Definition 1.4.2 Let $C \subset \mathbb{R}^n$ be convex and non empty, $C \subset D$. A function $f: D \rightarrow \mathbb{R}$ is called convex on C , if

$$\boxed{f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y)} \quad \forall x, y \in C, \quad \forall t \in [0, 1]$$

If it holds that

$$f((1 - t)x + ty) < (1 - t)f(x) + tf(y) \quad \forall x, y \in C, x \neq y \\ \forall t \in]0, 1[,$$

f is called strictly convex on C .



Beispiel 1.4.1 $f(x) = x$ is convex, $f(x) = x^2$ strictly convex.

Consider $f: D \rightarrow \mathbb{R}, D \subset \mathbb{R}^n$ open, non empty, $\Omega \subset D$ convex. If f is convex on Ω , the problem

$$\min_{x \in \Omega} f(x) \quad (\text{P})$$

is a *convex optimization problem*.

Theorem 1.4.1 Let (P) be a convex optimization problem. Then every local solution of (P) is also a global solution. The set of all solutions of (P) is convex.

Proof:

(i) Let \tilde{x} be a local solution, i.e., there exists $r > 0$ s.t.

$$f(x) \geq f(\tilde{x}) \quad \forall x \in \Omega \cap B(\tilde{x}, r). \quad (*)$$

We have to show:

$$f(y) \geq f(\tilde{x}) \quad \forall y \in \Omega.$$

Consider (arbitrary) $y \in \Omega$ and define $\tilde{x} + t(y - \tilde{x})$ for sufficiently small $t > 0$.

We have $\tilde{x} + t(y - \tilde{x}) \in \Omega \quad \forall t \in [0, 1]$ since

$$\tilde{x} + t(y - \tilde{x}) = (1 - t)\tilde{x} + ty \in \Omega, \text{ since } \Omega \text{ is convex.}$$

We further obtain $\tilde{x} + t(y - \tilde{x}) \in B(\tilde{x}, r)$ for t sufficiently small, i.e., $t \in ([0, t_0], t_0 > 0)$. Thus with (*) we find

$$f(\tilde{x}) \underset{\substack{\uparrow \\ (*)}}{\leq} f((1 - t)\tilde{x} + ty) \leq (1 - t)f(\tilde{x}) + tf(y).$$

Rearranging terms yields $f(\tilde{x}) \leq f(y)$.

(ii) Now let us consider two solution x, \tilde{x} with $f(\tilde{x}) = f(x) = \tilde{\alpha} = \min f$. Then

$$f((1 - t)\tilde{x} + tx) \leq (1 - t)f(\tilde{x}) + tf(x) = \tilde{\alpha} = \min$$

\Rightarrow hence, $(1 - t)\tilde{x} + tx$ is another solution. □

Theorem 1.4.2 $D \subset \mathbb{R}^n, \Omega \subset D$ convex, $\Omega \neq \emptyset, f: D \rightarrow \mathbb{R}$ strictly convex. If (P) has a solution \tilde{x} , then \tilde{x} is uniquely determined and a strict global minimum of f in Ω .

Proof: Assume that x, y are two (global! see previous result) minima of f . Then

$$f(x) = f(y) = \alpha = \min_{x \in \Omega} f(x).$$

Assume we had $x \neq y$. Then $z = \frac{1}{2}(x + y)$ yields a lower value than α , since

$$f(z) = f\left(\frac{1}{2}x + \frac{1}{2}y\right) \underset{\substack{\leq \\ \uparrow \\ \text{strict convexity}}}{<} \frac{1}{2}f(x) + \frac{1}{2}f(y) = \frac{1}{2}\alpha + \frac{1}{2}\alpha = \alpha.$$

Moreover $z \in \Omega$ which contradicts the optimality of x, y . Hence $x = y$ is a strict minimum \square

Beispiel 1.4.2

$$f(x) = \frac{1}{2}x^\top Hx + b^\top x$$

If H is positive definite, then f is strictly convex (exercise).

1.5 Some “classical” optimization problems

Example 1.5.1 (Nonlinear regression)

In linear regression, we seek an affine function $\eta(\xi) = x_1\xi + x_2$. More generally, we can search for a nonlinear function η of ξ , given by the nonlinear ansatz

$$\eta(\xi) = g(x_1, x_2, \xi)$$

or

$$\eta(\xi) = g(x_1, \dots, x_n, \xi)$$

with an unknown vector $x \in \mathbb{R}^n$, e.g.,

$$g = x_1 e^{\xi x_2} + x_3.$$

\Rightarrow minimization of

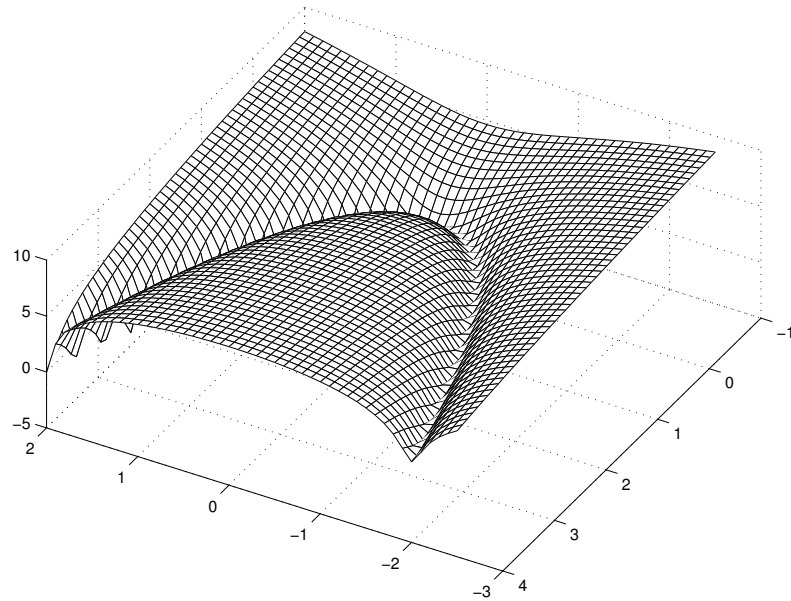
$$f(x) = \sum_{i=1}^m (\eta_i - g(x, \xi_i))^2$$

$$\text{Type:} \quad f(x) = \sum_{i=1}^m (f_i(x))^2 \quad f_i = \eta_i - g(\cdot, \xi_i).$$

Let us consider some pathological test example which are often used for testing algorithms.

Example 1.5.2 *Rosenbrock function („Banana shaped valley“)*

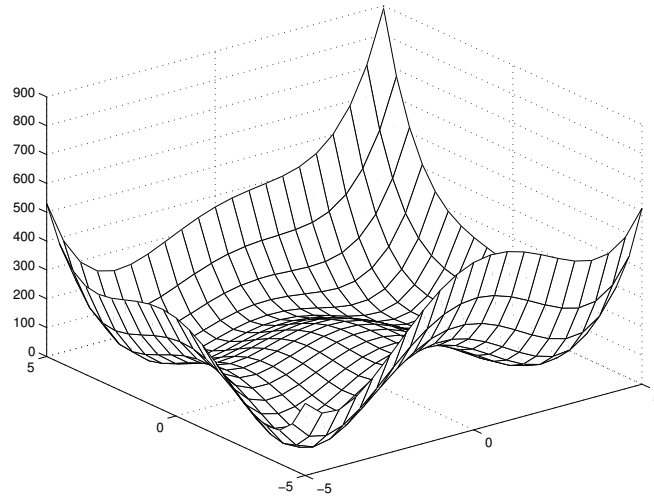
$$f(x_1, x_2) = \underbrace{100(x_2 - x_1^2)^2}_{\substack{\text{defines the} \\ \text{valley} \\ \text{(parabola)}}} + \underbrace{(1 - x_1)^2}_{\text{small tilt}}$$



Example 1.5.3 (*Himmelblau*)

$$f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

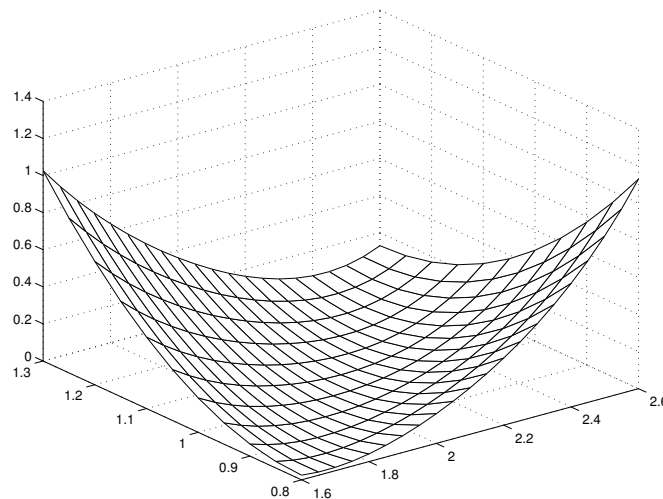
4 local minima which are also global minimum with objective value 0; 4 saddle points and a local maxima at $(-0.270845, -0.923039)^T$.



Example 1.5.4 (*Bazaraa-Shetty*)

$$f(x_1, x_2) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$$

Global minimum at $(2, 1)$. The Hessian is singular in $(2, 1)$ which can cause problems for numerical algorithms.



Example 1.5.5

$$f(x_1, \dots, x_5) = 2x_1^2 + 2x_2^2 + x_3^2 + x_4^2 + \frac{1}{2}x_5^2 - 4(x_1 + x_2) - 2(x_3 + x_4) - x_5 + 6.5$$

Global minimum at $\tilde{x} = (1, 1, 1, 1, 1)^\top$, $f(\tilde{x}) = 0$ (exercise).

Example 1.5.6 (*Dixon*)

$$f(x_1, \dots, x_{10}) = (1 - x_1)^2 + (1 - x_{10})^2 + \sum_{i=1}^9 (x_i^2 - x_{i+1})^2$$

Global minimum at $\tilde{x} = (1, \dots, 1)^\top$.

1.6 Numerical solution of optimization problems

In this course, we mainly discuss optimization methods which solve (P) numerically via an iterative methods - in some cases, these methods will find an exact solution in finitely many steps but generally, we aim for

$$\lim_{k \rightarrow \infty} x^{(k)} = \tilde{x}.$$

We will consider optimization problems with different properties (e.g., linear quadratic problems, nonlinear functionals with linear restrictions, general nonlinear problems but *linear* or *discrete* optimization problems).

1.7 Optimization software

1.7.1 Programming libraries

Recommended and available at TU Berlin:

- NAG-Library (Numerical Algorithms Group)
Fortran Codes
- minpack (public domain software)

1.7.2 Interactive programming languages

- MATLAB (MATrix LABoratory) commercial
- Scilab (SCientific, LABoratory) openly distributed via
INRIA, Paris
www.inria.fr

2 Derivative free optimization methods

Often the computation of the derivative of f is too expensive or – for non differentiable functions f – impossible such that methods which do not require derivative information have been suggested. Here, we briefly comment on two of them with the intention of solving the unconstrained problem

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

numerically.

2.1 Simplex method by Nelder and Mead

2.1.1 Basic constructions

Remark: The method is not related to the simplex algorithm known from linear programming!
Instead, it related to the

Definition 2.1.1 *Let $x^0, \dots, x^n \in \mathbb{R}^n$ be affine independent, i.e., $x^i - x^0$, $i = 1, \dots, n$ are linearly independent. The convex hull of x^0, \dots, x^n*

$$S = \left\{ \sum_{i=0}^n \lambda_i x^i \mid \lambda_i \geq 0, i = 0, \dots, n, \sum_{i=0}^n \lambda_i = 1 \right\}$$

is called (n -dimensional) simplex with nodes x^0, \dots, x^n .

- An initial simplex is provided at start.
- Find the vector with maximum objective value,

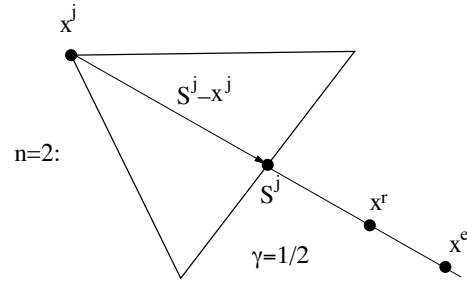
$$f(x^m) = \max \{f(x^0), \dots, f(x^n)\}$$

- Find a new vector with smaller objective value and replace x^m ersetzt.

This is done based on the following basic constructions:

Def.	$s^j = \frac{1}{n} \sum_{\substack{i=0 \\ i \neq j}}^n x^i$	centroid of the (remaining) nodes without x^j
-------------	---	---

Construction principles:



γ : reflection constant

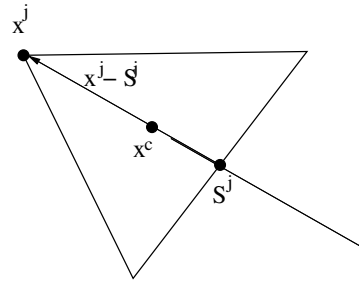
- **Reflection** of x^j at s^j

$$x^r = s^j + \gamma(s^j - x^j), \quad 0 < \gamma \leq 1$$

- The constructed x^r can be moved further “outwards”:

Expansion of x^r in direction of $s^j - x^j$ (i.e. in direction of $x^r - s^j$)

$$x^e = s^j + \beta(x^r - s^j), \quad \beta > 1 \quad \text{Expansion constant}$$

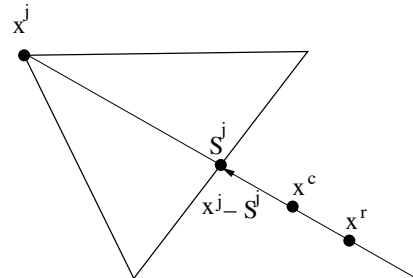


- **Contraction** (3 types)

(i) **Partial contraction (inside)** $x^c = s^j + \alpha(x^j - s^j)$ $0 < \alpha < 1$ contraction constant

(ii) **Partial contraction (outside)**

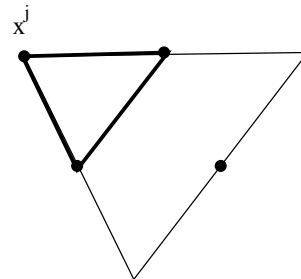
$$x^c = s^j + \alpha(x^r - s^j)$$



(iii) **Total contraction**

Replace all x^i except for x^j by

$$\hat{x}^i = x^i + \frac{1}{2}(x^j - x^i) = \frac{1}{2}(x^i + x^j)$$



2.1.2 Steps of the algorithm

The simplest variant is as follows:

Initially choose:

$\alpha \in (0, 1)$	contraction constant
$\beta > 1$	expansion constant
$\gamma \in (0, 1]$	reflection constant

We consider the following steps:

1. Choose an initial point $x^0 \in \mathbb{R}^n$, fix the remaining n nodes for construction of the initial simplex by

$$x^j = x^0 + e^j, j = 1, \dots, n,$$

where e^j denotes the j -th unit vector.

2. Compute the nodes with maximal and minimal objective value: x^m, x^l where

$$\begin{aligned} f(x^m) &= \max \{f(x^0), \dots, f(x^n)\} \\ f(x^l) &= \min \{f(x^0), \dots, f(x^n)\} \end{aligned}$$

and compute the centroid of the nodes without x^m

$$s^m = \frac{1}{n} \sum_{\substack{i=0 \\ i \neq m}}^n x^i$$

3. Reflection of x^m at centroid s^m

$$x^r = s^m + \gamma(s^m - x^m)$$

4. Construction of a new simplex

We distinguish the following cases

(i)

$$\boxed{f(x^r) < f(x^l)}$$

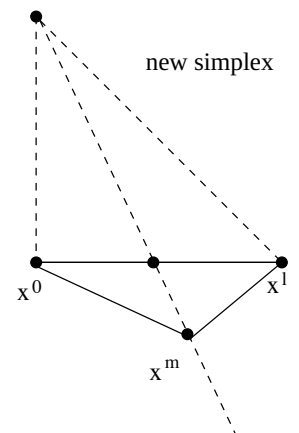
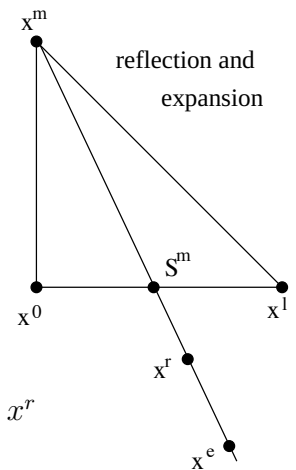
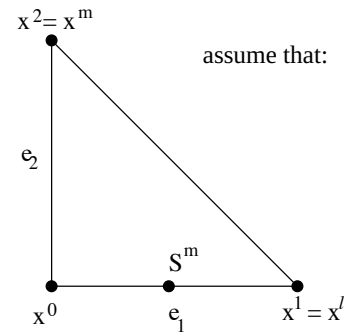
The direction was good so we try even more: expansion von x^r

$$x^e = s^m + \beta(x^r - s^m)$$

Replace x^m by the best of the two points:

$$\tilde{x}^m = \begin{cases} x^e, & f(x^e) < f(x^r) \\ x^r, & f(x^r) \leq f(x^e) \end{cases}$$

$$\underline{\underline{x^m := \tilde{x}^m}}$$



(ii)

$$\boxed{f(x^l) \leq f(x^r) \leq \max \{f(x^j), j \neq m\}}$$

Not bad but not good either – replace x^m by x^r

$$\underline{\underline{x^m := x^r}}$$

(iii)

$$\boxed{f(x^r) > \max \{f(x^j), j \neq m\}}$$

- If $f(x^r) \geq f(x^m)$: Partial contraction (inside)

$$x^c = s^m + \alpha(x^m - s^m)$$

- If $f(x^r) < f(x^m)$: Partial contraction (outside)

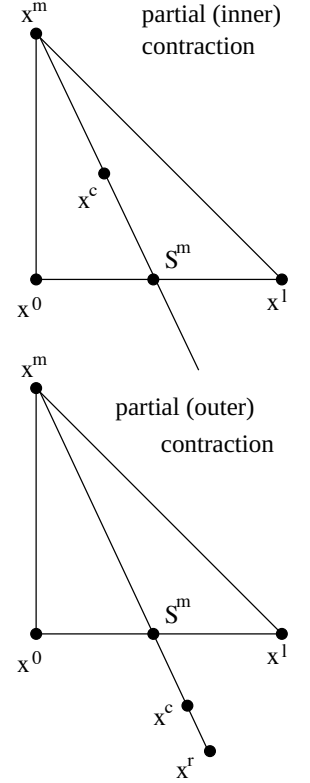
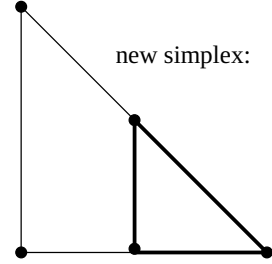
$$x^c = s^m + \alpha(x^r - s^m)$$

- If $f(x^c) < f(x^m)$, replace x^m by x^c

$$x^m := x^c$$

- If $f(x^c) \geq f(x^m)$, compute a total contraction w.r.t. x^l :

$$x^i := \frac{1}{2} (x^i + x^l), i \neq l$$



5. With the updated simplex (nodes $\{x^0, \dots, x^n\}$) go to Step 2.

The method constructs node sequences $\{x^{(k,0)}, \dots, x^{(k,n)}\}_{k=1}^{\infty}$ and ensures that

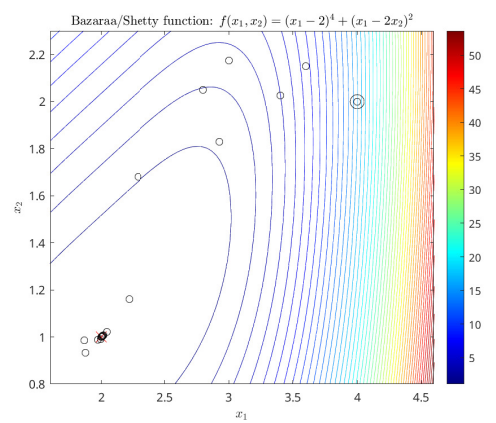
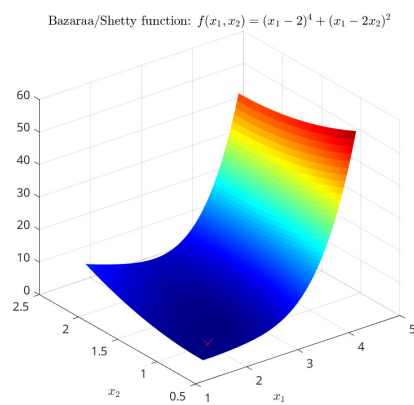
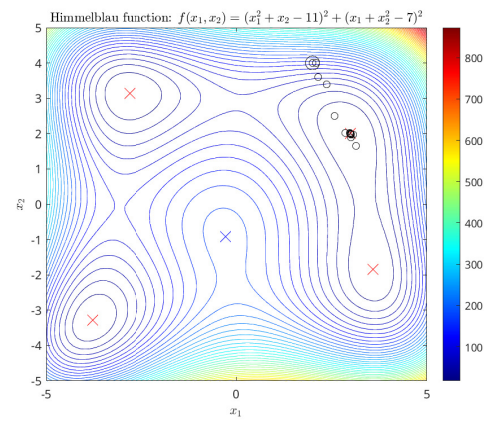
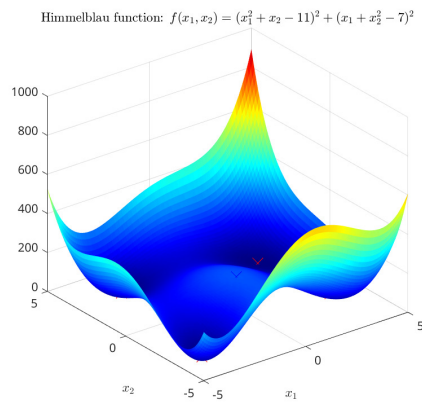
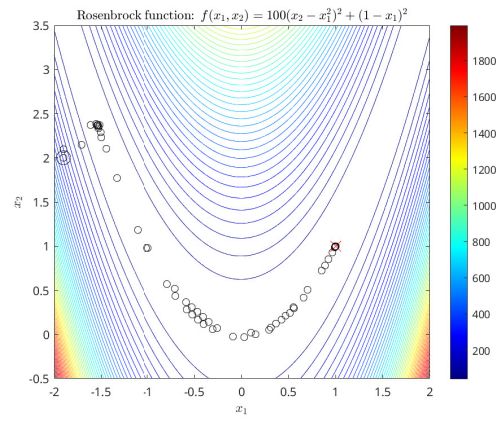
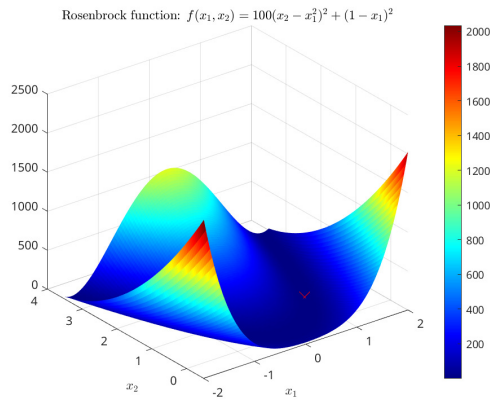
$$f(x^{(k+1,l)}) \leq f(x^{(k,l)})$$

One may consider $x^{(k,l)}$ as the current iterate.

Empirical studies recommend the following range of values: $0.4 \leq \alpha \leq 0.6$, $2 \leq \beta \leq 3$, $\gamma = 1$.

Available code: `fminsearch` (MATLAB®)
 `scipy.optimize.fmin` (SciPy Python)

Let us consider the MATLAB[®] implementation `fminsearch` which stops after 114/35/57 iterations, using 210/69/108 function evaluations.



2.2 Mutation selection methods

- random “mutation” of the current iterate
- selection of the “most useful” iterate

These methods belong to the class of stochastic search methods.

Basic idea:

1. Choose an initial vector $x^{(0)} \in \mathbb{R}^n$, set $k := 0$
2. Compute a new vector $v^{(k)}$ by random modification of $x^{(k)}$ (random numbers), e.g.,

$$\boxed{v_i^{(k)} = x_i^{(k)} + \delta_k \left(r_i^{(k)} - 0.5 \right)} \quad i = 1, \dots, n$$

$r_i^{(k)}$: random numbers taken from $[0, 1]$

δ_k : step size

3.

$$x^{(k+1)} = \begin{cases} v^{(k)} & \text{if } f(v^{(k)}) < f(x^{(k)}) \\ x^{(k)} & \text{else} \end{cases}$$

3 Unconstrained problems – theory

3.1 Optimality conditions

3.1.1 First order necessary conditions

Throughout chapter 3 we assume:

$$\begin{array}{ll} D \subset \mathbb{R}^n & \text{open, non empty} \\ f: D \rightarrow \mathbb{R} & \text{of certain smoothness} \end{array}$$

We consider the unconstrained optimization problem

$$(PU) \quad \boxed{\min_{x \in D} f(x)}$$

Theorem 3.1.1 (Fermat) *Let $\tilde{x} \in D$ be a local minimum of f and assume that f is differentiable in \tilde{x} . Then*

$$\boxed{\nabla f(\tilde{x}) = 0} \quad \text{first order necessary condition} \quad (3.1)$$

Proof: Known from Analysis I/II. □

Remark: Note that $f'(x): \mathbb{R}^n \rightarrow \mathbb{R}$ such that we can identify it with a row vector. Hence, $\nabla f(x)$ is a column vector and we have $\nabla f(x) = f'(x)^\top$.

Example 3.1.1

$$f(x) = \frac{1}{2}x^\top Hx + b^\top x \quad \text{symmetric } H \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n$$

It holds that

$$\nabla f(x) = Hx + b.$$

A solution of the problem

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

has to satisfy $Hx = -b$. If H is even positive definite, we conclude a) the existence of a solution (Example 1.3.1) and b) uniqueness. Then $\tilde{x} = -H^{-1}b$.

Example: Linear regression (Example 1.3.2 cont'd)

We found

$$H = 2 \begin{pmatrix} \sum_1^m \xi_i^2 & \sum_1^m \xi_i \\ \sum_1^m \xi_i & m \end{pmatrix}, \quad b = -2 \begin{pmatrix} \sum_1^m \xi_i \eta_i \\ \sum_1^m \eta_i \end{pmatrix}.$$

If H is positive definite, then the solution of the regression problem we have to solve the following linear system of equations

$$\begin{pmatrix} \sum_{i=1}^m \xi_i^2 \end{pmatrix} x_1 + \begin{pmatrix} \sum_{i=1}^m \xi_i \end{pmatrix} x_2 = - \sum_{i=1}^m \xi_i \eta_i$$

$$\sum_{i=1}^m \xi_i x_1 + m x_2 = - \sum_{i=1}^m \eta_i.$$

Compute the solution!

Definition 3.1.1 If f is differentiable in $\tilde{x} \in D$ and $\nabla f(\tilde{x}) = 0$, then \tilde{x} is called **stationary point** of f .

Remark: Optimization methods typically compute stationary points which are not necessarily local or global minima (maxima). Example: $f(x) = x^3$ at $x = 0$.

Example 3.1.2 Rosenbrock function: has exactly one stationary point at $\tilde{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$\nabla f(x) = \begin{pmatrix} -400 x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}.$$

If the objective function f is differentiable, (PU) is called **smooth** or **differentiable** optimization problem, otherwise **non smooth**, e.g.,

$$f(x) = \|x\| \quad \text{bei } x = 0.$$

Definition 3.1.2 f is **directionally differentiable** in $x \in D$ in the direction $h \in \mathbb{R}^n$ if the **directional derivative**

$$f'(x, h) := \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists. If this holds true for all $h \in \mathbb{R}^n$, f is called **directionally differentiable** at x .

Theorem 3.1.2 If \tilde{x} is a local minimum of (PU) and f is directionally differentiable at $\tilde{x} \in D$ then

$$\boxed{f'(\tilde{x}, h) \geq 0 \quad \forall h \in \mathbb{R}^n} \quad \text{“variational inequality”} \quad (3.2)$$

Proof: Since D is open there exists $\exists r > 0$ s.t.:

$$f(x) \geq f(\tilde{x}) \quad \forall x \in B(\tilde{x}, r).$$

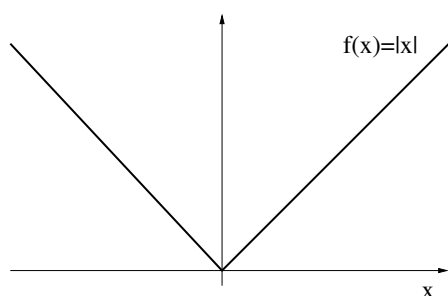
Let $h \in \mathbb{R}^n$ be arbitrary but fixed. Then $\tilde{x} + th \in B(\tilde{x}, r)$ for t sufficiently small such that

$$\begin{aligned} f(\tilde{x} + th) - f(\tilde{x}) &\geq 0 \\ \Rightarrow \frac{f(\tilde{x} + th) - f(\tilde{x})}{t} &\geq 0 \quad \Rightarrow \quad f'(\tilde{x}, h) \geq 0 \end{aligned}$$

□

(3.2) is intuitively clear: we have a local minimum in \tilde{x} , hence there cannot exist a descent direction!

Beispiel 3.1.1 $f(x) = |x|$ has a local minimum at $\tilde{x} = 0$.



f is not differentiable at $\tilde{x} = 0$ but the directional derivative exists:

$$\begin{aligned} \frac{f(th) - f(0)}{t} &= \frac{|th|}{t} = |h|, \quad t > 0 \\ \Rightarrow f'(0, h) &= |h| \geq 0 \quad \forall h \in \mathbb{R}. \end{aligned}$$

3.1.2 Second order necessary conditions

Theorem 3.1.3 Let f be twice continuously differentiable in a neighborhood of $\tilde{x} \in D$. If \tilde{x} is a local minimum of (PU), then, in addition to $\nabla f(\tilde{x}) = 0$ it holds that

$$\boxed{h^\top f''(\tilde{x})h \geq 0 \quad \forall h \in \mathbb{R}^n} \quad (3.3)$$

i.e., $f''(\tilde{x})$ is **positive semidefinite**.

Proof: Known from Analysis I/II. Sketch:

For arbitrary but fixed h , we define

$$F(t) = f(\tilde{x} + th).$$

$F \in C^2$ has a local minimum at $t = 0$ and its Taylor expansion reads:

$$\begin{aligned} F(t) &= F(0) + \underbrace{F'(0)}_{=0} t + \frac{1}{2} F''(\vartheta t) t^2 \\ \Rightarrow 0 &\leq \frac{F(t) - F(0)}{t^2} = \frac{1}{2} F''(\vartheta t) \end{aligned}$$

$t \downarrow 0$, continuity of $F'' \Rightarrow F''(0) = h^\top f''(\tilde{x})h \geq 0$. □

Example 3.1.3

$$\begin{aligned} f(x) &= \frac{1}{2} x^\top H x + b^\top x \\ f''(x) &= H \end{aligned}$$

If (PU) has a solution for this f , then H is to be positive semidefinite.

Beispiel 3.1.2 Rosenbrock function

$$\begin{aligned} f_{x_1} &= -400 x_1 (x_2 - x_1^2) - 2(1 - x_1) \\ f_{x_2} &= 200 (x_2 - x_1^2) \\ f_{x_1 x_1} &= -400 (x_2 - x_1^2) + 800 x_1 + 2 \\ f_{x_1 x_2} &= f_{x_2 x_1} = -400 x_1 \\ f_{x_2 x_2} &= 200 \\ \Rightarrow f''(1, 1) &= \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \text{ positive definite since } 802 > 0 \text{ and } \det(f''(1, 1)) > 0 \end{aligned}$$

3.1.3 Second order sufficient conditions

Recall that the previous conditions are not sufficient: $f(x) = x^3$ in $x = 0$). We therefore complement our analysis by second order sufficient conditions.

In the following, we say “ f is of class C^2 in U ”, or “in C^2 ”, if f is two times continuously differentiable in U .

Theorem 3.1.4 *Let $f \in C^2$ in a neighborhood of $\tilde{x} \in D$. If $\nabla f(\tilde{x}) = 0$ and*

$$\boxed{h^\top f''(z)h \geq 0 \quad \forall h \in \mathbb{R}^n} \quad (3.4)$$

for all $z \in B_\delta(\tilde{x})$ for some $\delta > 0$, then \tilde{x} is a local minimum of (PU).

Proof: Consider $x \in B_\delta(\tilde{x})$ and

$$\begin{aligned} f(x) - f(\tilde{x}) &= f'(\tilde{x})(x - \tilde{x}) + \underbrace{\frac{1}{2}(x - \tilde{x})^\top}_{h^\top} \underbrace{f''(\tilde{x} - \vartheta(x - \tilde{x})))}_{z} (x - \tilde{x}), \vartheta \in (0, 1) \\ &\geq 0 \quad \text{due to (3.4).} \\ &\Rightarrow \tilde{x} \text{ local minimum} \end{aligned}$$

□

Example 3.1.4 Linear regression

In this case, $f''(x) = H$ does not depend on x . If H is only positive semidefinite and \tilde{x} satisfies the necessary condition $H\tilde{x} + b = 0$, then \tilde{x} is a local minimum. If two data points ξ_i are different, then H is positive definite and we obtain existence and uniqueness.

Example 3.1.5

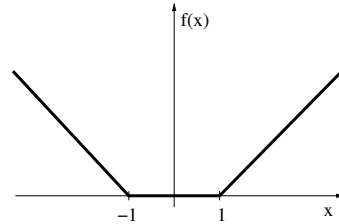
(i) $f(x) \equiv 0 \quad \forall x \in \mathbb{R}^n$

Every $x \in \mathbb{R}^n$ is a local minimum

(ii) $f(x) = \max\{0, |x| - 1\}$

All $x \in [-1, 1]$ are local minima,

$x = -1, x = 1$ are not covered by our theory ...



We can exclude these pathological examples by strenghtening the condition to obtain strict local minima:

Theorem 3.1.5 *Let $f \in C^2$ in a neighborhood of $\tilde{x} \in D$. If $\nabla f(\tilde{x}) = 0$, and $f''(\tilde{x})$ is positive definite, i.e.,*

$$h^\top f''(\tilde{x})h > 0 \quad \forall h \in \mathbb{R}^n, h \neq 0, \quad (3.5)$$

then $\exists r > 0, \alpha > 0$ s.t.

$$\boxed{f(x) \geq f(\tilde{x}) + \alpha \|x - \tilde{x}\|^2 \quad \forall x \in B(\tilde{x}, r)} \quad \text{“quadratic growth condition”}$$

In particular, \tilde{x} is a strict local minimum of (PU).

Proof: Analysis I/II. Sketch:

- We already know that (3.5) is equivalent to

$$h^\top f''(\tilde{x})h \geq \tilde{\alpha} \|h\|^2 \quad \forall h \in \mathbb{R}^n,$$

$$\alpha > 0.$$

- Then

$$\begin{aligned} f(x) &= f(\tilde{x}) + \underbrace{\nabla f(\tilde{x})^\top (x - \tilde{x})}_{=0} + \frac{1}{2}(x - \tilde{x})^\top f''(\tilde{x} + \vartheta(x - \tilde{x}))(x - \tilde{x}), \quad \vartheta \in (0, 1) \\ &= f(\tilde{x}) + \underbrace{\frac{1}{2}(x - \tilde{x})^\top f''(\tilde{x})(x - \tilde{x})}_{\geq \frac{1}{2}\tilde{\alpha}\|x - \tilde{x}\|^2} + \underbrace{\frac{1}{2}(x - \tilde{x})^\top [f''(\tilde{x} + \vartheta(x - \tilde{x})) - f''(\tilde{x})](x - \tilde{x})}_{|\cdot| \leq \frac{\tilde{\alpha}}{4}\|x - \tilde{x}\|^2, \text{ if } \|x - \tilde{x}\| \text{ is small} \\ (f \in C^2!)} \\ &\geq f(\tilde{x}) + \frac{\tilde{\alpha}}{4}\|x - \tilde{x}\|^2 \end{aligned}$$

$$\alpha := \frac{\tilde{\alpha}}{4}$$

□

Obviously, it holds that $h^\top f''(z)h \geq 0 \quad \forall h, \forall z \in B(\tilde{x}, r)$, hence (3.5) implies (3.4).

Example 3.1.6 *Linear regression with positive definite H*

Example 3.1.7 *Rosenbrock function in $\tilde{x} = [1, 1]^\top$.*

$$f''(1, 1) = \begin{pmatrix} 802 & -400 \\ 400 & 200 \end{pmatrix} \text{ is positive definite, s.t. } \tilde{x} \text{ is a strict local minimum.}$$

Example 3.1.8

$$f(x) = x^{2p}, \quad p \in \mathbb{N}, \quad x \in \mathbb{R}.$$

$\tilde{x} = 0$ is a local minimum, but the common rule " $f'(\tilde{x}) = 0 \wedge f''(\tilde{x}) > 0 \Rightarrow \text{local minimum}$ " only applies in the case $p = 1$

$$\begin{aligned} f'(x) &= 2p x^{2p-1} & \Rightarrow & f'(0) = 0 \\ f''(x) &= 2p(2p-1)x^{2p-2} & \Rightarrow & \begin{aligned} f''(0) &> 0 & \text{falls } p &= 1 \\ f''(0) &= 0 & \text{falls } p &> 1 \end{aligned} \end{aligned}$$

Theorem 3.1.5 is only applicable for $p = 1$, Theorem 3.1.4 for all p .

3.2 Convex optimization problems

We consider the convex problem

$$\begin{aligned} \text{(P)} \quad & \boxed{\min_{x \in \Omega} f(x)} \quad f: D \rightarrow \mathbb{R} \quad \text{convex } (D \text{ open}) \\ & \Omega \subset D \quad \text{convex, } \neq \emptyset, \text{ not necessarily open} \end{aligned}$$

Every local solution of (P) is thus is a global solution.

