# Nonlinear Optimization

**Lucia**, Alex, Max, + others

Lecture notes for the course "250063 VO Nonlinear optimization (2022W)"

by Univ.-Prof. Dr. Radu Ioan Bot

January 8, 2023

## Contents

# 1 First and second order optimality conditions

## Introductory Notions

Let $X \subseteq \mathbb{R}^n$ be a non empty set and $f : \mathbb{R}^n \to \mathbb{R}$. Consider the optimization problem

$$\min_{x \in X} f(x) \tag{1.1}$$

**Definition.** An element $x^* \in X$ is called a *local minimum* of (1.1) if there exists a neighborhood $U_\varepsilon(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| < \varepsilon\}$ s.t.

$$f(x^*) \leq f(x) \text{ for all } x \in U_\varepsilon(x^*) \cap X. \tag{1.2}$$

Here, and always in the following, the norm denotes the Euclidean norm on $\mathbb{R}^n$. If $f(x^*) \leq f(x)$ for all $x \in X$, then $x^*$ is called a *global minimum*.

In order to characterize local minima we introduce the Bouligand tangent cone to $X$ at a point $x_0 \in X$.

**Definition 1.1.** We define the *Bouligand tangent* cone to $X$ at a point $x_0 \in X$ as

$$T_X(x_0) = \left\{ d \in \mathbb{R}^n \,\middle|\, \exists\, (x^k)_{k\geq 1} \subseteq X, \exists\, (t_k)_{k\geq 1} \searrow 0, \text{ s.t. } \frac{x^k - x_0}{t_k} \to d \right\}.$$

**Remark.** (a) The Bouligand tangent cone $T_X(x_0)$ is not empty, since $0 \in T_X(x_0)$ (take $x^k = x_0,\ t_k = \frac{1}{k}$).

(b) The Bouligand tangent cone $T_X(x_0)$ is really a cone, as for all $d \in T_X(x_0),\ \lambda > 0$ we also have $\lambda d \in T_X(x_0)$ by multiplying the sequence $(t_k)_{k\geq 1}$ by $\frac{1}{\lambda}$.

(c) The Bouligand tangent cone $T_X(x_0)$ is not necessarily convex.

(d) Helpful (IMHO) intuition from this post: "The idea behind the Bouligand tangent cone is to have a model of $X$ that shows the directions we can move from a particular point $x_0$ and remain in $X$."

**Proposition 1.2.** *Let $x^*$ be a local minimum of (1.1) and $f$ be continuously differentiable in a neighborhood of $x^*$. Then it holds that*

$$\nabla f(x^*)^T d \geq 0 \quad \forall d \in T_X(x_0). \tag{1.3}$$

$\Rightarrow \exists$

*Proof.* Let $d \in T_X(x^*)$. Then there exist sequences $(x^k)_{k\geq 1} \subseteq X$, $(t_k)_{k\geq 1} \searrow 0$ such that

$$\frac{x^k - x^*}{t_k} \to d \text{ for } k \to \infty.$$

Thus also

$$x^k - x^* = \frac{x^k - x^*}{t_k} t_k \to d t_k \to 0 \quad (k \to \infty).$$

Let $U_\varepsilon(x^*)$ be a neighborhood of $x^*$ where $f$ is continuously differentiable and for which $f(x) \geq f(x^*)\ \forall x \in U_\varepsilon(x^*) \cap X$. Then there exists a $k_\varepsilon \in \mathbb{N}$ such that $x^k \in U_\varepsilon(x^*)\ \forall k \geq k_\varepsilon$. (local minimum)

By the mean value theorem, there exist $\xi^k \in (x^*, x^k) = \{\lambda x^* + (1-\lambda) x^k | \lambda \in [0,1]\}$ such that
convex

$$f(x^k) - f(x^*) = \nabla f(\xi^k)^T (x^k - x^*).$$

Let $\xi^k := \lambda_k x^* + (1-\lambda_k) x^k$. Moreover, for all $k \geq k_\varepsilon$ we have

$$\left\|\xi^k - x^*\right\| = \left\|\lambda_k x^* + (1-\lambda_k) x^k - x^*\right\| = |\lambda_k - 1| \left\|x^k - x^*\right\| \leq \left\|x^k - x^*\right\| \to 0,$$

and thus $\xi^k \to x^*$.

Since the gradient of $f$ is continuous by assumption, we have $\nabla f(\xi^k) \to \nabla f(x^*)$ and thus also

$$\nabla f(x^*)^T d = \lim_{k\to\infty} \frac{\nabla f(\xi^k)^T (x^k - x^*)}{t_k} = \lim_{k\to\infty} \frac{f(x^k) - f(x^*)}{t_k} \geq 0,$$

where the last inequality holds since $\frac{f(x^k)-f(x^*)}{t_k} \geq 0$ for all $k \geq 1$ □

**Definition 1.3.** For any cone $k \subseteq \mathbb{R}^n$ we denote the dual cone by

$$k^* = \left\{ s \in \mathbb{R}^n \middle| s^T d \geq 0\ \forall d \in k \right\}.$$

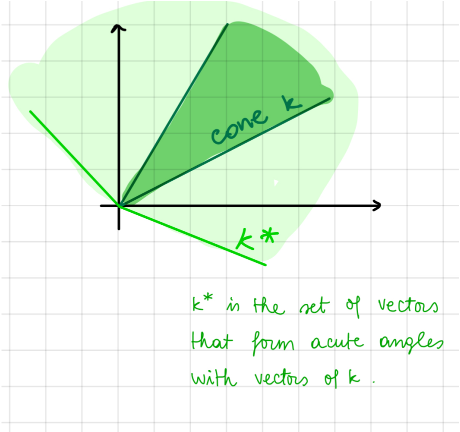
**Remark.** The dual cone $k^*$ is the set of vectors that form acute angles with vectors from $k$. Furthermore, (1.3) says that $\nabla f(x^*) \in (T_X(x^*))^*$.

**Remark.** a) If $X$ is a convex set then

$$T_X(x_0) = \operatorname{cl}(\operatorname{cone}(X - x_0))$$
(closure)

and

$$-(T_X(x_0))^* = N_X(x_0) = \left\{ s \in \mathbb{R}^n \middle| s^T (x - x_0) \leq 0\ \forall x \in X \right\}$$



k* is the set of vectors that form acute angles with vectors of k.

Figure 1: Sketch of a dual cone.

the so called *normal cone* to $X$ at $x_0$. In this case, (1.3) is equivalent to

$$\nabla f(x^*)^T(x - x^*) \geq 0 \text{ for all } x \in X.$$

For the details, see the exercise classes.

b) We claim that, if $x_0 \in \mathrm{int}(X)$, then $T_X(x_0) = \mathbb{R}^n$.

Indeed, one inclusion is obvious. For the other one, let $d \in \mathbb{R}^n$. Then, since $x_0 \in \mathrm{int}(X)$, there exists $k_0 \geq 0$ such that $x^k := x_0 + \frac{1}{k}d \in X \; \forall k \geq k_0$. Furthermore, define $t_k := \frac{1}{k} \searrow 0$. Naturally, $\frac{x^k - x_0}{t_k} = d \to d$. Thus $d \in T_X(x_0)$.

**Theorem 1.4.** *Let $x^*$ be a local minimum of (1.1) be continuously differentiable in a neighborhood $U_\varepsilon(x^*)$, then $x^*$ is a* critical point *of $f$. This means that*

$$\nabla f(x^*) = 0. \tag{1.4}$$

(Standard critical point)

*Proof.* Since $x^*$ is a local minimum of

$$\min_{x \in U_\varepsilon(x^*)} f(x),$$

by Prop. (1.2), we have $\nabla f(x^*)^T d \geq 0$ for all $d \in T_{U_\varepsilon(x^*)}(x^*) = \mathbb{R}^n$. Plugging in $d$ and $-d$, we obtain $\nabla f(x^*)^T d = 0$ for all $d \in \mathbb{R}^n$. Hence, we have $\nabla f(x^*)^T = 0$ and therefore $\nabla f(x^*) = 0$. □

Condition (1.4) is called *first order optimality condition*. It is a local statement: we care only about what happens in the neighborhood, not in the rest of the domain.

So far, we have considered unconstrained optimization. Now, let's switch to constrained optimization. In the following we will consider the general nonlinear optimization problem

constrained optimization

$$\begin{aligned} &\min f(x) \\ \text{s.t. } &g_i(x) \leq 0, i = 1, \dots, m \\ &h_j(x) = 0, i = 1, \dots, p \\ &x \in \mathbb{R}^n \end{aligned} \tag{1.5}$$

where $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}, i = 1, \dots, m, j = 1, \dots, p$ are continuously differentiable. The set

$$X = \left\{ x \in \mathbb{R}^n \middle| \begin{array}{l} g_i(x) \leq 0, \ i = 1 \dots, m, \\ h_j(x) = 0, \ j = 1, \dots, p \end{array} \right\} \tag{1.6}$$

is called the *feasible set* of (1.5).

$\hookrightarrow \min_{x \in X} f(x)$ .

## 1.1 The Linearized Tangent Cone

We introduce the linearized tangent cone as a "replacement" for the Bouligand tangent cone. This is because the linearized tangent cone can be easily determined, whereas the Bouligand tangent cone in general cannot. Our hope is that we can approximate the Bouligand tangent cone with the linearized tangent cone.

**Definition 1.5.** Let $x_0 \in X$.

a) The constraint $g_i(x) \leq 0$ is said to be *active* at $x_0$ if $g_i(x_0) = 0$. We define

$$\mathcal{A}(x_0) = \{i = 1, \dots, m | g_i(x_0) = 0\}$$

as the *set of active indices* at $x_0$. We also define the *set of inactive indices* at $x_0$ as

$$I(x_0) = \{1, \dots, m\} \setminus \mathcal{A}(x_0).$$

b) The set

$$T_{lin}(x_0) = \left\{ d \in \mathbb{R}^n \middle| \begin{array}{l} \boxed{\nabla g_i(x_0)^T d \leq 0} \ \forall i \in \mathcal{A}(x_0) \\ \boxed{\nabla h_j(x_0)^T d = 0} \ \forall j = 1, \dots, p \end{array} \right\}$$

is called the *linearized tangent cone* of $X$ at $x_0$.

Note that the linearized cone is really a cone (check with the definition of cone). A (maybe) surprising fact is that in many cases, $T_{lin}(x_0) = T_X(x_0)$, or at least, the two are not far from each other. We will study the relation between them in the next weeks.

**Remark 1.6.** [See exercise 4 from exercise session] It holds that $T_{lin}(x_0) = T_{X_{lin}}(x_0)$, where

$$X_{lin} = \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{l} g_i(x_0) + \nabla g_i(x_0)^T(x - x_0) \le 0, i = 1, \dots, m \\ h_j(x_0) + \nabla h_i(x_0)^T(x - x_0) = 0, j = 1, \dots, p \end{array} \right\}.$$

Hint: prove that $X_{lin}$ is convex and use the characterization of $T_X$ for $X$ convex. Also, note that $X_{lin}$ is a polyhedral set and is well-defined, because $x_0 \in X_{lin}$.

**Lemma 1.7.** *Let $x_0 \in X$. Then it holds that $T_X(x_0) \subseteq T_{lin}(x_0)$.*

*Proof.* Let $x_0 \in X$ and $d \in T_X(x_0)$. Then there exist sequences $(x^k)_{k\ge1} \subseteq X$, $(t_k)_{k\ge1} \searrow 0$ such that

$$\frac{x^k - x^*}{t_k} \to d \text{ for } k \to \infty.$$

We will first prove that for all $i \in \mathcal{A}(x_0)$, we have $\nabla g_i(x_0)^T d \le 0$. Indeed, let $i \in \mathcal{A}(x_0)$. By the mean value theorem, for all $k \ge 1$ there exists a $\xi^k \in (x_0, x^k)$ which fulfills

$$g_i(x^k) - g_i(x_0) = \nabla g_i(\xi^k)^T(x^k - x_0) \;\Rightarrow\; \nabla g_i(\xi^k)^T(x^k - x_0) = g_i(x^k) \le 0,$$

since $g_i(x_0) = 0$ because $i$ is active. Dividing by $t_k$ yields for all $k \ge 1$ (by continuity of the gradient)

$$0 \ge \nabla g_i(\xi^k)^T \frac{x^k - x_0}{t_k} \to \nabla g_i(x_0)^T d.$$

Next, we will prove, using the same argument, that for all $j = 1, \dots, p$, we have $\nabla h_j(x_0)^T d = 0$. Let $j \in \{1, \dots, p\}$. Then, by the mean value theorem, for all $k \ge 1$ we get an $\mu^k \in (x_0, x^k)$ such that

$$h_j(x^k) - h_j(x_0) = \nabla h_j(\mu^k)^T(x^k - x_0) \;\Rightarrow\; \nabla h_j(\mu^k)^T(x^k - x_0) = 0,$$

Start with $x_0 \in X$ and $d \in T_X(x_0)$ (then
define $d$ in terms of $T_{lin}(x_0)$ using $T_X(x_0)$

→ $d \in T_{lin}(x_0)$

since $h_j(x^k) = h_j(x_0) = 0$. Again by dividing by $t_k$, we get

$$0 = \nabla h_j(\mu^k)^T \frac{x^k - x_0}{t_k} \to \nabla h_j(x_0)^T d,$$

which shows that $d \in T_{lin}(x_0)$. □

**Example 1.8.** The equality $T_X(x_0) = T_{lin}(x_0)$ does not hold in general. Consider the optimization problem

$$\begin{aligned} &\min x_1 + x_2^2 \\ \text{s.t. } &g_1(x_1, x_2) = -x_2 \leq 0 \\ &g_2(x_1, x_2) = x_2 - x_1^3 \leq 0 \\ &x = (x_1, x_2) \in \mathbb{R}^2 \end{aligned}.$$

Naturally $x^* = (0, 0)$ is the unique global minimum. We have $\mathcal{A}(x^*) = \{1, 2\}$ and

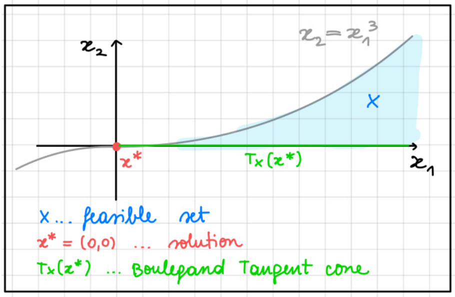

$x_2$
$x_2 = x_1^3$
$X$
$x^*$
$T_X(x^*)$
$x_1$
X ... feasible set
$x^* = (0,0)$ ... solution
$T_X(x^*)$ ... Bouligand Tangent cone

Figure 2: Visualization of Example 1.8

$$\begin{aligned} \nabla g_1(x_1, x_2) &= (0, -1)^T & \qquad \nabla g_1(0, 0) &= (0, -1)^T \\ \nabla g_2(x_1, x_2) &= (-3x_1^2, 1)^T & \qquad \nabla g_2(0, 0) &= (0, 1)^T \end{aligned}$$

so we get $T_{lin}(x^*) = \mathbb{R} \times \{0\}$. By drawing a picture, we easily see $T_X(x^*) = \mathbb{R}^+ \times \{0\}$. This implies $T_X(x^*) \subsetneq T_{lin}(x^*)$. In addition, since $\nabla f(x^*) = (1, 0)^T$, it holds

$$\nabla f(x^*)^T d = d_1 \geq 0, \; \forall d \in T_X(x^*),$$

but $\nabla f(x^*)^T d = d_1 \ngeq 0, \; \forall d \in T_{lin}(x^*)$. So we cannot replace $T_X(x^*)$ by $T_{lin}(x^*)$ in $(1.3)$ !