Characterization of convexity of f via derivatives:

Theorem 3.2.1 Let f be differentiable in D . Then f is convex in Ω if and only if

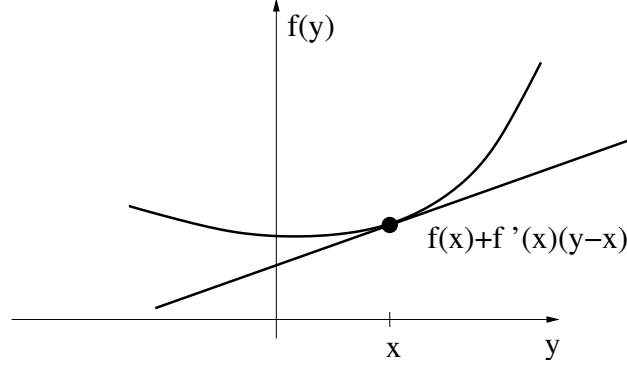
$$\boxed{f(y) \geq f(x) + f'(x)(y - x) \quad \forall x, y \in \Omega} \quad (3.6)$$

Furthermore, f is strictly convex in Ω if and only if

$$\boxed{f(y) > f(x) + f'(x)(y - x) \quad \forall x \neq y \in \Omega} \quad (3.7)$$

Proof: Exercise. □

Illustration for $n = 1$:



Theorem 3.2.2 Let f be differentiable in D and $\tilde{x} \in \Omega$. Then \tilde{x} is a solution of the convex optimization problem (P) if and only if

$$\boxed{f'(\tilde{x})(x - \tilde{x}) \geq 0 \quad \forall x \in \Omega} \quad \text{variational inequality} \quad (3.8)$$

Proof:

- (i) If \tilde{x} is a local solution, we consider an arbitrary $x \in \Omega$. Due to convexity of Ω , we conclude that $\tilde{x} + t(x - \tilde{x}) \in \Omega$. Moreover, since \tilde{x} is also a global solution, we know that $f(\tilde{x} + t(x - \tilde{x})) \geq f(\tilde{x})$ for all $t \in [0, 1]$. Dividing by t and considering the limit, we obtain $f'(\tilde{x})(x - \tilde{x}) \geq 0$.
- (ii) If (3.8) holds true, then convexity and (3.7) imply that

$$f(x) - f(\tilde{x}) \geq f'(\tilde{x})(x - \tilde{x}) \geq 0$$

□

Example 3.2.1 (Projection) Consider $C \subset \mathbb{R}^n$, non empty, closed and convex. Let $z \in \mathbb{R}^n, z \notin C$ be fixed.

Goal: find a point $Pr(z) \in C$ s.t. its distance to z is minimal \Rightarrow projection of z onto C

$$\min_{x \in C} f(x) = \min_{x \in C} \|x - z\|^2 = \min_{x \in C} \langle x - z, x - z \rangle$$

Note that f is coercive and we conclude the existence of a solution \tilde{x} . Moreover, f is differentiable with $\nabla f(x)^\top d = 2\langle x - z, d \rangle$. Since f is also convex (see characterization below), $\tilde{x} \in C$ is a solution if and only if $\langle \tilde{x} - z, x - \tilde{x} \rangle \geq 0$ for all $x \in C$.

Theorem 3.2.3 (Hahn-Banach) Let $C \subset \mathbb{R}^n$ be non empty, closed and convex, and let $z \notin C$. Then there exists $s \neq 0 \in \mathbb{R}^n$ which strictly separates C and $\{z\}$, i.e.,

$$\langle s, z \rangle > \sup_{x \in C} \langle s, x \rangle.$$

Proof: Let \tilde{x} be the solution of the optimization problem, i.e., the projection of z onto C . Then $\langle z - \tilde{x}, x - \tilde{x} \rangle \leq 0$ for all $x \in C$. With $s := z - \tilde{x} \neq 0$, we obtain for all $x \in C$

$$\langle z - \tilde{x}, x - \tilde{x} \rangle = \langle s, x - z + s \rangle = \langle s, x \rangle - \langle s, z \rangle + \|s\|^2 \leq 0$$

which shows the assertion. \square

Theorem 3.2.4 Let f be differentiable in D and strictly convex on Ω . Then $\tilde{x} \in \Omega$ is a unique strict global solution to (P) if and only if

$$f'(\tilde{x})(x - \tilde{x}) \geq 0 \quad \forall x \in \Omega.$$

Proof: Exercise

We can also characterize convexity via the second derivative.

Theorem 3.2.5 Let $D \subset \mathbb{R}^n$ be open, $\Omega \subset D$ non empty and convex, and let $f: D \rightarrow \mathbb{R}$ be twice continuously differentiable on D . Then:

- (i) If $f''(x)$ is positive semidefinite for all $x \in \Omega$, then f is convex on Ω . If Ω is open, then convexity of f on Ω implies positive semidefiniteness of $f''(x)$ for all $x \in \Omega$.
- (ii) If $f''(x)$ is positive definite $\forall x \in \Omega$, then f is strictly convex on Ω .

Proof: Let $x, y \in \Omega$, then there exists $t \in (0, 1)$ s.t.

$$f(y) - f(x) - f'(x)(y - x) = \frac{1}{2}(y - x)^\top f''(x + t(y - x))(y - x). \quad (*)$$

- (i) If $f''(x)$ is positive definite, then

$$f(y) - f(x) - f'(x)(y - x) \geq 0$$

Theorem 3.2.1 then yields convexity of f .

\Leftarrow : Assume that Ω is open and $x \in \Omega$ arbitrary. We have to show that $f''(x) \geq 0$. Consider arbitrary $d \in \mathbb{R}^n$. For $t \in \mathbb{R}, |t|$ sufficiently small, we have $x + td \in \Omega$ and with Theorem 3.2.1 we conclude that

$$\begin{aligned} f(x + td) &\geq f(x) + f'(x)(td) \\ f(x) &\geq f(x + td) + f'(x + td)(-td). \end{aligned}$$

Adding both inequalities, we obtain

$$[f'(x + td) - f'(x)](td) \geq 0$$

$$\begin{aligned}
\Rightarrow d^\top f''(x)d &= \lim_{t \rightarrow 0} \frac{1}{t} d^\top [f'(x+td) - f'(x)] d \\
&= \lim_{t \rightarrow 0} \underbrace{\frac{1}{t^2}}_{\geq 0} d^\top \underbrace{[f'(x+td) - f'(x)](td)}_{\geq 0} \geq 0
\end{aligned}$$

(ii) If $f''(x)$ is positive definite for all $x \in \Omega$, then $(*)$ implies

$$f(y) - f(x) - f'(x)(y - x) > 0 \quad \forall x, y \in \Omega, x \neq y.$$

By Theorem 3.2.1, it holds that f is strictly convex. □

An even stronger notion of convexity is:

Definition 3.2.1 If $D \subset \mathbb{R}^n, \Omega \subset D$ is non empty and convex, then $f: D \rightarrow \mathbb{R}$ is called **uniformly convex** on Ω , if there exists $\alpha > 0$ s.t.

$$\boxed{(1 - \lambda)f(x) + \lambda f(y) \geq f((1 - \lambda)x + \lambda y) + \lambda(1 - \lambda)\alpha \|x - y\|^2}$$

$\forall x, y \in \Omega, \lambda \in [0, 1].$

With this definition one can show that:

- If f is differentiable, then uniform convexity of f is equivalent to

$$f(y) - f(x) \geq f'(x)(y - x) + \alpha \|x - y\|^2 \quad \forall x, y \in \Omega$$

- For $f \in C^2$ uniform convexity follows from uniform positive definiteness of f'' , i.e.,

$$h^\top f''(x)h \geq \beta \|h\|^2 \quad \forall h \in \mathbb{R}^n,$$

where $\beta > 0$ does not depend on $x \in \Omega$.

If Ω is open, then uniform convexity of f on Ω implies uniform positive definiteness of f'' .

4 Unconstrained problems – numerical methods

4.1 Basics

Throughout this chapter, we consider numerical methods for solving the optimization problem

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

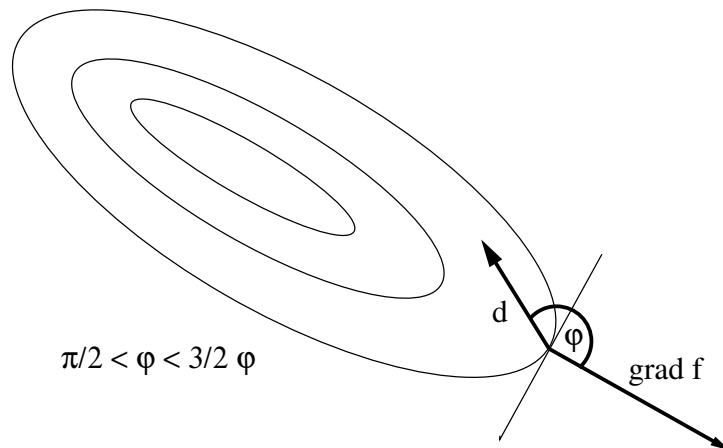
We already know that for a solution \tilde{x} it holds that

$$\nabla f(\tilde{x}) = 0 \tag{4.9}$$

and we can try to numerically solve this equation. However, this generally only yields a solution to this equation (i.e., a stationary point) but not necessarily a minimum. We are thus interested in numerical methods which solve (4.9) and simultaneously consider the minimization of (PU). Among these methods are descent methods which are iterative methods that consecutively decrease/minimize the objective value of f .

Definition 4.1.1 *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable in x , then $d \in \mathbb{R}^n$ is called **descent direction** if*

$$\nabla f(x)^\top d < 0$$



The definition makes sense taking into account the following result.

Lemma 4.1.1 *Let f be differentiable in x . If $d \in \mathbb{R}^n$ is a descent direction, then $\exists \bar{\sigma} > 0$ s.t. $\bar{\sigma} > 0$ mit $f(x + \sigma d) < f(x) \quad \forall \sigma \in (0, \bar{\sigma}]$.*

Proof: Due to

$$\nabla f(x)^\top d = \lim_{\sigma \rightarrow 0} \frac{f(x + \sigma d) - f(x)}{\sigma} < 0$$

for $\sigma > 0$ sufficiently small, it has to hold that $f(x + \sigma d) < f(x)$. □

Examples of descent directions:

Example 4.1.1

- If $\nabla f(x) \neq 0$ in x , then the **anti gradient**

$$\boxed{-\nabla f(x) \text{ is a descent direction}}$$

since: For $d = -\nabla f(x)$ we have

$$f'(x)d = \nabla f(x) \cdot (-\nabla f(x)) = -\|\nabla f(x)\|^2 < 0.$$

- If A is positive definite, then

$$\boxed{-A^{-1}\nabla f(x) \text{ is a descent direction}}$$

This follows since positive definiteness of A implies positive definiteness of A^{-1} .

Algorithm 4.1.1 (General descent method)

1. Choose an initial point $x^{(0)} \in \mathbb{R}^n, k := 0$.
2. Stop, if $\nabla f(x^{(k)}) = 0$
3. Compute descent direction $d = d^{(k)}$ and step size $\sigma = \sigma_k > 0$, such that

$$\begin{aligned} f(x^{(k)} + \sigma_k d^{(k)}) &< f(x^{(k)}), \\ x^{k+1} &:= x^{(k)} + \sigma_k d^{(k)} \end{aligned}$$

4. $k := k + 1$, go to 2.

Remarks:

- The stopping criterion is only of theoretical interest. Numerically, one typically uses $\|\nabla f(x^{(k)})\| < \varepsilon$ or $|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon_1 \wedge \|x^{(k+1)} - x^{(k)}\| < \varepsilon_2$, where $\varepsilon, \varepsilon_1, \varepsilon_2$ are positive stopping tolerances.

- Alternative:

$$\begin{aligned} f(x^{(k+1)}) - f(x^{(k)}) &\approx \sigma_k f'(x^{(k)}) d^{(k)} < \varepsilon_1 \\ \text{and } \|x^{(k+1)} - x^{(k)}\|_\infty &= \sigma_k \|d^{(k)}\|_\infty < \varepsilon_2 \end{aligned}$$

- Often the choice of σ is the main challenge

4.2 Newton's method

Newton's method for solving equations of the form $\nabla f(x) = 0$ is a common tool to find local extrema. If we set $F(x) := \nabla f(x)$, we obtain a multivariate vector valued function $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ to which we can apply the Newton method for solving

$$F(x) = 0.$$

Let us briefly recall the main idea. Given $x^{(k)}$, $F(x)$ locally (around $x^{(k)}$) behaves as $F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})$, s.t. we can search for a zero of this function. We thus solve

$$\boxed{F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) = 0} \quad \text{linear system of equations!} \quad (4.10)$$

and obtain a new approximation $x =: x^{(k+1)}$. If $F'(x^{(k)})$ is invertible, we obtain

$$x^{(k+1)} = x^{(k)} - F'(x^{(k)})^{-1} F(x^{(k)}). \quad (4.11)$$

For the convergence analysis of the method, let us assume that:

- (i) $F: \mathbb{R}^n \supset D \rightarrow \mathbb{R}^n$ is differentiable in D , D open, and it has a zero in $\tilde{x} \in D$.
- (ii) F' is Lipschitz continuous in D , i. e. $\exists L > 0$:

$$\boxed{\|F'(x) - F'(y)\| \leq L \|x - y\| \quad \forall x, y \in D}$$

- (iii) $F'(\tilde{x})^{-1}$ exists.

The convergence proof of Newton's method relies on the following (known) facts:

Lemma 4.2.1

$$\|F(x) - F(y) - F'(y)(x - y)\| \leq \frac{L}{2} \|x - y\|^2 \quad \forall x, y \in D.$$

(Consequence of the mean value theorem applied to $\varphi(t) = F(x + t(y - x))$.)

Lemma 4.2.2 *If $A \in \mathbb{R}^{n \times n}$ is a non singular matrix, $S \in \mathbb{R}^{n \times n}$ and $\|A^{-1}\| \|S\| < 1$, then $(A + S)^{-1}$ exists and*

$$\|(A + S)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|S\|}.$$

Lemma 4.2.3 *Let $G: \bar{B}(\tilde{x}, r) \rightarrow \mathbb{R}^n$ be a contraction, i.e., Lipschitz continuous with constant $L < 1$ in $\bar{B}(\tilde{x}, r)$. If \tilde{x} is a fixed point of G , then it is the only one in $\bar{B}(\tilde{x}, r)$. Starting with an arbitrary initial point $x^{(0)} \in \bar{B}(\tilde{x}, r)$, the sequence $x^{(k)}$ defined by*

$$x^{(k+1)} = G(x^{(k)})$$

converges to \tilde{x} and

$$\|x^{(k)} - \tilde{x}\| \leq L^k \|x^{(0)} - \tilde{x}\|.$$

(Follows with the Banach fixed point theorem.)

Theorem 4.2.1 (*Convergence of Newton's method*)

Under the above assumptions (i) - (iii) there exist $\delta > 0, c > 0$ s.t. for every $x^{(0)} \in B_\delta(\tilde{x})$, Newton's method defines a sequence $x^{(k)}$ converging to \tilde{x} with

$$\boxed{\|x^{(k+1)} - \tilde{x}\| \leq c \|x^{(k)} - \tilde{x}\|^2} \quad (\text{quadratic convergence}) \quad (4.12)$$

Sketch of the proof

a) With Lemma 4.2.2 one shows that

$$\|F'(x)^{-1}\| \leq 2 \|F'(\tilde{x})^{-1}\| \quad \forall x \in B(\tilde{x}, \delta_1)$$

b) This allows to conclude that

$$\begin{aligned} \|F'(x)^{-1} - F'(y)^{-1}\| &= \left\| \underbrace{F'(x)^{-1}}_{\|\cdot\| \leq 2\|F'(\tilde{x})^{-1}\|} \underbrace{(F'(y) - F'(x))}_{\leq L\|x-y\|} \underbrace{F'(y)^{-1}}_{\leq 2\|F'(\tilde{x})^{-1}\|} \right\| \\ &\leq 4L \|F'(\tilde{x})^{-1}\|^2 \|x - y\| \end{aligned}$$

c) From (4.11) Newton's method is a fixed point iteration for the function

$$\boxed{G(x) := x - F'(x)^{-1}F(x).}$$

G is a contraction in $B(\tilde{x}, \delta)$:

$$\begin{aligned} G(x) - G(y) &= \underbrace{x - y}_{=F'(x)^{-1}F'(x)(x-y)} - F'(x)^{-1}F(x) + F'(y)^{-1}F(y) \\ &= \underbrace{F'(x)^{-1}}_{\text{bounded by a)}} \underbrace{\{F(y) - F(x) - F'(x)(y - x)\}}_{\leq \frac{L}{2}\|x-y\| \|x-y\|} \\ &\quad + \underbrace{(F'(y)^{-1} - F'(x)^{-1})}_{\leq c\|x-y\| \text{ by b)}} \underbrace{\cdot F(y)}_{\text{small, if } y \text{ close to } \tilde{x}} \end{aligned}$$

One estimates to obtain

$$\|G(x) - G(y)\| \leq \frac{1}{2} \|x - y\| \quad \text{in } B(\tilde{x}, \delta)$$

for $\delta \leq \delta_1$ sufficiently small.

d) Now we apply the contraction Lemma 4.2.3. The iterates converge to \tilde{x} .
Quadratic convergence follows from

$$\begin{aligned} \|x^{(k+1)} - \tilde{x}\| &= \|x^{(k)} - F'(x^{(k)})^{-1}F(x^{(k)}) - \tilde{x}\| \\ &= \underbrace{\|F'(x^{(k)})^{-1}\|}_{\leq 2\|F'(\tilde{x})^{-1}\|} \underbrace{\|F(\tilde{x}) - F(x^{(k)}) - F'(x^{(k)})(\tilde{x} - x^{(k)})\|}_{\leq \frac{L}{2}\|\tilde{x} - x^{(k)}\|^2} \\ &\leq c \|x^{(k)} - \tilde{x}\|^2 \quad \text{with } \boxed{c = L \|F'(\tilde{x})^{-1}\|} \end{aligned}$$

□

Newton's method converges locally quadratic. Applied to the unconstrained optimization problem, this means that

$$\boxed{F(x) := \nabla f(x)}$$

and we thus demand

- f'' is Lipschitz continuous in a neighborhood of a minimum \tilde{x} of f
- $f''(\tilde{x})$ is positive definite
(This ensures the existence of $F'(\tilde{x})^{-1} = f''(\tilde{x})^{-1}$ and it complies with the optimization conditions.)

Theorem 4.2.2 *Under the previous assumptions, the Newton's method*

$$\boxed{x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1} \nabla f(x^{(k)})} \quad (4.13)$$

locally converges (quadratically) to \tilde{x} .

In a numerical realization, we do not explicitly compute the inverse of $f''(x^{(k)})$ but instead solve the linear system of equations

$$f''(x^{(k)}) (x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)}),$$

i.e. we compute a direction $d^{(k)}$ from

$$f''(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$

and defines

$$x^{(k+1)} := x^{(k)} + d^{(k)}.$$

For $d^{(k)}$ it holds

$$d^{(k)} = \underbrace{f''(x^{(k)})^{-1}}_{\text{pos. definite}} \underbrace{(-\nabla f(x^{(k)}))}_{\text{anti gradient}}$$

Thus $d^{(k)}$ is a descent direction, the so-called **Newton direction**. This however does not make the Newton method automatically a descent method - note that it always choose a step size $\sigma = 1$ which may be too large!

Instead, one may use a modified iteration procedure

$$x^{(k+1)} = x^{(k)} - \sigma_k f''(x^{(k)})^{-1} \nabla f(x^{(k)})$$

(damped Newton's method).

Remark: We can also interpret Newton's method as follows: The iteration (4.13) means

$$f''(x^{(k)}) (x^{(k+1)} - x^{(k)}) + \nabla f(x^{(k)}) = 0.$$

This is the the first order optimality condition for the solution of the quadratic optimization problem

$$\boxed{\min_{x \in \mathbb{R}^n} \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)})^\top f''(x^{(k)}) (x - x^{(k)})} \quad (\text{Q})_k}$$

If $f''(x^{(k)})$ is positive definite, we already know that there exists a unique solution which is given by $x^{(k+1)}$. Hence, Newton's method is equivalent to the solution of a sequence of quadratic optimization problems (if $f''(\tilde{x})$ is positive definite), it thus is a **sequential quadratic programming method – SQP-method**

One defines $(\text{Q})_k$ as follows:

$$\begin{aligned} & \min \nabla f(x^{(k)})^\top z + \frac{1}{2} z^\top f''(x^{(k)}) z \\ \text{and} \quad & x^{(k+1)} := x^{(k)} + z^{(k)} \end{aligned}$$

4.3 General descent methods – general properties

4.3.1 Efficient step sizes

If $d^{(k)}$ is a descent direction and σ_k is sufficiently small, we have $f(x^{(k)} + \sigma_k d^{(k)}) < f(x^{(k)})$. This however does not ensure convergence to a local minimum:

Example 4.3.1 $f(x) = x^2$, $d^{(k)} = -1$ for all $k \geq 0$, $x^{(0)} = 1$, and $\sigma_k = \left(\frac{1}{2}\right)^{k+2}$, $k = 0, 1, \dots$.
The sequence $x^{(k)}$ converges to $\frac{1}{2}$, the selected step sizes are too small.

If the descent methods starts with $x^{(0)}$, we only obtain smaller objective values and the iterates remain in $N(f, f(x^{(0)}))$.

Definition 4.3.1 Let $x \in N(f, f(x^{(0)}))$ and $d \in \mathbb{R}^n$ a descent direction. A step size σ is called **efficient**, if

$$\boxed{f(x + \sigma d) \leq f(x) - c \left(\frac{\nabla f(x) \cdot d}{\|d\|} \right)^2} \quad (4.14)$$

with a constant $c > 0$ independent of $x \in N(f, f(x^{(0)}))$.

Explanation: For $d = -\nabla f(x)$ the quadratic term is maximal. For $d \perp \nabla f(x)$ we do not obtain descent. The constant c is a minimal rate. Note: $d / \|d\|$ is a unit vector.

Given sequences $\{x^{(k)}\}, \{d^{(k)}\}$ with $\nabla f(x^{(k)})^\top d^{(k)} < 0$ and efficient step sizes σ_k , then (4.14) is satisfied with a constant $c > 0$ independent of k .

A specific form of efficiency is the principle of sufficient decrease: We demand the existence of constants $c_1, c_2 > 0$ independent of x and d s.t.

$$f(x + \sigma d) \leq f(x) + c_1 \sigma \nabla f(x)^\top d \quad (4.15)$$

(sufficiently fast descent)

and

$$\sigma \geq -c_2 \frac{\nabla f(x)^\top d}{\|d\|^2} \quad (4.16)$$

(minimal step size)

The two conditions imply efficiency with $c = c_1 c_2$, since

$$f(x + \sigma d) \leq f(x) - c_1 \left(c_2 \frac{\nabla f(x)^\top d}{\|d\|^2} \nabla f(x)^\top d \right) = f(x) - c_1 c_2 \left(\frac{\nabla f(x)^\top d}{\|d\|} \right)^2$$

Remark: Assuming Lipschitz continuity of $f'(x)$ on $N(f, f(x^{(0)}))$ one can prove existence of efficient step sizes, see Lemma 4.3.4 in [1].

4.3.2 Gradient related directions

If $N(f, f(x^{(0)}))$ is compact, the sequence of function values $\{f(x^{(k)})\}$ is bounded (from below). If the sequence of step sizes $\{\sigma_k\}$ is efficient, then

$$f(x^{(k+1)}) \leq f(x^{(k)}) - c \left(\frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|d^{(k)}\|} \right)^2$$

From monotonicity, we obtain convergence of the function values s.t. $(f(x^{(k+1)}) - f(x^{(k)})) \rightarrow 0$, $k \rightarrow \infty$, i.e.

$$\frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|d^{(k)}\|} \rightarrow 0, \quad k \rightarrow \infty. \quad (4.17)$$

This motivates to choose directions which allow to conclude that

$$\nabla f(x^{(k)}) \rightarrow 0, \quad k \rightarrow \infty. \quad (4.18)$$

Condition (4.17) can obviously hold independently of (4.18), for example if the direction $d^{(k)}$ asymptotically becomes orthogonal to $\nabla f(x^{(k)})$. This has to be excluded by staying uniformly bounded away from orthogonality: observe that

$$\begin{aligned} \cos(\nabla f(x^{(k)}), d^{(k)}) &= \frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|\nabla f(x^{(k)})\| \|d^{(k)}\|} =: \beta_k \\ \Rightarrow \quad \beta_k \|\nabla f(x^{(k)})\| &= \underbrace{\frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|d^{(k)}\|}}_{\rightarrow 0 \text{ for efficient step sizes}} \end{aligned}$$

This implies $\nabla f(x^{(k)}) \rightarrow 0$, if $-\beta_k \geq c > 0 \quad \forall k$.

Definition 4.3.2 Let $x \in N(f, f(x^{(0)}), d \in \mathbb{R}^n$. The direction d is called **gradient related** in x , if

$$-\nabla f(x)^\top d \geq c_3 \|\nabla f(x)\| \|d\| \quad (4.19)$$

with a constant $c_3 > 0$ which is independent of x and d .

It is called **strictly gradient related**, if additionally

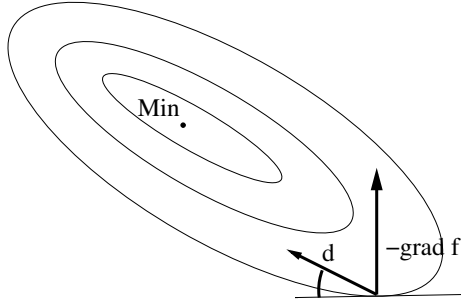
$$c_4 \|\nabla f(x)\| \geq \|d\| \geq \frac{1}{c_4} \|\nabla f(x)\| \quad (4.20)$$

with a constant $c_4 > 0$ which is independent of x and d .

Example 4.3.2 The anti gradient $d = -\nabla f$ is strictly gradient related since

$$\begin{aligned} -\nabla f(x)^\top d &= \|\nabla f(x)\|^2 = 1 \cdot \|\nabla f(x)\| \|d\| \\ \|\nabla f(x)\| &\stackrel{(\geq)}{=} \|d\| \stackrel{(\geq)}{=} \|\nabla f(x)\| \quad \text{i.e. } c_3 = c_4 = 1. \end{aligned}$$

Illustration of gradient related directions:



Guaranteed deviation of orthogonality to $-\nabla f(x)$ ensures sufficient decrease (if we do not have $\|d\| \rightarrow 0$).

For showing that the Newton direction is strictly gradient related, we assume:

(LUC) (Local uniform convexity)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $N(f, f(x^{(0)})) \subset D, \neq \emptyset$, open, convex, $f \in C^2$ on D . For $\alpha_1 > 0$ assume that

$$h^\top f''(x)h \geq \alpha_1 \|h\|^2 \quad \forall h \in \mathbb{R}^n, \forall x \in D,$$

i.e. uniform positive definiteness of f'' on D .

Without proof we conclude from (LUC):

Lemma 4.3.1

- $N(f, f(x^{(0)}))$ convex and compact
- $h^\top f''(x)h \geq \alpha_2 \|h\| \quad \forall h \in \mathbb{R}^n, \quad \forall x \in N(f, f(x^{(0)}))$
- $\|f''(x)\| \leq \alpha_2 \quad \quad \quad -'' -$
- $\|f''(x)^{(-1)}\| \leq \beta_2 := 1/\alpha_1 \quad \quad \quad -'' -$

- $\beta_1 \|h\|^2 \leq h^\top f''(x)^{(-1)} h \leq \beta_2 \|h\|^2, \quad \forall h \in \mathbb{R}^n, -''-$
- f is uniformly convex on D .

Example 4.3.3 (Newton direction)

Assume (LUC) is satisfied, $x \in N(f, f(x^{(0)}))$,

$$\begin{aligned} d &= -f''(x)^{(-1)} \nabla f(x) \\ \Rightarrow -\nabla f(x)^\top d &= \nabla f(x)^\top f''(x)^{(-1)} \nabla f(x) \geq \beta_1 \|\nabla f(x)\|^2 \quad (\text{Lemma 4.3.1}) \end{aligned}$$

And

$$\left. \begin{aligned} \|d\| &= \|-f''(x)^{(-1)} \nabla f(x)\| \leq \beta_2 \|\nabla f(x)\|^2 \\ \|\nabla f(x)\| &= \|-f''(x)d\| \leq \alpha_2 \|d\| \end{aligned} \right\} \Rightarrow (4.20)$$

i.e. strictly gradient related directions. (4.19) follows from

$$-\nabla f^\top d \stackrel{\text{see above}}{\geq} \beta_1 \|\nabla f\| \underbrace{\|\nabla f\|}_{\geq \frac{1}{\beta_2} \|d\|} \geq \frac{\beta_1}{\beta_2} \|\nabla f\| \|d\|.$$

4.3.3 General convergence results

The following assumptions will be used consistently throughout what follows:

- (LSC) For given $x^{(0)} \in \mathbb{R}^n$ the level set
 $N(f, f(x^{(0)})) = \{x | f(x) \leq f(x^{(0)})\}$
is compact.
- (FD) $f \in C^1$ on the open, convex set $D_0 \supset N(f, f(x^{(0)}))$.

This allows to show:

Theorem 4.3.1 Assume (LSC) and (FD), the search directions $d^{(k)}$ of the general descent method 4.1.1 are gradient related in $x^{(k)}$, the step sizes σ_k are efficient. If the algorithm does not terminate after finitely many steps, then $\nabla f(x^{(k)}) \rightarrow 0, k \rightarrow \infty$ and $\{x^{(k)}\}$ has an accumulation \tilde{x} .

For every such accumulation point, it holds that $\nabla f(\tilde{x}) = 0$.

The existence of an accumulation point for $\{x^{(k)}\}$ is of limited practical use. One is rather interested in: $x^{(k)} \rightarrow \tilde{x}$. Indeed, we have

Theorem 4.3.2 Additionally to the assumptions in Theorem 4.3.1, assume the general descent method 4.1.1 satisfies

- $d^{(k)}$ strictly gradient related,
- the step sizes $\{\sigma_k\}$ are bounded,
- the set of all zeros of ∇f in $N(f, f(x^{(0)}))$ is finite.

If the algorithm does not terminate after finitely many steps, then $x^{(k)}$ converges to a zero of ∇f .

Idea of the proof: Strictly gradient related, (4.20) \Rightarrow

$$\|x^{(k+1)} - x^{(k)}\| = \|\underbrace{\sigma_k}_{\leq \bar{\sigma}} d^{(k)}\| \leq c\bar{\sigma} \underbrace{\|\nabla f(x^{(k)})\|}_{\rightarrow 0, \text{Theorem 4.3.1}} \quad (*)$$

Due to (LSC) the set of all accumulation points H of $x^{(k)}$ is non empty. Moreover, for the distance it holds that

$$\boxed{d(x^{(k)}, H) < \varepsilon \quad \forall k > k_0} \quad (**)$$

Let \tilde{x} denote an arbitrary accumulation point of $x^{(k)}$. Since the set of accumulation points is finite, there is a ball $B(\tilde{x}, \rho)$ with $H \cap B(\tilde{x}, \rho) = \{\tilde{x}\}$. By (**) there exists l_0 with

$$\|x^{(l_0)} - \tilde{x}\| < \varepsilon$$

Due to (*) and $\nabla f(x^{(k)}) \rightarrow 0$ it also holds that

$$\begin{aligned} & \|x^{(l_0+1)} - x^{(l_0)}\| < \varepsilon \\ \Rightarrow & \|x^{(l_0+1)} - \tilde{x}\| < 2\varepsilon < \frac{\rho}{2} \quad \text{for } \varepsilon < \frac{\rho}{4} \end{aligned}$$

Hence $x^{(l_0+1)} \in B(\tilde{x}, \rho)$ and with (**) it follows

$$\|x^{(l_0+1)} - \tilde{x}\| < \varepsilon$$

By induction, we obtain $x^{(k)} \rightarrow \tilde{x}$. □

The previous results are general, but not very strong – in particular, they do not provide a convergence rate.

If (LUC), we do have uniform convexity in $N(f, f(x^{(0)}))$ and one can show that

$$\frac{\alpha_1}{2} \|x - \tilde{x}\|^2 \leq f(x) - f(\tilde{x}) \leq \frac{1}{2\alpha_1} \|\nabla f(x)\|^2 \quad (4.21)$$

in $N(f, f(x^{(0)}))$, where \tilde{x} is the only local minimum in $N(f, f(x^{(0)}))$, [1, Lemma 4.3.14]. At the same time, this is the global one.

(The first estimate follows from (LUC), Taylor expansion and $\nabla f(\tilde{x}) = 0$. The second estimate relies on an auxiliary quadratic optimization problem.)

This property is the basis for

Theorem 4.3.3 *Assumptions:*

- (LUC)

- $d^{(k)}$ gradient related in $x^{(k)}$
- $\{\sigma_k\}$ efficient.

If the algorithm does not terminate after finitely many steps, $\{x^{(k)}\}$ converges to the unique global minimum \tilde{x} of f .

There exists $q \in (0, 1)$ with

$$f(x^{(k)}) - f(\tilde{x}) \leq q^k (f(x^{(0)}) - f(\tilde{x})) \quad (4.22)$$

and

$$\|x^{(k)} - \tilde{x}\|^2 \leq \frac{2}{\alpha_1} q^k (f(x^{(0)}) - f(\tilde{x})) \quad k \geq 0. \quad (4.23)$$

Consequence: $\|x^{(k)} - \tilde{x}\| \leq C \sqrt{q^k} = C \tilde{q}^k$

Remark: $\{x^{(k)}\}$ behaves like a linearly convergent sequence since $\{x^{(k)}\}$ is called linearly convergent if

$$\|x^{(k+1)} - \tilde{x}\| \leq L \|x^{(k)} - \tilde{x}\|$$

with $0 < L < 1$. Then it holds

$$\|x^k - \tilde{x}\| \leq L^k \|x^{(0)} - \tilde{x}\|.$$

4.4 Line search methods

4.4.1 Exact step size

Given $x \in \mathbb{R}^n$ and a descent direction $d \in \mathbb{R}^n$, we are searching for a step size σ . Ideally, σ should be chosen such that

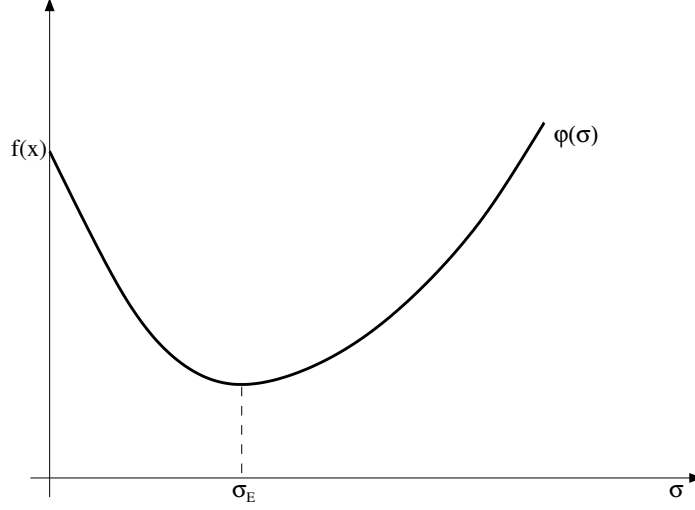
$$\min_{s \geq 0} f(x + sd) = \min_{s \geq 0} \varphi(s) = \varphi(\sigma).$$

This is however generally not possible (e.g., $f(x) = e^{(-x)}$) and would also require global optimization in \mathbb{R} . If (LUC) is satisfied, then $\varphi(s)$ will eventually become larger than $\varphi(0)$. Hence, $\varphi'(s) = \nabla f(x + sd)^\top d$ has a smallest positive zero σ_E .

Definition 4.4.1 The number σ_E with

$$\varphi'(s) \begin{cases} = 0 & , \text{ if } s = \sigma_E \\ < 0 & , \text{ if } s \in [0, \sigma_E) \end{cases}$$

is called **exact step size**.



It can be bounded from below as follows:

$$\begin{aligned}
0 &= \underset{\substack{\uparrow \\ \text{Def of } \sigma_E}}{\nabla f(x + \sigma_E d)^\top d} = \underset{\uparrow}{\nabla f(x)^\top d} + [\nabla f(x + \sigma_E d) - \nabla f(x)]^\top d \\
&\leq \underset{\substack{\uparrow \\ \text{Lipschitz cond}}}{\nabla f(x)^\top d} + \sigma_E L \|d\|^2 \\
&\Rightarrow \boxed{\sigma_E \geq \tilde{\sigma} = -\frac{\nabla f(x)^\top d}{L\|d\|^2}} \tag{4.24}
\end{aligned}$$

Moreover, one obtains a minimal descent

$$f(x + \sigma_E d) \leq f(x) + \frac{1}{2} \tilde{\sigma} \nabla f(x)^\top d \tag{4.25}$$

Thus $\sigma_E, \tilde{\sigma}$ are efficient. Unfortunately, σ_E is difficult to compute. An exception are quadratic functions

$$f(x) = \frac{1}{2} x^\top H x + b^\top x$$

for which σ_E is easily computable.

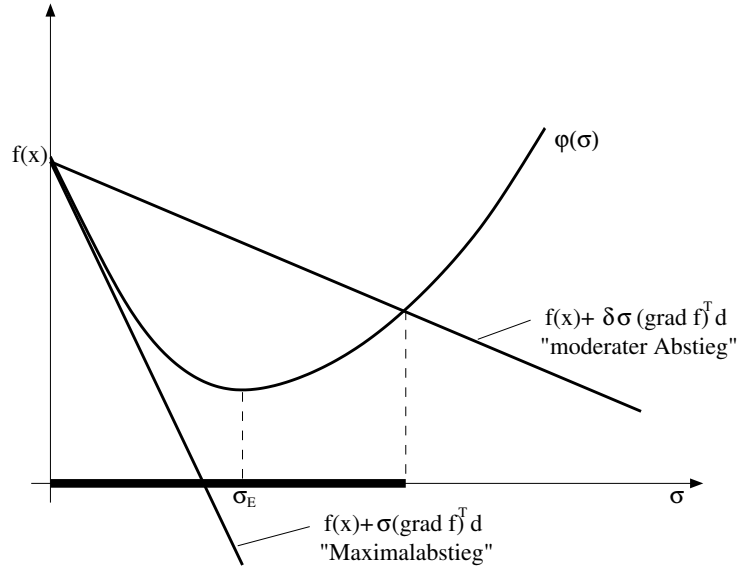
4.4.2 Armijo step size

Given: $x \in \mathbb{R}^n$, descent direction d .