**Lemma 1.9** (Lemma of Farkas). *Let be $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then the following are equivalent:*
*(a) The linear system $Ax = b$ has a solution $x \geqq 0$.*
*(b) For all $d \in \mathbb{R}^m$ with $A^T d \geqq 0$ it holds $b^T d \geq 0$.*

*Proof.* See exercise 6 from exercise session. □

**Theorem 1.10.** *Let $X$ be given as in (1.6) and $x_0 \in X$. It holds $-(T_{lin}(x_0))^* = N_{lin}(x_0)$ where*

$$N_{lin}(x_0) = \left\{ \sum_{i=1}^m \lambda_i \nabla g_i(x_0) + \sum_{j=1}^p \mu_j \nabla h_j(x_0) \;\middle|\; \begin{array}{l} \lambda_i \geq 0, i \in \mathcal{A}(x_0) \\ \lambda_i = 0, i \in I(x_0) \\ \mu_j \in \mathbb{R} \end{array} \right\}. \tag{1.7}$$

*Proof.* "$\supseteq$": Let $s \in N_{lin}(x_0)$ and $d \in T_{lin}(x_0)$. Then

$$(-s)^T d = -\sum_{i=1}^m \lambda_i \nabla g_i(x_0)^T d - \sum_{j=1}^p \mu_j \nabla h_j(x_0)^T d.$$

Since $\lambda_i \geq 0$ and $\nabla g_i(x_0)^T d \leq 0$ for all $i = 1, \ldots, m$ and $\nabla h_j(x_0)^T d = 0$ for all $j = 1, \ldots, p$, we have

$$(-s)^T d \geq 0,$$

which implies $-s \in (T_{lin}(x_0))^*$ and therefore $s \in -(T_{lin}(x_0))^*$.

"$\subseteq$": Take $s \in -(T_{lin}(x_0))^*$. Then we have, by definition, that $(-s)^T d \geq 0$ for all $d \in T_{lin}(x_0)$. Define $A \in \mathbb{R}^{n\times(|\mathcal{A}(x_0)|+2p)}$ as

$$A := (-\nabla g_{i_1}(x_0), \ldots, -\nabla g_{i_r}(x_0), \nabla h_1(x_0), \ldots, \nabla h_p(x_0), -\nabla h_1(x_0), \ldots, -\nabla h_p(x_0))$$

for $\mathcal{A}(x_0) = \{i_1, \ldots, i_r\}$. We have the following:

$$\begin{aligned} s \in -(T_{lin}(x_0))^* &\Leftrightarrow -s^T d \geq 0 \;\; \forall d \in T_{lin}(x_0) \\ &\Leftrightarrow -s^T d \geq 0 \;\; \forall d \in \mathbb{R}^n \text{ s.t. } A^T d \geqq 0 \end{aligned}$$

By Farkas' lemma, there exists a $\eta = (\lambda, \mu_1, \mu_2) \in \mathbb{R}_+^{|\mathcal{A}(x_0)|} \times \mathbb{R}_+^p \times \mathbb{R}_+^p$ such that

$$
\begin{aligned}
A\eta = -s \Leftrightarrow & \sum_{i=1}^{m} -\lambda_i \nabla g_i(x_0) + \sum_{j=1}^{p} (\mu_1)_j \nabla h_j(x_0) + \sum_{j=1}^{p} -(\mu_2)_j \nabla h_j(x_0) = -s \\
\Leftrightarrow s = & \sum_{i=1}^{m} \lambda_i \nabla g_i(x_0) + \sum_{j=1}^{p} \mu_j \nabla h_j(x_0)
\end{aligned}
$$

where $\lambda_i \geq 0$ for $i \in \mathcal{A}(x_0)$, $\lambda_i = 0$ for $i \in I(x_0)$ and $\mu_j \in \mathbb{R}$ for $j = 1, \ldots, p$. The second equivalence follows by setting $\mu = \mu_2 - \mu_1$. □

**Remark 1.11** (Life is not perfect). We know that $T_X(x_0) \subseteq T_{lin}(x_0)$. This implies

$$(T_{lin}(x_0))^* \subseteq (T_X(x_0))^*.$$

Indeed, the following holds:

$$s \in (T_{lin}(x_0))^* \Rightarrow s^T d \geq 0 \; \forall \; d \in T_{lin}(x_0) \Rightarrow s^T d \geq 0 \; \forall \; d \in T_X(x_0) \Rightarrow s \in (T_X(x_0))^*.$$

Since $(T_{lin}(x_0))^*$ has the representation (1.7), one would like to have for a local minimum $x^*$ the much nicer condition $\nabla f(x^*) \in (T_{lin}(x^*))^*$ instead of $\nabla f(x^*) \in (T_X(x^*))^*$. This is in general not the case! See example 1.8.

## 1.2 Karush–Kuhn–Tucker (KKT) conditions

**Definition 1.12.** a) The function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, given by

$$
\begin{aligned}
L(x, \lambda, \mu) =& f(x) + \lambda^T g(x) + \mu^T h(x) \\
=& f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)
\end{aligned} \tag{1.8}
$$

is called the *Lagrange function* associated to equation (1.5).

b) The conditions

$$
\begin{cases}
\nabla_x L(x, \lambda, \mu) = 0 & (1.9) \\
\lambda \geqq 0, \; g(x) \leqq 0, \; \lambda^T g(x) = 0 & (1.10) \\
h(x) = 0 & (1.11)
\end{cases}
$$

are called the *KKT optimality conditions* of (1.5).

It holds that:

$$
\begin{aligned}
\nabla_x L(x, \lambda, \mu) =& \nabla f(x) + \lambda^T \nabla g(x) + \mu^T \nabla h(x) \\
=& \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{p} \mu_j \nabla h_j(x).
\end{aligned}
$$

c) An element $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ fulfilling (1.9,1.10,1.11) is called a *KKT-point* of (1.5). The vectors $\lambda^*, \mu^*$ are called *Lagrange multipliers* associated with the restrictions $g(x^*) \leqq 0$ and $h(x^*) = 0$, respectively. The statement (1.10) is called *complementary condition* and is equivalent to

$$\lambda_i \geq 0, \; g_i(x^*) \leq 0, \; \lambda_i g_i(x^*) = 0 \text{ for } i = 1, \ldots, n.$$

$$\begin{aligned} &\min f(x) \\ \text{s.t. } &g_i(x) \leq 0, i = 1, \ldots, m \\ &h_j(x) = 0, i = 1, \ldots, p \\ &x \in \mathbb{R}^n \end{aligned}$$

1.5

$$X = \left\{ x \in \mathbb{R}^n \,\middle|\, \begin{array}{l} g_i(x) \leq 0, \; i = 1 \ldots, m, \\ h_j(x) = 0, \; j = 1, \ldots, p \end{array} \right\}$$

1, 6

# 2 First order necessary and sufficient optimality conditions

## 2.1 Optimality conditions under Abadie CQ

**Definition 2.1.** An element $x_0 \in X$, where $X$ is given by (1.6) is said to fulfill the *Abadie Constraint Qualification* (Abadie CQ) if $T_X(x_0) = T_{lin}(x_0)$.

Note that the Abadie CQ does not depend on $f$. Let's now convince ourselves that the Abadie CQ gives us very nice results. Later, we will be more critical and analyze when the Abadie CQ is satisfied.

**Theorem 2.2.** *Assume that a local minimum $x^*$ of (1.5) fulfills the Abadie CQ. Then there exist (not necessarily unique) Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT-point of (1.5).*

*Proof.* From Proposition (1.2), we know that $\nabla f(x^*) \in (T_X(x^*))^*$. Since $x^*$ fulfills the Abadie CQ, we also know that $(T_X(x^*))^* = (T_{lin}(x^*))^*$. Combining these two observations, we get

We have these two

$$-\nabla f(x^*) \in -(T_{lin}(x^*))^*.$$

$\Leftarrow$

From Theorem (1.10), we know that there exist $\lambda_i^* \geq 0$ for $i \in \mathcal{A}(x^*)$, $\lambda_i^* = 0$ for $i \in I(x^*)$, and $\mu_j^* \in \mathbb{R}$ for $j = 1, \ldots, p$ such that

$$\begin{aligned} &\sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = -\nabla f(x^*) \\ \Rightarrow\; &\sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) + \nabla f(x^*) = 0 \\ \Rightarrow\; &\nabla_x L(x^*, \lambda^*, \mu^*) = 0, \end{aligned}$$

where $L$ is the Lagrange function associated to (1.5). Moreover, we have $\lambda_i^* \geq 0$, $g_i(x^*) \leq 0$, $\lambda_i^* g_i(x^*) = 0$ for all $i = 1, \ldots, m$ (namely, if $i$ is active, we have $g_i(x^*) = 0$ and if $i$ is inactive, we have $\lambda_i^* = 0$). Furthermore, we already know from (1.5) that $h_j(x^*) = 0$ for all $j = 1, \ldots, p$. These last three observations imply that $(x^*, \lambda^*, \mu^*)$ is a KKT-point of (1.5). □

Theorem (2.2) is good news in the sense that it gives us a hint on how to find $x^*$. However, it has a major issue: it's suboptimal to assume something about $x^*$ (in particular, the Abadie CQ) before we even find it. We will fix this problem by finding friendlier conditions (meaning, conditions that we can use in practice) that imply the Abadie CQ.

**Corollary 2.3.** *Let $x^*$ be a local minimum of* (2.2. for linear systems)

$$\begin{aligned} &\min f(x) \\ \text{s.t. } &Ax \leq b, \text{ where } A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m \\ &Cx = d, \text{ where } C \in \mathbb{R}^{p\times n}, d \in \mathbb{R}^p, \end{aligned}$$

*which is (1.5) with $g(x) = Ax - b$ ad $h(x) = Cx - d$. Then there exist (not necessarily uniquely defined) Lagrange multipliers $\lambda^* \in \mathbb{R}^m$, $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT-point of the problem above.*

*Proof.* We only need to prove that, for this problem, the Abadie CQ is fulfilled. To do that, it is sufficient to prove that, in this case, we have $X = X_{lin}$. This implies $T_X(x^*) = T_{X_{lin}}(x^*) = T_{lin}(x^*)$ and the statement follows from Theorem (2.2).

Indeed, recall the definitions of $X$ (see (1.6)) and $X_{lin}$ (see Remark (1.6)). In our case we have

$$\begin{aligned} g_i(x) &= a_i^T x - b_i, \text{ for } i = 1, \dots, m \\ h_j(x) &= c_j^T x - d_j, \text{ for } j = 1, \dots, p. \end{aligned}$$

And furthermore we have for all $i = 1, \dots, m$ and $j = 1, \dots, p$:

$$\begin{aligned} g_i(x_0) + \nabla g_i(x_0)^T(x - x_0) &= a_i^T x_0 - b_i + a_i^T(x - x_0) = a_i^T x - b_i = g_i(x), \\ h_j(x_0) + \nabla h_j(x_0)^T(x - x_0) &= c_j^T x - d_j + c_j^T(x - x_0) = c_j^T x - d_j = h_j(x). \end{aligned}$$

Thus, we can conclude that $X = X_{lin}$. □

## 2.2 Optimality conditions under MFCQ

**Definition 2.4.** An element $x_0 \in X$, where $X$ is given by (1.6), is said to fulfill the *Mangasarian-Fromovitz Constraint Qualification* (MFCQ) if both of the following conditions are met:

1. the vectors $\nabla h_j(x_0), j = 1, \dots, p$ are linearly independent, and
2. there exist $d \in \mathbb{R}^n$ such that $\nabla g_i(x_0)^T d < 0$ for $i \in \mathcal{A}(x_0)$ and $\nabla h_j(x_0)^T d = 0$ for $j = 1, \dots, p$.

(what is $\mathcal{A}(x_0)$)

$$\mathcal{A}(x_0) = \{i = 1, \dots, m | g_i(x_0) = 0\}$$

**Lemma 2.5.** *Let $x_0$ be feasible for (1.5) s.t $x_0$ fulfills MFCQ, where $d \in \mathbb{R}^n$ is the vector for which condition 2 in MFCQ holds. Then there exists an $\varepsilon > 0$ and $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that*

*a)* $x$ *in continuously differentiable on* $(-\varepsilon, \varepsilon)$

*b)* $x(t) \in X \quad \forall t \in (0, \varepsilon)$

*c)* $x_0 = x(0)$

*d)* $d = \dot{x}(0)$

*Proof.* Let

$$\begin{aligned} H : \mathbb{R}^p \times \mathbb{R} &\to \mathbb{R}^p \\ (y, t) &\mapsto h(x_0 + td + \nabla h(x_0)^T y), \end{aligned}$$

where $h : \mathbb{R}^n \to \mathbb{R}^p$ is the same function as in (1.5) and

$$\nabla h(x) = \begin{pmatrix} \nabla h_1(x_0)^T \\ \vdots \\ \nabla h_p(x_0)^T \end{pmatrix} \in \mathbb{R}^{p \times n}$$

is the Jacobi matrix of $h$. Note that $H$ is well defined. It holds that:

(i) $H(0, 0) = h(x_0) = 0$, since $x_0$ is feasible

(ii) $\nabla_y H(y, t) = \nabla h(x_0 + td + \nabla h(x_0)^T y) \nabla h(x_0)^T$

(iii) $\nabla_y H(0, 0) = \nabla h(x_0) \nabla h(x_0)^T \in \mathbb{R}^{p \times p}$

We claim that $\nabla_y H(0, 0)$ is positive definite (and therefore invertible). Indeed, we have for all $s \in \mathbb{R}^p$:

$$\begin{aligned} s^T \nabla_y H(0, 0) s &= s^T \nabla h(x_0) \nabla h(x_0)^T s \\ &= (\nabla h(x_0)^T s)^T (\nabla h(x_0)^T s) \\ &= \left\| \nabla h(x_0)^T s \right\| \geq 0. \end{aligned}$$

Furthermore,

$$s^T \nabla_y H(0, 0) s = 0 \Leftrightarrow \nabla h(x_0)^T s = 0 \Leftrightarrow \sum_j s_j \nabla h_j = 0$$

and by condition $1$ in MFCQ we have that this is equivalent to $s = 0$. Thus the matrix is positive definite.

Now we have everything we need to use the implicit function theorem for $H(y,t) = 0$ at $(0,0)$. It says that there exists an $\varepsilon_0 > 0$ and $y : (-\varepsilon_0, \varepsilon_0) \to \mathbb{R}^p$ continuously differentiable such that $y(0) = 0$ and

$$H(y(t),t) = 0 \text{ for all } t \in (-\varepsilon_0, \varepsilon_0).$$

By differentiating this expression with the chain rule, we get

$$\nabla_y H(y(t),t)\dot{y}(t) + \nabla_t H(y(t),t) = 0 \text{ for all } t \in (-\varepsilon_0, \varepsilon_0).$$

In particular, if we plug in $t = 0$, this yields

$$\begin{aligned} &\nabla_y H(0,0)\dot{y}(0) + \nabla_t H(0,0) = 0 \\ \Leftrightarrow\ &\dot{y}(0) = -(\nabla y H(0,0))^{-1}\nabla_t H(0,0) \\ \Rightarrow\ &\dot{y}(0) = -(\nabla y H(0,0))^{-1}\nabla h(x_0)^T d = 0, \end{aligned}$$

where the last equality holds since $\nabla h(x_0)^T d = 0$ by condition $2$ of MFCQ. Now we construct $x$ using $y$. For all $t \in (-\varepsilon_0, \varepsilon_0)$, let

$$x(t) := x_0 + td + \nabla h(x_0)^T y(t).$$

This function is continuously differentiable because $y$ is. Furthermore, it holds that $x(0) = 0$ and

$$\begin{aligned} &x'(t) = d + \nabla h(x_0)^T \dot{y}(t) \\ \Rightarrow\ &x'(0) = d + \nabla h(x_0)^T \dot{y}(0) = d, \end{aligned}$$

so to prove the theorem, we only need to show *b)*. First, note that for all $t \in (0, \varepsilon_0)$, we have that

$$0 = H(y(t),t) = h(x_0 + td + \nabla h(x_0)^T y(t)) = h(x(t)).$$

Next, let $i \in \{1, \ldots, m\}$. On the one hand, if $g_i(x_0) < 0$, there exist $\varepsilon_i > 0$ such that $g_i(x(t)) < 0$ for all $t \in (0, \varepsilon_i)$. This is because $x(t) \to x(0) = x_0$ for $t \searrow 0$. On the other hand, if $g_i(x_0) = 0$, define

$$q_i(t) := g_i(x(t)).$$

Then $q'(t) = \nabla g_i(x(t))^T x'(t)$ and thus

$$q_i'(0) = \nabla g_i(x(0))^T x'(0) = \nabla g_i(x_0)^T d < 0,$$

where the last inequality follows by condition 2 in MFCQ. Moreover, we have

$$\begin{aligned} 0 > q_i'(0) &= \lim_{t \searrow 0} \frac{q_i(t) - q_i(0)}{t} \\ &= \lim_{t \searrow 0} \frac{g_i(x(t)) - g_i(x(0))}{t} \\ &= \lim_{t \searrow 0} \frac{g_i(x(t))}{t}. \end{aligned}$$

Therefore, there exists $\varepsilon_i > 0$ such that for all $t \in (0, \varepsilon_i)$ we have

$$\frac{g_i(x(t))}{t} < 0 \Leftrightarrow g_i(x(t)) < 0.$$

Now let $\varepsilon = \min\{\varepsilon_0, \ldots, \varepsilon_n\}$. Then for all $t \in (0, \varepsilon)$ it holds that

$$\begin{aligned} h_j(x_t) &= 0 \quad \forall j = 1, \ldots, p \\ g_i(x_t) &\leq 0 \quad \forall i = 1, \ldots, m \end{aligned}$$

and thus $x(t) \in X$. □ Q.E.D.

**Theorem 2.6.** *Assume that a local minimum $x^*$ of (1.5) fulfills the MFCQ. Then there exist (not necessarily unique) Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT-point of (1.5).*

This statement is very useful, because we can verify MFCQ, but Abadie CQ not so much.

*Proof.* We show that MFCQ $\Rightarrow$ Abadie CQ and use Theorem 2.2. It is sufficient to prove that $T_{lin}(x_0) \subseteq T_X(x_0)$ (because $\supseteq$ is always true). Let $d \in T_{lin}(x_0)$ and

$\bar{d} \in \mathbb{R}^n$ a solution of the the system in in MFCQ. Then:

$$\begin{aligned} \forall\, i \in \mathcal{A}(x_0)\ &: \nabla g_i(x^*)^T d \leq 0 \\ &\phantom{:}\ \nabla g_i(x^*)^T \bar{d} < 0 \\ \forall\, j = 1 \ldots p\ &: \nabla h_i(x^*)^T d = 0 \\ &\phantom{:}\ \nabla h_i(x^*)^T \bar{d} = 0 \end{aligned}$$

Now for all $\tau > 0$ fixed, let $d(\tau) = d + \tau\bar{d}$. Then the following holds

$$\begin{aligned} \forall\, i \in \mathcal{A}(x^*)\ &: \nabla g_i(x^*)^T d(\tau) < 0 \\ \forall\, j = 1 \ldots p\ &: \nabla h_i(x^*)^T d(\tau) = 0 \end{aligned}$$

This implies that $d(\tau)$ fulfills MFCQ for all $\tau > 0$. We choose an arbitrary $\tau > 0$. By Lemma (2.5), there is a $\varepsilon > 0$ and a continous differentiable $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that

$$\begin{cases} x(t) \in X \quad \forall t \in [0, \epsilon) \\ x(0) = x^* \\ \dot{x}(0) = d(\tau) \end{cases}$$

Now, let $t_k := \frac{\varepsilon}{k}$ for $k \geq 2$ and $x_k := x(t_k)$, which is in $X$ for $k \geq 2$. Then

$$\lim_{k\to\infty} \frac{x_k - x^*}{t_k} = \lim_{k\to\infty} \frac{x(t_k) - x(0)}{t_k} = \dot{x}(0) = d(\tau)$$

which gives $d(\tau) \in T_X(x^*)$. Because $d(\tau) \to d$ and $T_X(x^*)$ is closed, we get $d \in T_X(x^*)$. □

## 2.3 Optimality conditions under LICQ

**Definition 2.7.** An element $x_0 \in X$, where $X$ is given by (1.6), is said to fulfill the *Linear Independence Constraint Qualification* (LICQ) if the vectors

$$\{\nabla g_i(x_0)\}_{i\in\mathcal{A}(x_0)} \cup \{\nabla h_j(x_0)\}_{j=1}^{p}$$

are linearly independent.

**Theorem 2.8.** *Assume that a local minimum $x^*$ of (1.5) fulfills the LICQ. Then there* *exist uniquely defined Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT-point of (1.5).*

*Proof.* We show that LICQ $\Rightarrow$ MFCQ.

1. Note that

$$\{\nabla h_j(x^*)\}_{j=1}^p \subseteq \{\nabla g_i(x^*)\}_{i\in\mathcal{A}(x^*)} \cup \{\nabla h_j(x^*)\}_{j=1}^p.$$

   Since the RHS is linearly independent (by LICQ), so is the LHS, and this proves MFCQ(1).

2. Note that $|\mathcal{A}(x^*)|+p \leq n$, because the RHS in point 1 is linearly independent. Let $A \in \mathbb{R}^{n\times n}$ be a matrix such that

$$A = \begin{pmatrix} \nabla g_i(x^*)^T|_{i\in\mathcal{A}(x)} \\ \nabla h_j(x^*)^T|_{j=1,\dots,p} \\ \text{any rows which make A invertible} \end{pmatrix}$$

   is regular by completing the set of vectors to a basis of $\mathbb{R}^n$. Let

$$b = (\underbrace{-1,\dots,-1}_{|\mathcal{A}(x^*)|\text{ many}}, \underbrace{0,\dots,0}_{p\text{ many}}, \underbrace{\text{arbitrary numbers}}_{n-p-|\mathcal{A}(x^*)|\text{ many}})^T \in \mathbb{R}^n.$$

   Since $A$ is regular, there exists a unique $d \in \mathbb{R}^n$ such that $Ad = b$. This implies that:

$$\nabla g_i(x^*)d = -1 < 0 \ \ \forall i \in \mathcal{A}(x^*)$$
$$\nabla h_j(x^*)d = 0 \ \ \forall j = 1,\dots,p.$$

   Thus, MFCQ (2) holds and by Theorem 2.6, there exist Lagrange multipliers $\lambda^* \in \mathbb{R}^n_+$, $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT point of (1.5).

To prove uniqueness, note that for all inactive indices $i \in I(x^*)$, we have

$$g_i(x^*) < 0 \ \Rightarrow \ \lambda_i^* = 0.$$

Thus,

$$\nabla f(x^*) = -\sum_{i\in\mathcal{A}(x^*)} \lambda_i^* \nabla g_i(x^*) - \sum_{j=1}^p \mu_j^* \nabla h_j(x^*)$$

and by LICQ we know that $\lambda_i^*$ for $i \in \mathcal{A}(x^*)$ and $\mu_j$ for $j = 1, \dots, p$ are unique. □

So far, out of the constraint qualifications we have looked at, LICQ are the easiest ones to verify. There are problems where LICQ is *not* verified but MFCQ (or Abadie CQ) are, but such cases are pathological, in practice this never happens.

## 2.4 Optimality conditions for convex optimization problems

So far, we only had necessary optimality conditions. For the convex setting, we will prove a sufficient condition!

A convex optimization problem is a problem where both $f$ and the feasible set $X$ are convex. To this end, we consider (1.5) under the following assumptions:

(i) the function $f$ is convex,

(ii) the functions $g_i$, $i = 1, \dots, m$, are convex, and

(iii) the function $h$ is affine linear, so $h : \mathbb{R}^n \to \mathbb{R}^p$, $h(x) = Ax - b$, with $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$.

Then, (1.5) becomes

$$\boxed{\begin{aligned} &\min f(x) \\ \text{s.t. } & g_i(x) \leq 0,\ i = 1, \dots, m \\ & Ax = b \\ & x \in \mathbb{R}^n \end{aligned}} \tag{2.1}$$

Then the feasible set $X$ is convex!

**Definition 2.9.** The optimization problem (2.1) is said to fulfill *Slater's Constraint Qualification* (Slater CQ) if

$$\exists\, \underline{x' \in \mathbb{R}^n} \text{ s.t } \underline{g_i(x') < 0},\ i = 1, \dots, m \text{ and } \underline{Ax' = b}.$$

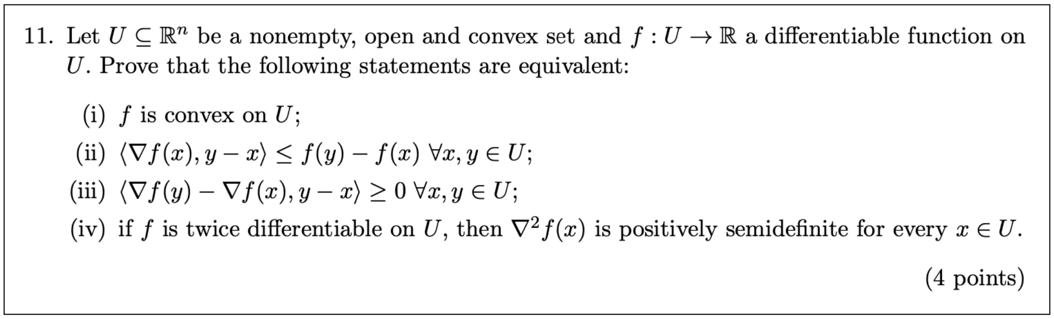
A very important observation is that, contrary to all the CQ we have considered so far, Slater CQ is a global CQ in the sense that it does not depend on an a priory given feasible element.

**Theorem 2.10.** *Let $x^*$ be a local (global) minimum of (2.1) and let Slater CQ be fulfilled. Then there exist (not necessarily unique) Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ such that $(x^*, \lambda^*, \mu^*)$ is a KKT point of (2.1).*

*Proof.* We will show that $x^*$ fulfills Abadie CQ, so, in particular, we will show $T_{lin}(x^*) \subseteq T_X(x^*)$. Let

$$d \in T_{lin}(x^*) = \left\{ \tilde{d} \in \mathbb{R}^n \middle| \begin{array}{l} \nabla g_i(x^*)^T \tilde{d} \leq 0, i \in \mathcal{A}(x^*) \\ A\tilde{d} = 0 \end{array} \right\}.$$

Let $x' \in X$ be such that $g_i(x') < 0 \; \forall i = 1, \dots, m$ and $Ax' = b$ (the existence of such an $x'$ is guaranteed by Slater CQ). Define $d' := x' - x^*$.