We require for $\sigma = \sigma_A$

$$\bullet \quad f(x + \sigma_A d) \leq f(x) + \delta \sigma_A \nabla f(x)^\top d \quad \text{sufficient decrease} \tag{4.26}$$

$$\bullet \quad \sigma_A \geq -c_2 \frac{\nabla f(x)^\top d}{\|d\|^2} \quad \text{efficiency} \tag{4.27}$$



Algorithm 4.4.1 (Armijo-Goldstein)

0. Fix constants

$$\begin{aligned} 0 < \delta < 1 & \quad \text{"flattening"} \\ \gamma > 0 & \quad \text{"efficiency constant"} \\ 0 < \beta_1 \leq \beta_2 < 1 \end{aligned}$$

1. Initial step size

$$\sigma_0 \geq -\gamma \frac{\nabla f(x)^\top d}{\|d\|^2}$$

$$j := 0$$

2. If

$$f(x + \sigma_j d) \leq f(x) + \delta \sigma_j \nabla f(x)^\top d,$$

then set $\sigma_A := \sigma_j$, stop.

3. Else reduce σ_j s.t.

$$\sigma_j \in [\beta_1 \sigma_j, \beta_2 \sigma_j]$$

$$j := j + 1, \text{ go to 2.}$$

With suitable assumptions (i.e. (LSC), (FD), (LUC)) the algorithm terminates after finitely many steps with a step size which satisfies (4.26) – (4.27) (cp. [1, Satz 4.4.3]).

The first condition is clear since $\sigma_j \leq \beta_2^j \sigma_0$, see sketch. For the second condition, let l denote the number of iteration steps.

If $l = 0$, then (4.27) is satisfied with $c_2 = \gamma$. If $l > 0$ then $s = \sigma_{l-1}$ is still outside the

desired region, i.e.,

$$\begin{aligned}
& \underbrace{f(x+sd) - f(x)}_{=\nabla f(x+\vartheta sd)^\top ds}_{0 < \vartheta < 1} > \delta s \nabla f(x)^\top d. \\
\Rightarrow \nabla f(x+\vartheta sd)^\top d &= \frac{1}{s} [f(x+sd) - f(x)] > \delta \nabla f(x)^\top d \quad | - \nabla f(x)^\top d \\
\Rightarrow -(1-\delta) \nabla f(x)^\top d &< [\nabla f(x+\vartheta sd) - \nabla f(x)]^\top d \leq \underset{\substack{\uparrow \\ \text{Lipsch.}}}{L\vartheta s \|d\|^2} \leq sL \|d\|^2 \\
\Rightarrow \boxed{s \geq -\frac{(1-\delta) \nabla f(x)^\top d}{L \|d\|^2}}
\end{aligned}$$

Due to $\sigma_A \geq \beta_1 s$ we obtain

$$\begin{aligned}
\sigma_A &\geq -\underbrace{\frac{\beta_1(1-\delta)}{L}}_{c_2} \frac{\nabla f(x)^\top d}{\|d\|^2} \\
c_2 &= \min \left\{ \gamma, \frac{\beta_1(1-\delta)}{L} \right\}
\end{aligned}$$

□

Remark: One may for example choose $\beta_1 = \beta_2 = \frac{1}{2}$.

For the choice of parameters: see, e.g., [1].

4.4.3 Powell step size

This method chooses σ such that

$$f(x + \sigma d) \leq f(x) + \delta \sigma \nabla f(x)^\top d \quad (\text{cf. Armijo}) \quad (4.28)$$

and

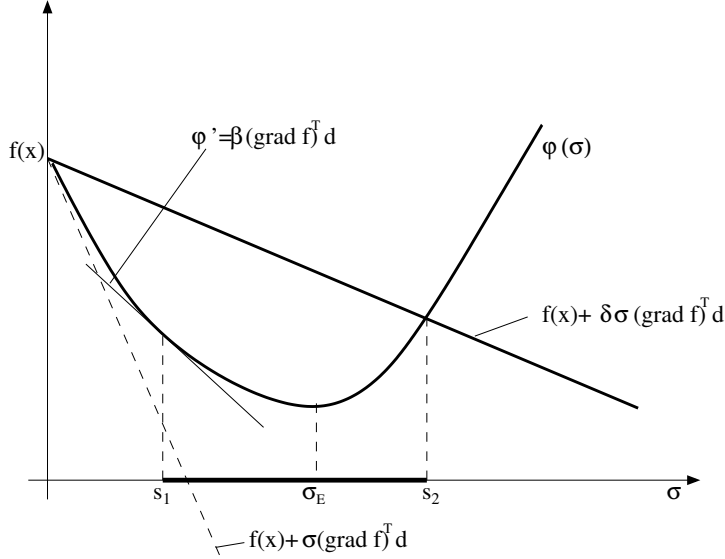
$$\nabla f(x + \sigma d)^\top d \geq \beta \nabla f(x)^\top d \quad (\text{minimal step size}) \quad (4.29)$$

with $0 < \delta < \beta < 1$.

Geometric interpretation $\varphi(s) := f(x + sd)$. Then it holds

$$\varphi'(s) = \nabla f(x + sd)^\top d.$$

Hence, the method computes $\sigma = \sigma_p$ as follows:



The existence of such a step size is shown in [1, Satz 4.4.5]. The computation is achieved by *nested intervals*.

For this we define

$$G_1(\sigma) = \begin{cases} \frac{f(x+\sigma d) - f(x)}{\sigma \nabla f(x)^\top d} & , \text{ for } \sigma > 0, \\ 1 & , \text{ for } \sigma = 0 \end{cases}$$

$$G_2(\sigma) = \frac{\nabla f(x + \sigma d)^\top d}{\nabla f(x)^\top d}$$

Then (4.28) $\Leftrightarrow G_1(\sigma) \geq \delta$ and (4.29) $\Leftrightarrow G_2(\sigma) \leq \beta$

Geometrically: \mathbb{R}_+ is split into 3 intervals $[0, s_1] \cup [s_1, s_2] \cup (s_2, \infty) =: I_1 \cup I_2 \cup I_3$ with $\varphi'(s_1) = \beta \nabla f(x)^\top d$ and $\varphi(s_2) = f(x) + s_2 \nabla f(x)^\top d$ with

$$\begin{aligned} G_1(\sigma) &\geq \delta \text{ and } G_2(\sigma) \geq \beta && \text{ in } I_1, \\ G_1(\sigma) &\geq \delta \text{ and } G_2(\sigma) \leq \beta && \text{ in } I_2, \\ G_1(\sigma) &\leq \delta \text{ and } G_2(\sigma) \leq \beta && \text{ in } I_3. \end{aligned}$$

Algorithm 4.4.2 (Powell)

1. Choose an initial step size $\sigma_0 > 0$, $j := 0$

(i) If $G_1(\sigma_0) \geq \delta$ and $G_2(\sigma_0) \leq \beta$: Done! $\sigma_p := \sigma_0$

(ii) If $\sigma_0 \in I_1$

$$a_0 := \sigma_0$$

$$b_0 := 2^l \sigma_0 \text{ with } \underline{\text{minimal}} \ l \in \mathbb{N}, \text{ s.t. } G_1(b_0) < \delta$$

Go to step 2.

(iii) If $\sigma_0 \in I_3$

$$b_0 = \sigma_0$$

$$a_0 = 2^{-l} \sigma_0 \text{ with } \underline{\text{minimal}} \ l \in \mathbb{N}, \text{ s.t. } G_2(a_0) > \beta \text{ and } G_1(a_0) \geq \delta$$

2. Mean $\sigma_j := \frac{1}{2}(a_j + b_j)$

(i) If $\sigma_j \in I_2$: done, $\sigma_p := \sigma_j$

(ii) If $\sigma_j \in I_1$: Set $a_{j+1} = \sigma_j, b_{j+1} = b_j$

(iii) If $\sigma_j \in I_3$:

$$a_{j+1} = a_j, b_{j+1} = \sigma_j$$

3. $j := j + 1$, go to step 2.

The Powell algorithm can also increase the step size, beginning with σ_0 . Hence, σ_0 can (generally) be arbitrary. Typical choices for β and δ are, e.g., $\delta = 0.1$ and $\beta = 0.9$.

Remarks:

- σ_p is (under suitable assumptions) obtained in finitely many steps (cf. [1, Satz 4.5.10])
- Assuming (LSC), (FD), (LUC) we have the following general result:
If σ_k is chosen according to Armijo/Powell, then the sequence $\{\sigma_k\}$ is efficient

4.5 Gradient descent (steepest descent)

Also called “method of steepest descent”. The direction is chosen as

$$d^{(k)} = -\nabla f(x^{(k)})$$

Remark: Easy to implement, but very slow.

Definition 4.5.1 Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $x \in \mathbb{R}^n$ arbitrary with $\nabla f(x) \neq 0$. Let $d \in \mathbb{R}^n$ denote the solution of

$$\min_{\|d\|=1} \nabla f(x)^\top d. \quad (4.30)$$

Every vector of the form $s = \lambda d, \lambda > 0$ is called **direction of steepest descent of f in x** .

Theorem 4.5.1 Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $x \in \mathbb{R}^n$ arbitrary with $\nabla f(x) \neq 0$. Then (4.30) has the unique solution $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$. In particular, s is a direction of steepest descent of f in x , if and only if $\exists \lambda > 0$ s.t. $s = -\lambda \nabla f(x)$.

Proof: Recall the Cauchy-Schwarz inequality: $|v^\top w| \leq \|v\| \|w\|$ with equality if and only if $v = \alpha w$. For $d \in \mathbb{R}^n, \|d\| = 1$ we thus have

$$\nabla f(x)^\top d \geq -\|\nabla f(x)\| \|d\| = -\|\nabla f(x)\|$$

with equality if and only if $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$. This shows the first assertion. The second assertion follows by the previous definition.

□

Algorithm 4.5.1

1. Choose $x^{(0)}, k := 0$, stopping criterion $\varepsilon > 0$.
2. If $\|\nabla f(x^{(k)})\| < \varepsilon$: Done
3. Compute

$$\begin{aligned}
 d^{(k)} &= -\nabla f(x^{(k)}) \\
 \sigma_k &\text{ as efficient step size (e.g. by Armijo)} \\
 x^{(k+1)} &:= x^{(k)} + \sigma_k d^{(k)} \\
 k &:= k + 1, \text{ goto 2.}
 \end{aligned}$$

□

Theorem 4.5.2 *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. If in Algorithm 4.5.1 we use the Armijo step size with $\sigma_0 = 1$ and $\beta = \beta_1 = \beta_2 < 1$, the method either terminates after a finite number of steps with a stationary point $x^{(k)}$ or it constructs a sequence $\{x^{(k)}\}$ s.t.*

1. For all k : $f(x^{(k+1)}) < f(x^{(k)})$
2. Every accumulation point of $x^{(k)}$ is a stationary point of f .

Proof: We only have to focus on $k \rightarrow \infty$. With $\sigma = 1, \beta = \beta_1 = \beta_2$ we know that $\sigma_k \in (0, 1]$ and

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)} + \sigma_k d^{(k)}) - f(x^{(k)}) \leq -\sigma_k \delta \|\nabla f(x^{(k)})\|.$$

This shows the first assertion. For the second, consider an accumulation point \bar{x} of $\{x^{(k)}\}$ and a subsequence $\{x^{(k)}\}_{k \in K}$ with $x^{(k)} \rightarrow \bar{x}$ for $K \ni k \rightarrow \infty$. Since $\{f(x^{(k)})\}$ is monotonically decreasing, there exists $\varphi \in \mathbb{R} \cup \{-\infty\}$ with $f(x^{(k)}) \rightarrow \varphi$ for $k \rightarrow \infty$. By continuity of f and $x^{(k)} \rightarrow \bar{x}$ for $K \ni k \rightarrow \infty$, we conclude that $f(x^{(k)}) \rightarrow f(\bar{x})$ for $K \ni k \rightarrow \infty$. Hence, $\varphi = f(\bar{x})$ and $f(x^{(k)}) \rightarrow f(\bar{x})$.

From the Armijo rule, we know that

$$f(x^{(0)}) - f(\bar{x}) = \sum_{k=0}^{\infty} (f(x^{(k)}) - f(x^{(k+1)})) \geq \delta \sum_{k=0}^{\infty} \sigma_k \|\nabla f(x^{(k)})\|^2$$

which implies that $\sigma_k \|\nabla f(x^{(k)})\|^2 \rightarrow 0$ for $k \rightarrow \infty$.

Assume now that $\nabla f(\bar{x}) \neq 0$. Then by continuity of ∇f and $x^{(k)} \rightarrow \bar{x}$ for $K \ni k \rightarrow \infty$ there exists $\ell \in K$ with

$$\|\nabla f(x^{(k)})\| \geq \frac{\|\nabla f(\bar{x})\|}{2} > 0 \quad \forall k \in K, k \geq \ell.$$

This yields $\sigma_k \rightarrow 0$ for $K \ni k \rightarrow \infty$. In particular, there exists $\ell' \in K, \ell' \geq \ell$ s.t. $\sigma_k \leq \beta$ for all $k \in K, k \geq \ell'$. Again with the Armijo rule this yields

$$f(x^{(k)} + \beta^{-1}\sigma_k d^{(k)}) - f(x^{(k)}) > -\delta\beta^{-1}\sigma_k \|\nabla f(x^{(k)})\|^2 \quad (4.31)$$

for all $k \in K, k \geq \ell'$. Consider then $\{t_k\}_{k \in K} = \{\beta^{-1}\sigma_k\}_{k \in K}$ and observe that $t_k \rightarrow 0$ for $K \ni k \rightarrow \infty$. By Taylor expansion, there exists $\tau_k \in [0, t_k]$ with

$$\begin{aligned} \lim_{K \ni k \rightarrow \infty} \frac{f(x^{(k)} + t_k d^{(k)}) - f(x^{(k)})}{t_k} &= \lim_{K \ni k \rightarrow \infty} \frac{t_k \nabla f(x^{(k)} + \tau_k d^{(k)})^\top d^{(k)}}{t_k} = -\|\nabla f(\bar{x})\|^2 \\ \lim_{K \ni k \rightarrow \infty} \|\nabla f(x^{(k)})\|^2 &= \|\nabla f(\bar{x})\|^2. \end{aligned}$$

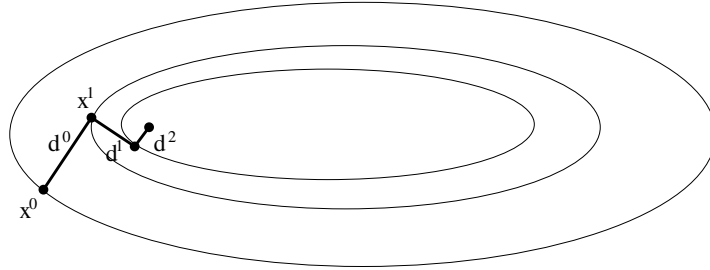
With (4.31) we obtain $0 < (1 - \delta)\|\nabla f(\bar{x})\|^2 \leq 0$ which shows that $\|\nabla f(\bar{x})\| \neq 0$ was false. \square

Disadvantage: The first steps may lead to fast descent but then it “can take a while”

Explanation: If we use the exact step size (which is generally hard), we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \sigma} f(\underbrace{x^{(k)} + \sigma d^{(k)}}_{x^{(k+1)}}) \Big|_{\sigma=\sigma_E} = \nabla f(x^{(k+1)})^\top d^{(k)} \\ &= -(d^{(k+1)})^\top d^{(k)} \\ \Rightarrow \quad (d^{(k+1)})^\top d^{(k)} &= 0 \end{aligned}$$

In flat valleys, the convergence is slow!



Theorem 4.5.3 *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be strictly convex and quadratic. Assume $x^{(k)}, \sigma_k$ are the iterates from Algorithm 4.5.1 with exact step size. Then*

$$\begin{aligned} f(x^{(k+1)}) - f(\bar{x}) &\leq \left(\frac{\lambda_{\max}(H) - \lambda_{\min}(H)}{\lambda_{\max}(H) + \lambda_{\min}(H)} \right)^2 (f(x^{(k)}) - f(\bar{x})) \\ \|x^{(k)} - \bar{x}\| &\leq \sqrt{\frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}}(H) \left(\frac{\lambda_{\max}(H) - \lambda_{\min}(H)}{\lambda_{\max}(H) + \lambda_{\min}(H)} \right)^k \|x^{(0)} - \bar{x}\| \end{aligned}$$

where $\bar{x} = -H^{-1}b$ is the global minimum of f and $\lambda_{\max}(H), \lambda_{\min}(H)$ denote maximal/minimal eigenvalues of H .

Remedy: Incorporate actual level sets!

4.6 Damped Newton's method

4.6.1 The algorithm

Descent direction = Newton direction

$$d^{(k)} = -f''(x^{(k)})^{-1} \nabla f(x^{(k)})$$

Algorithm 4.6.1 1. Initialize $x^{(0)} \in \mathbb{R}^n, k := 0$

2. If $\nabla f(x^{(k)}) = 0$. Done

3. Compute $d^{(k)}$ via

$$f''(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$

Step size σ_k : efficient (e.g. Armijo or Powell)

$$x^{(k+1)} := x^{(k)} + \sigma_k d^{(k)}$$

$k := k + 1$, goto 2.

4.6.2 Interpretation of Newton direction

Let $A = f''(x)$; A be symmetric positive definite and define an inner product and norm by

$$\begin{aligned} \langle x, y \rangle_A &:= x^\top A y \\ \|x\|_A &:= \sqrt{\langle x, x \rangle_A} = (x^\top A x)^{1/2} \end{aligned}$$

It then holds

Lemma 4.6.1 If $\nabla f(x) \neq 0$, the direction

$$\bar{d} = \frac{A^{(-1)} \nabla f(x)}{\|A^{(-1)} \nabla f(x)\|}$$

is the unique solution of the minimization problem

$$\begin{aligned} \min \quad & \nabla f(x)^\top d \\ \text{s.t.} \quad & \|d\|_A = 1. \end{aligned} \tag{4.32}$$

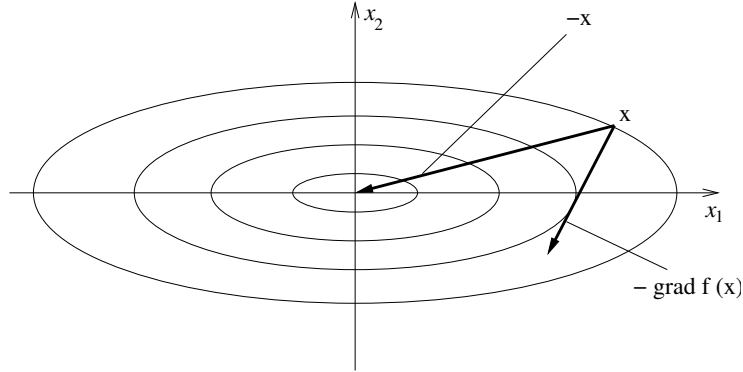
In other words, \bar{d} is the direction of steepest descent w.r.t. $\|\cdot\|_A$.

The benefit of choosing $-A^{(-1)} \nabla f$ instead of the gradient direction $-\nabla f$ is apparent if we consider the quadratic function

$$f(x) = \frac{1}{2} x^\top H x,$$

e.g., for $H = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ with $a, b > 0$. The level sets of $f(x)$ are ellipsoids of the form $ax_1^2 + bx_2^2 = r^2$. The gradient descent method yields a direction which does not go through the origin – the solution of $\min f(x)$.

On the other hand, with $-f''(x)^{-1}\nabla f(x) = -H^{-1}Hx = -x$ we obtain exactly the desired direction.



Consequence:

For a quadratic problem, damped Newton's method with exact step size would converge in one step.

4.6.3 Convergence of the method

Assume that (LUC) holds and that f'' is uniformly Lipschitz continuous on $N(f, f(x^{(0)}))$, i.e.,

$$\|f''(x) - f''(y)\| \leq L \|x - y\| \quad \forall x, y \in N(f, f(x^{(0)})). \quad (f2L)$$

This implies that the computed matrices $f''(x^{(k)})$ are positive definite and the method is well-posed.

By the general convergence result 4.3.3 the damped Newton's method converges linearly.

In fact, we even have:

Theorem 4.6.1 *Assume (LUC) holds and the step sizes σ_k are computed by Armijo or Powell, with the initial step size of damped Newton's method being $\sigma_0 = 1$. Further assume that $0 < \delta < 1/2$. Then for sufficiently large k , it holds that $\sigma_k = 1$.*

Proof: Cumbersome, see [1, Satz 474].

Consequence: After finitely many steps, damped Newton's method becomes Newton's method. Hereafter, it converges quadratically if (f2L) is satisfied, otherwise it converges with a super-linear rate.

Definition 4.6.1 $\{x^{(k)}\}$ converges super-linearly towards \tilde{x} , if

$$\lim_{n \rightarrow \infty} \frac{\|x^{(k+1)} - \tilde{x}\|}{\|x^{(k)} - \tilde{x}\|} = 0$$

Example: $x^{(k)} = q^k, |q| < 1$, converges linearly towards 0, not super-linearly. On the other hand, $x^{(k)} = \frac{q^k}{k!}$ converges super-linearly since

$$\frac{q^{k+1}}{(k+1)!} / \frac{q^k}{k!} = \frac{q}{k+1} \rightarrow 0, k \rightarrow \infty.$$

If the exact step size is used, it holds that $\sigma^k \rightarrow 1, k \rightarrow \infty$, and one can show quadratic convergence (cf. [1, Satz 4.6.4]).

Remark: (*Modification of the method*)

- If one uses $f''(x^{(0)})$ instead of $f''(x^{(k)})$ (simplified Newton's method), one obtains global (but linear) convergence.
- Recomputing (an approximation of) $f''(x^{(k)})$ after every n th step yields super-linear convergence.
- If derivatives are approximated by difference quotients, one obtains super-linear convergence if the discretization is sufficiently fine.

4.7 Variable metric and quasi Newton methods

General idea: based on information about $f''(x)$ (or approximations thereof), one uses a norm $\|\cdot\|_A$ benutzt, welche which incorporates the curvature of the level sets – as discussed previously for the quadratic case.

4.7.1 General procedure

Algorithm 4.7.1 (Variable metric)

1. Initialize $x^{(0)} \in \mathbb{R}^n, k := 0$
2. If $\nabla f(x^{(k)}) = 0$: stop
3. Compute:
 - symmetric positive definite matrix $A^{(k)}$
 - $d^{(k)} = -(A^{(k)})^{(-1)} \nabla f(x^{(k)})$
 - efficient step size σ_k
 - $x^{(k+1)} := x^{(k)} + \sigma_k d^{(k)}$
 - $k := k + 1$, goto 2.

Special cases: $A^{(k)} \equiv I$: gradient descent
 $A^{(k)} = f''(x^{(k)})$: damped Newton's method

In every step, one chooses the direction of steepest descent w.r.t. the norm $\|\cdot\|_{A^{(k)}}$.

4.7.2 Global convergence of variable metric methods

Assumption: uniform positive definiteness and boundedness of the matrices $A^{(k)}$.

Definition 4.7.1 A sequence $\{A^{(k)}\}$ of symmetric $n \times n$ matrices is called uniformly positive definite and bounded, if there exist constants $0 < \alpha_1 < \alpha_2$ s.t.

$$\alpha_1 \|x\|^2 \leq x^\top A^{(k)} x \leq \alpha_2 \|x\|^2 \quad \forall x \in \mathbb{R}^n$$

for all $k \in \mathbb{N}$.

(Equivalently: smallest eigenvalue $\lambda_1^{(k)} \geq \alpha_1$, largest eigenvalue $\leq \alpha_2$ or: smallest eigenvalue of $(A^k)^{-1} \geq \alpha_2^{-1}$, largest eigenvalue of $\leq \alpha_1^{-1}$)

It is not surprising that the above property for $\{A^{(k)}\}$ implies:

- (LSC), (FD) \Rightarrow directions $d^{(k)}$ are strictly gradient related
- similar convergence results as in Theorem 4.3.1 (accumulation points of $x^{(k)}$ with $\nabla f = 0$), 4.3.2 (convergence towards a zero of ∇f) and 4.3.3 (linear convergence)

4.7.3 Quasi Newton methods

We first start with the general idea.

Drawback of damped Newton's method: expensive computation of $f''(x^{(k)})$ in every step. Instead one is interested in computing a sequence of matrices $\{A^{(k)}\}$ s.t.:

- easy transition from $A^{(k)}$ to $A^{(k+1)}$
- $A^{(k)}$ is an approximation of $f''(x^{(k)})$

additionally, $A^{(k)}$ should be symmetric and positive definite.

Let us start by investigating the quadratic case

$$f(x) = \frac{1}{2} x^\top H x + b^\top x.$$

In this case $f''(x) = H$ and, hence

$$\begin{aligned} f''(x^{(k+1)}) (x^{(k+1)} - x^{(k)}) &= H (x^{(k+1)} - x^{(k)}) + b - b \\ &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), \quad k = 0, 1, 2, \dots \end{aligned}$$

If H is unknown and we only know the gradients of f as well as the vectors $x^{(0)}, \dots, x^{(n-1)}$, we obtain n linear systems

$$H (x^{(k+1)} - x^{(k)}) = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), \quad k = 0, \dots, n-1$$

which uniquely define the matrix H .

With this in mind, for $A^{(k)}$, initiated with a symmetric positive definite matrix $A^{(0)}$, we demand

$$\boxed{A^{(k+1)} (x^{(k+1)} - x^{(k)}) = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})}. \quad (4.33)$$

(4.33) is called *quasi Newton equation*

4.7.4 BFGS-Update

The quasi Newton equation does not have a unique solution but formulas have been developed which ensure a comparably easy computation of $A^{(k+1)}$. Most well-known: BFGS formula (named after Broyden, Fletcher, Goldfarb, Shanno).

We define:

$$\begin{aligned} x^{(k+1)} - x^{(k)} &=: s^{(k)} \\ \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) &=: y^{(k)} \end{aligned}$$

Given $A^{(k)}$ we obtain $A^{(k+1)}$ in two steps:

- First

$$\boxed{\tilde{A}^{(k)} := A^{(k)} - \frac{(A^{(k)} s^{(k)})(A^{(k)} s^{(k)})^\top}{(s^{(k)})^\top A^{(k)} s^{(k)}}} \quad (4.34)$$

If $A^{(k)}$ was symmetric and positive definite, then $\tilde{A}^{(k)}$ is symmetric and at least positive semi-definite. Moreover, it holds that

$$\tilde{A}^{(k)} s^{(k)} = 0,$$

such that \tilde{A}^k does not satisfy the quasi Newton equation. Obviously, it holds that

$$\text{rk}(A^{(k)} s^{(k)})(A^{(k)} s^{(k)})^\top = 1,$$

which is called (4.34) *symmetric rank one update*.

- Using another rank one update, one tries to compute a positive definite matrix:

$$A^{(k+1)} = \tilde{A}^{(k)} + \gamma_k w^{(k)} (w^{(k)})^\top$$

while, at the same time, satisfying the quasi Newton equation.

Quasi Newton equation:

$$A^{(k+1)} s^{(k)} = \underbrace{\tilde{A}^{(k)}}_{=0} + \gamma_k \overbrace{w^{(k)} (w^{(k)})^\top}^{\substack{\in \mathbb{R}, \\ =: \frac{1}{\gamma_k}}} s^{(k)} \stackrel{(!)}{=} y^{(k)}$$

$\Rightarrow w^{(k)}$ has to be a multiple of $y^{(k)}$ such that we choose $w^{(k)} = y^{(k)}$ and

$$\gamma_k = \frac{1}{(y^{(k)})^\top s^{(k)}}$$

Positive definiteness:

For the specific direction $s^{(k)}$ we obtain

$$0 < (s^{(k)})^\top A^{(k+1)} s^{(k)} = (s^{(k)})^\top y^{(k)} \quad (4.35)$$

One can show that $(s^{(k)})^\top y^{(k)} > 0$ implies positive definiteness of $A^{(k+1)}$ if $A^{(k)}$ was positive definite [1, Lemma 4.8.5].

Altogether:

$$A^{(k+1)} = A^{(k)} - \frac{A^{(k)} s^{(k)} (A^{(k)} s^{(k)})^\top}{(s^{(k)})^\top A^{(k)} s^{(k)}} + \frac{y^{(k)} (y^{(k)})^\top}{(y^{(k)})^\top s^{(k)}} \quad (4.36)$$

Since the sum of two rank one matrices is generally of rank two, this is called a *rank two update*.

Remark:

(4.35) is satisfied for a quadratic function if H is positive definite:

$$f(x) = \frac{1}{2} x^\top H x \quad \nabla f(x) = H x$$

$$\begin{aligned} \Rightarrow (y^{(k)})^\top s^{(k)} &= (\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}))^\top (x^{(k+1)} - x^{(k)}) \\ &= (H(x^{(k+1)} - x^{(k)}))^\top (x^{(k+1)} - x^{(k)}) \\ &\geq \alpha \|x^{(k+1)} - x^{(k)}\|^2 > 0. \end{aligned}$$

□

Additional remarks:

- Show that $\tilde{A}^{(k)} s^{(k)} = 0$ (without index k)

$$\begin{aligned} \tilde{A}s &= As - \frac{(As)(As)^\top}{s^\top As} s \\ &= \frac{1}{s^\top As} [(s^\top As) As - As \underbrace{s^\top A^\top s}_{=s^\top As, \text{ since } A \text{ is symmetric}}] \\ &= \frac{1}{s^\top As} [As \underbrace{\{s^\top As I - s^\top As I\}}_{=0}] = 0. \end{aligned}$$

- Show that matrix of type ss^\top has rank one:

$$\begin{aligned}
ss^\top = (s_i s_j) &= \begin{pmatrix} s_1 s_1 & s_1 s_2 & \dots & s_1 s_n \\ s_2 s_1 & s_2 s_2 & \dots & s_2 s_n \\ \vdots & & & \\ s_n s_1 & s_n s_2 & \dots & s_n s_n \end{pmatrix} \\
&= \begin{pmatrix} s_1 \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}, s_2 \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}, \dots, s_n \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix} \end{pmatrix}
\end{aligned}$$

Columns are multiples of $s \Rightarrow$ rank one.

4.7.5 BFGS for quadratic problems

We already know: for a quadratic function f , we obtain the exact step size σ_E via

$$\sigma_E = \frac{\nabla f(x)^\top d}{d^\top H d}$$

This will be used below.

Algorithm 4.7.2 (BFGS for (QU))

1. Initialize $x^{(0)}$, symmetric positive definite matrix $A^{(0)}$, $k := 0$.
2. If $\nabla f(x^{(0)}) = 0$: stop.
3. Compute

$$\begin{aligned}
d^{(k)} &= - (A^{(k)})^{(-1)} \nabla f(x^{(k)}) \\
\sigma_k &= \frac{\nabla f(x^{(k)})^\top d^{(k)}}{(d^{(k)})^\top H d^{(k)}} \\
x^{(k+1)} &= x^{(k)} + \sigma_k d^{(k)} \\
s^{(k)} &= x^{(k+1)} - x^{(k)} \\
y^{(k)} &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\
A^{(k+1)} &\text{ via BFGS update} \\
k &:= k + 1, \text{ goto 2.}
\end{aligned}$$

One typically computes $A^{(k+1)}$ via a slightly different alternative.

The following concept is of particular importance

Definition 4.7.2 (*H orthogonality*).

Let $H \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. The vectors $d^{(0)}, \dots, d^{(k)}, k < n$, are called **conjugate** or **orthogonal** w.r.t. H , if they are non-zero and further satisfy

$$\boxed{(d^{(i)})^\top H d^{(j)} = 0, \quad 0 \leq i < j \leq k.}$$

This happens for BFGS:

Theorem 4.7.1 Let H be symmetric and positive definite. Then the BFGS method for (QU) computes the minimum \tilde{x} of f in at most $m \leq n$ steps. If $m = n$, then $A^{(n)} = H$.

Proof idea: Let $\nabla f(x^{(0)}) \neq 0$ (otherwise we are done). Assume then that $x^{(1)}, y^{(1)}, s^{(1)}$ and $A^{(0)}$ are computed via the algorithm. Then $A^{(1)}$ is positive definite by [1, Lemma 4.8.5].

$$\begin{aligned} \underbrace{x^{(1)} - x^{(0)}}_{s^{(0)}} &= \sigma_0 d^{(0)}, \quad H s^{(0)} = \nabla f(x^{(1)}) - \nabla f(x^{(0)}) = y^{(0)} \\ \Rightarrow \nabla f(x^{(1)}) &= \nabla f(x^{(0)}) + H s^{(0)} = \nabla f(x^{(0)}) + \sigma_0 H d^{(0)} \\ \Rightarrow \nabla f(x^{(1)})^\top d^{(0)} &= \nabla f(x^{(0)})^\top d^{(0)} + \underbrace{\sigma_0 (d^{(0)})^\top H d^{(0)}}_{= -\nabla f(x^{(0)})^\top d^{(0)} \text{ by def. of } \sigma_0} = 0 \quad (*) \end{aligned}$$

Since the above implies $d^{(0)} = \sigma_0^{-1} s^{(0)}$, we obtain

$$(d^{(0)})^\top H d^{(1)} = \frac{1}{\sigma_0} \underbrace{(H s^{(0)})^\top}_{y^{(0)\top} \text{ s. o.}} \underbrace{d^{(1)}}_{(-A^{(1)})^{-1} \nabla f \text{ (BFGS method)}} = - \frac{\overbrace{(s^{(0)})^\top : \text{quasi Newton}}^{(y^{(0)})^\top (A^{(1)})^{-1}} \nabla f(x^{(1)})}{\sigma_0}$$

Now we utilize the quasi Newton equation for $A^{(1)}$:

$$\begin{aligned} (A^{(1)})^{-1} y^{(0)} &= s^{(0)} \\ \Rightarrow (d^{(0)})^\top H d^{(1)} &= - \frac{(s^{(0)})^\top \nabla f(x^{(1)})}{\sigma_0} = -\nabla f(x^{(1)})^\top d^{(0)} = 0. \quad ((*)) \end{aligned}$$

For $k = 1$ this shows:

$$\left. \begin{aligned} \text{(i)} \quad \nabla f(x^{(k)})^\top d^{(i)} &= 0 \\ \text{(ii)} \quad (A^{(k)})^{-1} y^{(i)} &= s^{(i)} \\ \text{(iii)} \quad (d^{(i)})^\top H d^{(k)} &= 0 \end{aligned} \right\} \text{ for } 0 \leq i < k$$

Induction $k \rightarrow k+1 \dots$ shows the assertion. □

Remark:

- The result is only true for exact arithmetic and if the exact step size is used
- alternatively, we could solve

$$H\tilde{x} + b = 0$$

with the Cholesky decomposition.

- One step of BFGS requires a similar amount of computations as one Cholesky decomposition. As a consequence, BFGS for (QU) does not pay off. It rather aims at nonlinear problems (for which it is implemented in MATLAB or NAGLIB).

4.7.6 BFGS for general nonlinear problems

The method is similar to the quadratic case. However, the exact step size σ_E is generally not available. One can show that

- the method converges linearly if efficient step sizes are used and (LUC) [1, Satz 4.8.12].
- If additionally (f2L) and the step sizes are computed by Armijo or Powell, one obtains super-linear convergence. The generated matrices $A^{(k)}$ are uniformly positive definite and bounded [1, Satz 4.8.13].

In general, we do not obtain $A^{(k)} \rightarrow f''(\tilde{x})$ but only

$$\lim_{k \rightarrow \infty} \frac{\| (A^{(k)} - f''(\tilde{x})) d^{(k)} \|}{\| d^{(k)} \|} \rightarrow 0.$$

4.8 Conjugate direction methods

4.8.1 CG for quadratic optimization problems

For BFGS we obtain H -orthogonal directions and convergence after at most n steps. Drawback: the matrices $A^{(k)}$ have to be stored which for large dimensions n can become problematic: for $n = 10000$ unknowns, we have to store 10^8 entries.

Moreover, matrices may have a particular structure which allows to compute matrix vector products efficiently without setting up the matrix. For example, the Hilbert matrix is defined by

$$H_{ij} = \frac{1}{i+j-1}, \text{ i.e., we have}$$

$$[Hv]_i = \sum_{j=1}^n H_{ij}v_j = \frac{v_i}{i}.$$

The idea of CG-methods (Conjugate Gradient) is to generate H -orthogonal directions without setting up $A^{(k)}$ explicitly.

Consider

$$(QU) \quad \min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^T H x + b^T x, \quad H \text{ symmetric positive definite}$$

Lemma 4.8.1 Let $d^{(0)}, d^{(1)}, \dots, d^{(n-1)}$ be conjugate directions. For every $x^{(0)} \in \mathbb{R}^n$ the iteration

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \sigma_k d^{(k)} \\ \sigma_k &= -\frac{\nabla f(x^{(k)})^\top d^{(k)}}{(d^{(k)})^\top H d^{(k)}} \quad (\text{exact step size}) \end{aligned}$$

computes the exact solution $x^{(n)} = -H^{-1}b$ after at most n steps.

PROOF.

$$\begin{aligned} \tilde{x} - x^{(0)} &= \sum_{i=0}^{n-1} \sigma_i d^{(i)} \quad (\text{Multiply from the left by } (d^{(i)})^\top \cdot H) \\ \Rightarrow \quad \sigma_k &= \frac{(d^{(k)})^\top H(\tilde{x} - x^{(0)})}{(d^{(k)})^\top H d^{(k)}} = -\frac{(d^{(k)})^\top (Hx^{(0)} + b)}{(d^{(k)})^\top H d^{(k)}} \\ &\dots = -\frac{(d^{(k)})^\top (Hx^{(k)} + b)}{(d^{(k)})^\top H d^{(k)}} = -\frac{(d^{(k)})^\top \nabla f(x^{(k)})}{(d^{(k)})^\top H d^{(k)}} \end{aligned}$$

□

Corollary 4.8.1 $x^{(k)}$ minimizes f on $\{x^{(k-1)} + \sigma d^{(k-1)} | \sigma \in \mathbb{R}\}$ as well as on $x^{(0)} + V_k$, where $V_k = \text{span}\{d^{(0)}, \dots, d^{(k-1)}\}$. In particular, it holds that

$$(d^{(i)})^\top \nabla f(x^{(k)}) = 0 \quad \text{for } i < k. \quad (*)$$

PROOF. It is sufficient to show (*).