11. Let $U \subseteq \mathbb{R}^n$ be a nonempty, open and convex set and $f : U \to \mathbb{R}$ a differentiable function on $U$. Prove that the following statements are equivalent:

   (i) $f$ is convex on $U$;

   (ii) $\langle \nabla f(x), y - x \rangle \leq f(y) - f(x) \; \forall x, y \in U$;

   (iii) $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0 \; \forall x, y \in U$;

   (iv) if $f$ is twice differentiable on $U$, then $\nabla^2 f(x)$ is positively semidefinite for every $x \in U$.

(4 points)

Figure 3: Exercise 11.

Due to Exercise 11 (ii) (see Figure 3), we have for $i \in \mathcal{A}(x^*)$:

$$\nabla g_i(x^*)^T d' = \nabla g_i(x^*)^T (x^* + d' - x^*) \leq g_i(x^* + d') - g_i(x^*) = g_i(x') - g_i(x^*) < 0,$$

where the last inequality holds because $g_i(x') < 0$ by Slater CQ and $g_i(x^*) = 0$ because $i \in \mathcal{A}(x^*)$. In addition, we have

$$Ad' = Ax' - Ax^* = b - b = 0.$$

So in particular, $d' \in T_{lin}(x^*)$. Now define $d(\tau) := d + \tau d'$ for all $\tau > 0$. Then $d(\tau) \in T_{lin}(x^*)$ for all $\tau > 0$, because

$$\nabla g_i(x^*)^T d(\tau) = \nabla g_i(x^*)^T d + \tau \nabla g_i(x^*)^T d' < 0 \;\text{ for all } i \in \mathcal{A}(x^*),$$

and

$$Ad(\tau) = Ad + \tau Ad' = 0.$$

Furthermore, note that

$$d = \lim_{\tau \to 0} d(\tau).$$

We will show that $d(\tau)$ is in $T_X(x^*)$ for all $\tau > 0$, as this finishes the proof because

$T_X(x^*)$ is closed. Indeed, define

$$x^k := x^* + \frac{1}{k} d(\tau) \text{ and } t_k := \frac{1}{k},$$

which gives

$$\frac{x^k - x^*}{t_k} = d(\tau) \to d(\tau) \text{ as } k \to \infty.$$

To finish the proof, it remains to show that

$$x^k \in X = \left\{ \tilde{x} \in \mathbb{R}^n \,\middle|\, \begin{array}{l} g_i(\tilde{x}) \le 0, i = 1, \ldots, m \\ A\tilde{x} - b = 0 \end{array} \right\}$$

for all $k$ large enough. First, we have the following for $k \ge 1$:

$$Ax^k = Ax^* + \frac{1}{k} Ad(\tau) = b,$$

because $x^* \in X$ and $d(\tau) \in T_{lin}(x^*)$. Next, let $i \in \mathcal{A}(x^*)$. Then, by the Mean Value Theorem, we can find for each $k \ge 1$ a $\xi^k \in (x^*, x^k)$ such that

$$g_i(x^k) = g_i(x^k) - g_i(x^*) = \nabla g_i(\xi^k)^T (x^k - x^*) = \frac{1}{k} \nabla g_i(\xi^k)^T d(\tau).$$

This is equivalent to

$$k g_i(x^k) = \nabla g_i(\xi^k)^T d(\tau).$$

Letting $k$ go to infinity, we get the following, where the first inequality holds because $d(\tau) \in T_{lin}(x^*)$:

$$0 > \nabla g_i(x^*)^T d(\tau) = \lim_{k \to \infty} \nabla g_i(\xi^k)^T d(\tau) = \lim_{k \to \infty} k g_i(x^k).$$

This implies that there exists $k_0^i \ge 1$ such that

$$k g_i(x^k) < 0 \quad \forall k \ge k_0^i,$$

and therefore

$$g_i(x^k) < 0 \quad \forall k \ge k_0^i.$$

Lastly, let $i \in I(x^*)$. We know that $g_i(x^*) < 0$ and $x^k \to x^*$ as $k \to \infty$. Combining these two observations, we get

$$0 > g_i(x^*) = \lim_{k\to\infty} g_i(x^k).$$

This implies that there exists $k_1^i \geq 1$ such that

$$g_i(x^k) < 0 \quad \forall k \geq k_1^i.$$

Take

$$k_0 := \max_{i=1,\dots,m} \{k_0^i, k_1^i\}.$$

Then $x^k \in X$ for $k \geq k_0$, which implies $d(\tau) \in T_X(x^*)$ and therefore $d \in T_X(x^*)$. $\square$

We now know the following:

$$\begin{aligned} \text{LICQ} \;\Rightarrow\; \text{MFCQ} \;&\Rightarrow\; \text{Abadie CQ} \\ \text{Slater CQ} \;&\Rightarrow\; \text{Abadie CQ}\,, \end{aligned}$$

but in general, there is no relation among LICQ, MFCQ and Slater CQ (see exercise session). Next we will show that in the convex setting the KKT optimality conditions are actually sufficient of optimality.

**Theorem 2.11.** *Let $(x^*, \lambda^*, \mu^*)$ be a KKT point of (2.1). Then $x^*$ is a local (global) minimum of (2.1).*

*Proof.* Let $x \in X$. Since $(x^*, \lambda^*, \mu^*)$ is a KKT point of (2.1), we have

$$\nabla f(x^*)^T = \left( -\sum_{i=1}^{n} \lambda_i^* (\nabla g_i(x^*))^T - (\mu^*)^T A \right).$$

Then by Exercise 11 (ii) (see Figure 3):

$$\begin{aligned} f(x) &\geq f(x^*) + \nabla f(x^*)^T (x - x^*) \\ &= f(x^*) + \left( -\sum_{i=1}^{n} \lambda_i^* (\nabla g_i(x^*))^T - (\mu^*)^T A \right) (x - x^*) \\ &= f(x^*) - \sum_{i\in\mathcal{A}(x^*)} \lambda_i^* (\nabla g_i(x^*))^T (x - x^*) - (\mu^*)^T (Ax - Ax^*), \end{aligned}$$

where the second equality follows because for $i$ inactive, we have $\lambda_i = 0$. Again by Exercise 11, we have

$$\nabla g_i(x^*)^T(x - x^*) \leq g_i(x) - g_i(x^*)$$

and furthermore it holds that $Ax - Ax^* = b - b = 0$. Thus, we have the following, where we use that $g_i(x^*) = 0$ for $i$ active:

$$\begin{aligned} f(x) &\geq f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^*(g_i(x) - g_i(x^*)) \\ &= f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^*(g_i(x)) \geq f(x^*), \end{aligned}$$

where the last inequality follows because for $i$ active, we have $\lambda_i^* \geq 0$ and $g_i(x) \leq 0$. □

# 3 Second order necessary and sufficient optimality conditions

## 3.1 The unconstrained case

**Theorem 3.1.** *Let $x^*$ be a local minimum of $f : \mathbb{R}^n \to \mathbb{R}$ and let $f$ be twice continuously differentiable in a neighborhood $U_\varepsilon(x^*)$ of $x^*$. Then the Hessian $\nabla^2 f(x^*) \in \mathbb{R}^{n\times n}$ is positive semidefinite.*

*Proof.* We assume that there exists $d \in \mathbb{R}^n$ such that $d^T\nabla^2 f(x^*)d < 0$. We know that the map $x \mapsto d^T\nabla^2 f(x)d$ is continuous at $x^*$ by assumption. Let $\alpha > 0$ be such that for all $t \in [0, \alpha]$:

$$x^* + td \in U_\varepsilon(x^*) \text{ and } d^T\nabla^2 f(x^* + td)d < 0. \tag{*}$$

Define the function $F : [0, \alpha] \to \mathbb{R}$ by $F(t) = f(x^* + td)$. Then the Taylor formula[1] applied to $F$ around $x_0 = 0$ implies that for all $t \in [0, \alpha]$ there is a $\xi_t \in (0, t)$ such that

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + \frac{1}{2}t^2 d^T\nabla^2 f(x^* + \xi_t d)d.$$

By Theorem (1.4) and (*), we have

$$f(x^* + td) = f(x^*) + \frac{1}{2}t^2 d^T\nabla^2 f(x^* + \xi_t d)d < f(x^*)$$

for all $t \in [0, \alpha]$, which contradicts the minimality of $x^*$. □

**Example 3.2.** We want to find the local minima of the following functions.

a) Take

$$f : \mathbb{R} \to \mathbb{R}, \; f(x) = -x^2.$$

First, inspired by Theorem (1.4), we solve the following equation

$$f'(x) = 0 \;\Leftrightarrow\; x^* = 0.$$

[1] Recall the *Taylor formula* for a twice differentiable function $f : I \to \mathbb{R}$ around a point $x_0 \in I$ for an interval $I \subseteq \mathbb{R}$. It says that for any $t \in I$, with $t \neq x_0$, there exist $\xi_t \in (x_0, t)$ such that

$$f(t) = f(x_0) + f'(x_0)(t - x_0) + \frac{1}{2}f''(\xi_t)(t - x_0)^2.$$

Then we check whether $f''(x^*) \geq 0$, but since $-2 < 0$, we conclude that $x^*$ is not a local minimum and therefore $f$ has no local minima.

b) Take
$$f : \mathbb{R}^2 \to \mathbb{R},\ f(x_1, x_2) = x_1^2 - x_2^3.$$
It holds
$$\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 \\ -3x_2^2 \end{pmatrix} \text{ and } \nabla^2 f(x_1, x_2) = \begin{pmatrix} 2 & 0 \\ 0 & -6x_2 \end{pmatrix}.$$
Similarly as before, we solve
$$\nabla f(x_1, x_2) = 0 \ \Leftrightarrow \ x^* = (0, 0)^T$$
and we check that
$$\nabla^2 f(x^*) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$
is positive semidefinite (because all its eigenvalues are nonnegative). But $x^*$ is still not a local minimum, since for all $\varepsilon > 0$ we can find the point $(0, \varepsilon/2) \in U_\varepsilon(0, 0)$ and
$$f(0, \varepsilon/2) = -\frac{\varepsilon^3}{8} < f(0, 0) = 0.$$

**Theorem 3.3.** *Let $f$ be twice continuous differentiable in a neighborhood $U_\varepsilon(x^*)$ of an element $x^* \in \mathbb{R}^n$ fulfilling*

*a)* $\nabla f(x^*) = 0$, *and*

*b)* $\nabla^2 f(x^*)$ *is positive definite.*

*Then $x^*$ is a strict local minimum of $f$, namely there exists a $\delta > 0$ such that*
$$f(x^*) < f(x) \quad \forall x \in U_\delta(x^*), x \neq x^*.$$

*Proof.* Since $\nabla^2 f(x^*)$ is positive definite, we know that its smallest eigenvalue $\lambda_{min}(\nabla^2 f(x^*))$ is strictly positive. Furthermore, we know from linear algebra that for all $d \in \mathbb{R}^n$ it holds
$$d^T \nabla^2 f(x^*) d \geq \lambda_{min}(\nabla^2 f(x^*)) \left\| d \right\|. \tag{*}$$

By the continuity of $\nabla^2 f(x^*)$, we can find a $\delta < \varepsilon$ (Q: what is $\varepsilon$?) such that for all $d \in \mathbb{R}^n$ with $\|d\| < \delta$ the following holds:

$$\left\|\nabla^2 f(x^* + d) - \nabla^2 f(x^*)\right\| < \frac{\lambda_{min}(\nabla^2 f(x^*))}{2}, \tag{**}$$

where the norm indicates the operator norm[2]. Take $d \in \mathbb{R}^n$ such that $\|d\| < \delta$. Define the function $F : [0,1] \to \mathbb{R}$ by $F(t) = f(x^* + td)$. Then plugging in $t = 1$ in the Taylor formula applied to $F$ around $x_0 = 0$ implies that there exists a $\xi_d \in (0,1)$ such that:

$$\begin{aligned} f(x^* + d) &= f(x^*) + \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^* + \xi_d d) d \\ &= f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^* + \xi_d d) d \\ &= f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{1}{2} d^T \nabla^2 f(x^* + \xi_d d) d - \frac{1}{2} d^T \nabla^2 f(x^*) d \\ &= f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{1}{2} d^T \Big(\nabla^2 f(x^* + \xi_d d) - \nabla^2 f(x^*)\Big) d. \end{aligned}$$

Note that by the Cauchy-Schwarz inequality, we have the following property for $d \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$:

$$\frac{1}{2} d^T A d \geq -\frac{1}{2} |d^T A d| \geq -\frac{1}{2} \|d\| \|Ad\| \geq -\frac{1}{2} \|d\| \|A\| \|d\| = -\frac{1}{2} \|d\|^2 \|A\|. \tag{***}$$

Combining (*),(***), (**) and using that $\|\xi_d d\| \leq \|d\| < \delta$, we get

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{1}{2} \lambda_{min}(\nabla^2 f(x^*)) \|d\|^2 - \frac{1}{2} \|d\|^2 \left\|\nabla^2 f(x^* + \xi_d d) - \nabla^2 f(x^*)\right\| \\ &\geq f(x^*) + \frac{1}{2} \lambda_{min}(\nabla^2 f(x^*)) \|d\|^2 - \frac{1}{4} \lambda_{min}(\nabla^2 f(x^*)) \|d\|^2 \\ &= f(x^*) + \frac{1}{4} \lambda_{min}(\nabla^2 f(x^*)) \|d\|^2. \end{aligned}$$

[2] The *operator norm* is defined as

$$\begin{aligned} \|\cdot\| : \mathbb{R}^{n \times n} &\to \mathbb{R} \\ A &\mapsto \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \end{aligned}$$

Plugging in $d := x - x^*$, we get for all $x \in U_\delta(x^*)$:

$$f(x) \geq f(x^*) + \frac{1}{4}\lambda_{min}(\nabla^2 f(x^*))||x - x^*||^2 > f(x^*)$$

for all $x \neq x^*$. □

**Remark.** We don't have an equivalence in Theorem (3.3): $f(x) = x^4$ has $f''(0) = 0$ but $0$ is still a strict local minimum of $f$.