It holds that $(d^{(i)})^\top \nabla f(x^{(i+1)}) = (d^{(i)})^\top (Hx^{(i+1)} + b) = (d^{(i)})^\top (\underbrace{Hx^{(i)} + b}_{=\nabla f(x^{(i)})}) + \sigma_i (d^{(i)})^\top H d^{(i)} =$

0 i.e. (*) holds true for $k = 1$ and for $i = k - 1$ if $k > 1$. It further holds that $\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = H(x^{(k+1)} - x^{(k)}) = \sigma_k H d^{(k)} \Rightarrow (d^{(i)})^\top (\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})) = 0$ for $k > i$ □

Algorithm 4.8.1 (Conjugate directions)

1. Initialize $x^{(0)}$, compute $d^{(0)} = -\nabla f(x^{(0)}) = -(Hx^{(0)} + b)$, $k := 0$
2. If $\nabla f(x^{(k)}) = 0 \rightarrow \text{stop}$
3. Compute

$$\sigma_k = \frac{\nabla f(x^{(k)})^\top \nabla f(x^{(k)})}{(d^{(k)})^\top H d^{(k)}}$$

$$x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$$

$$\nabla f(x^{(k+1)}) = Hx^{(k+1)} + b = \nabla f(x^{(k)}) + \sigma_k H d^{(k)}$$

$$\beta_k = \frac{\|\nabla f(x^{(k+1)})\|^2}{\|\nabla f(x^{(k)})\|^2}$$

$$d^{(k+1)} = -\nabla f(x^{(k+1)}) + \beta_k d^{(k)}.$$

Remark: σ_k corresponds to the exact step size since

$$\begin{aligned}\sigma_E &= -\frac{\nabla f(x^{(k)})^\top d^{(k)}}{(d^{(k)})^\top H d^{(k)}} = -\frac{\nabla f(x^{(k)})^\top (-\nabla f(x^{(k)}) + \beta_{k-1} d^{(k-1)})}{(d^{(k)})^\top H d^{(k)}} \\ &= \frac{\nabla f(x^{(k)})^\top \nabla f(x^{(k)})}{(d^{(k)})^\top H d^{(k)}}\end{aligned}$$

Theorem 4.8.1 (*Properties of the CG method*). As long as $\nabla f(x^{(k-1)}) \neq 0$ it holds:

(1) $d^{(k-1)} \neq 0$

(2)

$$\begin{aligned}V_k &:= \text{span}\{\nabla f(x^{(0)}), H\nabla f(x^{(0)}), \dots, H^{k-1}\nabla f(x^{(0)})\} \\ &= \text{span}\{\nabla f(x^{(0)}), \dots, \nabla f(x^{(k-1)})\} \\ &= \text{span}\{d^{(0)}, \dots, d^{(k-1)}\}\end{aligned}$$

(3) $d^{(0)}, \dots, d^{(k-1)}$ are conjugated

(4) $f(x^{(k)}) = \min_{z \in V_k} f(x^{(0)} + z)$

PROOF. For $k = 1$ this is clear. Assume the statement is true for $k - 1$. Let us define $g^{(k)} := \nabla f(x^{(k)})$. Then

$$\begin{aligned}g^{(k)} &= g^{(k-1)} + \sigma_{k-1} H d^{(k-1)} \\ \Rightarrow g^{(k)} &\in V_{k+1} \quad \text{and} \quad \text{span}\{g^{(0)}, \dots, g^{(k)}\} \subset V_{k+1}.\end{aligned}$$

By induction assumption $d^{(0)}, \dots, d^{(k-1)}$ are conjugated.

Corollary 4.8.1 \Rightarrow

$$(d^{(i)})^\top g^{(k)} = 0 \quad \text{for } i < k \quad (*)$$

$g^{(k)} \neq 0 \Rightarrow \{d^{(0)}, \dots, d^{(k-1)}, g^{(k)}\}$ and thus $\{g^{(0)}, \dots, g^{(k-1)}, g^{(k)}\}$ are linearly independent with dimension $k + 1$

$$\Rightarrow \text{span}\{g^{(0)}, \dots, g^{(k)}\} = V_{k+1}.$$

It holds $g^{(k)} + d^{(k)} = \beta_{k-1} d^{(k-1)} \in V_k$

$$\Rightarrow V_{k+1} = \text{span}\{d^{(0)}, \dots, d^{(k)}\} \Rightarrow (2).$$

Due to $g^{(k)} + d^{(k)} \in V_k$ it follows that $d^{(k)} \neq 0$ if $g^{(k)} \neq 0 \Rightarrow (1)$.

(3) follows by (longer) computations based on (*) and (2).

(4) follows with Lemma 4.8.1. □

4.8.2 Analysis of the CG method

For symmetric positive definite $H \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1 < \dots < \lambda_n$ the condition number is given by

$$\kappa(H) = \frac{\lambda_n}{\lambda_1}.$$

Applying the gradient descent method (with exact step size) to the quadratic optimization problem, one can show that the error in the energy norm $\|x\|_H := \sqrt{x^T H x}$ (cf., e.g., [7]) behaves as follows:

$$\|\tilde{x} - x^{(k+1)}\|_H \leq \left(\frac{\kappa(H) - 1}{\kappa(H) + 1} \right)^k \|\tilde{x} - x^{(0)}\|_H.$$

For the CG method, we obtain the following improved convergence estimate.

Theorem 4.8.2 *The approximation error of $\tilde{x} - x^{(k)}$ for the CG method can be bounded in the energy norm according to*

$$\|\tilde{x} - x^{(k)}\|_H \leq 2 \left(\frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)} + 1} \right)^k \|\tilde{x} - x^{(0)}\|_H$$

PROOF. From Theorem 4.8.1 we know

$$\|\tilde{x} - x^{(k)}\| \leq \|\tilde{x} - y\| \quad \forall y \in V_k. \quad (*)$$

Moreover, Theorem 4.8.1 also implies that $y \in V_k$ can be represented as linear combination of powers of H applied to $g^{(0)}$, i.e., there exists a polynomial P_{k-1} of degree $k-1$ s.t.

$$\begin{aligned} y &= x^{(0)} + P_{k-1}(H)g^{(0)} = x^{(0)} + P_{k-1}(H)(Hx^{(0)} + b) \\ &= x^{(0)} + HP_{k-1}(H)(x^{(0)} - \tilde{x}) \\ \Rightarrow \tilde{x} - y &= x - x^{(0)} - HP_{k-1}(H)(x^{(0)} - \tilde{x}) \\ &= \underbrace{(I + HP_{k-1}(H))}_{=: Q_k(H)}(\tilde{x} - x^{(0)}) \end{aligned}$$

with a polynomial $Q_k \in \mathcal{P}_k$ of degree k and $Q_k(0) = 1$.

Let $\{z_1, \dots, z_n\}$ denote an orthonormal system of eigenvectors of H , then

$$\begin{aligned} \tilde{x} - x^{(0)} &= \sum_{j=1}^n c_j z_j \\ \Rightarrow \tilde{x} - y &= \sum_{j=1}^n c_j Q_k(H) z_j = \sum_{j=1}^n c_j Q_k(\lambda_j) z_j \\ \Rightarrow \|\tilde{x} - y\|_H^2 &= \left[\sum_{j=1}^n c_j Q_k(\lambda_j) z_j \right]^\top H \left(\sum_{j=1}^n c_j Q_k(\lambda_j) z_j \right) \\ &= \sum_{j=1}^n \lambda_j c_j^2 Q_k^2(\lambda_j) \\ &\leq \min_{\substack{Q_k \in \mathcal{P}_k \\ Q_k(0)=1}} \max_{\lambda} |Q_k(\lambda)|^2 \underbrace{\sum_{j=1}^n \lambda_j c_j^2}_{= \|\tilde{x} - x^{(0)}\|_H^2}. \end{aligned}$$

By choosing Chebyshev polynomials of degree $\leq k$ and scaling their domain to $[\lambda_1, \lambda_n]$ one obtains the estimate

$$\alpha := \min_{\substack{Q_k \in \mathcal{P}_k \\ Q_k(0)=1}} \max_{1 \leq i \leq n} |Q_k(\lambda_i)| \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k$$

$$\text{with } \kappa = \kappa(H) = \frac{\lambda_n}{\lambda_1}.$$

4.8.3 Preconditioning

The previous results shows that we can expect fast convergence of the CG method if the condition number of H is small. The idea of *preconditioning* is to modify the problem such that the condition number of a modified system matrix is small, i.e., the contours of f should be as close as possible to circles.

In the following, we choose a positive definite B and consider the problem

$$\bar{H}\bar{x} = -b, \quad \text{with } \bar{x} = B^{-1}x \quad \text{and } \bar{H} = H \cdot B.$$

Attention: $\bar{H} = H \cdot B$ is not symmetric w.r.t. the Euclidean scalar product but w.r.t. $(\cdot, \cdot)_B =$ since

$$(x, HBy)_B = x^\top BHB y = (HBx)^\top By = (HB, y)_B.$$

The essential idea for the preconditioned CG method is to solve the auxiliary problem where the Euclidian scalar product (\cdot, \cdot) is replaced by $(\cdot, \cdot)_B$. Details are given in the book of Deuffhard and Hohmann.¹

For the corresponding approximation error, one can show:

$$\|\tilde{x} - x^{(k)}\|_H \leq 2 \left(\frac{\sqrt{\kappa(H \cdot B)} - 1}{\sqrt{\kappa(H \cdot B)} + 1} \right)^k \|\tilde{x} - x^{(0)}\|_A.$$

Preconditioning now requires to find a symmetric positive definite matrix B such that, on the one hand, products of the form By are easy to compute and, on the other hand, the condition number $\kappa(HB)$ is small.

Typical examples are

- (i) $B = D^{-1}$, where $D = \text{diag}(H)$ is a diagonal matrix with diagonal elements of H .
- (ii) Incomplete Cholesky decompositions of H .

4.8.4 CG methods for general nonlinear optimization problems

The method was first discussed by Fletcher and Reeves and is thus also called *Fletcher Reeves method*.

The steps are identical to the previous algorithm. Assuming knowledge of the exact step sizes $\sigma_k = \sigma_E$ one can show convergence (cf., e.g., [1, Satz 4.9.4]).

¹Deuffhard/Hohmann: *Numerische Mathematik 1*. de Gruyter, Berlin 1993.

4.9 Trust region methods

4.9.1 Motivation

Idea: so far, we computed $d^{(k)}$ and obtained σ^k by, e.g., one dimensional minimization
Now:

- (i) Replace f by a local approximation/model f_k , e.g.,
 - $f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\top d$
 - $f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)}) d$
- (ii) choose $\varrho_k > 0$ and define $B_{\varrho_k}(x^{(k)})$ as a trust region
- (iii) Compute $d^{(k)}$ as global solution to

$$\min_{\|d\| \leq \varrho_k} f_k(d). \quad (4.37)$$

Remark: If $f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\top d$

$$\text{it follows that } d^{(k)} = -\varrho_k \frac{\nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\|}$$

→ similar to the gradient descent method so that we do not expect good convergence behavior.

Requirements for a model f_k :

Given $x^{(k)} \in \mathbb{R}^n$, $\varrho_k > 0$ we demand

- (i) $f_k(0) = f(x^{(k)})$
- (ii) for $d^{(k)}$ solution to (4.37) it holds:

$$f_k(d^{(k)}) = f(x^{(k)}) \Rightarrow \nabla f(x^{(k)}) = 0$$

(stopping criterion!)

Example 4.9.1 $f \in C^2$ and $f_k(d) = f(x^{(k)}) + \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)}) d$ then $f_k(0) = f(x^{(k)})$.

If $f_k(d^{(k)}) = f(x^{(k)})$ then

$$f_k(d^{(k)}) = f(x^{(k)}) \leq f_k(d) \quad \forall d \in B_{\varrho_k}(0)$$

i.e., $\tilde{d} = 0$ solution to $\min_{\|d\| \leq \varrho_k} F(d)$ with

$$F(d) = \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)}) d$$

Optimality condition for a local minimum of F yields

$$0 = \nabla F(\tilde{d}) = \nabla f(x^{(k)}) + f''(x^{(k)}) \tilde{d} = \nabla f(x^{(k)}).$$

Selection of ϱ_k :

$$\text{compute } r_k = \frac{f(x^{(k)}) - f(x^{(k)} + d^{(k)})}{f(x^{(k)}) - f_k(d^{(k)})} = \frac{\text{actual descent}}{\text{model descent}}.$$

(The closer r_k is to 1, the more we trust in the region/radius.)

Details: choose $0 < \delta_1 < \delta_2 < 1$.

If

$$r_k \begin{cases} \in [\delta_1, \delta_2]: & x^{(k+1)} = x^{(k)} + d^{(k)}, \quad \varrho_k \text{ unchanged} \\ \geq \delta_2: & x^{(k+1)} = x^{(k)} + d^{(k)}, \quad \text{enlarge } \varrho_k \\ < \delta_1: & x^{(k+1)} = x^{(k)}, \quad \text{reduce } \varrho_k \end{cases}$$

4.9.2 Trust region Newton method

$$\min_{\|d\| \leq \varrho_k} \underbrace{f(x^{(k)}) + \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)}) d}_{=: f_k(d)} \quad (4.38)$$

Remark: except for the constraint, (4.38) coincides with $(Q)_k$, but (4.38) is always solvable while $(Q)_k$ requires positive definiteness of $f''(x)$. If ϱ_k is sufficiently small, then f_k is a good model and (4.38) leads to descent even if $f''(x^k)$ is not positive definite.

In the following, we consider

$$\min_{\|d\| \leq \varrho} \underbrace{c + b^\top d + \frac{1}{2} d^\top A d}_{=: \phi(d)}. \quad (4.39)$$

Define $g(d) := -\frac{1}{2}(\|d\|^2 - \varrho^2)$ then $\|d\| \leq \varrho \Leftrightarrow g(d) \geq 0$ and $\nabla g = -d$.

We introduce the Lagrangian

$$\phi_\lambda(d) = \phi(d) - \lambda g(d) = -\frac{1}{2} \varrho^2 \lambda + c + b^\top d + \frac{1}{2} d^\top (A + \lambda I) d.$$

Lemma 4.9.1 (cf. Chapter. 5)

Let $\tilde{d} \in \mathbb{R}^n$ with $\|\tilde{d}\| \leq \varrho$ and $\lambda \geq 0$, such that \tilde{d} is a (strict) global minimum of ϕ_λ and $\lambda g(\tilde{d}) = 0$. Then \tilde{d} is also a (strict) global minimum of (4.39).

PROOF. Let \tilde{d} be a global minimum of ϕ_λ and $\lambda g(\tilde{d}) = 0$

Hence, it follows that $\phi(\tilde{d}) = \phi_\lambda(\tilde{d}) \leq \phi_\lambda(d) = \phi(d) - \underbrace{\lambda g(d)}_{\geq 0} \leq \phi(d)$

(similarly for strict local minima). □

In Chapter 5 we will show

Lemma 4.9.2 Let $\tilde{d} \in \mathbb{R}^n$ be a global minimum of (4.39). Then there exists exactly one Lagrange multiplier $\lambda > 0$, s.t.:

$$\nabla \phi_\lambda(\tilde{d}) = \nabla \phi(\tilde{d}) + \lambda \tilde{d} = 0 \quad (\text{First order necessary conditions})$$

$$d^\top \phi''_\lambda(\tilde{d})d = d^\top (A + \lambda I)d \geq 0 \quad \forall d \in \mathbb{R}^n \quad (\text{Second order necessary conditions})$$

$$\lambda g(\tilde{d}) = 0 \quad (\text{Complementarity})$$

Theorem 4.9.1 Let $\tilde{d} \in \mathbb{R}^n$ with $\|\tilde{d}\| \leq \varrho$. Then:

\tilde{d} global solution to (4.39).

\Leftrightarrow there exists $\lambda \geq 0$ with

$$(i) \quad (A + \lambda I)\tilde{d} = -b.$$

$$(ii) \quad \lambda(\|\tilde{d}\| - \varrho) = 0, \text{ i.e., } \|\tilde{d}\| = \varrho \text{ if } \lambda > 0.$$

$$(iii) \quad A + \lambda I \text{ positive semidefinite.}$$

If $A + \lambda I$ is positive definite, \tilde{d} is uniquely determined.

PROOF. “ \Rightarrow ” (Lemma 4.9.2)

“ \Leftarrow ” (i)–(iii) satisfied with $\lambda \geq 0$

$$(ii) \Rightarrow \lambda g(\tilde{d}) = 0$$

$$\left. \begin{array}{l} (iii) \Rightarrow \phi_\lambda \text{ convex on } \mathbb{R}^n \\ (i) \Rightarrow \nabla \phi_\lambda(\tilde{d}) = (A + \lambda I)\tilde{d} + b = 0 \end{array} \right\} \Leftrightarrow$$

\tilde{d} global minimum of $\phi_\lambda \xrightarrow{\text{Lemma 4.9.1}} \tilde{d}$ global minimum of (4.39). □

W.r.t. the stopping criterion:

Lemma 4.9.3 Let $\tilde{d} \in \mathbb{R}^n$ be global solution to (4.39) with $\|\tilde{d}\| \leq \varrho$. Then

$$\phi(\tilde{d}) = c \quad \Leftrightarrow \quad b = 0 \quad \text{and} \quad A \text{ positive semidefinite}$$

$$\text{i.e., } f_k(d^{(k)}) = f(x^{(k)})$$

PROOF.

$$\phi(0) = c = \phi(\tilde{d}) \leq \phi(d) \Rightarrow 0 \text{ is solution to (4.39).}$$

Theorem (i) $\Rightarrow b = 0$, (ii) $\Rightarrow \lambda = 0$ (iii) $\Rightarrow A$ positive semidefinite

“ \Leftarrow ” obvious □

Application to (4.37), i.e.,

$$f_k = f(x^{(k)}) + \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)})d.$$

Let $d^{(k)}$ be global solution to (4.37), then

$$f_k(d^{(k)}) = f(x^{(k)}) \Leftrightarrow \nabla f(x^{(k)}) = 0 \quad \text{and} \quad f''(x^{(k)}) \text{ positive semidefinite}$$

i.e., $f_k(d^{(k)}) = f(x^{(k)})$ is an appropriate stopping criterion.

Lemma 4.9.4 (*Estimation of descent*).

Let $\tilde{d} \in \mathbb{R}^n$ with $\|\tilde{d}\| \leq \varrho$ global solution to (4.39). Then

$$c - \phi(\tilde{d}) \geq \frac{1}{2} \|b\| \min \left\{ \varrho, \frac{\|b\|}{\|A\|} \right\}.$$

PROOF. $d = 0$ is admissible $\phi(0) \geq \phi(\tilde{d})$

$\Rightarrow c - \phi(\tilde{d}) \geq c - \phi(0) = 0$ Hence, w.l.o.g., assume $b \neq 0$ and observe that

$$\begin{aligned} c - \phi(\tilde{d}) &= -b^\top d - \frac{1}{2} d^\top A d \\ &\geq -b^\top d - \frac{1}{2} \|A\| \|d\|^2 \end{aligned}$$

We distinguish the following two cases:

(i) $\varrho \|A\| \leq \|b\|$, choose $d = -\varrho \frac{b}{\|b\|}$

then $c - \phi(\tilde{d}) \geq \varrho \|b\| - \frac{\varrho^2}{2} \|A\| \geq \frac{\varrho}{2} \|b\|$

(ii) $\varrho \|A\| > \|b\|$, choose $d = -\frac{1}{\|A\|} b$

then $c - \phi(\tilde{d}) \geq \frac{\|b\|^2}{\|A\|} - \frac{1}{2} \frac{\|b\|^2}{\|A\|} = \frac{1}{2} \frac{\|b\|^2}{\|A\|}$
 $\Rightarrow c - \phi(\tilde{d}) \geq \frac{1}{2} \|b\| \min \left\{ \varrho, \frac{\|b\|}{\|A\|} \right\}.$

Algorithm: (Trust region Newton)

Given: $0 < \delta_1 < \delta_2 < 1$, $\sigma_1 \in]0, 1[$, $\sigma_2 > 1$, $\varrho_0 > 0$

1. Choose $x^{(0)} \in \mathbb{R}^n$

2. Compute global solution $d^{(k)}$ to

$$\min_{\|d\| \leq \varrho_k} \underbrace{f(x^{(k)}) + \nabla f(x^{(k)})^\top d + \frac{1}{2} d^\top f''(x^{(k)}) d}_{=: f_k(d)}$$

If $f(x^{(k)}) = f_k(d^{(k)}) \rightarrow \text{stop.}$

3. Compute

$$r_k = \frac{f(x^{(k)}) - f(x^{(k)} + d^{(k)})}{f(x^{(k)}) - f_k(d^{(k)})}.$$

If $r_k \geq \delta_1$ (successful)

- set $x^{(k+1)} = x^{(k)} + d^{(k)}$,
 - compute $\nabla f(x^{(k+1)}), f''(x^{(k+1)})$.
 - update ϱ_k :
 - if $r_k \in [\delta_1, \delta_2[$, choose $\varrho_{k+1} \in [\sigma_1 \varrho_k, \varrho_k]$
 - if $r_k \geq \delta_2$, choose $\varrho_{k+1} \in [\varrho_k, \sigma_2 \varrho_k]$.
 - Goto 2.
4. If $r_k < \delta_1$ (not successful)
- choose $\varrho_{k+1} \in]0, \sigma_1 \varrho_k]$,
 - set $x^{(k+1)} = x^{(k)}, \nabla f(x^{(k+1)}) = \nabla f(x^{(k)}), f''(x^{(k+1)}) = f''(x^{(k)})$
 - $k \rightarrow k + 1$, Goto 2.

5 Constrained problems – theory

5.1 Introductory examples

We consider

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in I \end{cases}$$

with index sets $I, \mathcal{E} \in \{1, \dots, n\}$ and $I \cap \mathcal{E} = \emptyset$.

$c_i(x) = 0, i \in \mathcal{E}$ are called equality constraints

$c_i(x) \geq 0, i \in I$ are called inequality constraints.

$\Omega = \{x \in \mathbb{R}^n \mid c_i(x) = 0, i \in \mathcal{E}, c_i(x) \geq 0, i \in I\}$ is called admissible set, and we can then write

$$\min_{x \in \Omega} f(x).$$

Notation: Let $x \in \Omega$, i.e. admissible, then we call the inequality constraint $i \in I$ active if $c_i(x) = 0$ and inactive, if $c_i(x) > 0$.

$\mathcal{A}(x) = \mathcal{E} \cup \{i \in I \mid c_i(x) = 0\}$ is called active set (in $x \in \Omega$).

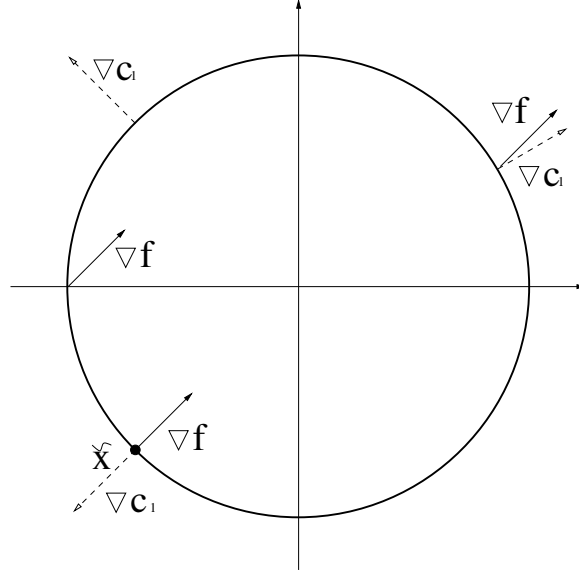
The following examples demonstrates challenges that arise for constrained optimization problems.

Example 5.1.1 (*One equality constraint*)

$$\min x_1 + x_2 \quad \text{s.t.} \quad x_1^2 + x_2^2 - 2 = 0$$

i.e.

$$f(x) = x_1 + x_2, \quad I = \emptyset \quad \mathcal{E} = \{1\} \quad \text{and} \quad c_1(x) = x_1^2 + x_2^2 - 2, \quad \nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \nabla c_1 = 2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$



The solution is $\tilde{x} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$.

In \tilde{x} it holds

$$\nabla f(\tilde{x}) = \lambda_1 \nabla c_1(\tilde{x}) \quad \text{with } \lambda_1 = -\frac{1}{2}. \quad (5.1)$$

To obtain (5.1) let us consider a first order Taylor approximation of the objective function and the constraints, respectively.

To ensure admissibility for small perturbations s , we demand:

$$\begin{aligned} c_1(x + s) &= 0, \text{ which means} \\ 0 &= c_1(x + s) \approx c_1(x) + \nabla c_1(x)^\top s = \nabla c_1(x)^\top s \end{aligned}$$

i.e., for admissibility we want

$$\nabla c_1(x)^\top s = 0. \quad (5.2)$$

On the other hand, for obtaining descent, we need

$$0 > f(x + s) - f(x) \approx \nabla f(x)^\top s \quad \text{i.e.}$$

in first order this yields

$$\nabla f(x)^\top s < 0. \quad (5.3)$$

Asssuming that (5.2) and (5.3) has to hold for arbitrary directions, i.e.

$$\nabla c_1^\top(x) d = 0 \quad \text{und } \nabla f(x)^\top s < 0,$$

then, based on the figure, we note that we can find a descent for all admissible x as long as ∇f and ∇c_1 are not parallel.

Let us now introduce the Lagrangian:

$$\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x).$$

It holds

$$\nabla_x \mathcal{L} = \nabla f(x) - \lambda_1 \nabla c_1(x)$$

and, hence, (5.1) is equivalent to the existence of a $\bar{\lambda}$ such that in $x = \tilde{x}$ we have

$$\nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) = 0.$$

Example 5.1.2 (One inequality constraint)

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0$$

i.e., in this case $\mathcal{E} = \emptyset$, $I = \{1\}$ and $c_1 = 2 - x_1^2 - x_2^2$, $\nabla c_1 = -2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Again, the solution is $\tilde{x} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$, and it holds

$$\nabla f(\tilde{x}) = \tilde{\lambda}_1 \nabla c_1(\tilde{x}) \quad \text{with} \quad \tilde{\lambda}_1 = \frac{1}{2}.$$

The condition for descent reads

$$\nabla f(x)^\top s < 0. \tag{5.4}$$

For admissibility, it has to hold

$$0 \leq c_1(x + s) \approx c_1(x) + \nabla c_1(x)^\top s,$$

i.e., in first order

$$c_1(x) + \nabla c_1(x)^\top s \geq 0. \tag{5.5}$$

Let us analyze when both (5.4) and (5.5) are true. For this purpose, we distinguish the following cases.

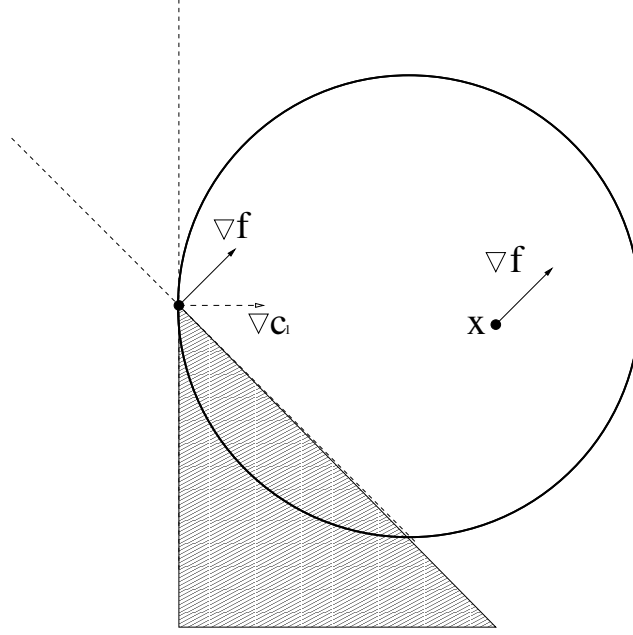
Case 1: x is inactive, i.e., $c_1(x) > 0$. Then (5.5) $\forall s \in \mathbb{R}^n$ with $\|s\|$ sufficiently small. For (5.4) we may choose $s = -\alpha \nabla f(x)$ with $\alpha > 0$ sufficiently small.

Case 2: x active, i.e., $c_1(x) = 0$
and it has to hold $\nabla f(x)^\top s < 0$ and $\nabla c_1(x)^\top s \geq 0$.

Admissibility directions that satisfy both conditions lie within the colored cone which is empty only if

$$\nabla f(x) = \lambda_1 \nabla c_1(x) \tag{5.6}$$

with positive λ_1 , i.e., $\lambda_1 \geq 0$.



For the Lagrangian, we again obtain

$$\nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) = 0 \quad \text{and} \quad \tilde{\lambda} \geq 0.$$

Moreover

$$\tilde{\lambda}_1 c_1(\tilde{x}) = 0 \quad (\text{complementarity condition}). \quad (5.7)$$

In case 1 we had $c_1(\tilde{x}) > 0$, i.e., $\tilde{\lambda}_1 = 0$ and $\nabla f(\tilde{x}) = 0$.

In case 2 it may hold $\lambda_2 \geq 0$ such that we obtain (5.6).

Example 5.1.3 (Two inequality constraints)

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0, \quad x_2 \geq 0$$

i.e., $I = \{1, 2\}$

$$\nabla c_1 = -2x \quad \nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \nabla c_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We obtain the solution $\tilde{x} = (-\sqrt{2}, 0)^\top$. In first order, the condition for descent again reads

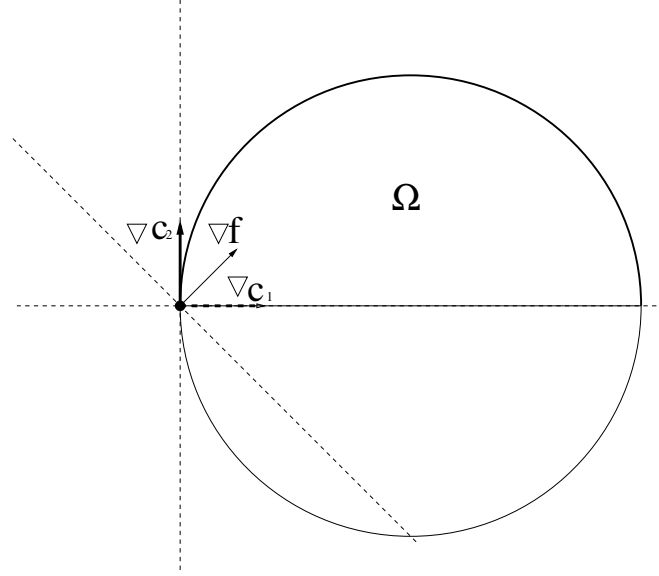
$$\nabla c_i(x)^\top d \geq 0, \quad i \in I \cap \mathcal{A}(\tilde{x}) \quad \text{and} \quad \nabla f(x)^\top d < 0. \quad (5.8)$$

As we see in the figure, in \tilde{x} we cannot find a direction d s.t. (5.8) holds true.

Lagrangian: $\mathcal{L}(x, \lambda_1, \lambda_2) = f(x) - \lambda_1 c_1(x) - \lambda_2 c_2(x)$.

Similar to (5.7) we require the necessary optimality condition

$$\nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) = 0 \quad \text{for} \quad \tilde{\lambda} \geq 0. \quad (5.9)$$



Here, $\tilde{\lambda} \geq 0$ is to be understood componentwise, i.e., $\tilde{\lambda} \geq 0 \Leftrightarrow \tilde{\lambda}_i \geq 0, i = 1, 2$. The complementarity condition reads

$$\tilde{\lambda}_1 c_1(\tilde{x}) = 0 \quad \text{and} \quad \tilde{\lambda}_2 c_2(\tilde{x}) = 0. \quad (5.10)$$

In $\tilde{x} = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$ it holds $\nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\nabla c_1(\tilde{x}) = \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix}$, $\nabla c_2(\tilde{x}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, i.e., with $\tilde{\lambda} = \begin{pmatrix} \frac{1}{2\sqrt{2}} \\ 1 \end{pmatrix}$ we obtain (5.9).

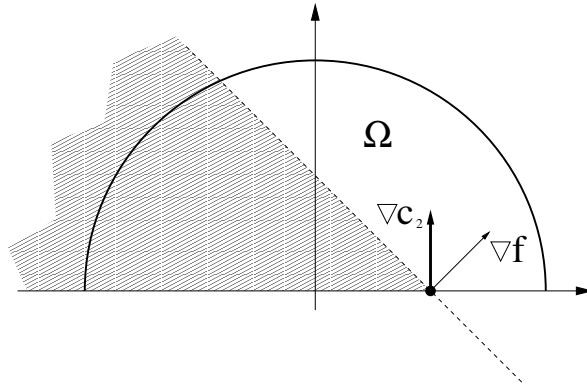
Now let us consider $x = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}$ where both restrictions are active and we have

$$\nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \nabla c_1(x) = \begin{pmatrix} -2\sqrt{2} \\ 0 \end{pmatrix} \quad \nabla c_2(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

i.e., $\nabla f = -\frac{1}{2\sqrt{2}}\nabla c_1(x) + 1 \cdot \nabla c_2(x)$, s.t. $\lambda_1 < 0$, i.e., the second condition in (5.9) is not satisfied.

In $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ only the constraint c_2 is active. Condition for descent is

$$\nabla c_2^\top d \geq 0 \quad \text{and} \quad \nabla f(x)^\top d < 0.$$



This is satisfied for all directions in the colored cone, for example for $d = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$.

For (5.9) it had to hold

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \stackrel{!}{=} \lambda_1 \begin{pmatrix} -2 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (5.11)$$

We know that c_1 is inactive, i.e., $c_1(x) > 0$, and (5.10) implies $\lambda_1 = 0$. Hence, there does not exist λ_2 , s.t. (5.11) is satisfied.

5.2 Tangent cone and constraint qualifications

Notation: Let $x \in \Omega$, i.e., admissible, then the sequence $\{z^{(k)}\}$ is called admissible approximation of x , if $\lim_{k \rightarrow \infty} z^{(k)} = x$ and $z^{(k)} \in \Omega$ for k sufficiently large.

Definition 5.2.1 $d \in \mathbb{R}^n$ is called tangent in $x \in \Omega$, if there exist an admissible approximation $\{z^{(k)}\}$ of x and a sequence $\{t_k\} \subset \mathbb{R}^+$ with $t_k \rightarrow 0$ s.t.

$$\lim_{k \rightarrow \infty} \frac{z_k - x}{t_k} = d.$$

The set $T_\Omega(x) = \{d \in \mathbb{R}^n \mid d \text{ is tangent in } x \in \Omega\}$ is called tangent cone.

Remarks:

1. $K \subset \mathbb{R}^n$ is called cone, if for $x \in K$ it also holds that $\alpha x \in K$ for all $\alpha > 0$.
2. Let $d \in T_\Omega(x)$. For $\alpha > 0$ define $\tilde{t}_k = \frac{t_k}{\alpha}$ s.t. $\tilde{t}_k \rightarrow 0$ and

$$\frac{z^{(k)} - x}{\tilde{t}_k} = \frac{\alpha z^{(k)} - \alpha x}{t_k} \longrightarrow \alpha d, \text{ i.e. } \alpha d \in T_\Omega(x),$$

and $T_\Omega(x)$ is a cone.

Based on the tangent cone, we obtain the following necessary optimality condition:

Theorem 5.2.1 Let $\tilde{x} \in \Omega$ be solution to (P) , then

$$\nabla f(\tilde{x})^\top d \geq 0 \quad \forall d \in T_\Omega(\tilde{x}).$$

PROOF. Let $d \in T_\Omega(\tilde{x})$. Then, there exists $z^{(k)}$ with $z^{(k)} \in \Omega$ for k sufficiently large and $z^{(k)} \rightarrow \tilde{x}$ for $k \rightarrow \infty$. Moreover,

$$d = \lim_{k \rightarrow \infty} \frac{z^{(k)} - \tilde{x}}{t_k}, \text{ i.e. } z^{(k)} = \tilde{x} + t_k d + o(\|t_k d\|).$$

Since $t_k > 0$ for $k \in \mathbb{N}$ this yields

$$0 \leq \frac{1}{t_k} (f(z^{(k)}) - f(\tilde{x})) = \frac{1}{t_k} \nabla f(\tilde{x})^\top t_k d + \frac{1}{t_k} o(t_k)$$

which shows the assertion when $k \rightarrow \infty$. □

For the derivation of descent directions, it is essential to study linearizations of the constraints. For this, let us define:

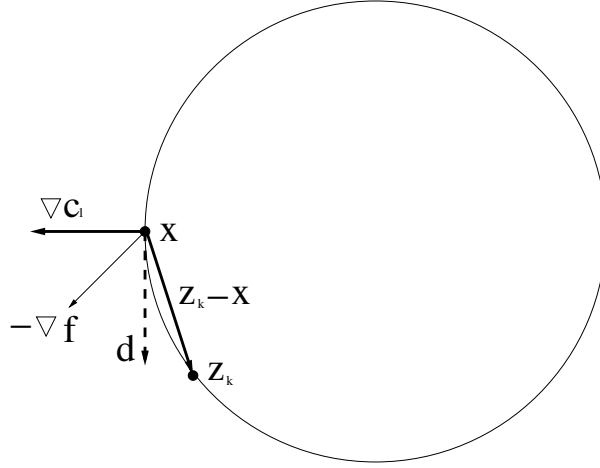
Definition 5.2.2 Let $x \in \Omega$ and $\mathcal{A}(x)$ denote the active set. Then

$$L_{\Omega}(x) = \left\{ d \in \mathbb{R}^n \mid d^{\top} \nabla c_i(x) = 0, \forall i \in \mathcal{E} \quad \text{and} \quad d^{\top} \nabla c_i(x) \geq 0 \forall i \in I \cap \mathcal{A}(x) \right\}$$

is called linearizing cone of Ω in x .

Remark: $T_{\Omega}(x)$ is independent of the specific algebraic specification of Ω , contrary to $L_{\Omega}(x)$.

Example 5.2.1 Once more, let us consider $\min x_1 + x_2$ s.t. $x_1^2 + x_2^2 - 2 = 0$.



Consider $x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$

Choose $z^{(k)} = \begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} \\ -\frac{1}{k} \end{pmatrix}$ and $t_k = \|z^{(k)} - x\|$

Then: $\|z^{(k)}\|^2 = 2 - \frac{1}{k^2} + \frac{1}{k^2} = 2$ i.e., $z^{(k)} \in \Omega \quad \forall k \in \mathbb{N}$. Since $z^{(k)} \rightarrow x$, $z^{(k)}$ is an admissible approximation of x , and the figure shows

$$\frac{z^{(k)} - x}{\|z^{(k)} - x\|} \rightarrow d \quad \text{with} \quad d = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Moreover, it holds $f(z^{(k+1)}) > f(z^{(k)}) \quad \forall k \geq 2$, i.e., x cannot be a solution.

Alternatively: approximate x “from above” by $\tilde{z}^{(k)} = \begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} \\ 1/k \end{pmatrix}$, again $\|\tilde{z}^{(k)}\|^2 = 2 \quad \forall k \in \mathbb{N}$.

One can show that f decreases along $\{z^{(k)}\}$ and that admissible directions are of the form $d = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}$, $\alpha \geq 0$. Altogether, we obtain $T_{\Omega}(x) = \left\{ \begin{pmatrix} 0 \\ d_2 \end{pmatrix} \mid d_2 \in \mathbb{R} \right\}$. For the linearizing cone, we obtain

$$d \in L_{\Omega}(x) \quad \text{if} \quad 0 = \nabla c_1(x)^{\top} d = 2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{\top} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = -2\sqrt{2}d_1 \quad \text{i.e.,} \quad L_{\Omega}(x) = T_{\Omega}(x).$$

On the other hand, we may also characterize Ω as follows

$$\Omega = \{x \mid c_1(x) = 0\} \quad \text{with} \quad c_1(x) = (x_1^2 + x_2^2 - 2)^2 = 0.$$

Then $d \in L_\Omega(x)$ if and only if

$$0 = \nabla c_1(x)^\top d = \begin{pmatrix} 4(x_1^2 + x_2^2 - 2)x_1 \\ 4(x_1^2 + x_2^2 - 2)x_2 \end{pmatrix}^\top \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}^\top \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}.$$

This holds for arbitrary $d \in \mathbb{R}^2$ such that $L_\Omega(x) = \mathbb{R}^2$.

Example 5.2.2 Reconsider Example 5.1.2, i.e.,

$$\min x_1 + x_2 \quad \text{s.t.} \quad 2 - x_1^2 - x_2^2 \geq 0.$$

For $x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$, every admissible approximation from Example 5.2.1 is admissible and, moreover, there exist infinitely many more, e.g.,

$$z_k = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix} + \frac{1}{k} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad \text{with} \quad w_1 > 0.$$

For $k \geq (w_1^2 + w_2^2)/(2\sqrt{2}w_1)$, it follows that $z_k \in \Omega$, i.e., for $t_k = \frac{1}{k}$ the direction $d = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ with $w_1 > 0$ is contained in the tangent cone. Altogether we obtain

$$T_\Omega(x) = \left\{ \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \in \mathbb{R}^2 \mid d_1 \geq 0 \right\}.$$

For the directions in the linearizing cone we have

$$d \in L_\Omega(x) \Leftrightarrow 0 \leq \nabla c_1(x)^\top d = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}^\top \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = 2\sqrt{2}d_1$$

i.e., $d_2 \in \mathbb{R}$ is arbitrary and $d_1 \geq 0$, and we have $T_\Omega(x) = L_\Omega(x)$.

Example 5.2.3 Consider an admissible set characterized by

$$\begin{aligned} c_1(x) &= 1 - x_1^2 - (x_2 - 1)^2 \geq 0 \\ c_2(x) &= -x_2 \geq 0 \end{aligned}$$

i.e., $\Omega = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$.

Hence, every approximating sequence of $x = 0$ has to satisfy $z_k = 0$ for k sufficiently large so we conclude $T_\Omega(x) = \{0\}$. For directions in the linearizing cone it holds $d \in L_\Omega(x)$, if

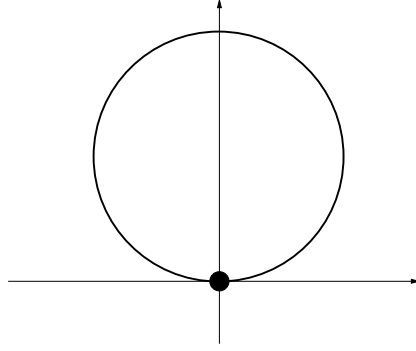
$$\begin{aligned} \nabla c_1(x)^\top d &\geq 0 \quad \text{and} \quad \nabla c_2(x)^\top d \geq 0 \quad \text{i.e.} \\ -2x^\top d &= -2 \cdot 0 \cdot d \geq 0 \quad \text{and} \quad \begin{pmatrix} 0 \\ -1 \end{pmatrix}^\top d = -d_2 \geq 0. \end{aligned}$$

Consequently, $L_\Omega(0) = \{d \in \mathbb{R}^n \mid d_2 \leq 0\} \neq T_\Omega(x)$. □

For the derivation of necessary optimality conditions, the directions from the tangent cone $T_\Omega(x)$ have to be put in relation to those from the linearizing cone. In more detail, we aim at $T_\Omega(x) = L_\Omega(x)$ (cf. Lemma 5.4.1). For this purpose, we need a certain kind of regularity condition to be satisfied. For numerical methods, the following one is the most common:

Definition 5.2.3 (LICQ)

Let $x \in \Omega$. The regularity condition (LICQ) (= linear independence constraint qualification) is satisfied, if $\{\nabla c_i(x), i \in \mathcal{A}(x)\}$ are linearly independent.



5.3 First order necessary optimality conditions

Similar to the introductory examples, we begin by defining the Lagrangian:

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup I} \lambda_i c_i(x).$$

We may now formulate first order necessary conditions as follows:

Theorem 5.3.1 *Let \tilde{x} be a solution to (P) and assume that f and c_i are continuously differentiable. If (LICQ) holds in \tilde{x} , then there exists a vector of Lagrange multipliers $\tilde{\lambda}$ with components $\tilde{\lambda}_i$, $i \in \cup I$ s.t.:*

$$\nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) = 0 \tag{1}$$

$$c_i(\tilde{x}) = 0 \quad \text{for all } i \in \mathcal{E} \tag{2}$$

$$c_i(\tilde{x}) \geq 0 \quad \text{for all } i \in I \tag{3}$$

$$\tilde{\lambda}_i \geq 0 \quad \text{for all } i \in I \tag{4}$$

$$\tilde{\lambda}_i c_i(\tilde{x}) = 0 \quad \text{for all } i \in \mathcal{E} \cup I. \tag{5}$$

Remarks:

1. The conditions (1)–(5) are called Karush-Kuhn-Tucker (KKT) conditions.
2. (5) is a complementarity condition. With (5) it follows that $\tilde{\lambda}_i = 0$ if c_i for $i \in I$ is inactive, i.e. (1) in Theorem 5.3.1 can be expressed as

$$0 = \nabla f(\tilde{x}) - \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i \nabla c_i(\tilde{x}).$$

For numerical methods, one often requires a stronger complementarity condition:

Definition 5.3.1 *Let \tilde{x} be solution to (P) and let $\tilde{\lambda}$ satisfy (1)–(5). Then $\tilde{\lambda}$ satisfies a strict complementarity condition if for all $i \in I$ it additionally holds*

$$c_i(x) = 0 \Rightarrow \lambda_i > 0.$$

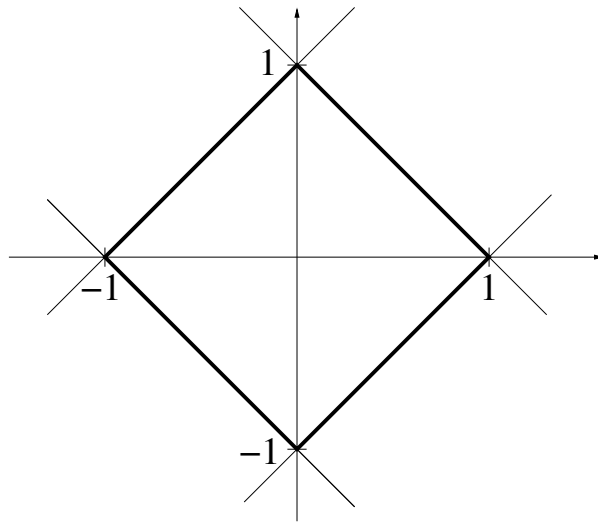
In other words, $\tilde{\lambda}_i > 0$ for all $i \in I \cap \mathcal{A}(\tilde{x})$.

Example 5.3.1

$$\min \left(x_1 - \frac{3}{2} \right)^2 + \left(x_2 - \frac{1}{2} \right)^4$$

s.t.

$$\begin{aligned} 1 - x_1 - x_2 &\geq 0 \\ 1 - x_1 + x_2 &\geq 0 \\ 1 + x_1 - x_2 &\geq 0 \\ 1 + x_1 + x_2 &\geq 0. \end{aligned}$$



From the figure we see that $\tilde{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. In \tilde{x} the constraints c_1 and c_2 are active. It holds

$$\nabla f = \begin{pmatrix} 2\left(x_1 - \frac{3}{2}\right) \\ 4\left(x_2 - \frac{1}{2}\right)^3 \end{pmatrix}, \quad \nabla f(\tilde{x}) = \begin{pmatrix} -1 \\ -\frac{1}{2} \end{pmatrix}, \quad \nabla_{c_1}(\tilde{x}) = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \nabla_{c_2}(\tilde{x}) = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

s.t. we obtain

$$\nabla f(\tilde{x}) = \frac{3}{4} \begin{pmatrix} -1 \\ -1 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

For $\tilde{\lambda} = \left(\frac{3}{4}, \frac{1}{4}, 0, 0 \right)$, all KKT conditions are satisfied.

Example 5.3.2

$$\begin{aligned} \min f(x) = |x|^2 \quad & \text{s.t.} \quad x_1 + x_2 + x_3 = 3 \\ & 2x_1 - x_2 + x_3 \leq 5 \end{aligned}$$

i.e.,

$$\begin{aligned}
c_1 &= x_1 + x_2 + x_3 - 3 = 0 \\
c_2 &= 5 - 2x_1 + x_2 - x_3 \geq 0 \\
\mathcal{L}(x, \lambda) &= x_1^2 + x_2^2 + x_3^2 - \lambda_1(x_1 + x_2 + x_3 - 3) \\
&\quad - \lambda_2(5 - 2x_1 + x_2 - x_3) \\
\nabla_x \mathcal{L}(x, \lambda) &\stackrel{!}{=} \Leftrightarrow \left. \begin{aligned} 2x_1 - \lambda_1 + 2\lambda_2 &= 0 \\ 2x_2 - \lambda_1 - \lambda_2 &= 0 \\ 2x_3 - \lambda_1 + \lambda_2 &= 0 \end{aligned} \right\} \quad (5.12)
\end{aligned}$$

Moreover, it has to hold $\lambda_2(5 - 2x_1 + x_2 - x_3) = 0$. Assume $\lambda_2 \neq 0 \Rightarrow$

$$\begin{aligned}
x_1 &= \frac{1}{2} \lambda_1 - \lambda_2 \\
x_2 &= \frac{1}{2} \lambda_1 + \frac{1}{2} \lambda_2 \\
x_3 &= \frac{1}{2} \lambda_1 - \frac{1}{2} \lambda_2
\end{aligned}$$

Inserting this in the active constraints, we obtain:

$$\begin{aligned}
5 &= \lambda_1 - 2\lambda_2 - \frac{1}{2} \lambda_2 - \frac{1}{2} \lambda_2 \\
&= \lambda_1 - 3\lambda_2 \Rightarrow \lambda_1 = 5 + 3\lambda_2 \\
3 &= \frac{1}{2} \lambda_1 - \lambda_2 + \frac{1}{2} \lambda_1 + \frac{1}{2} \lambda_2 + \frac{1}{2} \lambda_1 - \frac{1}{2} \lambda_2 \\
&= \frac{3}{2} \lambda_1 - \lambda_2 \\
\Rightarrow 3 &= \frac{15}{2} + \frac{9}{2} \lambda_2 - \lambda_2 = \frac{15}{2} + \frac{7}{2} \lambda_2 \\
\Rightarrow \lambda_2 &= -\frac{9 \cdot 2}{2 \cdot 7} = -\frac{9}{7} < 0.
\end{aligned}$$

This is a contradiction since it must hold $\lambda_2 \geq 0$. Hence, $\lambda_2 = 0$ and with (5.12) we obtain $x_1 = x_2 = x_3 = \frac{\lambda_1}{2}$ and, since $c_1 = 0$, it follows that $x_1 = x_2 = x_3 = 1$

$$\lambda_1 = 2 \quad \text{i.e.} \quad \tilde{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{satisfies the KKT conditions.}$$

5.4 Proof of Theorem 5.3.1

The proof is carried out in three steps. First, we establish a connection between tangent and linearizing cone. For this purpose, let us introduce the following notation:

$$A(\tilde{x})^\top = [\nabla c_i(\tilde{x})]_{i \in \mathcal{A}(\tilde{x})}$$

i.e., the matrix $A(\tilde{x})$ contains the gradients of the active constraints as its rows.

Lemma 5.4.1 *Let \tilde{x} be admissible, then:*

- (1) $T_\Omega(\tilde{x}) \subset L_\Omega(\tilde{x})$
- (2) *If (LICQ) holds true in \tilde{x} , then we even have*

$$T_\Omega(\tilde{x}) = L_\Omega(\tilde{x}).$$

PROOF. Without loss of generality, we may assume that $c_i(x)$, $i = 1, \dots, m$ denote the active constraints.

Proof of (1):

Let $d \in T_\Omega(\tilde{x})$ then there exist $\{z^{(k)}\}$ and $\{t_k\}$ s.t.

$$\lim_{k \rightarrow \infty} \frac{z^{(k)} - \tilde{x}}{t_k} = d,$$

with $t_k > 0$ for all k . Let $i \in \mathcal{E}$ then by Taylor expansion (with remainder term) there exists $\theta \in (0, 1)$ s.t.:

$$0 = \frac{1}{t_k} c_i(z^{(k)}) = \frac{1}{t_k} c_i(\tilde{x} + (z^{(k)} - \tilde{x})) = \nabla c_i(\tilde{x} + \theta(z^{(k)} - \tilde{x})) \frac{z^{(k)} - \tilde{x}}{t_k}$$

i.e., for $k \rightarrow \infty$ we obtain $\nabla c_i(\tilde{x})^\top d = 0$.

For $i \in \mathcal{A}(\tilde{x}) \cap I$ we similarly arrive at

$$0 \leq \nabla c_i(\tilde{x})^\top d.$$

Altogether, we conclude that $d \in L_\Omega(\tilde{x})$.

Proof of (2): (LICQ) is satisfied, hence, $A \in \mathbb{R}^{m \times n}$ has full row rank m .

For the kernel of A , it holds $\dim(\ker(A)) = n - \text{rk}(A)$.

Let $Z \in \mathbb{R}^{n \times n-m}$ be a basis matrix for the null space, i.e., the columns of Z span the kernel of A :

$$A(\tilde{x})Z = 0$$

with $\text{rk}(Z) = n - m$.

For an arbitrary $d \in L_\Omega(\tilde{x})$ and $\{t_k\} \subset \mathbb{R}^+ \setminus \{0\}$ with $t_k \rightarrow 0$, let us define the system of equations

$$R : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n \quad \text{via}$$

$$R(z, t) = \begin{pmatrix} c(z) - tA(\tilde{x})d \\ Z^T(z - \tilde{x} - td) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We will utilize the implicit function theorem to show that there exists a sequence $\{z^{(k)}\}$ that is an admissible approximation of \tilde{x} with $R(z^{(k)}, t_k) = 0$ and $\frac{z^{(k)} - \tilde{x}}{t_k} \rightarrow d$, i.e., $d \in T_\Omega(\tilde{x})$.

Note that

$$R(\tilde{x}, 0) = \begin{pmatrix} c(\tilde{x}) - 0 \cdot A(\tilde{x})d \\ Z^\top(\tilde{x} - \tilde{x} - 0 \cdot d) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Moreover, R is continuously differentiable in a neighborhood of $(\tilde{x}, 0)$ with

$$[D_z R]_{i\text{-th column}} = \begin{cases} \nabla c_i^\top, & i \in \{1, \dots, m\} \\ [Z^\top]_{i\text{-th column}}, & i \in \{m+1, \dots, n\} \end{cases}$$

$$\text{i.e., } D_z R(\tilde{x}, 0) = \begin{pmatrix} A \\ Z^\top \end{pmatrix}.$$

We have $\text{rk} \left(\begin{pmatrix} A \\ Z^\top \end{pmatrix} \right) = n$ such that the implicit function theorem guarantees the existence of $z(t)$ with $z(0) = \tilde{x}$ and $R(z(t), t) = 0$ for t sufficiently small, i.e., $z^{(k)} = z(t_k) \rightarrow \tilde{x}$ and $R(z_k, t_k) = 0$ for k sufficiently large.

Hence, it holds that $c(z_k) = t_k A(\tilde{x})d$ and, since $d \in L_\Omega(\tilde{x})$, for all $i \in \mathcal{E}$:

$$c_i(z^{(k)}) = t_k \nabla c_i(\tilde{x})^\top d = 0$$

and for $i \in \mathcal{A}(\tilde{x}) \cap I$

$$c_i(z^{(k)}) = t_k \nabla c_i(\tilde{x})^\top d \geq 0$$

implying that $z^{(k)}$ is admissible for k sufficiently large. Taylor expansion (with remainder term) yields $\theta \in (0, 1)$ s.t. for sufficiently large k :

$$\begin{aligned} 0 &= \frac{1}{t_k} R(z^{(k)}, t_k) = \frac{1}{t_k} \begin{pmatrix} c(\tilde{x} + (z^{(k)} - \tilde{x})) - t_k A(\tilde{x})d \\ Z^\top(z^{(k)} - \tilde{x} - t_k d) \end{pmatrix} \\ &= \begin{pmatrix} A(\tilde{x} + \theta(z^{(k)} - \tilde{x})) \\ Z^\top \end{pmatrix} \frac{z^{(k)} - \tilde{x}}{t_k} - \begin{pmatrix} A(\tilde{x})d \\ Z^\top \end{pmatrix} d. \end{aligned}$$

Since $\begin{pmatrix} A(\tilde{x}) \\ Z^\top \end{pmatrix}$ is regular, this implies

$$\frac{z^{(k)} - \tilde{x}}{t_k} \rightarrow d$$

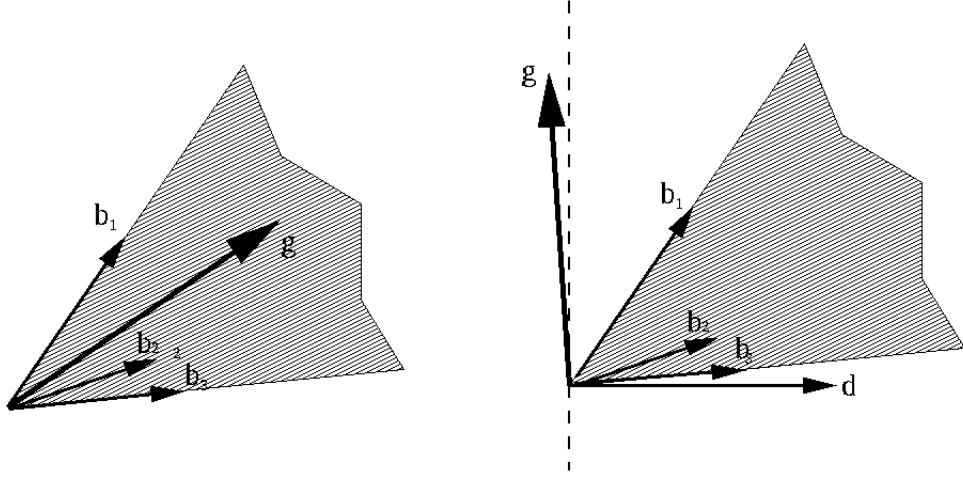
i.e., $d \in T_\Omega(\tilde{x})$. □

Another essential step/tool is the Farkas' Lemma:

Lemma 5.4.2 *Let $K = \{By + Cw \mid y \in \mathbb{R}^m, y \geq 0, w \in \mathbb{R}^p\}$ mit $B \in \mathbb{R}^{(n,m)}$, $C \in \mathbb{R}^{n \times p}$ be fixed. Then for every $g \in \mathbb{R}^n$ either $g \in K$, or there exists $d \in \mathbb{R}^n$ s.t.*

$$g^\top d < 0, \quad B^\top d \geq 0, \quad C^\top d = 0.$$

The proof of this lemma can be found in, e.g., [2]. For $n = 2$, $m = 3$, $C = 0$, the statement is visualized in the following figure.



Now we apply the Farkas' Lemma to the particular statement from 5.3.1.

Let $N = \{ \sum_{i \in \mathcal{A}(\tilde{x})} \lambda_i \nabla c_i, \text{ and } \lambda_i \geq 0 \text{ for } i \in \mathcal{A}(\tilde{x}) \cap I \}$ and $g = \nabla f(\tilde{x})$, then either

$$\begin{aligned} \nabla f(\tilde{x}) &= \sum_{i \in \mathcal{A}(\tilde{x})} \lambda_i \nabla c_i(\tilde{x}) = A(\tilde{x})^\top \lambda, \\ \text{with } \lambda_i &\geq 0 \text{ for } i \in \mathcal{A}(\tilde{x}) \cap I \end{aligned} \quad (\text{I})$$

or there exists $d \in \mathbb{R}^n$ with

$$\begin{aligned} \nabla f(\tilde{x})^\top d &< 0, \\ \nabla c_i^\top d &= 0, \quad i \in \mathcal{E} \quad \text{and} \quad \nabla c_i^\top d \geq 0, \quad i \in \mathcal{A}(\tilde{x}) \cap I \end{aligned}$$

i.e.,

$$\nabla f(\tilde{x})^\top d < 0 \quad \text{and} \quad d \in L_\Omega(\tilde{x}). \quad (\text{II})$$

By assumption, $\tilde{x} \in \Omega$ and (LICQ) holds true. Then with Theorem 5.2.1 we obtain $\nabla f(\tilde{x})^\top d \geq 0 \quad \forall d \in T_\Omega(\tilde{x}) = L_\Omega(\tilde{x})$ s.t. (II) is not true and it has to hold (I) instead.

Define

$$\tilde{\lambda}_i = \begin{cases} \lambda_i, & i \in \mathcal{A}(\tilde{x}) \\ 0, & \text{else,} \end{cases}$$

then $\tilde{\lambda} \geq 0$ and the complementarity condition is satisfied.

5.5 Second order optimality conditions

In the follows, we assume that f and $c_i, i \in I \cup \mathcal{E}$ are two times continuously differentiable.

Let \tilde{x} be a critical point and $\tilde{\lambda}$ s.t. the KKT conditions are satisfied. Then we define the *critical cone*

$$C(\tilde{x}, \tilde{\lambda}) = \{w \in L_\Omega(\tilde{x}) \mid \nabla c_i(\tilde{x})^\top w = 0 \quad \forall i \in \mathcal{A}(\tilde{x}) \cap I \quad \text{with} \quad \tilde{\lambda}_i > 0\}$$

i.e., it holds

$$\omega \in C(\tilde{x}, \tilde{\lambda}) \Leftrightarrow \begin{cases} \nabla c_i(\tilde{x})^\top \omega = 0 & \forall i \in \mathcal{E} \\ \nabla c_i(\tilde{x})^\top \omega = 0 & \forall i \in \mathcal{A}(\tilde{x}) \cap I \text{ with } \tilde{\lambda}_i > 0 \\ \nabla c_i(\tilde{x})^\top \omega \geq 0 & \forall i \in \mathcal{A}(\tilde{x}) \cap I \text{ with } \tilde{\lambda}_i = 0 \end{cases}$$

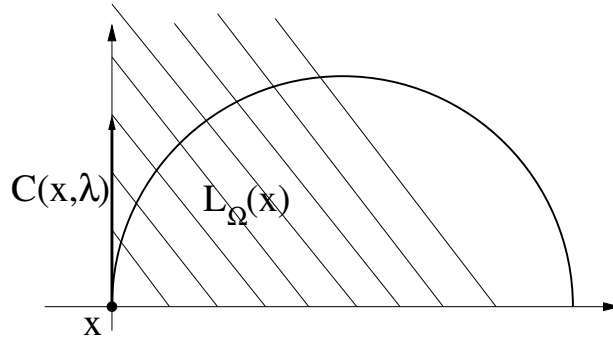
$\tilde{\lambda}_i = 0$ für alle inaktiven Nebenbedingungen, also folgt für $\omega \in C(\tilde{x}, \tilde{\lambda})$

$$\lambda_i \nabla c_i(\tilde{x})^\top \omega = 0 \quad \forall i \in \mathcal{E} \cup I,$$

and with the first KKT condition, we obtain

$$\omega^\top \nabla f(\tilde{x}) = \sum_{i \in \mathcal{E} \cup I} \tilde{\lambda}_i \omega^\top \nabla c_i(\tilde{x}) = 0.$$

In other words, the critical cone $C(\tilde{x}, \tilde{\lambda})$ contains directions of the linearizing cone for which first order derivatives do not allow to conclude about decrease/increase of f .



Example 5.5.1 Consider

$$\begin{aligned} \min x_1 \quad s.t. \quad & x_2 \geq 0 \\ & 1 - (x_1 - 1)^2 - x_2^2 \geq 0. \end{aligned}$$

A solution is $\tilde{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$,

$$\mathcal{A}(\tilde{x}) = \{1, 2\} \quad \nabla c_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \nabla c_2(x) = \begin{pmatrix} -2(x_1 - 1) \\ -2x_2 \end{pmatrix}$$

$$\nabla c_2(0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

$$\nabla f = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 0 \cdot \nabla c_1 + \frac{1}{2} \nabla c_2$$

i.e., $\tilde{\lambda} = \left(0, \frac{1}{2}\right)^\top$.

(LICQ) is satisfied, i.e., $\tilde{\lambda}$ is uniquely determined.

$$\begin{aligned} L_\Omega(\tilde{x}) &= \{d \in \mathbb{R}^2 \mid \nabla c_1^\top d \geq 0 \text{ and } \nabla c_2^\top d \geq 0\} \\ &= \{d \in \mathbb{R}^2 \mid d_1 \geq 0 \text{ and } d_2 \geq 0\} \\ C(\tilde{x}, \tilde{\lambda}) &= \{w \mid \nabla c_1^\top w \geq 0 \text{ and } \nabla c_2^\top w = 0\} \\ &= \{w \mid w_1 = 0 \text{ and } w_2 \geq 0\}. \end{aligned}$$

The following result shows that the Hessian of the Lagrangian has a non negative curvature along critical directions.

Theorem 5.5.1 (*Second order necessary condition*)

Let \tilde{x} be a local solution to (P). Let (LICQ) hold true and let $\tilde{\lambda}$ be the Lagrange multiplier such that the KKT conditions are satisfied. Then

$$\omega^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) \omega \geq 0 \quad \forall \omega \in C(\tilde{x}, \tilde{\lambda}).$$

PROOF. $\omega \in C(\tilde{x}, \tilde{\lambda}) \subset L_\Omega(\tilde{x})$, i.e., as in the proof of Lemma 5.4.1 we can construct an admissible approximation $\{z^{(k)}\}$ of \tilde{x} and a sequence $\{t_k\}$ s.t.

$$\lim_{k \rightarrow \infty} \frac{z^{(k)} - \tilde{x}}{t_k} = \omega$$

or, alternatively,

$$z^{(k)} - \tilde{x} = t_k \omega + o(t_k).$$

Moreover, it has been shown that

$$c_i(z^{(k)}) = t_k \nabla c_i(\tilde{x})^\top \omega \quad \forall i \in \mathcal{A}(\tilde{x}).$$

We then obtain

$$\mathcal{L}(z^{(k)}, \tilde{\lambda}) = f(z^{(k)}) - \sum_{i \in \mathcal{E} \cup I} \tilde{\lambda}_i c_i(z^{(k)}) = f(z^{(k)}) - t_k \sum_{i \in \mathcal{E} \cup I} \lambda_i \nabla c_i(\tilde{x})^\top \omega = f(z^{(k)}),$$

Second order Taylor expansion of $\mathcal{L}(z_k, \tilde{\lambda})$ around \tilde{x} with remainder term yields $\theta \in (0, 1)$ s.t.:

$$\begin{aligned} f(z^{(k)}) &= \mathcal{L}(z^{(k)}, \tilde{\lambda}) = \mathcal{L}(\tilde{x}, \tilde{\lambda}) + (z^{(k)} - \tilde{x})^\top \nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) \\ &\quad + \frac{1}{2} (z_k - \tilde{x})^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x} + \theta(\tilde{x} - z^{(k)}), \tilde{\lambda}) (z^{(k)} - \tilde{x}) \\ &\Rightarrow 0 \leq \frac{1}{t_k^2} (f(z^{(k)}) - f(\tilde{x})) = \frac{1}{2} \left(\frac{z^{(k)} - \tilde{x}}{t_k} \right)^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x} + \theta(\tilde{x} - z^{(k)}), \tilde{\lambda}) \frac{z^{(k)} - \tilde{x}}{t_k}. \end{aligned}$$

Considering the limit $k \rightarrow \infty$ shows the statement. \square

Theorem 5.5.2 Let $\tilde{x} \in \mathbb{R}^n$ be admissible s.t. there exists a Lagrange multiplier $\tilde{\lambda}$ satisfying the KKT conditions. Moreover, assume there exists $\sigma > 0$ s.t.

$$w^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) w \geq \sigma \|w\|^2 \quad \forall w \in C(\tilde{x}, \tilde{\lambda}).$$

Then \tilde{x} is a strict local solution to (P) and there exists $\tilde{\sigma} > 0$ s.t. the following growth condition is satisfied

$$f(z) \geq f(\tilde{x}) + \tilde{\sigma} \|z - \tilde{x}\|^2 \tag{5.13}$$

for all $z \in B_r(\tilde{x}) \cap \Omega$ and a suitably chosen $r > 0$.

PROOF. Let $\{z^{(k)}\}$ be an arbitrary admissible approximation of \tilde{x} . We will show that it holds

$$f(z^{(k)}) \geq f(\tilde{x}) + \frac{\sigma}{4} \|z^{(k)} - \tilde{x}\|^2 \quad (5.14)$$

if k is sufficiently large. If (5.14) was false, then there exists an admissible approximation $\{z^{(k)}\}$ of \tilde{x} with

$$f(z^{(k)}) < f(\tilde{x}) + \frac{\sigma}{4} \|z^{(k)} - \tilde{\lambda}\|^2 \quad \text{for } k \text{ sufficiently large.} \quad (5.15)$$

For $d^{(k)} := \frac{z^{(k)} - \tilde{x}}{\|z^{(k)} - \tilde{x}\|}$ it holds that $\|d^{(k)}\| = 1$, i.e., at least for a subsequence (again denoted by the index k) there exists $d \in \mathbb{R}^n$ with

$$d = \lim_{k \rightarrow \infty} d^{(k)}.$$

This implies $d \in T_{\Omega}(\tilde{x}) \subset L_{\Omega}(\tilde{x})$, see Lemma 5.4.1.

Claim:

$$d \in C(\tilde{x}, \tilde{\lambda}). \quad (5.16)$$

The KKT conditions imply

$$\mathcal{L}(z^{(k)}, \tilde{\lambda}) = f(z^{(k)}) - \sum_{i \in \mathcal{A}(\tilde{x})} \underbrace{\lambda_i c_i(z^{(k)})}_{\geq 0} \leq f(z^{(k)}).$$

If we had $d \notin C(\tilde{x}, \tilde{\lambda})$, there would exist $j \in \mathcal{A}(\tilde{x}) \cap I$ s.t.

$$\tilde{\lambda}_j \nabla c_j(\tilde{x})^\top d > 0, \quad (5.17)$$

and for the remaining indices $i \in \mathcal{A}(\tilde{x})$ we obtained

$$\tilde{\lambda}_i \nabla c_i(\tilde{x})^\top d \geq 0.$$

Since

$$z^{(k)} = \tilde{x} + \|z^{(k)} - \tilde{x}\| d + o(\|z^{(k)} - \tilde{x}\|). \quad (5.18)$$

Taylor expansion leads to

$$\begin{aligned} \tilde{\lambda}_j c_j(z^{(k)}) &= \underbrace{\tilde{\lambda}_j c_j(\tilde{x})}_{=0} + \lambda_j \nabla c_j(\tilde{x})^\top (z^{(k)} - \tilde{x}) + o(\|z^{(k)} - \tilde{x}\|) \\ &\stackrel{(5.18)}{=} \|z^{(k)} - \tilde{x}\| \tilde{\lambda}_j \nabla c_j(\tilde{x})^\top d + o(\|z^{(k)} - \tilde{x}\|) \end{aligned} \quad (5.19)$$

$$\begin{aligned} \Rightarrow \mathcal{L}(z^{(k)}, \tilde{\lambda}) &= f(z^{(k)}) - \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i c_i(z^{(k)}) \leq f(z^{(k)}) - \tilde{\lambda}_j c_j(z^{(k)}) \\ &= f(z^{(k)}) - \|z^{(k)} - \tilde{x}\| \tilde{\lambda}_j \nabla c_j(\tilde{x})^\top d + o(\|z^{(k)} - \tilde{x}\|). \end{aligned}$$

On the other hand, from the necessary conditions we conclude

$$\mathcal{L}(z^{(k)}, \tilde{x}) = f(\tilde{x}) + o(\|z^{(k)} - \tilde{x}\|).$$

Altogether, we arrive at

$$f(z^{(k)}) \geq f(\tilde{x}) + \|z^{(k)} - \tilde{x}\| \underbrace{\tilde{\lambda}_j \nabla c_j(\tilde{x})^\top d}_{>0} + o(\|z^{(k)} - \tilde{x}\|).$$

Due to (5.17) there exists k_0 s.t. for $\tau = \tilde{\lambda}_j \nabla c_j(\tilde{x})^\top d > 0$ it holds

$$\frac{f(z^{(k)}) - f(\tilde{x})}{\|z^{(k)} - \tilde{x}\|} \geq \frac{\tau}{2} > 0 \quad \text{for all } k \geq k_0$$

contradicting (5.15). Hence, we can conclude that (5.16) is true, i.e., $d \in C(\tilde{x}, \tilde{\lambda})$ and $\|d\| = 1$ implying

$$d^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) d \geq \sigma.$$

Moreover, it holds that

$$\mathcal{L}(z^{(k)}, \tilde{\lambda}) = f(z^{(k)}) - \sum_{i \in \mathcal{E} \cup I} \underbrace{\tilde{\lambda}_i c_i(z^{(k)})}_{\geq 0} \leq f(z^{(k)}).$$

Taylor expansion of $\mathcal{L}(z^{(k)}, \tilde{\lambda})$ around \tilde{x} as in the proof of Theorem 5.5.1 and taking into account

$$z^{(k)} - \tilde{x} = \|z^{(k)} - \tilde{x}\| d + o(\|z^{(k)} - \tilde{x}\|)$$

yields

$$\begin{aligned} f(z^{(k)}) &\geq \mathcal{L}(z^{(k)}, \tilde{\lambda}) = f(\tilde{x}) + \underbrace{(z^{(k)} - \tilde{x})^\top \nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda})}_{=0} \\ &\quad + \frac{1}{2} (z^{(k)} - \tilde{x})^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) (z^{(k)} - \tilde{x}) + o(\|z^{(k)} - \tilde{x}\|^2) \\ &= f(\tilde{x}) + \frac{1}{2} \underbrace{d^\top \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) d}_{\geq \sigma} \|z^{(k)} - \tilde{x}\|^2 + o(\|z^{(k)} - \tilde{x}\|^2) \\ &= f(\tilde{x}) + \frac{\sigma}{2} \|z^{(k)} - \tilde{x}\|^2 + o(\|z^{(k)} - \tilde{x}\|^2) \\ &= f(\tilde{x}) + \frac{\sigma}{4} \|z^{(k)} - \tilde{x}\|^2 + \underbrace{\frac{\sigma}{2} \|z^{(k)} - \tilde{x}\|^2 + o(\|z^{(k)} - \tilde{x}\|^2)}_{\geq 0 \text{ for } k \text{ sufficiently large}}, \end{aligned}$$

which, once more, contradicts (5.15). Hence, for any admissible approximation $\{z^{(k)}\}$ of \tilde{x}

$$f(z^{(k)}) \geq f(\tilde{x}) + \frac{\sigma}{4} \|z^{(k)} - \tilde{x}\|^2.$$

□

Remark: If we have *strict complementarity*, for $i \in \mathcal{A}(\tilde{x}) \cap I$, it holds $\lambda_i > 0$ and

$$C(\tilde{x}, \tilde{\lambda}) = \left\{ d \in \mathbb{R}^n \mid \nabla c_i(\tilde{x})^\top d = 0 \quad \forall i \in \mathcal{A}(\tilde{x}) \right\}.$$

Let

$$A(\tilde{x}) = \left(\nabla c_i(\tilde{x})^\top \right)_{i \in \mathcal{A}(\tilde{x})}$$

then $C(\tilde{x}, \tilde{\lambda}) = \ker(A(\tilde{x}))$, i.e., the sufficient optimality conditions read

$$w^\top \nabla_{xx} \mathcal{L}(\tilde{x}, \tilde{\lambda}) w \geq \sigma \|w\|^2 \quad \forall w \in \ker(A(\tilde{x})). \quad (5.20)$$

Let $\dim(\ker(A(\tilde{x}))) = l$ and $Z \in \mathbb{R}^{n \times l}$ be a matrix spanning the null space of $A(\tilde{x})$, i.e., $w = Zz \quad \forall z \in \mathbb{R}^l$, then

$$w^\top \nabla_{xx} \mathcal{L}(\tilde{x}, \tilde{\lambda}) w = z^\top Z^\top \nabla_{xx} \mathcal{L}(\tilde{x}, \tilde{\lambda}) Z z,$$

i.e. (5.20) is equivalent to the positive definiteness of $Z^\top \nabla_{xx} \mathcal{L}(\tilde{x}, \tilde{\lambda}) Z$ on \mathbb{R}^l .

This condition is easy to check numerically which is why strict complementarity is often assumed in practice.