## 3.2 The constrained case

We consider problem 1.5 under the assumption that $f, g_i, h_j$ are twice continuously differentiable. For a KKT point (which we assume exists) $(x^*, \lambda^*, \mu^*)$ of (1.5), define the following two subsets of $\mathcal{A}(x^*)$:

$$\begin{aligned} \mathcal{A}_0(x^*) &= \{i \in \mathcal{A}(x^*) : \lambda_i^* = 0\} \text{ the set of \textit{weak active indices}, and} \\ \mathcal{A}_>(x^*) &= \{i \in \mathcal{A}(x^*) : \lambda_i^* > 0\} \text{ the set of \textit{strong active indices}.} \end{aligned}$$

We consider the following subset of $T_{lin}(x^*)$.

$$T_2(x^*) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \nabla g_i(x^*)^T d = 0, i \in \mathcal{A}_>(x^*) \\ \nabla g_i(x^*)^T d \leq 0, i \in \mathcal{A}_0(x^*) \\ \nabla h_j(x^*)^T d = 0, j = 1, \ldots, p \end{array} \right\}$$

**Theorem 3.4.** *Let $x^*$ be a local minimum of 1.5 which fulfills (LICQ). Let $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ be the (according to Theorem (2.8)) uniquely defined Lagrange multipliers such that $(x^*, \lambda^*, \mu^*)$ is a KKT point of (1.5). Then it holds for all $d \in T_2(x^*)$:*

$$d^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d \geq 0,$$

*in other words, $\nabla^2_{xx} L(x^*, \lambda^*, \mu^*)$ is positive semidefinite on $T_2(x^*)$.*

*Proof.* Let $d \in T_2(x^*), d \neq 0$. We split $\mathcal{A}_0(x^*)$ into

$$\begin{aligned} \mathcal{A}_0^<(x^*) &= \{i \in \mathcal{A}_0(x^*) : \nabla g_i(x^*)^T d < 0\} \\ \mathcal{A}_0^=(x^*) &= \{i \in \mathcal{A}_0(x^*) : \nabla g_i(x^*)^T d = 0\}. \end{aligned}$$

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

$$\tilde{g}(x) := \begin{pmatrix} g_i(x)|_{i \in I(x^*)} \\ g_i(x)|_{i \in \mathcal{A}_0^<(x^*)} \end{pmatrix} \text{ and } \tilde{h}(x) := \begin{pmatrix} h_j(x)|_{i=1,\ldots,p} \\ g_i(x)|_{i \in \mathcal{A}_>(x^*)} \\ g_i(x)|_{i \in \mathcal{A}_0^=(x^*)} \end{pmatrix}.$$

Furthermore, we define

$$\tilde{X} := \{x \in \mathbb{R}^n : \tilde{g}(x) \leqq 0, \tilde{h}(x) = 0\} \subseteq X.$$

Then, $x^* \in \tilde{X}$ (this is easy to check). Furthermore, $x^*$ fulfills (MFCQ) for $\tilde{X}$ and the $d$ fixed above, because

a) The set
$$\{\nabla h_j(x^*)\}_{j=1}^p \cup \{\nabla g_i(x^*)\}_{i \in \mathcal{A}_>(x^*)} \cup \{\nabla g_i(x^*)\}_{i \in \mathcal{A}_0^=(x^*)}$$
is linearly independent, as it is a subset of
$$\{\nabla h_j(x^*)\}_{j=1}^p \cup \{\nabla g_i(x^*)\}_{i \in \mathcal{A}(x^*)}$$
which is linearly independent by LICQ.

b) $$\begin{cases} \nabla g_i(x^*)^T d < 0 & \forall i \in \mathcal{A}_0^<(x^*) \\ \nabla h_j(x^*)^T d = 0 & \forall j = 1, \ldots, p \quad \leftarrow d \in T_2(x^*) \\ \nabla g_i(x^*)^T d = 0 & \forall i \in \mathcal{A}_>(x^*) \quad \leftarrow d \in T_2(x^*) \\ \nabla g_i(x^*)^T d = 0 & \forall i \in \mathcal{A}_0^=(x^*) \quad \leftarrow \text{definition of } \mathcal{A}_0^=(x^*) \end{cases}$$

Therefore, by Lemma (2.5) there is a $\varepsilon > 0$ and $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ with $x$ twice continuous differentiable (plug in the proof of Lemma (2.5) that the Hessian is continuous) such that $x(0) = x^*$, $\dot{x}(0) = d$ and $x(t) \in \tilde{X}$ for all $t \in [0, \varepsilon)$. Define

$$\varphi : (-\varepsilon, \varepsilon) \to \mathbb{R},\ \varphi(t) = L(x(t), \lambda^*, \mu^*),$$

where $L$ is the Lagrangian. Note that $\varphi$ is twice continuously differentiable because $f$ and $x$ are. Then we have

$$\begin{aligned} &\dot{\varphi}(t) = \nabla_x L(x(t), \lambda^*, \mu^*)^T \dot{x}(t), \text{ and} \\ &\ddot{\varphi}(t) = \dot{x}(t)^T \nabla_{xx} L(x(t), \lambda^*, \mu^*)^T \dot{x}(t) + \nabla_x L(x(t), \lambda^*, \mu^*) \ddot{x}(t). \end{aligned}$$

Since $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$, this gives

$$\dot{\varphi}(0) = 0 \text{ and } \ddot{\varphi}(0) = d^T \nabla_{xx} L(x^*, \lambda^*, \mu^*) d$$

Our goal is to show that $\ddot{\varphi}(0) \geq 0$. Note the following four observations:

(i) We have

$$\sum_{i \in \mathcal{A}_>(x^*)} \lambda_i^* g_i(x(t)) = 0$$

because $g_i(x(t)) = 0$ for every $i \in \mathcal{A}_>(x^*)$, since $x(t) \in \tilde{X}$.

(ii) We have

$$\sum_{i \in \mathcal{A}_0(x^*)} \lambda_i^* g_i(x(t)) = 0$$

because $\lambda_i^* = 0$ for every $i \in \mathcal{A}_0(x^*)$.

(iii) We have

$$\sum_{i \in I(x^*)} \lambda_i^* g_i(x(t)) = 0$$

because $\lambda_i^* = 0$ for every $i \in I(x^*)$, by definition of a KKT point and of $I(x^*)$.

(iv) We have

$$\sum_{j=1}^{p} \mu_i^* h_j(x(t)) = 0$$

because $h_j(x(t)) = 0$ for every $j = 1, \ldots, p$, since $x(t) \in \tilde{X}$.

Therefore, by definition of the Lagrangian we have for all $t \in [0, \varepsilon)$ that

$$\begin{aligned} \varphi(t) &= L(x(t), \lambda^*, \mu^*) \\ &= f(x(t)) + \sum_{i \in \mathcal{A}_>(x^*)} \lambda_i^* g_i(x(t)) + \sum_{i \in \mathcal{A}_0(x^*)} \lambda_i^* g_i(x(t)) + \sum_{i \in I(x^*)} \lambda_i^* g_i(x(t)) + \sum_{j=1}^{p} \mu_i^* h_j(x(t)) \\ &= f(x(t)) \qquad (*) \end{aligned}$$

Now, because $x(0) = x^*$ is a local minimum of $f$ with respect to $X$, there exists $\delta > 0$ such that

$$f(z) \geq f(x^*) \text{ for all } z \in U_\delta(x^*) \cap X.$$

Since $x$ is continuous and (clearly) $x(0) = x^* \in U_\delta(x^*) \cap \tilde{X}$, there exists $\alpha \in (0, \varepsilon)$ such that for all $t \in [0, \alpha)$ it holds that

$$x(t) = x(0 + t) \in U_\delta(x^*) \cap \tilde{X}.$$

Then for all $t \in [0, \alpha)$ the following holds:

$$f(x(t)) \geq f(x^*),$$

which by (*) is equivalent to

$$\varphi(t) \geq \varphi(0) \text{ for all } t \in [0, \alpha). \tag{**}$$

Assume that $\ddot{\varphi}(0) < 0$. Then, since $\varphi$ is continuous, there exist $\beta \in (0, \alpha)$ such that $\ddot{\varphi}(t) < 0$ for all $t \in [0, \beta)$. Applying Taylor's formula yields a $t_\beta \in (0, \beta)$ such that

$$\varphi(\beta) = \varphi(0) + \beta\dot{\varphi}(0) + \frac{1}{2}\beta^2\ddot{\varphi}(t_\beta) = \varphi(0) + \frac{1}{2}\beta^2\ddot{\varphi}(t_\beta) \Rightarrow \varphi(\beta) < \varphi(0),$$

because $\dot{\varphi}(0) = 0$ and $\ddot{\varphi}(t_\beta) < 0$. This is a contradiction to (**). □

The following result provides a second order *sufficient* optimality condition.

**Theorem 3.5.** *Let $(x^*, \lambda^*, \mu^*)$ be a KKT-point of (1.5) such that for all $d \in T_2(x^*) \setminus \{0\}$, it holds*

$$d^T \nabla_{xx} L(x^*, \lambda^*, \mu^*) d > 0,$$

*i.e., $\nabla_{xx} L(x^*, \lambda^*, \mu^*)$ is positive definite on $T_2(x^*)$. Then $x^*$ is a strict local minimum of (1.5), namely there exists $\varepsilon > 0$ such that*

$$f(x) > f(x^*) \; \forall x \in U_\varepsilon(x^*) \cup X, x \neq x^*.$$

*Proof.* Assume that $x^*$ is not a strict local minimum. Then we have for all $k \geq 1$:

$$\exists \, x^k \in U_{\frac{1}{k}}(x^*) \cap X : x^k \neq x^*, f(x^k) \leq f(x^*).$$

We define for all $k \geq 1$:

$$d^k := \frac{1}{||x^k - x^*||}(x^k - x^*).$$

Then, $\|d_k\| = 1$ for all $k \geq 1$ and, as $(d^k)_{k \geq 1}$ is a bounded sequence, there is a

$d^* \in \mathbb{R}^n$ with $\|d^*\| = 1$ and a subsequence

$$(d^{k_l})_{l\geq 1} \subseteq (d^k)_{k\geq 1} \text{ such that } \lim_{l\to\infty} d^{k_l} = d^*.$$

We will show that $d^* \in T_2(x^*) \setminus \{0\}$ and that $(d^*)^T \nabla_{xx} L(x^*, \lambda^*, \mu^*)d \leq 0$, which is a contradiction and will prove the theorem.

First, we will prove that $d^* \in T_{lin}(x^*)$. Let $j = 1, \ldots, p$. Then, by the Mean Value Theorem, we know that for all $k \geq 1$ there exists $\xi^k \in (x^*, x^k)$ such that the following holds (because $x^k, x^* \in X$):

$$\begin{aligned} 0 &= h_j(x^k) - h_j(x^*) = \nabla h_j(\xi^k)^T (x^k - x^*) \\ \Rightarrow\ 0 &= \nabla h_j^T(\xi^k) \frac{(x^k - x^*)}{\|(x^k - x^*)\|} \end{aligned}$$

So in particular, we have the following for all $l \geq 1$:

$$0 = \nabla h_j^T(\xi^{k_l}) \frac{(x^{k_l} - x^*)}{\|(x^{k_l} - x^*)\|} \to \nabla h_j(x^*)^T d^*$$

and so we get $\nabla h_j(x^*)^T d^* = 0$. Now let $i \in \mathcal{A}(x^*)$. Then, again by the Mean Value Theorem, we know that for all $k \geq 1$ there exists $\xi^k \in (x^*, x^k)$ such that the following holds:

$$0 \geq g_i(x^k) = g_i(x^*) + \nabla g_i(\xi^k)^T (x^k - x^*) = \nabla g_i(\xi^k)^T (x^k - x^*).$$

The same argument as above proves that $\nabla g_i(x^*)^T d^* \leq 0$.

Next, we will prove that $d^* \in T_2(x^*) \setminus \{0\}$. Since $\|d^*\| = 1$, we know that $d^* \neq 0$. We only need to show that for all $i \in \mathcal{A}_>(x^*)$, we have

$$\nabla g_i(x^*)^T d = 0.$$

Assume that there is a $\tilde{i} \in \mathcal{A}_>(x^*)$ such that $\nabla g_{\tilde{i}}(x^*)^T d^* < 0$. First, we note that by the Mean Value Theorem, we know that for all $k \geq 1$ there exists $\xi^k \in (x^*, x^k)$ such that the following holds:

$$0 \geq f(x^k) - f(x^*) = \nabla f(\xi^k)^T (x^k - x^*),$$

and again by the same argument as above, we conclude that

$$\nabla f(x^*)^T d^* \leq 0.$$

Recall:

(i) by definition of $\mathcal{A}_0(x^*)$ and KKT point, we have $\lambda_i = 0$ for $i \in \mathcal{A}_0(x^*)$ and $i \in I(x^*)$,

(ii) we already showed above that $\nabla h_j(x^*)^T d^* = 0$ for $j = 1, \dots, p$,

(iii) we also showed above that $\nabla g_i(x^*)^T d^* \leq 0$ for all $i \in \mathcal{A}(x^*)$, and in particular for all $i \in \mathcal{A}_>(x^*)$.