Example 5.5.2 $\min -x^2 + 1 \quad \text{s.t. } x \geq -1, x \leq \frac{1}{2}.$

Global minimum in $x = -1$, local minimum in $x = \frac{1}{2}$. In $\tilde{x} = \frac{1}{2}$ we have

$$\mathcal{L}(\tilde{x}, \lambda_1, \lambda_2) = -x^2 + 1 - \lambda_1(x + 1) - \lambda_2\left(\frac{1}{2} - x\right).$$

In $\tilde{x} = \frac{1}{2}$, the constraint c_1 is inactive, i.e., $\lambda_1 = 0$

$$KKT \Rightarrow \nabla_x \mathcal{L} = -2\tilde{x} + \lambda_2 \stackrel{!}{=} 0 \Rightarrow \lambda_2 = 1$$

$$\text{i.e., } C(\tilde{x}, \tilde{\lambda}) = \left\{ d \in \mathbb{R} \mid c'_2(\tilde{x}) d = 0 \right\} = \{0\}$$

i.e., sufficient condition $d \mathcal{L}_{xx}(\tilde{x}, \tilde{\lambda}) d \geq \sigma d^2 \quad \forall d \in \mathbb{R} \text{ with } d = 0$

i.e., the condition is satisfied for every σ .

Example 5.5.3

$$\begin{aligned} \min & -\frac{1}{2}\sqrt{x_1} - \frac{1}{2}x_2 \quad \text{s.t.} \quad x_1 \geq 0 \\ & x_2 \geq 0 \\ & 1 - x_1 - x_2 \geq 0 \end{aligned}$$

$$\begin{aligned} \mathcal{L}(x, \lambda) &= -\frac{1}{2}\sqrt{x_1} - \frac{1}{2}x_2 - \lambda_1 x_1 - \lambda_2 x_2 - \lambda_3(1 - x_1 - x_2) \\ \nabla_x \mathcal{L}(x, \lambda) &= \begin{pmatrix} -\frac{1}{4\sqrt{x_1}} - \lambda_1 + \lambda_3 \\ -\frac{1}{2} - \lambda_2 + \lambda_3 \end{pmatrix} = 0 \end{aligned}$$

KKT system:

$$\begin{aligned}
-\frac{1}{4} \frac{1}{\sqrt{x_1}} - \lambda_1 + \lambda_3 &= 0 \\
-\frac{1}{2} - \lambda_2 + \lambda_3 &= 0 \\
x_1 \geq 0, x_2 \geq 0, 1 - x_1 - x_2 &\geq 0 \\
\lambda_1 x_1 &= 0 \\
\lambda_2 x_2 &= 0 \\
\lambda_3(1 - x_1 - x_2) &= 0.
\end{aligned}$$

One solution is given by

$$\tilde{x} = \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix}, \quad d.h. \quad \lambda_1 = 0, \quad \lambda_2 = 0 \ll, \quad \lambda_3 = \frac{1}{2},$$

i.e., we have strict complementarity.

Sufficient condition:

$$\begin{aligned}
\nabla_{xx}^2 \mathcal{L}(x, \lambda) &= \begin{pmatrix} \frac{1}{8} \frac{1}{x_1^{3/2}} & 0 \\ 0 & 0 \end{pmatrix} \\
\nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}) &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{is not positive definite on } \mathbb{R}^2
\end{aligned}$$

but we have

$$\begin{aligned}
C(\tilde{x}, \tilde{\lambda}) &= \left\{ d \in \mathbb{R}^2 \mid \nabla c_3^\top d = -d_1 - d_2 = 0 = d_1 + d_2 \right\} \\
&= \left\{ d \in \mathbb{R}^2 \mid d_1 = -d_2 \right\}.
\end{aligned}$$

Since it holds $d_1^2 = d_2^2$, we obtain

$$d^\top \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} d = d_1^2 = \frac{1}{2} d_1^2 + \frac{1}{2} d_2^2 = \frac{1}{2} \|d\|^2,$$

i.e., the sufficient condition is satisfied for $\sigma = \frac{1}{2}$.

5.6 Problems with box constraints

Here, we analyze the case $\Omega = \left\{ x \in \mathbb{R}^n \mid v_i \leq x_i \leq w_i, 1 \leq i \leq n \right\}$, i.e., a box in \mathbb{R}^n .

Let $v = (v_1, \dots, v_n)^\top$ and $w = (w_1, \dots, w_n)^\top$. We assume that

$v < w$ (componentwise) and consider the problem

$$\min_{v \leq x \leq w} f(x), \quad \text{with } f: D \supset \Omega \rightarrow \mathbb{R} \text{ sufficiently smooth.} \quad (5.21)$$

i) transformation to standard form:

$$\begin{aligned} v \leq x \leq w &\Leftrightarrow x - v \geq 0 \quad \text{und} \quad w - x \geq 0 \\ &\Leftrightarrow \begin{pmatrix} I \\ -I \end{pmatrix} x + \begin{pmatrix} -v \\ w \end{pmatrix} \geq 0 \Leftrightarrow Gx - v \geq 0 \end{aligned}$$

due to $v < w$ at least one of the constraints (or both) have to be inactive, hence, the gradients of the active constraints are linearly independent and the Lagrange multiplier is uniquely determined.

ii) Necessary optimality conditions

$$\begin{aligned} \mathcal{L}(x, \lambda) &= f(x) - \sum_{i=1}^n \lambda_i^u (x_i - v_i) - \sum_{i=1}^n \lambda_i^o (-x_i + w_i) \\ \mathcal{L}_{x_i} = 0 &\Leftrightarrow \frac{\partial f}{\partial x_i} - \lambda_i^u + \lambda_i^o = 0, \quad \lambda_i^u, \lambda_i^o \geq 0 \end{aligned}$$

λ_i^u : multiplier for the lower bound

λ_i^o : multiplier for the upper bound

Additionally: one of the multipliers is always zero, one may (but does not have to) be strictly positive.

In more detail:

$$\begin{aligned} \frac{\partial f}{\partial x_i} = 0 &\Rightarrow \lambda_i^u = \lambda_i^o = 0 \\ \frac{\partial f}{\partial x_i} \geq 0 &\Rightarrow \lambda_i^u = \frac{\partial f}{\partial x_i}, \lambda_i^o = 0 \\ \frac{\partial f}{\partial x_i} \leq 0 &\Rightarrow \lambda_i^o = -\frac{\partial f}{\partial x_i}, \lambda_i^u = 0. \end{aligned}$$

Conclusion:

$$\begin{aligned} \lambda_i^u &= \left[\frac{\partial f}{\partial x_i} \right]^+ \\ \lambda_i^o &= \left[\frac{\partial f}{\partial x_i} \right]^-. \end{aligned}$$

iii) The variational inequality: Ω is convex s.t. we know

$$\nabla f(\tilde{x})^\top (x - \tilde{x}) \geq 0 \quad \forall x \in \Omega \quad \text{i.e.} \quad v \leq x \leq w$$

$$\text{Choose } x \in \Omega \text{ s.t. } x_j = \begin{cases} \tilde{x}_j, & j \neq i \\ z, & j = i \end{cases}, \text{ with } z \in \mathbb{R} \text{ s.t. } v_i \leq z \leq w_i.$$

It then follows

$$\begin{aligned} \frac{\partial f}{\partial x_i} \tilde{x}_i &\leq \frac{\partial f}{\partial x_i} z \quad \forall z \in v_i \leq z \leq w_i \\ \text{i.e.} \quad \tilde{x}_i &= \arg \min_{v_i \leq z \leq w_i} \frac{\partial f}{\partial x_i} z. \end{aligned}$$

Conclusion

$$\begin{aligned}\frac{\partial f}{\partial x_i} > 0 &\Rightarrow \tilde{x}_i = v_i \\ \frac{\partial f}{\partial x_i} < 0 &\Rightarrow \tilde{x}_i = w_i.\end{aligned}$$

For $\frac{\partial f}{\partial x_i} = 0$ we do not obtain additional information.

iv) Second order conditions:

Note that $\nabla_{xx}\mathcal{L}(\tilde{x}, \tilde{\lambda}) = f''(\tilde{x})$

$$\begin{aligned}C(\tilde{x}, \tilde{\lambda}) &= \left\{ d \in \mathbb{R}^n \mid \nabla c_i^T d = 0 \quad \forall i \in \mathcal{A}(\tilde{x}) \right. \\ &\quad \left. \text{with } \tilde{\lambda}_i > 0 \quad \text{and } \nabla c_i^T d \geq 0, \quad \text{if } \tilde{\lambda}_i = 0 \right\} \\ &= \left\{ d \in \mathbb{R}^n \mid d_i = 0, \quad \text{if } \left| \frac{\partial f}{\partial x_i} \right| \neq 0, \right. \\ &\quad \left. \text{else } d_i \geq 0, \quad \text{if } \tilde{x}_i = v_i, \quad d_i \leq 0, \quad \text{if } \tilde{x}_i = w_i \right\}.\end{aligned}$$

We require the stronger condition

$$\begin{aligned}d^\top f''(\tilde{x})d &\geq \alpha \|d\|^2 \quad \forall d \in \mathbb{R}^n \\ \text{with } d_i &= 0 \quad \text{if } \left| \frac{\partial f}{\partial x_i} \right| \neq 0.\end{aligned}\tag{5.22}$$

5.7 Further regularity conditions

For linear affine constraints

$$c_i(x) = a_i^T x + b_i$$

with $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, the linearizing cone $L_\Omega(x)$ appropriately characterizes the set Ω . In particular, we have the following result.

Lemma 5.7.1 *Let $\tilde{x} \in \mathbb{R}$ be such that all active constraints are linear affine. Then*

$$L_\Omega(\tilde{x}) = T_\Omega(\tilde{x}).$$

PROOF. We already know that $T_\Omega(\tilde{x}) \subset L_\Omega(\tilde{x})$. Consider $w \in L_\Omega(\tilde{x})$. By definition, we have

$$L_\Omega(\tilde{x}) = \left\{ d \in \mathbb{R}^n \mid a_i^\top d = 0 \quad \forall i \in \mathcal{E} \quad \text{and} \quad a_i^\top d \geq 0 \quad \forall i \in \mathcal{A}(\tilde{x}) \cap I \right\}.$$

Let $i \in I \setminus \mathcal{A}(\tilde{x})$, i.e., c_i is inactive, $c_i(\tilde{x}) > 0$. Then there exists $\bar{t} > 0$ s.t.

$$c_i(\tilde{x} + tw) > 0 \quad \forall t \in [0, \bar{t}]$$

and c_i remains inactive.

Let us define

$$z^{(k)} := \tilde{x} + \frac{\bar{t}}{k} w$$

so that for $i \in I \cap \mathcal{A}(\tilde{x})$ we obtain

$$\begin{aligned} c_i(z^{(k)}) &= c_i(z^{(k)}) - c_i(\tilde{x}) = a_i^\top (z^{(k)} - \tilde{x}) \\ &= \frac{\bar{t}}{k} a_i^\top w \geq 0 \quad \text{by definition of } L_\Omega(\tilde{x}). \end{aligned}$$

Similarly, for $i \in \mathcal{E}$

$$c_i(z^{(k)}) = c_i(z^{(k)}) - c_i(\tilde{x}) = \frac{\bar{t}}{k} a_i^\top w = 0.$$

Hence, $\{z^{(k)}\}$ is an admissible approximation of \tilde{x} . Moreover, it holds that

$$\lim_{k \rightarrow \infty} \frac{z^{(k)} - \tilde{x}}{\bar{t}/k} = \lim_{k \rightarrow \infty} \frac{(\bar{t}/k)w}{\bar{t}/k} = w$$

and, consequently, $w \in T_\Omega(\tilde{x})$. □

Definition 5.7.1 *The Mangasarian-Fromovitz constraint qualification (MFCQ) is satisfied if there exists $w \in \mathbb{R}^n$ s.t.*

$$\begin{aligned} \nabla c_i(\tilde{x})^\top w &> 0 \quad \forall i \in \mathcal{A}(\tilde{x}) \cap I \\ \nabla c_i(\tilde{x})^\top w &= 0 \quad \forall i \in \mathcal{E} \end{aligned}$$

and the vectors $\{\nabla c_i(\tilde{x}), i \in \mathcal{E}\}$ are linearly independent.

Remarks:

(1) One can show that (MFCQ) implies $L_\Omega(\tilde{x}) = T_\Omega(\tilde{x})$, i.e., (ACQ).

(2) If (LICQ) is satisfied there exists $w \in \mathbb{R}^2$ with

$$\begin{aligned} \nabla c_i(\tilde{x})^\top w &= 1 \quad \forall i \in \mathcal{A}(\tilde{x}) \cap I \\ \nabla c_i(\tilde{x})^\top w &= 0 \quad \forall i \in \mathcal{E}, \end{aligned}$$

i.e., (LICQ) \Rightarrow (MFCQ), (LICQ) is a stronger condition.

(3) In case of (MFCQ) the Lagrange multipliers are not necessarily unique.

A sufficient condition for (MFCQ) is the global Slater condition:

Definition 5.7.2 Let $D \subset \mathbb{R}^n$ be open and convex, $-c_i$, $i \in I$ convex and differentiable on D , c_i , $i \in \mathcal{E}$ linear affine, i.e., there exist $a_i \in \mathbb{R}^n$, $i = 1, \dots, m$, $b \in \mathbb{R}^m$ s.t.

$$c_i(x) = a_i^\top x + b_i.$$

The global Slater constraint qualification (SCQ) is satisfied if additionally $\{a_i, i = 1, \dots, m\}$ are linearly independent and there exists $v \in \mathbb{R}^n$ with

$$\begin{aligned} c_i(v) &= 0 \quad \forall i \in \mathcal{E} \\ c_i(v) &> 0 \quad \forall i \in I. \end{aligned}$$

Corollary: Let (SCQ) be satisfied, then in every $\tilde{x} \in \Omega$ (MFCQ) is satisfied, i.e., in every $\tilde{x} \in \Omega$ it holds that $L_\Omega(\tilde{x}) = T_\Omega(\tilde{x})$.

PROOF. Let $\tilde{x} \in \Omega$ be arbitrary, then for all $i \in \mathcal{E}$

$$\nabla c_i(\tilde{x})^\top (v - \tilde{x}) = a_i^\top (v - \tilde{x}) = 0.$$

For $i \in I$ $-c_i$ is convex and

$$\nabla c_i(\tilde{x})^\top (v - \tilde{x}) \geq c_i(v) - c_i(\tilde{x})$$

if $i \in I \cap \mathcal{A}(\tilde{x})$ it follows $c_i(\tilde{x}) = 0$, i.e., for $d = v - \tilde{x}$ it holds that

$$\begin{aligned} \nabla c_i(\tilde{x})^\top d &= 0, \quad i \in \mathcal{E} \\ \nabla c_i(\tilde{x})^\top d &\geq c_i(v) > 0, \quad i \in I \cap \mathcal{A}(\tilde{x}). \end{aligned}$$

□

5.8 Geometric interpretation of the necessary optimality conditions

Let us characterize the optimality conditions independently of the algebraic specification of the admissible set. For this purpose re-consider

$$\min_{x \in \Omega} f(x).$$

Definition 5.8.1 The normal cone to the tangent cone in $x \in \Omega$ is defined via

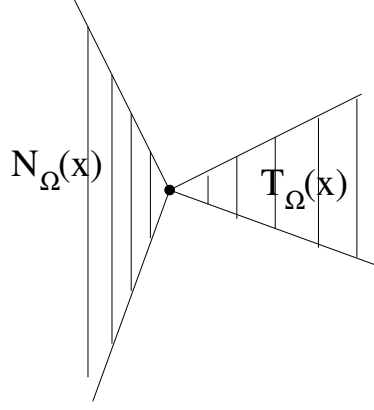
$$N_\Omega(x) = \left\{ v \in \mathbb{R}^n \mid v^\top w \leq 0 \quad \forall w \in T_\Omega(x) \right\}.$$

Every $v \in N_\Omega(x)$ is called normal vector.

Let \tilde{x} be local solution to (P), then it holds

$$\begin{aligned} &\nabla f(\tilde{x})^\top d \geq 0 \quad \forall d \in T_\Omega(\tilde{x}) \\ \Leftrightarrow &-\nabla f(\tilde{x})^\top d \leq 0 \quad \forall d \in T_\Omega(\tilde{x}) \\ \Leftrightarrow &\nabla f(\tilde{x}) \in N_\Omega(\tilde{x}) \end{aligned}$$

i.e., we obtain



Theorem 5.8.1 *Let \tilde{x} be local solution to (P), then*

$$-\nabla f(\tilde{x}) \in N_{\Omega}(\tilde{x}).$$

Remark: If $\tilde{x} \in \text{int } \Omega$ then $T_{\Omega}(\tilde{x}) = \mathbb{R}^n$, i.e., $N_{\Omega}(\tilde{x}) = \{0\}$ and we obtain $\nabla f(\tilde{x}) = 0$. Moreover, Theorem 5.8.1 indicates that $\nabla f(\tilde{x})$ is related to linear combinations of the active constraints. Indeed, it holds:

Lemma 5.8.1 *Let (LICQ) be satisfied in the local solution \tilde{x} to (P). Then $N_{\Omega}(\tilde{x}) = -N$ where N is defined via*

$$N := \left\{ \sum_{i \in \mathcal{A}(\tilde{x})} \lambda_i \nabla c_i(\tilde{x}), \lambda_j \geq 0 \quad \text{für } j \in \mathcal{A}(\tilde{x}) \cap I \right\}.$$

PROOF. With the Farkas' lemma, we obtain

$$\begin{aligned} g \in N &\Rightarrow g^{\top} d \geq 0 \quad \forall d \in L_{\Omega}(\tilde{x}) = T_{\Omega}(\tilde{x}) \\ \text{i.e. } g \in -N &\Rightarrow g^{\top} d \leq 0 \quad \forall d \in T_{\Omega}(\tilde{x}) \\ \Rightarrow N_{\Omega}(\tilde{x}) &= -N. \end{aligned}$$

□

From Lemma 5.8.1 we conclude that $-\nabla f(\tilde{x}) \in -N$, i.e., there exist multipliers $\tilde{\lambda}_i$ s.t.

$$-\nabla f(\tilde{x}) = - \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i c_i(\tilde{x})$$

respectively

$$\nabla f(\tilde{x}) - \sum_{i \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_i c_i(\tilde{x}) = 0,$$

i.e., we again obtain the KKT conditions.

5.9 Lagrange multipliers and sensitivity

Here, we will study the role of Lagrange multipliers associated with active inequality constraints with regard to the importance of individual constraints.

For this purpose, let us consider

$$\begin{aligned} & \min f(x) \\ & \text{s.t. } c_i(x) \geq 0, \quad i \in I. \end{aligned} \quad (\text{P})$$

If c_i is inactive, we have $c_i(\tilde{x}) > 0$ and

$$c_i(z) > 0 \quad \forall z \in B_\varepsilon(\tilde{x}) \cap \Omega.$$

Due to the complementarity condition, this implies $\tilde{\lambda}_i = 0$, i.e., the Lagrange multipliers shows that the inequality c_i is not relevant.

Assume now that $c_i(\tilde{x}) = 0$, i.e., $i \in \mathcal{A}(\tilde{x})$. We consider a small perturbation, i.e., instead of $c_i(x)$ we consider

$$\hat{c}_i(x) = c_i(x) + \varepsilon \|\nabla c_i(\tilde{x})\| \geq 0.$$

Let $\tilde{x}(\varepsilon)$ be solution to (P_ε) where c_i has been replaced by \hat{c}_i ersetzt wurde. We assume that for ε sufficiently small, we have $\mathcal{A}(\tilde{x}) = \mathcal{A}(\tilde{x}(\varepsilon))$ and, consequently, $\hat{c}_i(\tilde{x}(\varepsilon)) = 0$ i.e.,

$$-\varepsilon \|\nabla c_i(\tilde{x})\| = c_i(\tilde{x}(\varepsilon)) - \underbrace{c_i(\tilde{x})}_{=0} \approx (\tilde{x}(\varepsilon) - \tilde{x})^\top \nabla c_i(\tilde{x}).$$

Indices $j \in \mathcal{A}(\tilde{x})$ with $j \neq i$ are not perturbed, i.e., for these we obtain

$$0 = c_j(\tilde{x}(\varepsilon)) - c_j(\tilde{x}) \approx (\tilde{x}(\varepsilon) - \tilde{x})^\top \nabla c_j(\tilde{x}).$$

Altogether, the KKT condition imply

$$\begin{aligned} f(\tilde{x}(\varepsilon)) - f(\tilde{x}) & \approx (\tilde{x}(\varepsilon) - \tilde{x})^\top \nabla f(\tilde{x}) = (\tilde{x}(\varepsilon) - \tilde{x})^\top \sum_{j \in \mathcal{A}(\tilde{x})} \tilde{\lambda}_j \nabla c_j(\tilde{x}) \\ & \approx (\tilde{x}(\varepsilon) - \tilde{x})^\top \tilde{\lambda}_i \nabla c_i(\tilde{x}) \approx -\varepsilon \|\nabla c_i(\tilde{x})\| \tilde{\lambda}_i. \end{aligned}$$

For $\varepsilon \searrow 0$, we obtain

$$\left. \frac{df(\tilde{x}(\varepsilon))}{d\varepsilon} \right|_{\varepsilon=0} = -\|\nabla c_i(\tilde{x})\| \tilde{\lambda}_i.$$

We thus conclude that if $\tilde{\lambda}_i \|\nabla c_i\|$ is large, the value of the objective function increases significantly. On the other hand, if $\tilde{\lambda}_i = 0$, small perturbations of c_i will not change the value of the objective function drastically. We thus define:

Definition 5.9.1 *Let \tilde{x} be solution to (P) and the KKT conditions are satisfied. An inequality constraint c_i is called strongly active if $i \in \mathcal{A}(\tilde{x})$ and $\lambda_i > 0$. c_i is called weakly active if $i \in \mathcal{A}(\tilde{x})$ and $\lambda_i = 0$.*

5.10 Duality

In this section, we restrict ourselves to inequality constraints and focus on

$$\begin{aligned} & \min f(x) \\ & \text{s.t. } c_i(x) \geq 0, \quad 1 \leq i \leq m. \end{aligned} \quad (\text{P})$$

Let us define $c(x) := (c_1(x), \dots, c_m(x))^\top$ s.t. we obtain the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) - \lambda^\top c(x) \quad \text{mit } \lambda \in \mathbb{R}^m.$$

Let us introduce the dual objective function $q: \mathbb{R}^m \rightarrow \mathbb{R}$ via

$$q(\lambda) := \inf_x \mathcal{L}(x, \lambda). \quad (5.23)$$

In many cases, specific values of λ may result in $\inf_x \mathcal{L}(x, \lambda) = -\infty$ s.t. we define

$$D := \{\lambda \in \mathbb{R}^m \mid q(\lambda) > -\infty\}$$

to be the domain of q .

Computing the global minimum of (5.23) generally is difficult but here we focus on convex f and $-c_i$ with $\lambda_i \geq 0$ which results in a convex $\mathcal{L}(\cdot, \lambda)$. Hence, all local minima are global minima and the required computations simplify since we only have to check necessary optimality conditions.

We now define the *dual problem* to (P)

$$\begin{aligned} & \max q(\lambda) \\ & \text{s.t. } \lambda \geq 0. \end{aligned} \quad (\text{P}_D)$$

Example 5.10.1

$$\begin{aligned} & \min \frac{1}{2} (x_1^2 + x_2^2) \\ & \text{s.t. } x_1 - 1 \geq 0. \end{aligned}$$

Geometrically, we obtain the solution $\tilde{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$$\mathcal{L}(x, \lambda) = \frac{1}{2} (x_1^2 + x_2^2) - \lambda_1 (x_1 - 1).$$

Evaluation of the KKT conditions yields

$$\nabla f(\tilde{x}) = \tilde{x} \stackrel{!}{=} \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

i.e., $\tilde{\lambda}_1 = 1$ and the optimal value of the objective function is

$$f(\tilde{x}) = \frac{1}{2} (1 + 0) = \frac{1}{2}.$$

For fixed λ_1 , \mathcal{L} is convex in $x \in \mathbb{R}^2$ and we obtain the global minimum by studying $\nabla_x \mathcal{L} = 0$, i.e.,

$$\begin{aligned} x_1 - \lambda_1 &= 0 \\ x_2 &= 0. \end{aligned}$$

Inserting this into \mathcal{L} leads to

$$\begin{aligned} q(\lambda_1) &= \frac{1}{2} \lambda_1^2 - \lambda_1(\lambda_1 - 1) \\ &= -\frac{1}{2} \lambda_1^2 + \lambda_1 = \lambda_1 \left(1 - \frac{1}{2} \lambda_1 \right). \end{aligned}$$

The dual problem now reads

$$\max_{\lambda_1 \geq 0} \left(-\frac{1}{2} \lambda_1^2 + \lambda_1 \right)$$

with solution $\lambda_1 = 1$ and $q(\lambda_1) = \frac{1}{2}$, i.e., we obtain the exact same value for λ_1 and the same optimal value as for the primal problem.

In the following, we study the connection between (P) and (P_D) in more detail.

Lemma 5.10.1 $q(\lambda)$ is concave and its domain D is convex.

PROOF. Let $\lambda^1, \lambda^2 \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ then

$$\mathcal{L}(x, (1 - \alpha)\lambda^1 + \alpha\lambda^2) = (1 - \alpha)\mathcal{L}(x, \lambda^1) + \alpha\mathcal{L}(x, \lambda^2).$$

Since $\inf(f(x) + g(x)) \geq \inf f(x) + \inf g(x)$ it follows

$$q((1 - \alpha)\lambda^1 + \alpha\lambda^2) \geq \alpha q(\lambda^1) + (1 - \alpha)q(\lambda^2). \quad (5.24)$$

For $\lambda^1, \lambda^2 \in D$ this implies

$$q((1 - \alpha)\lambda^1 + \alpha\lambda^2) > -\infty,$$

i.e., $(1 - \alpha)\lambda^1 + \alpha\lambda^2 \in D$ and D is convex. Similarly, concavity of q follows with (5.24).

Lemma 5.10.2 For every admissible $\bar{x} \in \Omega$ to (P) and every $\bar{\lambda} \geq 0$ it holds that

$$q(\bar{\lambda}) \leq f(\bar{x}).$$

PROOF.

$$q(\bar{\lambda}) = \inf_x (f(x) - \bar{\lambda}^\top c(x)) \leq f(\bar{x}) - \underbrace{\bar{\lambda}^\top c(\bar{x})}_{\geq 0} \leq f(\bar{x}).$$

□

In the following, we exploit the KKT conditions for (P) which read

$$\begin{aligned}\nabla f(\bar{x}) - c'(\bar{x})^\top \bar{\lambda} &= 0 \\ c(\bar{x}) &\geq 0 \\ \bar{\lambda} &\geq 0 \\ \bar{\lambda}_i c_i(\bar{x}) &= 0, \quad i = 1, \dots, m.\end{aligned}$$

Here, $c'(\bar{x})^\top = [\mathcal{D}c(\bar{x})]^\top = (\nabla c_1(\bar{x}), \dots, \nabla c_m(\bar{x}))$.

Under certain assumptions, the optimal Lagrange multipliers for (P) are solutions to the dual problem (P_D) . More precisely:

Theorem 5.10.1 *Let \bar{x} be solution to (P) and f and $-c_i(x)$ convex functions on \mathbb{R}^n . Then every $\bar{\lambda}$ s.t. $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions is a solution to the dual problem (P_D) .*

PROOF. $(\bar{x}, \bar{\lambda})$ satisfy the KKT conditions, $\bar{\lambda} \geq 0$

$\Rightarrow \mathcal{L}(\cdot, \bar{\lambda})$ is convex and differentiable

$$\begin{aligned}\Rightarrow \mathcal{L}(x, \bar{\lambda}) &\geq \mathcal{L}(\bar{x}, \bar{\lambda}) + \underbrace{\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda})^\top (x - \bar{x})}_{=0} = \mathcal{L}(\bar{x}, \bar{\lambda}) \\ &\quad \uparrow \\ &\quad \text{convexity}\end{aligned}$$

$$\Rightarrow q(\bar{\lambda}) = \inf_x \mathcal{L}(x, \bar{\lambda}) = \mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x}) - \bar{\lambda}^\top c(\bar{x}) = f(\bar{x}). \quad (5.25)$$

On the other hand, with Lemma 5.10.2 we obtain $q(\lambda) \leq f(\bar{x})$ for all $\lambda \geq 0$, i.e., $\bar{\lambda}$ is solution to (P_D) . \square

Under additional assumptions, the solution to the dual problem can be used to characterize the solution of (P).

Theorem 5.10.2 *Let f and $-c_i$, $i = 1, \dots, m$ be convex and continuously differentiable on \mathbb{R}^n and \bar{x} a solution of (P) s.t. (LICQ) is satisfied. Let further $\hat{\lambda}$ be the solution to (P_D) , where the infimum of $\mathcal{L}(x, \hat{\lambda})$ is attained in \hat{x} . Moreover, let $\mathcal{L}(\cdot, \hat{\lambda})$ be strictly convex. Then $\bar{x} = \hat{x}$, i.e., the unique solution to (P) and $f(\bar{x}) = \mathcal{L}(\hat{x}, \hat{\lambda})$.*

PROOF. Assume $\bar{x} \neq \hat{x}$. (LICQ) is satisfied in $\bar{x} \Rightarrow$ there exists $\bar{\lambda}$ s.t. the KKT conditions are satisfied. Hence, with Theorem 5.10.1 it follows that in addition to $\hat{\lambda}$ also $\bar{\lambda}$ is solution to (P_D) and due to (5.25) it holds

$$\mathcal{L}(\bar{x}, \bar{\lambda}) = q(\bar{\lambda}) = q(\hat{\lambda}) = \mathcal{L}(\hat{x}, \hat{\lambda}).$$

By assumption $\hat{x} = \arg \min_x \mathcal{L}(x, \bar{\lambda})$ s.t. it has to hold

$$\nabla_x \mathcal{L}(\hat{x}, \hat{\lambda}) = 0$$

and, by strict convexity of $\mathcal{L}(x, \hat{\lambda})$, it follows that

$$\mathcal{L}(\bar{x}, \hat{\lambda}) - \mathcal{L}(\hat{x}, \hat{\lambda}) > \nabla_x \mathcal{L}(\hat{x}, \hat{\lambda})^\top (\bar{x} - \hat{x}) = 0.$$

Hence,

$$\begin{aligned} \mathcal{L}(\bar{x}, \hat{\lambda}) &> \mathcal{L}(\hat{x}, \hat{\lambda}) = \mathcal{L}(\bar{x}, \bar{\lambda}) \\ \Rightarrow \quad f(\bar{x}) - \hat{\lambda}^\top c(\bar{x}) &> f(\bar{x}) - \bar{\lambda}^\top c(\bar{x}) \\ \text{i.e.} \quad -\hat{\lambda}^\top c(\bar{x}) &> \bar{\lambda}^\top c(\bar{x}) = 0 \end{aligned}$$

contradicting $\hat{\lambda} \geq 0$ and $c(\bar{x}) \geq 0$, i.e., our assumption was wrong and, thus, $\hat{x} = \bar{x}$. \square

For numerical purposes, the following dual problem (introduced by Wolfe) is advantageous:

$$\left. \begin{array}{l} \max_{x, \lambda} \mathcal{L}(x, \lambda) \\ \text{s.t.} \quad \nabla_x \mathcal{L}(x, \lambda) = 0 \\ \lambda \geq 0. \end{array} \right\} \quad (\text{P}_{DW})$$

The connection between (P) and (P_{DW}) is as follows:

Theorem 5.10.3 *Let f and $-c_i$, $1 \leq i \leq m$ be convex and continuously differentiable on \mathbb{R}^n . Further, let $(\bar{x}, \bar{\lambda})$ be a solution to (P) for which (LICQ) holds true. Then $(\bar{x}, \bar{\lambda})$ also solves (P_{DW}).*

PROOF. KKT conditions $\Rightarrow \nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = 0$ and $\bar{\lambda} \geq 0$ and $\mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x})$.

Hence, for every (x, λ) it follows that the constraint of (P_{DW}) are satisfied:

$$\begin{aligned} \mathcal{L}(\bar{x}, \bar{\lambda}) &= f(\bar{x}) \geq f(\bar{x}) - \underbrace{\lambda^\top c(\bar{x})}_{\geq 0} \\ &= \mathcal{L}(\bar{x}, \lambda) \geq \mathcal{L}(x, \lambda) + \underbrace{\nabla_x \mathcal{L}(x, \lambda)}_{\substack{\uparrow \\ \text{convexity}}}(\bar{x} - x) = \mathcal{L}(x, \lambda). \end{aligned}$$

\square

Example 5.10.2 (linear optimization)

$$\min c^\top x, \quad \text{s.t.} \quad Ax - b \geq 0$$

with $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ dual objective function:

$$\begin{aligned} q(\lambda) &= \inf_x \mathcal{L}(x, \lambda) = \inf_x (c^\top x - \lambda^\top (Ax - b)) \\ &= \inf_x [(c - A^\top \lambda)^\top x + b^\top \lambda]. \end{aligned}$$

If $c - A^\top \lambda \neq 0$, set $x = -(c - A^\top \lambda)r$, with $r \in \mathbb{R}$ s.t.

$$(c - A^\top \lambda)^{-\top} x + b^\top \lambda = -r \|c - A^\top \lambda\|^2 + b^\top \lambda \xrightarrow{r \rightarrow \infty} -\infty$$

i.e.,

$$\inf_x [(c - A^\top \lambda)^\top x + b^\top \lambda] = -\infty.$$

Hence, when maximizing we do not have to consider the case $c - A^\top \lambda \neq 0$. We obtain the dual problem

$$\begin{cases} \max b^\top \lambda \\ \text{s.t.} & A^\top \lambda = c \\ & \lambda \geq 0. \end{cases}$$

It further holds

$$\begin{aligned} \mathcal{L}(x, \lambda) &= (c - A^\top \lambda)^\top x + b^\top \lambda \\ \nabla_x \mathcal{L}(x, \lambda) &= c - A^\top \lambda. \end{aligned}$$

The Wolfe dual problem thus becomes

$$\begin{aligned} \max_{x, \lambda} & (c - A^\top \lambda)^\top x + b^\top \lambda \\ \text{s.t.} & \quad c - A^\top \lambda = 0, \lambda \geq 0 \end{aligned}$$

respectively,

$$\begin{aligned} \max & b^\top \lambda \\ \text{s.t.} & \quad c - A^\top \lambda = 0 \\ & \quad \lambda \geq 0. \end{aligned}$$

Example 5.10.3 (linear quadratic optimization)

$G \in \mathbb{R}^{n \times n}$ symmetric positive definite, $A \in \mathbb{R}^{m \times n}$

$$\begin{aligned} \min & \frac{1}{2} x^\top G x + c^\top x \\ \text{s.t.} & \quad A x - b \geq 0. \end{aligned} \tag{P}$$

Here, the dual objective function is given by

$$q(\lambda) = \inf_x \mathcal{L}(x, \lambda) = \inf_x \frac{1}{2} x^\top G x + c^\top x - \lambda^\top (A x - b),$$

G is positive definite, i.e., $\mathcal{L}(\cdot, \lambda)$ is strictly convex and the minimum is attained for $\nabla_x \mathcal{L}(x, \lambda) = 0$, i.e.,

$$\begin{aligned} G x + c - A^\top \lambda &= 0 \\ \Leftrightarrow x &= G^{-1}(A^\top \lambda - c). \end{aligned}$$

Inserting yields

$$\begin{aligned}
q(\lambda) &= \frac{1}{2}(A^\top \lambda - c)^\top \underbrace{G^{-1}G}_{I} G^{-1}(A^\top \lambda - c) \\
&\quad + c^\top G^{-1}(A^\top \lambda - c) - \lambda^\top A G^{-1}(A^\top \lambda - c) + \lambda^\top b \\
q(\lambda) &= -\frac{1}{2}(A^\top \lambda - c)^\top G^{-1}(A^\top \lambda - c) + \lambda^\top b.
\end{aligned}$$

The Wolfe dual problem becomes

$$\begin{aligned}
\max_{\lambda, x} \quad & \frac{1}{2} x^\top G x + c^\top x - \lambda^\top (A x - b) \\
\text{s.t.} \quad & G x + c - A^\top \lambda = 0 \\
& \lambda \geq 0.
\end{aligned}$$

Using the equality constraint

$$(c - A^\top \lambda)^\top x = -x^\top G x$$

we obtain

$$\begin{aligned}
\max_{\lambda, x} \quad & -\frac{1}{2} x^\top G x + \lambda^\top b \\
\text{s.t.} \quad & \begin{cases} G x + c - A^\top \lambda = 0 & \lambda = 0 \\ \lambda \geq 0. \end{cases}
\end{aligned}$$

5.11 Outlook: numerical solution of nonlinear optimization problems with constraints

In chapter 6, we will study problems with linear constraints which are important as only in this case there exist truly efficient numerical algorithms. Moreover, these problems form the basis for more general problems with nonlinear constraints.

In chapter 7, we consider SQP methods for solving nonlinear constrained problems. In every iteration step, we compute $(x^{(k)}, \lambda^{(k)})$ as a solution to a linear quadratic problem of the form

$$\begin{aligned}
\min_p \quad & \frac{1}{2} p^\top \nabla_{xx} \mathcal{L}(x^{(k)}, \lambda^{(k)}) p + \nabla f(x^{(k)})^\top p \\
\text{s.t.} \quad & \nabla c_i(x^{(k)})^\top p + c_i(x^{(k)}) = 0, \quad i \in \mathcal{E} \\
& \nabla c_i(x^{(k)})^\top p + c_i(x^{(k)}) \geq 0, \quad i \in I.
\end{aligned}$$

In chapter 8 we discuss penalty and augmented Lagrangian methods. By introducing a penalty term, we will transform constrained problems into a sequence of unconstrained problems. In case of equality constraints, we may for example consider

$$f(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x). \quad (5.26)$$

Here, $\mu > 0$ denotes the so-called penalty parameter. For the computation of the solution of (P), (5.26) will be computed for a sequence of increasing penalty parameters $\{\mu_k\}$.

Augmented Lagrangian methods have better numerical properties (in general). For these methods, the modified objective function combines properties of the Lagrangian and the penalty objective function (5.26). More precisely, we consider

$$\mathcal{L}_A(x, \lambda; \mu) = f(x) - \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(x).$$

The numerical method then provides approximations for the solution \tilde{x} of (P) as well as for the Lagrange multiplier.

6 Problems with linear constraints - methods

6.1 Quadratic optimization problems

6.1.1 Problems with equality constraints

We consider

$$\begin{array}{ll} \min_{x \in \mathbb{R}^n} & \frac{1}{2} \langle x, Qx \rangle + \langle q, x \rangle \\ \text{s.t.} & Ax = b \end{array} \quad (\text{QE})$$

$Q \in \mathbb{R}^{n \times n}$ symmetric, $A \in \mathbb{R}^{m \times n}$, $m \leq n$.

Assumption:

- $\text{rk}(A) = m$ (full rank)
- $d^\top Q d \geq \alpha \|d\|^2 \quad \forall d \in \ker(A)$

In this case, (QE) has exactly one solution \tilde{x} with exactly one corresponding Lagrange multiplier which fulfill

$$\mathcal{A} \begin{pmatrix} \tilde{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix} \quad \text{with} \quad \mathcal{A} = \begin{pmatrix} Q & -A^\top \\ A & 0 \end{pmatrix} \quad (6.27)$$

Sometimes the above system is rewritten based on the transformation $\tilde{x} = x + p$, $h = Ax - b$, $g = q + Qx$ s.t.

$$\begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} -p \\ \lambda \end{pmatrix} = \begin{pmatrix} q \\ h \end{pmatrix}$$

with the invertible KKT matrix

$$K = \begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix}$$

Possible numerical methods for solving the KKT system are

- direct factorization of K (often too expensive)

- Schur complement methods (for detail, see [2])
- nullspace method, i.e., elimination of the constraints.

We will focus on the last method, i.e., we will rely on a nullspace matrix for A in order to eliminate the constraint $Ax = b$. We proceed in three steps.

1. Compute a nullspace matrix
2. Compute \tilde{x} by solving a “free” optimization problem
3. Compute λ

Step 1: QR decomposition of A^\top

Compute a unitary/orthogonal matrix H and an upper triangular matrix R with

$$\boxed{HA^\top = \begin{pmatrix} R \\ 0 \end{pmatrix}} \quad \begin{array}{l} H \in \mathbb{R}^{n \times n} \\ R \in \mathbb{R}^{m \times m} \end{array}$$

We decompose H according to

$$\boxed{H = \begin{pmatrix} Y^\top \\ Z^\top \end{pmatrix} \begin{array}{l} \} m \text{ rows} \\ \} n - m \text{ rows} \end{array}} \quad \begin{array}{l} \text{rows of } Y: \text{ first } m \text{ rows of } H, \\ \text{rows of } Z: \text{ last } (n - m) \text{ rows of } H. \end{array}$$

Both, Y and Z have to be of full rank since $\text{rk}(H) = n$. Hence, the columns of Y and Z span all of \mathbb{R}^n and every $x \in \mathbb{R}^n$ can be uniquely represented as

$$x = Yx_y + Zx_z = H^\top \begin{pmatrix} x_y \\ x_z \end{pmatrix}.$$

This particular holds true for $x = d \in \ker(A)$, and, thus

$$0 = Ad = A(Yd_y + Zd_z) = AH^\top \begin{pmatrix} d_y \\ d_z \end{pmatrix} = (R^\top, 0) \begin{pmatrix} d_y \\ d_z \end{pmatrix} = R^\top d_y.$$

Consequently, all $d \in \ker(A)$ are parametrized by

$$d = Zd_z, d_z \in \mathbb{R}^{n-m} \text{ arbitrary.}$$

\Rightarrow The matrix Z computed as above is a nullspace matrix.

We have represented x as follows:

$$x = \underbrace{Yx_y}_{\Rightarrow \in (\ker A)^\perp} + \underbrace{Zx_z}_{\in \ker A}.$$

Step 2a: let \tilde{x} be the (unknown) solution. Parts of \tilde{x} can be computed immediately:

$$\begin{aligned} \tilde{x} &= Y\tilde{x}_y + Z\tilde{x}_z \\ \Rightarrow b &= A\tilde{x} = AY\tilde{x}_y + 0 \\ b &= AY\tilde{x}_y = R^\top \tilde{x}_y & \boxed{R^\top \tilde{x}_y = b} \\ \Rightarrow \tilde{x}_y &= (R^\top)^{-1}b \end{aligned}$$

(which is easily possible by solving the (triangular) system of equations $R^\top \tilde{x}_y = b$)

Step 2b: “ $\Omega =$ specific solution of $Ax + b$ plus general solution of $Ax = 0$ ”

Specific solution: $Y\tilde{x}_y =: w$. Then $Aw = b$.

Hence:

$$\begin{aligned} x &\in \Omega \\ \Updownarrow \\ x &= w + Zz, \quad z \in \mathbb{R}^{n-m}. \end{aligned}$$

General solution: $Zz, z \in \mathbb{R}^{n-m}$

\Rightarrow **Reduced problem:**

$$\min_{z \in \mathbb{R}^{n-m}} f(w + Zz) = \frac{1}{2}(w + Zz)^\top Q(w + Zz) + q^\top (Zz + w).$$

Neglecting constant terms (which are not relevant for the optimization) this can be simplified to:

$$f = \frac{1}{2} z^\top \underbrace{Z^\top QZ}_{\tilde{Q}} z + \underbrace{\langle Z^\top Qw, z \rangle + \langle Z^\top q, z \rangle}_{=\tilde{q}^\top z}$$

\Rightarrow new form

$$\boxed{\min_{z \in \mathbb{R}^{n-m}} F(z) = \frac{1}{2} z^\top \tilde{Q} z + \tilde{q}^\top z} \quad (\text{QG})_r$$

$$\begin{aligned} \text{with } \tilde{Q} &:= Z^\top QZ \\ \tilde{q} &:= Z^\top Qw + Z^\top q. \end{aligned}$$

The above assumptions imply that \tilde{Q} is positive definite which renders the problem uniquely solvable. Necessary condition for \tilde{z} :

$$\nabla F(\tilde{z}) = 0 \quad \Leftrightarrow \quad \tilde{Q}\tilde{z} = -\tilde{q},$$

where \tilde{z} takes the role of \tilde{x}_z from above, i.e.,

$$Z^\top QZ\tilde{x}_z = -Z^\top q - Z^\top Qw = -Z^\top q - Z^\top QY\tilde{x}_y.$$

How do solve this system efficiently? Instead of $Z^\top QZ$ and $Z^\top q$ let us first consider the full matrix H :

- Compute

$$\boxed{-Hq =: \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \begin{matrix} \leftarrow \mathbb{R}^m \\ \leftarrow \mathbb{R}^{n-m} \end{matrix} =: h}$$

- Compute

$$B := HQH^\top = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

.....

Auxiliary computations:

$$HQH^\top = \begin{pmatrix} Y^\top \\ Z^\top \end{pmatrix} Q(Y, Z) = \begin{pmatrix} Y^\top \\ Z^\top \end{pmatrix} (QY, QZ) = \begin{pmatrix} Y^\top QY & Y^\top QZ \\ Z^\top QY & Z^\top QZ \end{pmatrix}$$

.....

$$\begin{aligned} \Rightarrow \quad B_{11} &= Y^\top QY & B_{12} &= Y^\top QZ \\ B_{21} &= Z^\top QY & B_{22} &= Z^\top QZ \quad \text{und} \quad h_1 = -Y^\top q \\ & & &= \tilde{Q} & h_2 &= -Z^\top q. \end{aligned}$$

All terms relevant for computing \tilde{x}_z are contained in H . The initial equation

$$\underbrace{Z^\top QZ}_{B_{22}} \tilde{x}_z = \underbrace{-Z^\top q}_{h_2} - \underbrace{Z^\top QY}_{B_{21}} \tilde{x}_y$$

now reads as

$$B_{22}\tilde{x}_z = h_2 - B_{21}\tilde{x}_y$$

Solution by, e.g., Cholesky decomposition.

\Rightarrow we finally obtain \tilde{x}_z, \tilde{x}_y and set

$$\tilde{x} := Y\tilde{x}_y + Z\tilde{x}_z.$$

Step 3: computation of the multiplier λ

$$Q\tilde{x} - A^\top \lambda = -q.$$

We insert \tilde{x} as given above

$$\tilde{x} = H^\top \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix}$$

and premultiply the equation with H .

\Rightarrow

$$\underbrace{HQH^\top}_B \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix} - \underbrace{HA^\top}_{\begin{pmatrix} R \\ 0 \end{pmatrix}} \lambda = \underbrace{-Hq}_h$$

i.e.,

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \tilde{x}_y \\ \tilde{x}_z \end{pmatrix} - \begin{pmatrix} R \\ 0 \end{pmatrix} \lambda = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

$$\Rightarrow \quad \boxed{R\lambda = -h_1 + B_{11}\tilde{x}_y + B_{12}\tilde{x}_z}.$$

which is easy to solve since R is upper triangular. We have obtained the final result!

6.1.2 Problems with inequality constraints - the active set method

We consider

$$\boxed{\begin{array}{ll} \min_{x \in \mathbb{R}^n} & \frac{1}{2} \langle x, Qx \rangle + \langle q, x \rangle \\ \text{s.t.} & Ax = b \text{ and } Gx \geq r \end{array}} \quad (\text{QLI})$$

with $Q = Q^\top \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{m \times n}$, $m \leq n$, $G \in \mathbb{R}^{p \times n}$. The admissible set is given by

$$\Omega = \{x \in \mathbb{R}^n \mid Ax = b, Gx \geq r\}.$$

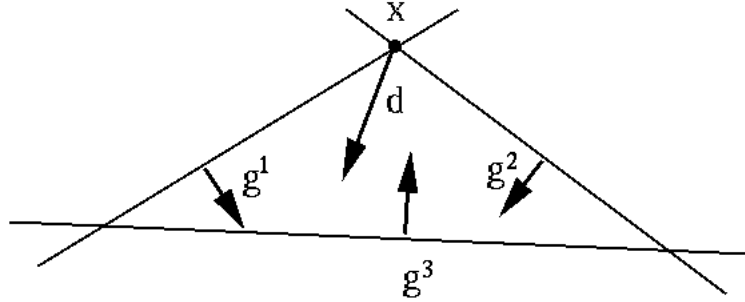
For the linearizing cone we obtain

$$L_\Omega(\tilde{x}) = \{d \in \mathbb{R}^n \mid Ad = 0, \langle g^j, d \rangle \geq 0 \quad \forall j \in J(\tilde{x})\}.$$

Let us introduce the following notation: we collect all vectors $g^i, i \in J(x)$ in a matrix $G(x)$ (with rows given by $(g^i)^\top$). It then holds

$$\boxed{d \text{ is an admissible direction in } x \iff Ad = 0, G(x)d \geq 0}$$

Geometrical visualization: (3 inequality constraints $\langle g^i, x \rangle \geq r^i, i = 1, 2, 3$)



$$\left. \begin{array}{l} \langle g^1, x \rangle = r^1 \\ \langle g^2, x \rangle = r^2 \end{array} \right\} \text{ active} \quad J(x) = \{1, 2\}$$

$$\left. \begin{array}{l} \langle g^3, x \rangle > r^3 \end{array} \right\} \text{ inactive} \quad G(x) = \begin{pmatrix} (g^1)^\top \\ (g^2)^\top \end{pmatrix}$$

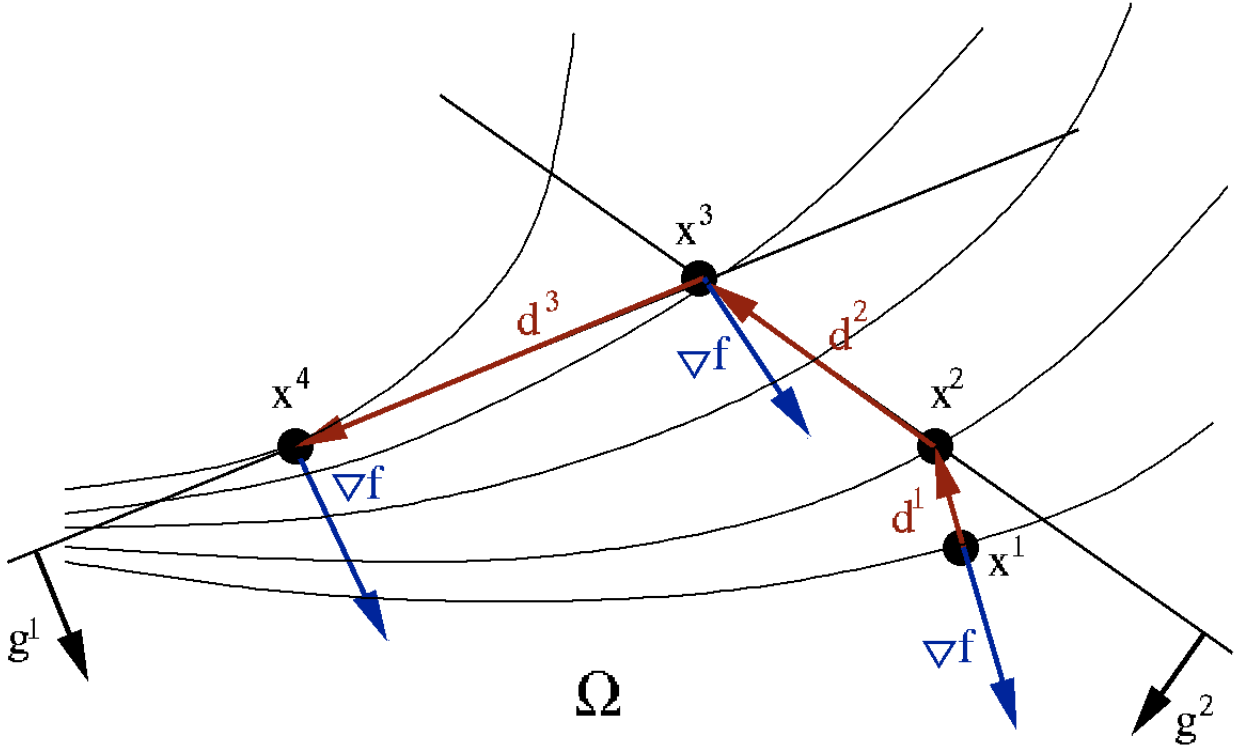
From the figure, we anticipate: for $x + td$ staying admissible for t sufficiently small, we demand

$$\langle g^1, d \rangle \geq 0 \quad \wedge \quad \langle g^2, d \rangle \geq 0.$$

The third constraint – which is inactive – is not relevant here.

Before stating the mathematical details of the so-called active set strategy, let us motivate the essential steps based on some illustrations.

We consider the following situation:



Step 1:

Initial vector $x^{(1)}$ in the interior of Ω . We may ignore both constraints and start with a method for unconstrained optimization until one (or multiple) restrictions become active.

Step 2:

We have reached the boundary of Ω in $x^{(2)}$ – the constraints g^2 has become active.

- If we had $g^2 \parallel \nabla f$, then

$$\nabla f - \mu g^2 = 0.$$

If, in this case, we additionally had $\mu \geq 0$. **Stop.** In our case, this is not true as ∇f is no linear combination of g^2 .

- We continue our search in a subspace characterized by $\langle g^2, d \rangle = 0$, i.e., we look for iterates of the form

$$x^{(2)} + \{d \mid \langle g^2, d \rangle = 0\}$$

Step 3:

In our case, the method reaches a point $x^{(3)}$ in which an additional constraints becomes active, namely $\langle g^1, x \rangle$.

- Apparently, $x^{(3)}$ is not optimal yet since

$$(*) \quad \begin{aligned} \nabla f &= \mu^1 g^1 + \mu^2 g^2 \quad \text{with} \quad \mu^1 > 0 \quad \text{but} \quad \mu^2 < 0 \\ 0 &= \nabla f - \mu^1 g^1 - \mu^2 g^2 \end{aligned}$$

- Which direction should we follow?

For d we require:

$$\begin{array}{c} \left. \begin{array}{l} \langle g^1, d \rangle \geq 0 \\ \langle g^2, d \rangle \geq 0 \end{array} \right\} \text{ for staying admissible} \\ \langle \nabla f, d \rangle < 0 \\ \updownarrow \qquad \qquad \text{for achieving descent .} \\ \langle -\nabla f, d \rangle > 0. \end{array}$$

(*) \Rightarrow

$$\langle \nabla f, d \rangle = \underbrace{\mu^1}_{>0} \underbrace{\langle g^1, d \rangle}_{\geq 0} + \underbrace{\mu^2}_{<0} \underbrace{\langle g^2, d \rangle}_{\geq 0}$$

Since $\langle \nabla f, d \rangle$ should become negative, with large absolute value, our strategy consists in choosing $\langle g^1, d \rangle = 0$ but

$$\langle g^2, d \rangle > 0$$

\Rightarrow We **deactivate constraint 2** and keep constraint 1 active.

Step 4:

Search in the affine space with $\langle g^1, d \rangle = 0$ which, in our case, leads to $x^{(4)}$ which is the solution since here it holds that $\nabla f = \mu^1 g^1$ with $\mu^1 > 0$, and the optimality conditions are satisfied

$$\begin{aligned} 0 &= \nabla f - \mu^1 g^1 - 0 \cdot g^2 \\ \langle g^1, x^{(4)} \rangle &= r^1, \quad \langle g^2, x^{(4)} \rangle > r^2 \\ \left(\langle g^2, x^{(4)} \rangle - r^2 \right) \cdot 0 &= 0. \end{aligned}$$

Conclusion:

- Stop, if the necessary conditions $\mu^i \geq 0$ are satisfied
- Activate constraints which are found/hit throughout the algorithm
- Deactivate constraints if the multipliers become negative (details will follow below).
- Search is performed in linear affine subspaces, i.e., we perform optimization with equality constraints.

Let us formulate the method in a mathematically more precise manner.

- Current iterate: $x^{(k)}$
 $J_k := J(x^{(k)})$: set of active indices (here $\langle g^i, x \rangle = r^i, i \in J_k$)
 $p_k = |J(x^k)|$: number of active indices
 $G_k = G(x^k)$: matrix consisting of $g^i, i \in J_k$ (precisely: with rows g_i^\top)
 $B_k = \begin{pmatrix} A \\ G_k \end{pmatrix}$: parametrizes the currently active linear equality constraints

- Starting from $x^{(k)}$ we set up a *quadratic optimization problem* (with equality constraints)

$$\boxed{\begin{array}{ll} \min_{d \in \mathbb{R}^n} & \frac{1}{2} \langle Qd, d \rangle + \langle Qx^{(k)} + q, d \rangle \\ \text{s.t.} & B_k d = 0 \end{array}} \quad (Q_k)$$

Remark: Up to a constant depending on $x^{(k)}$, the objective function of (Q_k) is $f(x^{(k)} + d)$.
Result:

- direction $d^{(k)}$
- multipliers $\mu_j^{(k)}, j \in J_k; \lambda_i^{(k)}$ (for equality constraints)
- we extend by $\mu_j^{(k)} = 0, j \notin J_k$

$$\begin{aligned} \text{Altogether: } \lambda^{(k)} &\in \mathbb{R}^m \quad (m \text{ equalities}) \\ \mu^{(k)} &\in \mathbb{R}^{p_k} \quad \text{respectively, } \tilde{\mu}^{(k)} \in \mathbb{R}^p \quad (\text{extended by zeros}) \end{aligned}$$

Assumptions for the method:

- B_k is always of maximal rank, i.e., linear independency of $a^i, i = 1, \dots, m; g^j, j \in J_k$
- positive definiteness of Q on $\ker(B_k)$.

The assumptions guarantee unique solvability of (Q_k) and the corresponding multipliers $\lambda^{(k)}, \mu^{(k)}$ are uniquely determined.

Admissible set for (Q_k) :

$$\Omega_k = \{d \in \mathbb{R}^n \mid Ad = 0, G_k d = 0\} \subset L(\Omega, x^{(k)}).$$

Consequently, every $d \in \Omega_k$ is an admissible direction.

The next step is obtained by evaluation of the **necessary optimality conditions** for (Q_k) :

$$\nabla f(x^{(k)} + d^{(k)}) - A^\top \lambda^{(k)} - G_k^\top \mu^{(k)} = 0$$

i.e.,

$$\boxed{Q(d^{(k)} + x^{(k)}) + q - A^\top \lambda^{(k)} - G_k^\top \mu^{(k)} = 0.} \quad (\dagger)$$

From here, we (uniquely) obtain $\lambda^{(k)}, \mu^{(k)}$.

We distinguish the following three cases:

Case 1 $\boxed{d^{(k)} = 0 \quad \text{and} \quad \mu^{(k)} \geq 0.}$

Then $\nabla f(x^{(k)}) - A^\top \lambda^{(k)} - G_k^\top \mu^{(k)} = 0$, and $x^{(k)}$ satisfies the KKT conditions. Due to convexity, these are sufficient for optimality:

$x^{(k)}$ is solution to (QU) : **Stop**

Case 2 $\boxed{d^{(k)} = 0 \quad \text{but} \quad \mu^{(k)} \not\geq 0.}$

Here, there exists *at least one* $j \in J_k$ with $\mu_j^{(k)} < 0$. As in our example, we then deactivate one constraints, ideally, we choose $j \in J_k$ with

$$\mu_j^{(k)} = \min\{\mu_i^{(k)}, i \in J_k\}$$

Deactivation is done by redefinition of the active set:

$$\boxed{\tilde{J}_k := J_k \setminus \{j\} .}$$

We now set up and solve (\tilde{Q}_k) . The method will ensure that the deactivated restrictions are not violated.

Result: $\tilde{d}^{(k)}, \tilde{\mu}^{(k)}, \tilde{\lambda}^{(k)}$

(For Q to remain positive definite on (the unknown space) $\ker(\tilde{B}_k)$, we demand the stronger condition of positive definiteness on the larger subspace $\ker(A)$.)

We will show: $\boxed{\tilde{d}^{(k)} \neq 0} .$

Note: we had $d^{(k)} = 0$

$$\Rightarrow Qx^{(k)} + q - A^\top \lambda^{(k)} - G_k^\top \mu^{(k)} = 0$$

$$Q\tilde{d}^{(k)} + Qx^{(k)} + q - A^\top \tilde{\lambda}^{(k)} - \tilde{G}_k^\top \tilde{\mu}^{(k)} = 0$$

If $\tilde{d}^{(k)} = 0$ then

$$A^\top \lambda^{(k)} + G_k^\top \mu^{(k)} = A^\top \tilde{\lambda}^{(k)} + \tilde{G}_k^\top \tilde{\mu}^{(k)}$$

$$\Rightarrow \mu_j^{(k)} g^j = \text{linear combination of } d^i, g^i$$

$$\Rightarrow \text{since } \mu_j^{(k)} \neq 0 \quad \text{the same holds true for } g^j \text{ which contradicts } B_k \text{ being maximal rank.}$$

Still left to show:

Case 3 $\boxed{d^{(k)} \neq 0.}$ Here, $d^{(k)}$ is a **descent direction**.

Proof: The necessary conditions for (\dagger) and $x^{(k)}$ imply

$$\begin{aligned} \nabla f(x^{(k)}) &= Qx^{(k)} + q = -Qd^{(k)} + A^\top \lambda^{(k)} + G_k^\top \mu^{(k)} \\ &= -Qd^{(k)} + B_k^\top \begin{pmatrix} \lambda^{(k)} \\ \mu^{(k)} \end{pmatrix} \end{aligned}$$

Taking the inner product with $d^{(k)}$ yields

$$\begin{aligned} \Rightarrow \langle \nabla f(x^{(k)}), d^{(k)} \rangle &= -\underbrace{\langle d^{(k)}, Qd^{(k)} \rangle}_{<0 \text{ due to definiteness}} + \left\langle \begin{pmatrix} \lambda^{(k)} \\ \mu^{(k)} \end{pmatrix}, \underbrace{B_k d^{(k)}}_{=0} \right\rangle \\ &< 0 . \end{aligned}$$

Remark. In case 2, we still have to show that the direction $\tilde{d}^{(k)}$ is admissible, i.e., also for the deactivated restriction j it holds that

$$\langle g^j, \tilde{d}^{(k)} \rangle \geq 0.$$

Indeed, this is true since with $d^{(k)} = 0$ and (\dagger) we find:

$$0 = \nabla f(x^{(k)}) - \sum_{i=1}^m a^i \lambda_i^{(k)} - \sum_{\substack{i \neq j \\ i \in J_k}} g^i \mu_i^{(k)} - \mu_j^{(k)} g^j.$$

We take the inner product with $\tilde{d}^{(k)}$. From case 3, we know that $\tilde{d}^{(k)}$ is a descent direction. Moreover, we demanded $A\tilde{d}^{(k)} = 0$ and $\langle g^i, \tilde{d}^{(k)} \rangle = 0$, $i \in \tilde{J}_k$. Thus

$$\underbrace{\langle \nabla f(x^{(k)}), \tilde{d}^{(k)} \rangle}_{\substack{<0 \\ \text{(descent)}}} - \underbrace{\mu_j^{(k)}}_{\substack{<0 \\ \text{(Case 2)}}} \langle g^j, \tilde{d}^{(k)} \rangle = 0$$

$$\Rightarrow \boxed{\langle g^j, \tilde{d}^{(k)} \rangle > 0.}$$

This confirms our intuition from the geometric example: with regard to the j th restriction, $\tilde{d}^{(k)}$ is pointing into the interior of Ω .

Altogether we obtain the following

Algorithm 6.1.1 (Active set strategy for (QU))

1. Compute initial vector $x^{(0)} \in \Omega$, set $k := 0$.
2. Set up (Q_k) and determine the direction $d^{(k)}$, multipliers $\lambda^{(k)}$, $\mu^{(k)}$.
3. If $d^{(k)} = 0$ and $\mu^{(k)} \geq 0$: **Stop**; $x^{(k)}$ is the solution.
4. If $d^{(k)} = 0$ and $\mu^{(k)} \not\geq 0$, perform a deactivation step:
 - Determine $\mu_j^{(k)} = \min\{\mu_i^{(k)}, i \in J_k\}$
 - $J_k := J_k \setminus \{j\}$
 - delete the associated row j in G_k
 - solve the modified problem (Q_k) . The result always satisfies $d^{(k)} := \tilde{d}^{(k)} \neq 0$
5. Now it holds $d^{(k)} \neq 0$. Compute a step length σ_k (details below) and set

$$x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}.$$

It remains to discuss the choice of the step length σ_k .

Computation of the step size

We assume that $d^{(k)}$ is an admissible direction. The new iterate is given by

$$x^{(k+1)} = x^{(k)} + \tau d^{(k)}$$

with a specific $\tau > 0$. How do we choose τ ?

- **Maximal descent**

$$\begin{aligned}
f(x^{(k)} + td^{(k)}) &= f(x^{(k)}) + \underbrace{t \nabla f(x^{(k)}) d^{(k)}}_{= -t \langle d^{(k)}, Qd^{(k)} \rangle} + \frac{1}{2} t^2 \langle d^{(k)}, Qd^{(k)} \rangle \\
&= -t \langle d^{(k)}, Qd^{(k)} \rangle \quad \text{cf. final result of case 3.} \\
&= f(x^{(k)}) + \underbrace{\left(\frac{1}{2} t^2 - t \right) \langle d^{(k)}, Qd^{(k)} \rangle}_{\text{minimal for } t = 1}
\end{aligned}$$

\Rightarrow For maximal descent, we should choose $\tau = 1$.

- **Admissibility of $x^{(k+1)}$**

For the last iterate $x^{(k)}$, we had

$$\begin{aligned}
\langle a^i, x^{(k)} \rangle &= b_i \quad i = 1, \dots, m \\
\langle g^i, x^{(k)} \rangle &= r_i \quad i \in J_k \quad (\text{active inequalities.}) \\
\langle g^i, x^{(k)} \rangle &> r_i \quad i \notin J_k \quad (\text{inactive inequalities.})
\end{aligned}$$

The choice of $d^{(k)}$ ensures for all $t \geq 0$

$$\begin{aligned}
\langle a^i, x^{(k)} + td^{(k)} \rangle &= b_i \quad \text{since } \langle a^i, d^{(k)} \rangle = 0 \\
\langle g^i, x^{(k)} + td^{(k)} \rangle &= r_i \quad \forall i \in J_k \setminus \{j\}, \quad \text{analogously} \\
\langle g^j, x^{(k)} + td^{(k)} \rangle &> r_j \quad \text{since for the deactivated restriction we have } \langle g^j, d^{(k)} \rangle > 0.
\end{aligned}$$

Hence, we only have to take care of the inactive constraints! Apparently, we only have a bound for t , if there exists at least one $j \notin J_k$ with

$$\langle g^j, d^{(k)} \rangle < 0.$$

Here, we require

$$\langle g^j, d^{(k)} \rangle + t \langle g^j, d^{(k)} \rangle \geq r_j, \quad \text{Maximum of } t : \text{ for equality}$$

for this specific j at most

$$t = \frac{r_j - \langle g^j, x^{(k)} \rangle}{\langle g^j, d^{(k)} \rangle}.$$

\Rightarrow Maximal admissible step size:

$$\begin{aligned}
\tau_k &= \min_{j \in I_k} \left\{ \frac{r_j - \langle g^j, x^{(k)} \rangle}{\langle g^j, d^{(k)} \rangle} \right\} \quad \text{if } I_k \neq \emptyset \\
&\quad \text{where } I_k = \{j / \langle g^j, d^{(k)} \rangle < 0\}. \\
\tau_k &:= \infty \quad \text{if } I_k = \emptyset \quad (\text{no restriction necessary}).
\end{aligned}$$

Insgesamt: $\boxed{\sigma_k = \min(1, \tau_k)}$.

This describes the entire method. It holds

Theorem 6.1.1 *Let Q be positive definite on $\ker(A)$ and for all $x \in \Omega$ assume that the matrix $B(x) = \begin{pmatrix} A \\ G(x) \end{pmatrix}$ has maximal rank. Then the method computes the solution of (QLI) in finitely many steps.*

Proof: We already discussed the well-posedness of the method.

The method consists in solving quadratic optimization problems of the form

$$\min f(x) \quad \text{s.t.} \quad Ax = b, \quad \langle g^j, x \rangle = r_i \quad \forall i \in J \quad (1)$$

where J is an arbitrary subset of $\{1, \dots, p\}$ characterizing possible active inequality constraints. We do not exactly solve (1) during the method but obtain optimality systems for (1), such as

$$Qx + q - A^\top \lambda - G(x)^\top \mu = 0, \quad (2)$$

where x denotes the unique solution of (1). There are only finitely many subsets of J , thus only finitely many different problems (1), as well as finitely many such solutions x which only allow for setting up finitely many systems (2).

Let us now assume that the method has not finished in step k , i.e., we follow **step 5** with $d^{(k)} \neq 0$. Then:

(i) $I_k = \emptyset$. Here, $\sigma_k = 1$, hence

$$x^{(k+1)} = x^{(k)} + d^{(k)}$$

and from (†) we know

$$Q(\underbrace{x^{(k)} + d^{(k)}}_{x^{(k+1)}}) + q - A^\top \lambda^{(k)} - G_k^\top \mu^{(k)} = 0,$$

such that $x^{(k+1)} =: x$ satisfies (2) ($\lambda^{(k)}$ and $\mu^{(k)}$ are uniquely determined). This does not imply that $x^{(k+1)}$ is optimal since we may obtain $\mu^{(k)} \not\geq 0$. Thus $x^{(k+1)}$ is one of the finitely many solutions of (1).

(ii) $I_k \neq \emptyset$. Here, we distinguish two “subcases”:

- $\tau_k \geq 1$: Then $\sigma_k = \min(1, \tau_k) = 1$. As before – i.e., $x^{(k+1)}$ is one of the solutions of (1)
- $\tau_k < 1$: All active constraints remain active, but at least one additional restriction is added such that the cardinality of the active constraints increases.
We assumed: The set $\{a^i\}_{i=1, \dots, m} \cup \{g^j\}$, $j \in J_k$ is always linearly independent. There can only occur $n - m$ such increases (at the beginning, we had at least m linearly independent vectors and every time a new one is added).
 \Rightarrow after at most $n - m$ iterations we have $I_{k+i} = \emptyset$ or $\tau_{k+i} \geq 1 \Rightarrow$ new solution of (2).

Moreover, $d^{(k)}$ is *descent direction* and we certainly find

$$f(x^{(k+j)}) < f(x^{(k)}).$$

Thus, the occurring $x^{(k+j)}$ all are different and by finiteness of the possibilities for (1) and (2), the method has to stop after finitely many steps.

□

6.2 Equality constraints, nonlinear objective

We follow the idea for a quadratic objective function: we “eliminate” the equality constraints by help of a nullspace matrix. We consider

$$\boxed{\begin{array}{l} \min f(x) \\ Ax = b \end{array}} \quad (\text{PLG})$$

with $f: \mathbb{R}^n \rightarrow \mathbb{R}$, not necessarily quadratic. Hence $\min_{x \in \Omega} f(x)$ with $\Omega = \{x \in \mathbb{R}^n / Ax = b\}$. Again, denote by $w \in \Omega$ a specific solution of $Ax = b$ and $Z: \mathbb{R}^l \rightarrow \ker(A)$ a nullspace matrix. Then, we solve the unconstrained problem

$$\boxed{\min_{z \in \mathbb{R}^l} F(z) := f(w + Zz)} \quad (6.1)$$

The computation of a nullspace matrix is independent of f and only depends on Ω . Hence, we can follow the same steps as before (QR decomposition, etc.).

The free optimization problem (6.1) can now be solved (assuming smoothness of f) with any method for unconstrained optimization. We could stop here but there are some more interesting subtleties! In particular, let us discuss the analogy of minimizing f and F .

Consider a standard descent method.

$$\begin{aligned} \text{For } f: x^{(k+1)} &= x^{(k)} + \sigma_k d^{(k)} \\ F: z^{(k+1)} &= z^{(k)} + \sigma_k v^{(k)} \end{aligned}$$

If $v^{(k)}$ is a descent direction for F in $z^{(k)}$, then for $d^{(k)} := Zv^{(k)}$ we obtain

$$\begin{aligned} \nabla f(x^{(k)})^\top d^{(k)} &= \nabla f(x^{(k)})^\top Zv^{(k)} \\ &= (Z^\top \nabla f(x^{(k)}))^\top v^{(k)} \\ &= \nabla F(z^{(k)})^\top v^{(k)} < 0, \end{aligned}$$

hence, $d^{(k)}$ is a descent direction for f . Moreover

$$x^{(k+1)} = w + Zz^{(k+1)} = \underbrace{w + Zz^{(k)}}_{x^{(k)}} + \underbrace{\sigma_k Zv^{(k)}}_{d^{(k)}} = x^{(k)} + \sigma_k d^{(k)}.$$

Consequence: We can proceed within the x variables and do not have to use the z variables.

Algorithm 6.2.1 (Reduced descent method)

1. Compute $x^{(0)} \in \Omega$, nullspace matrix Z , $k := 0$.
 2. If $\underbrace{Z^\top \nabla f(x^{(k)})}_{\text{reduced gradient}} = 0$: **Stop**.
 3. Otherwise, compute descent direction $\boxed{d^{(k)} := Zv^{(k)}}$, efficient step size σ_k and
$$x^{(k+1)} := x^{(k)} + \sigma_k d^{(k)}$$
- $k := k + 1$, goto 2.

We need $Z, v^{(k)}$ and the correspondences

$$\begin{array}{ccccccc} F(z^{(k)}), & F(z^{(k)} + \sigma_k v^{(k)}), & \nabla F(z^{(k)}), & \nabla F(z^{(k)} + \sigma_k v^{(k)}), \\ \Downarrow & \Downarrow & \Downarrow & \Downarrow \\ f(x^{(k)}) & f(x^{(k)} + \sigma_k d^{(k)}) & Z^\top \nabla f(x^{(k)}) & Z^\top \nabla f(x^{(k)} + \sigma_k d^{(k)}) \end{array}$$

It looks like the computation of $d^{(k)} = Zv^{(k)}$ requires $v^{(k)}$ and we could not work in the x variables.

For concrete methods this looks differently!

Example 6.2.1

Reduced gradient method:

$$\begin{aligned} v^{(k)} &:= -\nabla F(z^{(k)}) = -Z^\top \nabla f(x^{(k)}) \\ \Rightarrow \quad &\boxed{d^{(k)} = Zv^{(k)} = -ZZ^\top \nabla f(x^{(k)})} \end{aligned}$$

Special case: $Z = P$, projection matrix onto $\ker(A)$, \Rightarrow

Projected gradient method: $Z = P$

$$\begin{aligned} d^{(k)} &= -ZZ^\top \nabla f(x^{(k)}) = -ZZ \nabla f(x^{(k)}) \\ &\boxed{d^{(k)} = -Z \nabla f(x^{(k)})} \end{aligned}$$

Variable metric method (reduced):

Sequence $\{A^{(k)}\}$ of positive definite matrices;

$$\begin{aligned} v^{(k)} &= -(A^{(k)})^{-1} \nabla F(z^{(k)}) = -(A^{(k)})^{-1} Z^\top \nabla f(x^{(k)}) \\ \Rightarrow \quad &\boxed{d^{(k)} = Zv^{(k)} = -Z(A^{(k)})^{-1} Z^\top \nabla f(x^{(k)})} \end{aligned}$$

Special case: reduced Newton method:

$$\begin{aligned} A^{(k)} &:= F''(z^{(k)}) = \underbrace{Z^\top f''(x^{(k)}) Z}_{\text{reduced Hessian}} \\ \Rightarrow \quad &\boxed{d^{(k)} = -Z(Z^\top f''(x^{(k)}) Z)^{-1} Z^\top \nabla f(x^{(k)})}. \end{aligned}$$

Analog considerations can be carried out for the reduced BFGS method.

A nice application of nonlinear optimization with linear equality constraints:

Nonlinear regression with cubic splines

Measurements $(\xi_i, \eta_i), i = 1, \dots, m$

Ansatz $\eta(\xi) = g(x, \xi)$

Goal find x but first: appropriate ansatz for g .
Idea: splines with coefficients x .

Let $\xi_1 < \xi_2 < \dots < \xi_m$.

We cover the interval $[\xi_1, \xi_m]$ by *nodes*

$$\begin{aligned}\tau_0 &< \tau_1 < \dots < \tau_N, \\ \tau_0 &\leq \xi_1, \xi_m \leq \tau_N.\end{aligned}$$

Reuirements:

- On $[\tau_i, \tau_{i+1}]$, $g =: g_i(x, \xi)$ is a cubic polynomial in ξ
 - $g(x, \cdot) \in C^2[\tau_0, \tau_N]$
- $\Rightarrow g, g', g''$ have to be continuous in the nodes.