Then we have the following:

$$\begin{aligned} 0 &\geq \nabla f(x^*)^T d^* \\ &= - \sum_{i \in \mathcal{A}_0(x^*)} \lambda_i \nabla g_i(x^*)^T d^* - \sum_{i \in \mathcal{A}_>(x^*)} \lambda_i \nabla g_i(x^*)^T d^* - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*)^T d^* - \sum_{j=1}^{p} \mu_j \nabla h_j(x^*)^T d^* \\ &= - \sum_{i \in \mathcal{A}_>(x^*)} \lambda_i \nabla g_i(x^*)^T d^* \geq \lambda_{\tilde{i}} \nabla g_{\tilde{i}}(x^*)^T d^* > 0 \end{aligned}$$

which gives a contradiction. The second to last inequality follows because $\lambda_i \nabla g_i(x^*)^T d^* \leq 0$ for all $i \in \mathcal{A}_>(x^*)$.

Last, we will show that $(d^*)^T \nabla_{xx} L(x^*, \lambda^*, \mu^*) d^* \leq 0$. By Taylor's formula, we know that for all $k \geq 1$ there exists $\xi^k \in (x^*, x^k)$ such that:

$$L(x^k, \lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*) + \nabla_x L(x^*, \lambda^*, \mu^*)^T (x^k - x^*) + \frac{1}{2}(x^k - x^*)^T \nabla_{xx} L(\xi^k, \lambda^*, \mu^*)(x^k - x^*).$$

Since $(x^*, \lambda^*, \mu^*)$ is a KKT point, it holds that $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ and therefore

$$L(x^k, \lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*) + \frac{1}{2}(x^k - x^*)^T \nabla_{xx} L(\xi^k, \lambda^*, \mu^*)(x^k - x^*). \qquad (*)$$

Furthermore, again since $(x^*, \lambda^*, \mu^*)$ is a KKT point, it holds that $L(x^*, \lambda^*, \mu^*) = f(x^*)$ and therefore

$$L(x^*, \lambda^*, \mu^*) = f(x^*) \geq f(x^k) \geq f(x^k) + \sum_{i=1}^{m} \lambda_i^* g_i(x^* k) + \sum_{j=1}^{p} \mu_j^* h_h(x^*) = L(x^k, \lambda^*, \mu^*).$$

Then, from (*) and by dividing by $\left\| x^k - x^* \right\|$, we get

$$0 \geq \frac{1}{2} \frac{(x^k - x^*)^T}{\|x^k - x^*\|} \nabla_{xx} L(\xi^k, \lambda^*, \mu^*) \frac{x^k - x^*}{\|x^k - x^*\|}.$$

So in particular, we have the following for all $l \geq 1$:

$$0 \geq \frac{1}{2} \frac{(x^{k_l} - x^*)^T}{\|x^{k_l} - x^*\|} \nabla_{xx} L(\xi^{k_l}, \lambda^*, \mu^*) \frac{x^{k_l} - x^*}{\|x^{k_l} - x^*\|} \;\rightarrow\; (d^*)^T \nabla_{xx} L(x^*, \lambda^*, \mu^*) d^*,$$

which shows that $(d^*)^T \nabla_{xx} L(x^*, \lambda^*, \mu^*) d^* \leq 0$. □

# 4 A general descent algorithm

In the following, we will discuss a general descent algorithm for solving *unconstrained* optimization problems of the form

$$\min_{x \in \mathbb{R}^n} f(x), \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

**Definition 4.1.** A vector $d \in \mathbb{R}^n$ is called a *descent direction* of $f$ at $x \in \mathbb{R}^n$ if

$$\exists\, \bar{t} > 0 \text{ s.t. } f(x + td) < f(x) \quad \forall\, t \in (0, \bar{t}]. \tag{4.2}$$

The descent direction is a local object (i.e., it is defined at a point $x \in \mathbb{R}^n$). Furthermore, observe that we do not only require $f(x + td) < f(x)$ on one point, but on a whole interval.

**Lemma 4.2.** *Let $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$. Then $d$ is a descent direction of $f$ at $x$.*

*Proof.* Recall, by the definition of directional derivative, that

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

Therefore, there exists $\bar{t} > 0$ such that, for all $t \in (0, \bar{t}]$, we have

$$\frac{f(x + td) - f(x)}{t} < 0 \Leftrightarrow f(x + td) < f(x).$$

□

**Remark.** Asking for $\nabla f(x)^T d < 0$ is equivalent to asking for the angle between $-\nabla f(x)$ and $d$ to be (strictly) acute, as we have:

$$\cos(\sphericalangle(-\nabla f(x), d)) = -\frac{\nabla f(x)^T d}{\|\nabla f(x)\| \, \|d\|} > 0.$$

**Example 4.3.** The condition $\nabla f(x)^T d < 0$ is not necessary for $d$ being a descent direction. Take, for example, $f(x) = -x^2$ and $x = 0$. Then every $d \in \mathbb{R} \setminus \{0\}$ is a descent direction of $f$ at $x = 0$ (see Figure 4), but $f'(0)d = 0$.

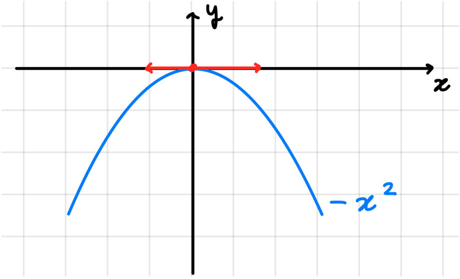


Figure 4: Every $d \in \mathbb{R} \setminus \{0\}$ is a descent direction of $f$ at $x = 0$

**Remark 4.4.** Let $x$ be *not* a critical point for $f$, meaning $\nabla f(x) \neq 0$. Then $d := -\nabla f(x)$ is a descent direction of $f$ at $x$, since

$$\nabla f(x)(-\nabla f(x)) = -\left\|\nabla f(x)\right\|^2 < 0.$$

Moreover, if $B \in \mathbb{R}^{n\times n}$ is a positive definite matrix, then $d := -B\nabla f(x)$ is also a descent direction of $f$ at $x$, since