We define $[\tau_i, \tau_{i+1}]$

$$\begin{aligned}g_i(x, \tau) &= \frac{1}{\tau_{i+1} - \tau_i} (\gamma_{i+1}(\tau - \tau_i)^3 + \gamma_i(\tau_{i+1} - \tau)^3) + \beta_i(\tau - \tau_i) + \alpha_i \\ \gamma_0 &= \gamma_N = 0 \\ \gamma_i &= g_i''(x, \tau_i)/6 \quad \text{“moments”} \\ i &= 1, \dots, N-1.\end{aligned}$$

For this choice, g'' is automatically continuous but this does not mean that g and g' are continuous as well. This requirement yields additional constraints for

$$x = (\alpha_0, \dots, \alpha_{N-1}, \beta_0, \dots, \beta_{N-1}, \dots, \gamma_1, \dots, \gamma_{N-1}).$$

In particular: continuity of g' : $\Delta\tau_i := \tau_{i+1} - \tau_i$

$$\begin{aligned}g_i'(x, \tau_i) &= g_{i-1}'(x, \tau_i) \\ 3\gamma_i\Delta\tau_i + \beta_i &= \beta_{i-1} + 3\gamma_i\Delta\tau_{i-1} \\ \Rightarrow \beta_i &= \beta_{i-1} + 3\gamma_i(\Delta\tau_{i-1} - \Delta\tau_i)\end{aligned}$$

\Rightarrow In fact, only β_0 is free

Continuity of g :

$$\begin{aligned}g_i(x, \tau_i) &= g_{i-1}(x, \tau_i) \\ \gamma_i(\Delta\tau_i)^2 + \alpha_i &= \gamma_i(\Delta\tau_{i-1})^2 + \alpha_{i-1} + \beta_{i-1}\Delta\tau_{i-1} \\ \Rightarrow \alpha_i &= \alpha_{i-1} + \gamma_i((\Delta\tau_{i-1})^2 - (\Delta\tau_i)^2) + \beta_{i-1}\Delta\tau_{i-1}.\end{aligned}$$

Again, only α_0 is free. Altogether we obtain

$$\begin{aligned}\min f(x) &= \sum_{i=1}^m (\eta_i - g(x, \xi_i))^2 \\ \text{s.t. } \beta_i &= \beta_{i-1} + 3\gamma_i(\Delta\tau_i - \Delta\tau_{i-1}) \\ \alpha_i &= \alpha_{i-1} + \beta_{i-1}\Delta\tau_{i-1} + \gamma_i(\Delta\tau_{i-1}^2 - \Delta\tau_i^2) \\ i &= 1, \dots, N-1.\end{aligned}$$

This problem can be reduced to a *free optimization problem* in the variables:

$$(\alpha_0, \beta_0, \gamma_1, \dots, \gamma_{N-1})^T.$$

6.3 Inequality constraints, nonlinear objective

Let us consider

$$\begin{array}{l} \min f(x) \\ Ax = b, \quad Gx \geq r \end{array}$$

Basic idea: Taylor approximation of f up to second order \leadsto quadratic objective function, solution by the previously discussed methods, followed by a new approximation of f : *SQP method*.

We start with a short excursion on the *idea of the SQP method*:

We consider two optimization problems in parallel

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

Ω : convex set.

The first order necessary conditions for x^* read:

$$\nabla f(x^*) = 0 \qquad \langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in \Omega.$$

The equation on the left is a nonlinear system of equations which we can solve by, e.g., the Newton method. On the right hand side, we are faced with a variational inequality which we have not discussed how to solve. Let us thus only write down the Newton method for the problem on the left:

$$(*) \qquad \nabla f(x^{(k)}) + f''(x^{(k)})(x - x^{(k)}) = 0.$$

The solution is $x = x^{(k+1)}$. For the variational inequality, we do not have an appropriate analogy yet. Note that $(*)$ is simply the first order necessary condition for the optimization problem

$$(**) \qquad \min_{x \in \mathbb{R}^n} \frac{1}{2} \langle x - x^{(k)}, f''(x^{(k)})(x - x^{(k)}) \rangle + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle.$$

Hence, it does not matter if we solve the problem $(**)$ or the equation $(*)$. However, while $(*)$ does not have an obvious analogy for constrained optimization, we can simply generalize $(**)$: we simply additionally incorporate the constraint $x \in \Omega$:

$$\min_{x \in \Omega} \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2} \langle x - x^{(k)}, f''(x^{(k)})(x - x^{(k)}) \rangle.$$

Solution: $x = x^{(k+1)}$.

This is exactly how we proceed with the initial problem. Here, it holds that

$$\Omega = \{x \in \mathbb{R}^n \mid Ax = b, \quad Gx \geq r\}$$

is convex. We iterate as follows:

$x^{(k)}$ is given. We then set up:

$$\begin{array}{ll} \min & \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2} \langle x - x^{(k)}, f''(x^{(k)})(x - x^{(k)}) \rangle \\ \text{s.t.} & Ax = b, \quad Gx \geq r \end{array}$$

(QP_k)

Solution: $x^{(k+1)}$.

Then $k := k + 1$, and we continue with a new iterate.

For convergence, as for the Newton method for $(*)$, we have to impose additional assumptions: we require invertibility of $f''(x^{(k)})$ which can be ensured by $f \in C^2$ and $f''(x^*)$ regular. Since we are searching for a minimum, we require $f''(x^*)$ to be positive semidefinite which, by regularity, implies positive definiteness of $f''(x^*)$! This is helpful for (QP_K) since: if $\det(f''(x^*)) > 0$, then $\det(f''(x^{(k)}))$ for $x^{(k)}$ close to x^* , and in this case (QP_k) has exactly one solution!

For the numerical realization, we rather use the formulation

$$d = x - x^{(k)} \Leftrightarrow x = x^{(k)} + d.$$

From $Ax^{(k)} = b$ it follows $Ad = 0$ and thus

$$(QP_k) \Leftrightarrow \boxed{\begin{array}{l} \min \langle \nabla f(x^{(k)}), d \rangle + \frac{1}{2} \langle d, f''(x^{(k)})d \rangle \\ \text{s.t. } Ad = 0, Gx^{(k)} + Gd \geq r. \end{array}}$$

The problem serves as a computational tool for obtaining a search direction d . We may now use a “full” step $x^{(k+1)} = x^{(k)} + d^{(k)}$ (Newton method) or use a step size control.

Remark: Instead of $f''(x^{(k)})$, as for the variable metric methods, one may use appropriate approximations $A^{(k)}$, see, e.g., [1]. We stick with f'' .

Assumptions for the well-posedness of the algorithm

- $f''(x^{(k)})$ is positive definite on $\ker(A)$
 \Rightarrow existence of exactly one solution (QP_k)
- $B(x) = \begin{pmatrix} A \\ G(x) \end{pmatrix}$ always has full rank
 \Rightarrow multipliers λ, μ are unique.

Optimality conditions for (QP_k) :

$$f''(x^{(k)})d^{(k)} + \nabla f(x^{(k)}) - A^\top \lambda^{(k+1)} - G^\top \mu^{(k+1)} = 0 \quad (6.2)$$

$$\mu^{(k+1)} \geq 0, \langle \mu^{(k+1)}, Gx^{(k)} + Gd^{(k)} - r \rangle = 0. \quad (6.3)$$

when solving (QP_k) we distinguish the following two cases:

Case 1 $d^{(k)} = 0$.

No change, we expect $x^{(k)}$ to be optimal. Indeed, (6.2–6.3) imply

$$\begin{aligned} \nabla f(x^{(k)}) - A^\top \lambda^{(k+1)} - G^\top \mu^{(k+1)} &= 0, \mu^{(k+1)} \geq 0 \\ \langle \mu^{(k+1)}, Gx^{(k)} - r \rangle &= 0. \end{aligned}$$

$\Rightarrow x^{(k)}$ satisfies the optimality conditions \Rightarrow STOP
 (optimality follows by sufficient conditions).

Fall 2 $d^{(k)} \neq 0$.

Then $d^{(k)}$ is a descent direction since

$$\begin{aligned}\nabla f(x^{(k)}) &= -f''(x^{(k)})d^{(k)} + A^\top \lambda^{(k+1)} + G^\top \mu^{(k+1)} \mid \cdot d^{(k)} \\ \langle \nabla f, d^{(k)} \rangle &= -\underbrace{\langle d^{(k)}, f'' d^{(k)} \rangle}_{>0} + \underbrace{\langle A d^{(k)}, \lambda^{(k+1)} \rangle}_{=0} + \underbrace{\langle G d^{(k)}, \mu^{(k+1)} \rangle}_{\leq 0} \\ &< 0 \quad \Rightarrow \quad \text{descent direction} \quad \text{see below.}\end{aligned}$$

With regard to $\langle G d^{(k)}, \mu^{(k+1)} \rangle$: the constraints read

$$\langle g^i, x^{(k)} \rangle + \langle g^i, d^{(k)} \rangle \geq r_i.$$

For the inactive indices $i \notin J(d^{(k)})$ we have $\mu_i^{(k+1)} = 0$. For the active indices we have $\mu_i^{(k+1)} \geq 0$ as well as

$$\begin{aligned}\langle g^i, d^{(k)} \rangle &= \underbrace{r_i - \langle g^i, x^{(k)} \rangle}_{\leq 0, \text{ since } x^{(k)} \text{ was admissible}},\end{aligned}$$

Thus $\langle g^i, d^{(k)} \rangle \leq 0$. Altogether this implies $\langle G d^{(k)}, \mu^{(k+1)} \rangle \leq 0$.

Computation of step sizes

Since $d^{(k)}$ is admissible, we can at least chose $x^{(k+1)} = x^{(k)} + 1 \cdot d^{(k)}$. The maximal step size thus satisfies $\tau_k \geq 1$. Typically: $\sigma_k = 1$ (classical SQP) or step size control.

As we have seen, for the choice $\sigma_k \equiv 1$, the SQP method is a generalization of the Newton method and often called that way. Under classical assumptions, it inherits its local quadratic convergence.

Assumptions: (Let \tilde{x} be a local minimum).

(i) $f \in C^2$ in a ball $B(\tilde{x}, \delta)$ around \tilde{x}

(ii) f'' is Lipschitz on $B(\tilde{x}, \delta)$, i.e.,

$$\|f''(x) - f''(y)\| \leq L\|x - y\| \quad \forall x, y \in B(\tilde{x}, \delta)$$

(iii) $B(\tilde{x})$ has maximal rank

(iv) Positive definiteness:

$$d^\top f''(\tilde{x})d \geq \alpha\|d\|^2 \quad \forall d : Ad = 0, G(\tilde{x})d = 0$$

(Second order sufficient condition)

(v) If $\langle g^i, \tilde{x} \rangle = r_i$ then $\tilde{\mu}_i > 0$ for the corresponding multiplier (condition of *strict complementarity*)

Remark: we always have $\tilde{\mu}_i > 0 \Rightarrow \langle g^i, \tilde{x} \rangle = r_i$. Strict complementarity implies the reverse direction.

Satz 6.3.1 Under assumptions (i)-(v), the SQP method converges locally quadratic, i.e.,

$$\begin{aligned}\|x^{(k+1)} - \tilde{x}\| + \|\lambda^{(k+1)} - \tilde{\lambda}\| + \|\mu^{(k+1)} - \tilde{\mu}\| \\ \leq c(\|x^{(k)} - \tilde{x}\|^2 + \|\lambda^{(k)} - \tilde{\lambda}\|^2 + \|\mu^{(k)} - \tilde{\mu}\|^2).\end{aligned}$$

7 Problems with nonlinear constraints - methods

7.1 The Lagrange-Newton method

We consider the general problem

$$\begin{aligned} \min f(x) \\ h(x) = 0 \\ g(x) \geq 0. \end{aligned} \tag{PNI}$$

Here, we generally assume:

- $f, g, h \in C^2$
- \tilde{x} is a *regular* local solution
- $\tilde{\lambda}, \tilde{\mu}$ are corresponding Lagrange multipliers.

We obtain $\tilde{x}, \tilde{\lambda}, \tilde{\mu}$ by the Karush-Kuhn-Tucker conditions which require:

$$\begin{aligned} \nabla f(\tilde{x}) - h'(\tilde{x})^\top \tilde{\lambda} - g'(\tilde{x})^\top \tilde{\mu} &= 0 \\ h(\tilde{x}) &= 0 \\ \tilde{g}(\tilde{x}) &= 0. \end{aligned}$$

Here, in \tilde{g} we collect the active inequalities, i.e.,

$$\tilde{g}(x) = (g_j(x))_{j \in J(\tilde{x})}.$$

The inactive constraints can locally be neglected. By solving this system, we want to obtain $\tilde{x}, \tilde{\lambda}, \tilde{\mu}$. For notational purposes, we define

$$z = \begin{pmatrix} x \\ \lambda \\ \nu \end{pmatrix}, \quad F(z) = \begin{pmatrix} \nabla f(x) - h'(x)^\top \lambda - \tilde{g}'(x)^\top \nu \\ h(x) \\ \tilde{g}(x) \end{pmatrix}$$

with $\nu = (\mu_j)_{j \in J(x)}$. $\tilde{\nu} := (\tilde{\mu}_j)_{j \in J(\tilde{x})}$. For $\tilde{z}^\top = (\tilde{x}^\top, \tilde{\lambda}^\top, \tilde{\nu}^\top)$ this yields

$$F(\tilde{z}) = 0.$$

It seems natural to apply Newton's method to approximately solve for \tilde{z} . However, this requires a priori knowledge of the indices in $J(\tilde{x})$, hence we have to know the active inequality constraints! For the convergence of the method, we need

- (8.1.2) $f, g, h \in C^{2,1}$ and $F'(\tilde{z})$ is nonsingular. The matrix $F'(z)$ is of the form

$$F'(z) = \begin{pmatrix} \mathcal{L}_{xx}(x, \lambda, \nu) & h'(x)^\top & \tilde{g}'(x)^\top \\ h'(x) & 0 & 0 \\ \tilde{g}'(x) & 0 & 0 \end{pmatrix}.$$

This matrix is of type

$$\mathcal{A} = \begin{pmatrix} Q & -A^\top \\ A & 0 \end{pmatrix},$$

for which it holds: If Q is positive definite on $\ker(A)$ and A has maximal rank, then \mathcal{A} is invertible. In our case

$$Q = \mathcal{L}_{xx}, \quad A = \begin{pmatrix} h'(x) \\ g'(x) \end{pmatrix},$$

such that $F'(\tilde{z})$ is nonsingular if

- (8.1.3) (LICQ) is satisfied, i.e., the gradients $\nabla h_i(\tilde{x})$, $\nabla g_j(\tilde{x})$ of the active constraints are linearly independent,
- (8.1.5) and a second order sufficient condition is satisfied:

$$d^\top \mathcal{L}_{xx}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) d \geq \alpha \|d\|^2$$

for all d with $h'(\tilde{x})d = 0$ and $\tilde{g}'(\tilde{x})d = 0$.

We further require:

- (8.1.4) *strict complementarity*: $g_j(\tilde{x}) = 0 \Rightarrow \mu_j > 0$.

These assumptions ensure local quadratic convergence of the Newton method:

Based on the initial vector $z^{(0)} = (x^{(0)}, \lambda^{(0)}, \nu^{(0)})$ compute

$$z^{(k+1)} = z^{(k)} - F'(z^{(k)})^{-1} F(z^{(k)}),$$

i.e., we have to solve the linear system

$$F'(z^{(k)})(z - z^{(k)}) = -F(z^{(k)})$$

for $z^{(k+1)}$. This leads to:

$$\begin{pmatrix} \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}, \nu^{(k)}) & -h'(x^{(k)})^\top & -\tilde{g}'(x^{(k)})^\top \\ h'(x^{(k)}) & 0 & 0 \\ \tilde{g}'(x^{(k)}) & 0 & 0 \end{pmatrix} \begin{pmatrix} x - x^{(k)} \\ \lambda - \lambda^{(k)} \\ \nu - \nu^{(k)} \end{pmatrix} \\ = - \begin{pmatrix} \nabla f(x^{(k)}) - h'(x^{(k)})^\top \lambda^{(k)} - \tilde{g}'(x^{(k)})^\top \nu^{(k)} \\ h(x^{(k)}) \\ \tilde{g}(x^{(k)}) \end{pmatrix}.$$

Some terms cancels and we arrive at

$$\begin{aligned} \nabla f(x^{(k)}) + \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}, \nu^{(k)})(x - x^{(k)}) - h'(x^{(k)})^\top \lambda - \tilde{g}'(x^{(k)})^\top \nu &= 0 \\ h(x^{(k)}) + h'(x^{(k)})(x - x^{(k)}) &= 0 \\ \tilde{g}(x^{(k)}) + \tilde{g}'(x^{(k)})(x - x^{(k)}) &= 0. \end{aligned} \tag{7.4}$$

These (almost) are the necessary optimality conditions for the linear quadratic problem

$$\begin{aligned} \min_x \quad & \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)})^\top \mathcal{L}_{xx}(\dots)(x - x^{(k)}) \\ \text{s.t.} \quad & h(x^{(k)}) + h'(x^{(k)})(x - x^{(k)}) = 0 \\ & g(x^{(k)}) + g'(x^{(k)})(x - x^{(k)}) \geq 0, \end{aligned} \tag{Q1}_k$$

if we can show that the inactive constraints $j \notin J(\tilde{x})$ are also inactive for this problem, i.e., do not become relevant. To some extent, the Newton method (7.4) and the SQP method $(Q1)_k$ are equivalent.

Again, let us introduce the direction $d = x - x^{(k)}$ and consider:

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & \frac{1}{2} \langle d, \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}, \mu^{(k)})d \rangle + \langle \nabla f(x^{(k)}), d \rangle \\ \text{s.t.} \quad & h(x^{(k)}) + h'(x^{(k)})d = 0 \\ & g(x^{(k)}) + g'(x^{(k)})d \geq 0 \end{aligned} \quad (Q2)_k$$

with solution $d^{(k)} = x^{(k+1)} - x^{(k)}$ with new multipliers $\mu^{(k+1)}, \lambda^{(k+1)}$. We obtain

Algorithm 7.1.1 Lagrange-Newton method for (PNI)

1. $k := 0, z^{(0)} = (x^{(0)}, \lambda^{(0)}, \mu^{(0)})$
2. Solve $(Q2)_k \rightarrow d^{(k)}; \lambda^{(k+1)}, \mu^{(k+1)}$
3. STOP if $d^{(k)} = 0$
4. $x^{(k+1)} = x^{(k)} + d^{(k)}, \text{ goto } 2.$

Remarks:

1. The method converges locally quadratic! This makes sense as (under assumptions) it is equivalent to the Newton method (if the active constraints are known and strict complementarity is given). Strict complementarity: active constraints remain active in a specific neighborhood since the multipliers remain positive).
2. Due to the nonlinearity of the constraints, the iterates are generally not admissible, i.e., $x^{(k)} \notin \Omega$. Moreover, we only have local convergence which motivates modifications of the method!

7.2 Sequential quadratic programming

As mentioned before, the classical Lagrange-Newton method is only locally convergent. Moreover, the computation of \mathcal{L}_{xx} may be too expensive which motivates modifications of the method:

• **Approximations $A^{(k)}$ of $\mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}, \mu^{(k)})$**

Similar to variable metric methods, one uses symmetric positive definite matrices $A^{(k)}$ and computes directions by solving:

$$\begin{aligned} \min_d \quad & \langle \nabla f(x^{(k)}), d \rangle + \frac{1}{2} \langle d, A^{(k)}d \rangle \\ \text{s.t.} \quad & h(x^{(k)}) + h'(x^{(k)})d = 0 \\ & g(x^{(k)}) + g'(x^{(k)})d \geq 0. \end{aligned} \quad (QP)_k$$

Here, we assume that the admissible set Ω_k is nonempty. A sufficient condition is given in [1, Satz 8.3.1]: h_i affine linear, linear independent gradients, Slater-type conditions.

Under standard assumptions, one can then show the existence of exactly one solution of $(QP)_k$ as well as the uniform boundedness of the sequences $d^{(k)}, \alpha^{(k)}, \mu^{(k)}$ [1, Satz 8.3.2].

• **step size control**

We do not take a full Newton step but set

$$x^{(k+1)} = x^{(k)} + \sigma d^{(k)} .$$

It should be noted:

- a) The generated $x^{(k)}$ are generally *not* admissible.
- b) $d^{(k)}$ is in general no descent direction. (Theoretically, it could happen that $x^{(k)}$ provides an optimal value because it is inadmissible. This may cause increase of admissible step lengths.)

We try to avoid this by introducing so-called *merit functions*

Definition 7.2.1 ϕ is called merit function if:

- If $\tilde{x} \in \Omega$ is a local solution of (PNE), then \tilde{x} is a local (free) minimum of ϕ .
- $d^{(k)}$ is descent direction for ϕ .

Popular choices for merit functions are of the following type:

$$\phi = \phi(x; \beta, \gamma) := f(x) + \sum_{j=1}^p \beta_j g_j(x)_- + \sum_{i=1}^m \gamma_i |h_i(x)|$$

with constants $\beta_j \geq 0$, $\gamma_i \geq 0$. ϕ is not differentiable!

$$g_j(x)_- := -\min\{0, g_j(x)\} = -\frac{g_j(x) - |g_j(x)|}{2} .$$

If $x \in \Omega$, then $g_j(x) \geq 0$, hence $g_j(x)_- = 0$ and $h_i(x) = 0$, thus $|h_i(x)| = 0$. This yields

$$x \in \Omega \Rightarrow f(x) = \phi(x; \beta, \gamma) .$$

For $x \notin \Omega$ we obtain $\phi > f$. In that sense, $\sum \beta_j (g_j)_-$ and $\sum \gamma_i |h_i|$ are penalty terms which penalizes violations of the constraints:

$$\underbrace{\sum_{i=1}^m \gamma_i |h_i(x)| + \sum_{j=1}^p \beta_j g_j(x)_-}_{\text{penalty term}} .$$

γ_i, β_j penalty parameter

The function ϕ is called (*exact*) *penalty function*. We have the following important

Theorem 7.2.1 Under assumptions 8.1.2–8.1.5 it holds: If \tilde{x} is a local minimum of (PNU) and

$$\beta_j > \mu_j, \quad \gamma_i > |\lambda_i|$$

$j = 1, \dots, p$, $i = 1, \dots, m$, then \tilde{x} is a strict local minimum of $\phi(x; \beta, \gamma)$. Here, λ, μ denote the Lagrange multipliers for \tilde{x} .

By some cumbersome estimates one can show:

We fix $\varepsilon > 0$ and $\delta \in (0, 1)$. If the parameters β_j and γ_i are sufficiently large then

$$\beta_j \geq \mu_j^{(k+1)} + \varepsilon, \quad \gamma_i \geq |\lambda_i^{(k+1)}| + \varepsilon,$$

and for sufficiently small $\sigma > 0$ it holds

$$\phi(x^{(k)} + \sigma d^{(k)}, \beta, \gamma) \leq \phi(x^{(k)}; \beta, \gamma) - \sigma \delta [\langle d^{(k)}, A^{(k)} d^{(k)} \rangle + \varepsilon \|g(x^{(k)})_-\|_1 + \varepsilon \|h(x^{(k)})\|_1],$$

i.e., the merit function can indeed be reduced. Moreover, we have

$$\begin{aligned} |h_i(x^{(k)} + \sigma d^{(k)})| &\leq (1 - \delta\sigma) |h_i(x^{(k)})| \quad i = 1, \dots, m \\ g_j(x^{(k)} + \sigma d^{(k)}) &\geq (1 - \delta\sigma) g_j(x^{(k)}) \quad j = 1, \dots, p, \end{aligned}$$

i.e., the inadmissibility is reduced in every step. This is the basis for the SQP method which we do not discuss in detail here but rather refer to the exposition in [1, Section 8.2.3].

8 Penalty, barrier and augmented Lagrangian methods

8.1 The quadratic penalty method

Idea: replace the optimization problems by a penalty function which is composed of

- the original objective function
- an additional term which is positive if the constraints are violated and otherwise zero.

We first restrict ourselves to equality constraints.

$$\begin{aligned} \min f(x) \quad & f: \mathbb{R}^n \rightarrow \mathbb{R} \\ \text{s.t.} \quad & h(x) = 0 \quad h: \mathbb{R}^n \rightarrow \mathbb{R}^m. \end{aligned} \tag{PNE}$$

Consider the quadratic penalty functions

$$Q(x, \mu) = f(x) + \frac{1}{2\mu} \sum_{i=1}^m h_i^2(x)$$

with $\mu > 0$ a penalty parameter.

Idea: consider a sequence $\{\mu^{(k)}\} \subset \mathbb{R}$ with $\mu^{(k)} \searrow 0$ and compute the minimum $x^{(k)}$ of $Q(x, \mu^{(k)})$ with a method for unconstrained optimization.

Example 8.1.1 $\min x_1 + x_2 \quad \text{s.t.} \quad x_1^2 + x_2^2 - 2 = 0$

Here the solution is $\tilde{x} = (-1, -1)^\top$ and the penalty function is

$$Q(x, \mu) = x_1 + x_2 + \frac{1}{2\mu} (x_1^2 + x_2^2 - 2)^2.$$

Remark:

1. Penalty function for the general case:

$$Q(x, \mu) := f(x) + \frac{1}{2\mu} \sum_{i=1}^m h_i^2(x) + \frac{1}{2\mu} \sum_{j=1}^p [g_j(x)]_-^2$$

with $[x]_- = -\min\{x, 0\}$

2. In contrast to the merit function from Section 7, $Q(x, \mu)$ is a quadratic function which is continuously differentiable (though with a general discontinuous second derivative).

Quadratic penalty methods

1. Initialization: choose $\mu^{(0)} > 0$, $\tau^{(0)} > 0$ und $x_s^{(k)}$, $k = 0$
2. Inner iteration: initial value $x_s^{(k)}$

$$(p_k) \quad \begin{cases} \text{compute } x^{(k)} & \text{as an approximate minimum} \\ \text{of } Q(x, \mu^{(k)}) & \\ \text{s.t. } \|\nabla Q(x, \mu^{(k)})\| \leq \tau^{(k)}. & \end{cases}$$

3. Choose $\mu^{(k+1)} \in (0, \mu^{(k)})$
4. Choose $x_s^{(k+1)}$ (e.g., $x_s^{(k+1)} = x^{(k)}$), $k \rightarrow k + 1$ Go to 2.

Remarks:

1. Effective choice of $\mu^{(k)}$ should be adaptive, if computation of $\min Q(x, \mu^{(k)})$ is difficult, choose $\mu^{(k+1)} = 0.7\mu^{(k)}$ else choose $\mu^{(k+1)} = 0.1\mu^{(k)}$.
2. For equality constraints, $Q(x, \mu^{(k)})$ is twice continuously differentiable, hence, standard methods are applicable. For small $\mu^{(k)}$, we may run into ill conditioning for $Q''(x, \mu^{(k)})$.

Theorem 8.1.1 *Let $x^{(k)} = \arg \min Q(x, \mu^{(k)})$ be the exact global minimum of $Q(x, \mu^{(k)})$ and $\mu^{(k)} \searrow 0$. Then every accumulation point $\{x^{(k)}\}$ is a solution to (PNE).*

PROOF. Let $\tilde{x} \in \Omega = \{x \in \mathbb{R}^n \mid h(x) = 0\}$ be global solution to (PNE), i.e.,

$$f(\tilde{x}) \leq f(x) \quad \forall x \in \Omega,$$

then

$$Q(x^{(k)}, \mu^{(k)}) \leq Q(\tilde{x}, \mu^{(k)})$$

$$\begin{aligned}
\Rightarrow \quad f(x^{(k)}) + \frac{1}{2\mu^{(k)}} \sum_{i=1}^m h_i(x^{(k)})^2 &\leq f(\tilde{x}) + \frac{1}{2\mu} \sum_{i=1}^m h_i^2(\tilde{x}) = f(\tilde{x}) \\
&\Rightarrow \sum_{i=1}^m h_i(x^{(k)})^2 \leq \mu_k(f(\tilde{x}) - f(x^{(k)})) .
\end{aligned} \tag{8.1}$$

Let x^* be an accumulation point of $\{x^{(k)}\}$, i.e., there exists a subsequence $\{x^{(k')}\}$ with $x^{(k')} \rightarrow x^*$, then

$$\sum_{i=1}^m h_i(x^{(k')})^2 \rightarrow \sum_{i=1}^m h_i^2(x^*) = 0$$

$$\Rightarrow h_i(x^*) = 0 \text{ for } 1 \leq i \leq m, \text{ i.e., } x^* \in \Omega$$

$$(8.1) \Rightarrow f(x^*) \leq f(\tilde{x}) \quad \square$$

Theorem 8.1.1 assumes that for every subproblem a global minimum $x^{(k)}$ can be obtained. The following result allows for inexact minimization of $Q(x, \mu^{(k)})$:

Theorem 8.1.2 *Let $x^{(k)}$ be chosen such that $\|\nabla Q(x^{(k)}, \mu^{(k)})\| \leq \tau^{(k)}$ and $\tau^{(k)} \rightarrow 0$. Further, let $\mu^{(k)} \searrow 0$.*

All accumulation points \tilde{x} of $\{x^{(k)}\}$, for which $\nabla h_i(\tilde{x})$ $1 \leq i \leq m$ are linearly independent, are KKT points. Let $\{x^{(k')}\}$ be a subsequence with $x^{(k')} \rightarrow \tilde{x}$, then

$$\lim_{k' \rightarrow \infty} \frac{h_i(x^{(k')})}{\mu^{(k')}} = \lambda_i$$

with Lagrange multiplier λ_i^ .*

PROOF.

$$\begin{aligned}
Q(x, \mu) &= f(x) + \frac{1}{2\mu} \sum_{i=1}^m h_i^2(x) \\
\nabla Q &= \nabla f + \sum_{i=1}^m \frac{h_i}{\mu} \nabla h_i .
\end{aligned} \tag{8.2}$$

Moreover, we have

$$\begin{aligned}
\left\| \sum_{i=1}^m h_i(x^{(k')}) \nabla h_i(x^{(k')}) \right\| &\leq \mu^{(k')} \|\nabla Q(x^{(k')}, \mu^{(k')})\| + \mu^{(k')} \|\nabla f(x^{(k')})\| \\
&\leq \mu^{(k')} (\tau^{(k')} + \|\nabla f(x^{(k')})\|) \xrightarrow[k' \rightarrow \infty]{} 0
\end{aligned}$$

$$\Rightarrow \sum_{i=1}^m h_i(\tilde{x}) \nabla h_i(\tilde{x}) = 0$$

$$\Rightarrow h_i(\tilde{x}) = 0 \quad 1 \leq i \leq m \quad \text{due to linear independency of } \nabla h_i(\tilde{x}), \quad 1 \leq i \leq m, \text{ i.e., } \tilde{x} \in \Omega.$$

Define $\lambda^{(k')} := -\frac{h(x^{(k')})}{\mu^{(k')}}}$

$$(8.2) \Rightarrow \sum_{i=1}^m \frac{h_i}{\mu^{(k')}} \nabla h_i(x^{(k')}) = - \sum_{i=1}^m \lambda_i^{(k')} \nabla h_i(x^{(k')}) = -h'(x^{(k')})^\top \lambda^{(k')} \\ = \nabla Q(x^{(k')}, \mu^{(k')}) - \nabla f(x^{(k')})$$

for k' sufficiently large, we have $\text{rk}(h'(x^{(k')})) = m$

$\Rightarrow h'(x^{(k')})h'(x^{(k')})^\top$ regular

$\Rightarrow h'(x^{(k')})h'(x^{(k')})^\top \lambda^{(k')} = -h'(x^{(k')})[\nabla Q - \nabla f(x^{(k')})]$

and $\lambda^{(k')} = -(h'(x^{(k')})h'(x^{(k')})^\top)^{-1}h'(x^{(k')})[\nabla Q - \nabla f]$

$\xrightarrow{k' \rightarrow \infty} \tilde{\lambda} = (h'(\tilde{x})h'(\tilde{x})^\top)^{-1}h'(\tilde{x})\nabla f(\tilde{x})$

and (8.2) $\Rightarrow \nabla f(\tilde{x}) - h'(\tilde{x})^\top \tilde{\lambda} = 0$. □

Remark:

1. In contrast to Theorem 8.1.1, $\{x^{(k)}\}$ is linearly convergent to stationary/KKT points.

We may use $\frac{h_i(x^{(k)})}{\mu^{(k)}}$ as estimates for Lagrange multipliers.

2.

$$Q''(x, \mu^{(k)}) = f''(x) + \sum_{i=1}^m \underbrace{\frac{h_i}{\mu^{(k)}}}_{\approx \lambda^{(i)}} h_i'' + \frac{1}{\mu^{(k)}} h'^\top(x) h'(x) \\ \approx \underbrace{\mathcal{L}(x, \lambda^*)}_{\text{well conditioned}} + \underbrace{\frac{1}{\mu^{(k)}} h'^\top(x) h'(x)}_{\text{ill conditioned}},$$

i.e., the computation of the Newton direction becomes more ill conditioned with increasing k , i.e., the computation becomes inaccurate.

8.2 The logarithmic barrier method

Consider (PNE)

$$\begin{aligned} & \min f(x) \\ & \text{s.t. } g(x) \geq 0 \end{aligned}$$

let $\Omega^\circ = \{x \in \mathbb{R}^n \mid g_j(x) > 0, 1 \leq j \leq p\} \neq \emptyset$.

Barrier functions are

- ∞ outside of Ω° ,
- smooth inside of Ω° ,

- for $x^{(k)} \rightarrow x$ with $x \in \partial\Omega^\circ$ it holds $\lim P(x) = \infty$.

Most important example:

$$P(x, \mu) = f(x) - \mu \sum_{j=1}^p \log(g_j(x)).$$

Comparison

$$Q(x, \mu) = f(x) + \frac{1}{2\mu} \sum_{j=1}^p [g_j(x)]_-$$

Example 8.2.1

$$\begin{aligned} & \min x \\ & s.t. \quad x \geq 0 \quad 1 - x \geq 0 \\ & P(x, \mu) = x - \mu \log x - \mu \log(1 - x) \end{aligned}$$

Remark:

1. As for penalty functions, one can show that $P''(x, \mu)$ becomes ill conditioned.
2. Algorithm is similar to the penalty method.
3. Under suitable assumptions one can show
 - i) $\exists x(\mu)$ continuously differentiable for μ sufficiently smooth s.t. $x(\mu)$ is local minimum of $P(x, \mu)$ and $\lim_{\mu \searrow 0} x(\mu) = \tilde{x}$, is local solution of (PNE).
 - ii) $\lambda_i(\mu) = \frac{\mu}{g_i(x(\mu))} \xrightarrow{\mu \searrow 0} \tilde{\lambda}$ the Lagrange multiplier.
 - iii) The trajectory $\mathcal{C}_p = \{x(\mu) \mid \mu > 0\}$ is called central path (or primal central path).
 - iv) Handling of equality constraints

$$B(x; \mu) := f(x) - \mu \sum_{j=1}^p \log(g_j(x)) + \frac{1}{2\mu} \sum_{i=1}^m h_i^2(x).$$

8.3 Augmented Lagrangian methods

Motivation: $Q(x, \mu) = f(x) + \frac{1}{2\mu} \sum h_i(x)^2$.

In general, minima $x^{(k)}$ of $Q(x, \mu^{(k)})$ do not satisfy $x^{(k)} \in \Omega$ since, as seen in the proof of Theorem 8.1.2, $h_i(x^{(k)}) \approx -\mu^{(k)} \tilde{\lambda}_i$ for $1 \leq i \leq m$.

Idea: Modify Q to get $x^{(k)}$ “closer” to Ω (thereby avoiding $\mu^{(k)} \searrow 0$)

Consider the augmented Lagrangian:

$$\mathcal{L}_A(x, \lambda; \mu) := f(x) - \sum_{i=1}^m \lambda_i h_i + \frac{1}{2\mu} \sum_{i=1}^m h_i^2$$

$$\nabla_x \mathcal{L}_A = \nabla f(x) - \sum_{i=1}^m \left(\lambda_i - \frac{h_i}{\mu} \right) \nabla h_i.$$

Construct an algorithm with

- $\mu^{(k)}$ and $\lambda^{(k)}$ fixed
- $x^{(k)} = \arg \min \mathcal{L}_A(x, \lambda^{(k)}; \mu^{(k)})$.

In analogy to Theorem 8.1.2 we may expect

$$\tilde{\lambda}_i \approx \lambda_i^{(k)} - \frac{h_i(x^{(k)})}{\mu^{(k)}} \quad (8.3)$$

or $h_i(x^{(k)}) \approx -\mu^{(k)}(\lambda_i^* - \lambda_i^{(k)})$ (i.e., if $\lambda_i^{(k)}$ close to λ_i^* , then $x^{(k)}$ “closer” to Ω than with the penalty method.

Update of $\lambda_i^{(k)}$ via (8.3):

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} - \frac{h_i(x^{(k)})}{\mu^{(k)}}$$

- Augmented Lagrangian (equality constraints)
 1. Initialize $\mu^{(0)} > 0$, $\tau^{(0)} > 0$, $x_s^{(0)}, \lambda^{(0)}$, $k = 0$.
 2. Choose an initial value $x_s^{(k)}$ and minimize $\mathcal{L}(x, \lambda^{(k)}; \mu^{(k)})$ s.t.

$$\|\nabla_x \mathcal{L}(x^{(k)}, \lambda^{(k)}; \mu^{(k)})\| \leq \tau^{(k)}.$$

3. $\lambda_i^{(k+1)} = \lambda_i^{(k)} - \frac{h_i(x^{(k)})}{\mu^{(k)}}$, $\mu^{(k+1)} \in (0, \mu^{(k)})$
 $x_s^{(k+1)} = x^{(k)}$, $k \rightarrow k+1$ go to 2

Example 8.3.1

$$\begin{cases} \min & x_1 + x_2 \\ \text{s.t.} & x_1^2 + x_2^2 - 2 = 0 \end{cases}$$

$$\mathcal{L}_A = x_1 + x_2 - \lambda(x_1^2 + x_2^2 - 2) + \frac{1}{2\mu} (x_1^2 + x_2^2 - 2)^2$$

$$\tilde{x} = (-1, -1)^\top, \tilde{\lambda} = -0.5.$$

For $\mu^{(k)} = 1$ and $\lambda^{(k)} = -0.4$ we obtain

$$\arg \min Q(x, 1) \approx (-1.1, -1.1)$$

but $\arg \min \mathcal{L}_A(x, -0.4, 1) \approx (-1.02, -1.02)$.