$$\nabla f(x)^T(-B\nabla f(x)) = -\nabla f(x)^T B\nabla f(x) < 0.$$

**Algorithm 1** Line search algorithm

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, set $k := 0$
2: If $x^k$ fulfills a stopping criterion, stop
3: Find a descent direction $d^k$ of $f$ at $x^k$
4: Find a step size $t_k > 0$ such that $f(x^k + t_k d^k) < f(x^k)$
5: Set $x^{k+1} := x^k + t_k d^k$, set $k := k + 1$ and go to Step 2
```

The line search algorithm has two degrees of freedom: the choice of descent direction (Step 3) and the choice of step size (Step 4).

**Example 4.5.** This example is temporarily here to fix the numbering of the theorems (the line search algorithm should be Algorithm 4.5).

**Example 4.6.** Unfortunately, not every step size is good. Take, for example,

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = x^2.$$

Then $x^* = 0$ is a global minimum. For all $x > 0$, take $d = -1$ as descent direction of $f$ at $x$. Define, for all $k \geq 0$, $t_k = 2^{-(k+2)}$. Let $x^0 = 1$. Then we have for all $k \geq 0$:

$$\begin{aligned} x^{k+1} &= x^k - t_k = \ldots = x^0 - t_0 - t_1 - \ldots - t_k \\ &= x^0 - \sum_{i=0}^{k} \frac{1}{2^{i+2}} = \frac{1}{2} + \frac{1}{2^{k+2}} \to \frac{1}{2}\ (k \to \infty). \end{aligned}$$

But $\frac{1}{2}$ is not the global minimum!

**Definition 4.7 (Step size strategy).** (a) A set-valued mapping

$$T : \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows (0, +\infty)$$

which assigns to each pair $(x, d) \in \mathbb{R}^n \times \mathbb{R}^n$ a set of step sizes $T(x, d) \subseteq (0, +\infty)$ is called *step size strategy*.

(b) A step size strategy is called *well-defined* if, for every $(x, d) \in \mathbb{R}^n \times \mathbb{R}^n$ fulfilling $\nabla f(x)^T d < 0$, it holds that $T(x, d) \neq \emptyset$.

(c) A step size strategy is called *efficient* if there exists $\theta > 0$ such that, for every $(x, d) \in \mathbb{R}^n \times \mathbb{R}^n$ where $d$ is a descent direction of $f$ at $x$, it holds

$$f(x + td) \leq f(x) - \theta \left( \frac{f(x)^T d}{\|d\|} \right)^2 \quad \forall t \in T(x, d).$$

In this case, every step size $t \in T(x, d)$ is called efficient.

**Remark.** For intuition (as far as I understood this, pls correct me if I'm wrong), the pair $(x, d)$ will be the current iterate and we will choose a step size in $T(x, d)$ to reach the next iterate. The double arrow in (a) indicates that the values of $T$ are sets. Condition (b) guarantees, that if $d$ is a descent direction, the strategy gives us some $t \in \mathbb{R}_+$, which doesn't have to be the case if $d$ is not a descent direction. The constant $\theta > 0$ in (c) does not depend on $(x, d)$.

In the following, we assume (the worst case scenario) that Algorithm 1 does not stop after finitely many iterations. This means that we generate an infinite sequence of iterates.

**Theorem 4.8.** *Let $f$ be continuously differentiable and $(x^k)_{k \geq 0}$ a sequence generated by Algorithm 1 such that*

(a) *the so-called* angle condition *holds, i.e.*

$$\sphericalangle(\nabla f(x^k), d^k) \notin (\pi/2, 3/2\pi)$$

$$\exists\, c > 0 \; s.t. \; \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \, \|d^k\|} \geq c \quad \forall\, k \geq 0, \tag{4.3}$$

(b) *the sequence of step sizes $t_k$ is efficient.*

*Then, every accumulation point of $(x^k)_{k\geq 0}$ is a critical point of $f$.*

**Remark.** The angle condition states that $\sphericalangle(-\nabla f(x^k), d^k)$ stays uniformly away from 90° (Insert picture, see video on moodle minute 26). Furthermore, we are not claiming that the whole sequence $(x^k)_{k\geq 0}$ converges (that is, in general, not the case) - but only that every convergent subsequence of $(x^k)_{k\geq 0}$ converges to a critical point of $f$. Lastly, we are not assuming that $(x^k)_{k\geq 0}$ *has* an accumulation point. For that to be the case, we would need to assume that $f$ is coercive (but this is just a fun fact and not relevant for our statement).

*Proof.* Since $t_k$ is efficient, we have for all $k \geq 0$:

$$f(x^{k+1}) = f(x^k + t_k d^k) \leq f(x^k) - \theta \left( \frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)^2 .$$

By the angle condition, we have for all $k \geq 0$:

$$\left( \frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)^2 \geq c^2 \left\| \nabla f(x^k) \right\|^2 .$$

Combining these two observations, we get for all $k \geq 0$:

$$f(x^{k+1}) \leq f(x^k) - c^2\theta \left\| \nabla f(x^k) \right\|^2 \leq f(x^k). \tag{*}$$

$\Rightarrow$ non-increasing

Therefore, the sequence $(f(x^k))_{k\geq 0}$ is non-increasing.

Let us choose an accumulation point of the sequence $(x^k)_{k\geq 0}$, meaning that there exists a subsequence $(x^{k_l})_{k\geq 0)}$ such that $x^{k_l} \to x^*$ $(l \to \infty)$, which implies, by continuity of $f$, that

$$f(x^{k_l}) \to f(x^*) \; (l \to \infty).$$

Combining this last observation with the fact that $(f(x^k))_{k\geq 0}$ is non-increasing, we can conclude (by Exercise 18) that

$$f(x^k) \to f(x^*) \; (k \to \infty).$$

From (*), we get for all $k \geq 0$:

$$c^2\theta \left\|\nabla f(x^k)\right\|^2 \leq f(x^k) - f(x^{k+1}) \to 0 \ (k \to \infty).$$

Since the LHS is non-negative, we get

$$\left\|\nabla f(x^k)\right\| \to 0 \ (k \to \infty) \text{ and therefore, } \left\|\nabla f(x^{k_l})\right\| \to 0 \ (l \to \infty).$$

By continuity of the gradient, we know that $\left\|\nabla f(x^{k_l})\right\| \to \|\nabla f(x^*)\| \ \ (l \to \infty)$ and we can conclude $\nabla f(x^*) = 0$. □

**Remark 4.9.** (a) The angle condition is fulfilled for $d^k = -\nabla f(x^k)$ and every $c \in (0, 1]$.

(b) How to choose an "optimal" step size? One could, for example, consider the *minimization rule*, which consists in choosing $t = t_{\min}$ such that

$$f(x + t_{\min} d) = \min_{t>0} f(x + td).$$

This rule is, under certain assumptions, well-defined and efficient. However, the step size cannot always be calculated explicitly. One exception is for

$$f : \mathbb{R}^n \to \mathbb{R}\,,\, f(x) = \frac{1}{2}x^T Ax - b^T x,$$

with $A \in \mathbb{R}^{n\times n}$ symmetric and positive definite and $b \in \mathbb{R}^n$. Indeed, for $x, d \in \mathbb{R}^n$ with $\nabla f(x)^T d < 0$, the step size given by the minimization rule is well-defined, efficient and can be explicitly calculated (see Exercise 21):

$$t_{\min} = -\frac{\nabla f(x)^T d}{d^T Ad}.$$

# 5 Step Size Strategies

In this section, we will discuss three "popular" step size strategies for $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable and

$$x, d \in \mathbb{R}^n \text{ such that } \nabla f(x)^T d < 0. \tag{5.1}$$

## 5.1 The Wolfe-Powell step size strategy

Let $\sigma \in (0, \frac{1}{2})$ and $\rho \in [\sigma, 1)$. The Wolfe-Powell step size strategy consists in finding $t > 0$ such that

$$f(x + td) \leq f(x) + \sigma t \nabla f(x)^T d \tag{5.2}$$

and

$$\nabla f(x + td)^T d \geq \rho \nabla f(x)^T d. \tag{5.3}$$

To this end, we define $\phi(t) := f(x + td)$ (keep in mind, $x, d$ are fixed vectors). It follows that $\phi'(t) = \nabla f(x + td)^T d$ and $\phi'(0) = \nabla f(x)^T d < 0$. Then we can reformulate the conditions above:

$$\begin{aligned}
(5.2) \;&\Leftrightarrow\; \phi(t) \leq \phi(0) + \sigma t \phi'(0) \\
(5.3) \;&\Leftrightarrow\; \phi'(t) \geq \rho \phi'(0)
\end{aligned}$$

$\phi(t) = f(x + td)$

Insert and explain picture, see video on moodle 1h6min to 1h14min.

The step size $t$ is chosen in a set on which the following two conditions are satisfied:

- the graph of $\phi$ lies below the Armijo-Goldstein line, and
- the graph of $\phi$ is decreasing less steep than at 0 (and less/ less equal steep than the Armijo-Goldstein line) or is even increasing.

**Definition.** Let $x \in \mathbb{R}^n$. Define the lower level set of $f$ at $x$ as:

$$\mathcal{L}(x) = \{z \in \mathbb{R}^n : f(z) \leq f(x)\}$$

**Definition.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ and $x^0 \in \mathbb{R}^n$. For $x \in \mathcal{L}(x^0)$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, let

$$\boxed{T_{WP}(x, d) = \{t > 0 : f(x + td) \leq f(x) + \sigma t \nabla f(x)^T d \text{ and } \nabla f(x + td)^T d \geq \rho \nabla f(x)^T d\}}$$

be the set of Wolfe-Powell step size strategies in $x$ in direction $d$.

**Theorem 5.1.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ and $x^0 \in \mathbb{R}^n$. For $x \in \mathcal{L}(x^0)$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, let $T_{WP}$ be the set of Wolfe-Powell step size strategies in $x$ in direction $d$. Then it holds*

*(a) If $f$ is bounded from below, then $T_{WP}(x, d)$ is nonempty (or, equivalently, the strategy is well-defined).*

*(b) If $\nabla f$ is Lipschitz-continuous on $\mathcal{L}(x^0)$, then there exist $\theta > 0$ such that*

TWP Beltraeut

$$f(x + td) \leq f(x) - \theta \left( \frac{\nabla f(x)^T d}{\|d\|} \right)^2 \text{ for all } t \in T_{WP}(x, d)$$

*(or, equivalently, the strategy is efficient).*

*Proof.* Recall the notation

$$\begin{aligned} \phi(t) &= f(x + td) \text{ and} \\ \psi(t) &= f(x) + \sigma t \nabla f(x)^T d. \end{aligned} \tag{*}$$

Let $x \in \mathcal{L}(x^0)$, then, by definition, we have $f(x) \leq f(x^0)$. Furthermore, let $d \in \mathbb{R}^n$ be such that $\nabla f(x)^T d < 0$.

(a) It suffices to show that there exists $t > 0$ such that

$$\begin{aligned} \phi(t) &\leq \psi(t) \text{ and} \\ \phi'(t) &\geq \rho \phi'(0). \end{aligned}$$

Note that $\rho \phi'(0) > \phi'(0)$, since $\rho < 1$ and $\phi'(0) < 0$. Indeed, we have the following (since $\sigma \in (0, \frac{1}{2})$:

$$\phi'(0) = \nabla f(x)^T d < \sigma \nabla f(x)^T d = \psi'(0).$$

Therefore, we have $(\psi - \phi)'(0) > 0$ and, by definition of the derivative:

$$\lim_{t \searrow 0} \frac{(\psi - \phi)(t) - (\psi - \phi)(0)}{t} > 0.$$

Plugging in (*), we have:

$$\lim_{t \searrow 0} \frac{(\psi - \phi)(t)}{t} > 0 \implies \exists\, t_0 > 0 \text{ s.t. } \forall\, t \in (0, t_0) :\ \psi(t) > \phi(t).$$

Let $t^*$ be the first $t > 0$ at which $\phi$ and $\psi$ intersect:

$$t^* := \min\{t > 0 \text{ s.t. } \phi(t) = \psi(t)\}.$$

Note that $t^*$ exists, since $\lim_{t\to\infty} \psi(t) = -\infty$ and $\phi$ is bounded from below (because $f$ is bounded from below). Then we have:

$$\phi'(t^*) = \lim_{\substack{t \to t^* \\ t < t^*}} \frac{\phi(t) - \phi(t^*)}{t - t^*} \geq \lim_{\substack{t \to t^* \\ t < t^*}} \frac{\psi(t) - \psi(t^*)}{t - t^*} = \psi'(t^*)$$

where the inequality above holds since $\phi(t^*) = \psi(t^*)$, $\phi(t) < \psi(t)$ for $t < t^*$ and $t - t^* < 0$. Moreover, we have:

$$\psi'(t^*) = \sigma \nabla f(x)^T d \geq \rho \nabla f(x)^T d = \rho \phi'(0).$$

Therefore, $t^* \in T_{WP}(x, d)$.

(b) Let $t \in T_{WP}(x, d)$. Then we have

$$\phi(t) = f(x + td) \leq \psi(t) = f(x) + t\sigma \nabla f(x)^T d < f(x) \leq f(x^0),$$

which implies that $x + td \in \mathcal{L}(x^0)$. Furthermore, since $\rho < 1$ and $\nabla f(x)^T d < 0$:

$$\begin{aligned} 0 &< (\rho - 1)\nabla f(x)^T d \\ &= \rho \nabla f(x)^T d - \nabla f(x)^T d \\ &\leq \nabla f(x + td)^T d - \nabla f(x)^T d \\ &\leq (\nabla f(x + td) - \nabla f(x))^T d. \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(\nabla f(x + td) - \nabla f(x))^T d \leq \|\nabla f(x + td) - \nabla f(x)\| \, \|d\| .$$

Since $\nabla f$ is Lipschitz continuous with constant $L > 0$ on $\mathcal{L}(x^0)$,

$$\|\nabla f(x + td) - \nabla f(x)\| \, \|d\| \leq L \|td\| \, \|d\| = Lt \|d\|^2 .$$

In conclusion, we obtain

$$t > \frac{(\rho - 1)\nabla f(x)^T d}{L\left\|d\right\|^2}.$$

Define $\theta := \frac{\sigma(1-\rho)}{L}$. Then we have

$$\begin{aligned} f(x+td) &\leq f(x) + \sigma t \nabla f(x)^T d \\ &\leq f(x) + \frac{\sigma(\rho-1)}{L}\left(\frac{\nabla f(x)^T d}{\left\|d\right\|^2}\right)^2 \\ &= f(x) - \theta\left(\frac{\nabla f(x)^T d}{\left\|d\right\|^2}\right)^2. \end{aligned}$$

□

## 5.2 The strong Wolfe-Powell step size strategy

In this section, we want to add an upper bound on the increase of $\phi$ (insert picture). Let $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1]$, $x \in \mathcal{L}(x^0)$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$. The strong Wolfe-Powell step size strategy consists in finding $t > 0$ such that

$$f(x+td) \leq f(x) + \sigma t \nabla f(x)^T d \tag{5.4}$$

and

$$|\nabla f(x+td)^T d| \leq -\rho \nabla f(x)^T d. \tag{5.5}$$

(difference) absolute value

We can reformulate the conditions above:

$$\begin{aligned} (5.4) \;&\Leftrightarrow\; \phi(t) \leq \phi(0) + \sigma t \phi'(0) = \psi(t) \\ (5.5) \;&\Leftrightarrow\; |\phi'(t)| \geq \rho\phi'(0) \;\Leftrightarrow\; \rho\phi'(0) \leq \phi'(t) \leq -\rho\phi'(0). \end{aligned}$$

standard

$\phi(t) = f(x+td)$

**Definition.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ and $x^0 \in \mathbb{R}^n$. For $x \in \mathcal{L}(x^0)$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, let

$$\boxed{T_{SWP}(x,d) = \{t > 0 : f(x+td) \leq f(x) + \sigma t \nabla f(x)^T d \text{ and } |\nabla f(x+td)^T d| \geq -\rho \nabla f(x)^T d\}}$$

be the set of strong Wolfe-Powell step size strategies in $x$ in direction $d$.

**Theorem 5.2.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ and $x^0 \in \mathbb{R}^n$. For $x \in \mathcal{L}(x^0)$ and $d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, let $T_{SWP}$ be the set of strong Wolfe-Powell step size strategies in $x$ in direction $d$. Then it holds*

(a) *If $f$ is bounded from below, then $T_{SWP}(x, d)$ is nonempty (or, equivalently, the strategy is well-defined).*

(= bounded by above because of absolute value)

(b) *If $\nabla f$ is Lipschitz-continuous on $\mathcal{L}(x^0)$, then there exist $\theta > 0$ such that*

⇒ $T_{SWP}$ is efficient

$$f(x + td) \leq f(x) - \theta \left( \frac{\nabla f(x)^T d}{\|d\|} \right)^2 \text{ for all } t \in T_{SWP}(x, d)$$

*(or, equivalently, the strategy is efficient).*

*Proof.* As in the proof of Theorem (5.1), recall the notation

$$\phi(t) = f(x + td) \text{ and}$$
$$\psi(t) = f(x) + \sigma t \nabla f(x)^T d.$$

(a) As in the proof of Theorem (5.1), there exists

$$t^* := \min\{t > 0 \text{ s.t. } \phi(t) = \psi(t)\}$$

with $\phi'(t^*) \geq \psi'(t^*)$ and of course, $\phi(t^*) = \psi(t^*)$. We first consider the case where $\phi'(t^*) \leq 0$. In this case, we have

$$|\nabla f(x + t^* d)^T d| = |\phi'(t^*)| = -\phi'(t^*) \leq -\psi'(t^*) = -\sigma\phi'(0) \leq -\rho\phi'(0).$$

Therefore, $t^* \in T_{SWP}(x, d)$. In the case where $\phi'(t^*) > 0$, we have $\nabla f(x)^T d = \phi'(0) < 0$ and therefore there exists $t^{**} \in (0, t^*)$ such that $\phi(t^{**}) = 0$ (compare with Figure (??)). Therefore,

$$\phi(t^{**}) < \psi(t^{**}) \;\Leftrightarrow\; f(x + t^{**} d) < f(x) + \sigma t^{**} \nabla f(x)^T d.$$

Hence, we have

$$|\nabla f(x + t^{**} d)^T d| = |\phi'(t^{**})| = 0 \leq -\rho\phi'(0) = -\phi\nabla f(x)^T d,$$

which implies $t^{**} \in T_{SWP}(x, d)$.

(b) Follows from Theorem (5.1), since $T_{SWP}(x,d) \subseteq T_{WP}(x,d)$.

□

## 5.3 The (backtracking) Armijo rule

In the following, we will introduce a step size strategy which is easy to implement but not efficient. Let $\sigma \in (0,1)$ and $\beta \in (0,1)$. For $x, d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, the Armijo rule consists in choosing

$$\boxed{t := \max\{\beta^l : l = 0,1,2,\dots\} \text{ s.t. } f(x+td) \le f(x) + \sigma t \nabla f(x)^T d.}$$

**Theorem 5.3.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0,1)$ and $\beta \in (0,1)$. Then for every $x, d \in \mathbb{R}^n$ such that $\nabla f(x)^T d < 0$, there exists $l \ge 0$ such that*

$$f(x + \underline{\beta^l} d) \le f(x) + \sigma \underline{\beta^l} \nabla f(x)^T d.$$

*Or, equivalently, the Armijo rule is well-defined.*

*Proof.* According to the proof of Theorem (5.1), there exists $t_0 > 0$ such that for all $t \in (0, t_0)$, we have

$$\phi(t) = f(x+td) < f(x) + \sigma \nabla f(x)^T d = \psi(t).$$

Since $\beta \in (0,1)$, we conclude that there exists $l \ge 0$ such that $\beta^l \in (0, t_0)$. □

**Remark 5.4.** The Armijo rule is not efficient. Take, for example,

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = \frac{x^2}{2},\ x^0 = -3,\ \sigma = \frac{1}{2},\ d^k := \frac{1}{2^k}\ \forall\, k \ge 0, \text{ and } x^{k+1} := x^k + t_k d^k,$$

where $t_k$ is the Armijo step size. It holds that $t_k = 1$ for all $k \ge 0$ and $x^k = -3 + 1 + \frac{1}{2} + \ldots + \frac{1}{2^{k-1}} = -1 - \frac{1}{2^{k-1}}$ for all $k \ge 1$. Indeed, we have

$$\frac{(x+td)^2}{2} \le \frac{x^2}{2} + \frac{1}{2}tx + d \ \Leftrightarrow\ t^2 d^2 \le -txd \ \Leftrightarrow\ t \le \frac{x}{d} \text{ for } xd < 0.$$

Carrying out the calculations for $k = 0$ and $k = 1$, we obtain the following:

$$x^0 = -3,\ d^0 = 1 \ \Rightarrow\ t^0 \le 3 \ \Rightarrow\ \beta^l \le 3 \ \Rightarrow\ t^0 = 1$$
$$x^1 = -3 + 1 = -2,\ d^1 = \frac{1}{2} \ \Rightarrow\ t^1 \le 2 \cdot 2 = 4 \ \Rightarrow\ \beta^l \le 4 \ \Rightarrow\ t^1 = 1$$

By induction, one can prove the claims above on $t^k$ and $x^k$. Furthermore, it can be shown by contradiction that there exists no $\theta > 0$ such that for all $(x, d)$ with $xd < 0$ and for $t = 1$ we have

$$\frac{(x + td)^2}{2} \leq \frac{x^2}{2} - \theta\left(\frac{xd}{d}\right)^2 = \frac{x^2}{2} - \theta x^2.$$

# 6 The gradient method

In the following, we consider $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable.

**Algorithm 2** Gradient method, algorithm of steepest descent

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, $\sigma \in (0,1)$, $\beta \in (0,1)$, $\varepsilon \geq 0$ and set $k := 0$
2: If $\left\|\nabla f(x^k)\right\| \leq \varepsilon$, stop
3: Set $d^k := -\nabla f(x^k)$
4: Find a step size $t_k := \max\{\beta^l : l = 0, 1, \ldots\}$ with $f(x^k + t_k d^k) < f(x^k) + \sigma t_k \nabla f(x^k)^T d^k$ (Armijo rule)
5: Set $x^{k+1} := x^k + t_k d^k$, set $k := k+1$ and go to Step 2
```

For the analysis of the gradient method, we will assume the worst case scenario of $\varepsilon = 0$ and the algorithm not stopping after finitely many steps.

**Example 6.1.** This example is temporarily here to fix the numbering of the theorems (the gradient method should be Algorithm 6.1).

We will need the following lemma for the proof of Theorem 6.3.

**Lemma 6.2.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x, d \in \mathbb{R}^n$, $(x^k)_{k \geq 0} \subseteq \mathbb{R}^n$, $(d^k)_{k \geq 0} \subseteq \mathbb{R}^n$, $(t^k)_{k \geq 0} \subseteq \mathbb{R}$ such that $x^k \to x$, $d^k \to d$, $t^k \searrow 0$ as $k \to \infty$. Then it holds:*

$$\lim_{k \to \infty} \frac{f(x^k + t_k d^k) - f(x^k)}{t_k} = \nabla f(x)^T d.$$

*Proof.* For all $k \geq 0$, we can apply the Mean Value Theorem. Then there exists $\xi^k \in (x^k, x^k + t_k d^k)$ such that

$$\begin{aligned} & f(x^k + t_k d^k) - f(x^k) = \nabla f(\xi^k)^T (t_k d^k) \\ \Leftrightarrow \quad & \frac{f(x^k + t_k d^k) - f(x^k)}{t_k} = \nabla f(\xi^k)^T (d^k) \to \nabla f(x)^T d \ \ (k \to \infty). \end{aligned}$$

Here, we used that $\nabla f(\xi^k) \to \nabla f(x)$. This holds by continuity of the gradient and because $\xi^k \to x$, since $x^k \to x$. □

**Theorem 6.3** (Convergence of gradient method)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then every limit (accumulation) point of the sequence $(x^k)_{k \geq 0}$ generated by Algorithm (2) is a critical point of $f$.*

*Proof.* Let $x^* \in \mathbb{R}^n$ be a limit point of $(x^k)_{k\geq 0}$, i.e., there exist a subsequence $(x^{k_l})_{l\geq 0}$ such that $x^{k_l} \to x^*$ as $l \to \infty$. Assume that $x^*$ is not a critical point of $f$, i.e., $\nabla f(x^*) \neq 0$. Then we have for all $k \geq 0$:

$$f(x^{k+1}) = f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k \nabla f(x^k)^T d^k = f(x^k) - \sigma t_k \left\|\nabla f(x^k)\right\|^2 \leq f(x^k). \tag{*}$$

Therefore, $(f(x^k))_{k\geq 0}$ is non-increasing.

Since $f$ is continuous, we have $f(x^{k_l}) \to f(x^*)$ as $l \to \infty$. By exercise 18, it follows that $f(x^k) \to f(x^*)$ as $k \to \infty$. Hence, using (*), we get for all $k \geq 0$:

$$0 \leq \sigma t_k \left\|\nabla f(x^k)\right\|^2 \leq f(x^k) - f(x^{k+1}) \to 0 \ \ (k \to \infty).$$

This implies that $t_k \left\|\nabla f(x^k)\right\|^2 \to 0$ as $(k \to \infty)$, and thus $t_{k_l} \left\|\nabla f(x^{k_l})\right\|^2 \to 0$ as $(l \to \infty)$. Because $\|\nabla f(x^*)\|^2 \neq 0$ by assumption, we can conclude that $t_{k_l} \to 0$ as $(l \to \infty)$.

Now, take $t_{k_l} := \beta^{m_{k_l}}$. By the Armijo rule, we have for all $l \geq 0$:

$$\begin{aligned} & f(x^{k_l} + \beta^{m_{k_l}-1} d^{k_l}) > f(x^{k_l}) + \sigma \beta^{m_{k_l}-1} \nabla f(x^{k_l})^T d^{k_l} \\ \Leftrightarrow \ \ & \frac{f(x^{k_l} + \beta^{m_{k_l}-1} d^{k_l}) - f(x^{k_l})}{\beta^{m_{k_l}-1}} > \sigma \nabla f(x^{k_l})^T d^{k_l} \end{aligned} \tag{**}$$

Note that this holds because, by the Armijo rule, $t_{k_l} = \beta^{m_{k_l}}$ is the first number for which the inequality $f(x^{k_l} + \beta^{m_{k_l}} d^{k_l}) \leq f(x^{k_l}) + \sigma \beta^{m_{k_l}} \nabla f(x^{k_l})^T d^{k_l}$ is fulfilled, so for $\beta^{m_{k_l}-1}$, the inequality is *not* fulfilled.

Now we will use Lemma (6.2). Observe the following:

$$\begin{aligned} & \beta^{m_{k_l}-1} = \frac{1}{\beta} t_{k_l} \to 0 \ \ (l \to \infty), \\ & x^{k_l} \to x^* \ \ (l \to \infty), \\ & d^{k_l} = -\nabla f(x^{k_l}) \to -\nabla f(x^*) \ \ (l \to \infty). \end{aligned}$$

Taking the limit in (**), we have

$$\begin{aligned} & \nabla f(x^*)^T(-\nabla f(x^*)) \geq \sigma \nabla f(x^*)^T(-\nabla f(x^*)) \\ \Leftrightarrow \ \ & -\left\|\nabla f(x^*)\right\|^2 \geq -\sigma \left\|\nabla f(x^*)\right\|^2 \\ \Leftrightarrow \ \ & 1 \leq \sigma, \end{aligned}$$

which is a contradiction. Thus, $\nabla f(x^*) = 0$. □

**Remark.** Note that in Theorem (6.3), we do not assume that $(x^k)_{k \geq 0}$ has any limit points. We only show that *if* it has limit points, then they are all critical points of $f$.

# 7 The gradient method for convex optimization problems

In this section, we will analyze the convergence properties of the gradient method for solving the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \tag{7.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function with a $L_{\nabla f}$-Lipschitz continuous gradient, meaning that there exist $L_{\nabla f} > 0$ such that:

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

In the following, we will assume that

$$\arg\min\ f := \{x^* \in \mathbb{R}^n : x^* = \inf_{x\in\mathbb{R}^n} f(x)\} \subseteq \mathbb{R}^n$$

is nonempty (this need not always the be the case, take for example $f(x) = x$) and we will define

$$f^* := \inf_{x\in\mathbb{R}^n} f(x) \in \mathbb{R}.$$

**Lemma 7.1** (Descent lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable with a $L_{\nabla f}$-Lipschitz continuous gradient. Then*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L_{\nabla f}}{2} \|y - x\|^2 \quad \forall\, x, y \in \mathbb{R}^n.$$

*Proof.* Let $x, y \in \mathbb{R}^n$ and define $\phi : [0, 1] \to \mathbb{R}^n$ by $\phi(t) = f(x + t(y - x))$. Then we have for all $t \in [0, 1]$: $\phi'(t) = \nabla f(x + t(y - x))^T (y - x)$. From the fundamental theorem of differentiation and integration, we get

$$\phi(1) - \phi(0) = \int_0^1 \phi'(t) dt \ \Leftrightarrow\ f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt.$$

This implies the following:

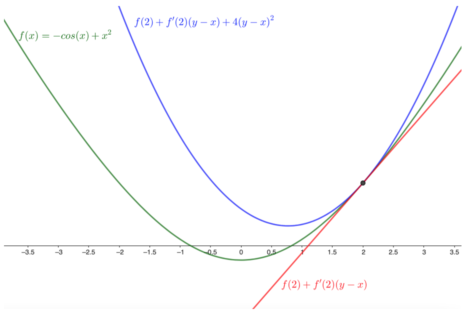
$$\begin{aligned}
f(y) - f(x) - \nabla f(x)^T(y-x) &= \int_0^1 \nabla f(x+t(y-x))^T(y-x) - \nabla f(x)^T(y-x)\, dt \\
&= \int_0^1 \Big(\nabla f(x+t(y-x)) - \nabla f(x)\Big)^T (y-x)\, dt \\
&\leq \int_0^1 \left|\Big(\nabla f(x+t(y-x)) - \nabla f(x)\Big)^T (y-x)\right| dt \\
&\leq \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\| \, \|y-x\| \; dt \\
&\leq \int_0^1 L_{\nabla f} t \, \|y-x\|^2 \; dt \\
&= L_{\nabla f} \|y-x\|^2 \int_0^1 t\, dt = \frac{L_{\nabla f}}{2} \|y-x\|^2 .
\end{aligned}$$

□

**Remark.** We know from Exercise 11, that under the assumptions of Lemma (7.1) we have

$$f(y) \geq f(x) + \nabla f(x)^T(y-x).$$

This means that the graph of $f$ is "squeezed" between a linear and a quadratic function.



**Theorem 7.2** (Convergence)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable with a $L_{\nabla f}$-Lipschitz continuous gradient, $\arg\min f \neq \emptyset$, $x^0 \in \mathbb{R}^n$ be an arbitrary starting point and $\gamma \in (0, \frac{1}{L_{\nabla f}})$. For the sequence $(x^k)_{k\geq 0}$ generated by the gradient method with constant step size $\gamma$: $x^{k+1} := x^k - \gamma \nabla f(x^k)$ for all $k \geq 0$, it holds*

*(a)* $f(x^k) - f(x) \leq \frac{1}{2\gamma k}\left( \|x^0 - x\|^2 - \left\|x^k - x\right\|^2 \right)$ *for all* $x \in \mathbb{R}^n$ *and* $k \geq 1$

*(b)* $0 \leq f(x^k) - f^* \leq \dfrac{dist^2_{\arg\min f(x^0)}}{2\gamma k}$ *for all* $k \geq 1$

*(c)* $(x^k)_{k\geq 0}$ *converges to an element of* $\arg\min f$ *(a global minimum of* $f$*).*

*Proof.* (a) Let $k \geq 0$. Then we have by Exercise 11 (aka, the gradient inequality):

$$f(x) - f(x^k) \geq \nabla f(x^k)^T(x - x^k). \qquad \text{(*)}$$

Furthermore, we have by Lemma (7.1):

$$\begin{aligned} &f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L_{\nabla f}}{2}\left\|x^{k+1} - x^k\right\|^2 \\ \Leftrightarrow &f(x^k) - f(x^{k+1}) \geq \nabla f(x^k)^T(x^k - x^{k+1}) - \frac{L_{\nabla f}}{2}\left\|x^{k+1} - x^k\right\|^2 \qquad \text{(**)} \end{aligned}$$

Adding (*) and (**), we get

$$f(x) - f(x^{k+1}) \geq \nabla f(x^k)^T(x - x^{k+1}) - \frac{L_{\nabla f}}{2}\left\|x^{k+1} - x^k\right\|^2.$$

Substituting $\frac{1}{\gamma}(x^k - x^{k+1})$ for $\nabla f(x^k)$, we get

$$f(x) - f(x^{k+1}) \geq \frac{1}{\gamma}(x^k - x^{k+1})^T(x - x^{k+1}) - \frac{L_{\nabla f}}{2}\left\|x^{k+1} - x^k\right\|^2.$$

Next, we make use of the identity

$$\|a\|^2 + \|b\|^2 + 2a^Tb = \|a + b\|^2 \;\Leftrightarrow\; 2(-a)^Tb = \|a\|^2 + \|b\|^2 - \|a + b\|^2.$$

By setting $a = x^{k+1} - x^k$ and $b = x - x^{k+1}$, we get

$$\begin{aligned} &f(x) - f(x^{k+1}) \geq \frac{1}{2\gamma}(\left\|x^{k+1} - x\right\|^2 + \left\|x - x^{k+1}\right\|^2 - \left\|x - x^k\right\|^2) - \frac{L_{\nabla f}}{2}\left\|x^{k+1} - x^k\right\|^2 \\ \Leftrightarrow &2\gamma(f(x^{k+1}) - f(x)) \leq \left\|x^k - x\right\|^2 - \left\|x^{k+1} - x\right\|^2 + (\gamma L_{\nabla f} - 1)\left\|x^{k+1} - x^k\right\|^2. \end{aligned}$$

Since $\gamma L_{\nabla f} - 1 < 0$, we get for all $k \geq 0$ and for all $x \in \mathbb{R}^n$:

$$2\gamma(f(x^{k+1}) - f(x)) \leq \left\|x^k - x\right\|^2 - \left\|x^{k+1} - x\right\|^2. \qquad (7.2)$$

For the next steps, we will use a telescopic argument. Let $K \geq 1$. We write (7.2) for $k = 0, \dots, K-1$ and sum up the $K$ resulting inequalities. This gives:

$$2\gamma\left(\sum_{k=0}^{K-1} f(x^{k+1}) - Kf(x)\right) \leq \left\|x^0 - x\right\|^2 - \left\|x^K - x\right\|^2. \tag{7.3}$$

For every $k \geq 0$, set $x := x^k$ in (7.2). This gives

$$\begin{aligned} & 2\gamma(f(x^{k+1}) - f(x^k)) \leq -\left\|x^{k+1} - x^k\right\|^2 \leq 0 \\ \Leftrightarrow\ & 2\gamma((k+1)f(x^{k+1}) - kf(x^k)) - 2\gamma f(x^{k+1}) \leq 0 \end{aligned} \tag{*}$$

Write (*) for $k = 0, \dots, K-1$ and sum up the $K$ inequalities. This gives

$$2\gamma K f(x^K) - 2\gamma \sum_{k=0}^{K-1} f(x^{k+1}) \leq 0 \tag{7.4}$$

Adding (7.3) and (7.4), we get for all $x \in \mathbb{R}^n$ and for all $K \geq 1$:

$$\begin{aligned} & 2\gamma(Kf(x^K) - Kf(x)) \leq \left\|x^0 - x\right\|^2 - \left\|x^K - x\right\|^2 \\ \Leftrightarrow\ & f(x^K) - f(x) \leq \frac{1}{2\gamma K}\left( \left\|x^0 - x\right\|^2 - \left\|x^K - x\right\|^2 \right) \end{aligned}$$

(b) Take $x := x^* \in \arg\min f \neq \emptyset$ in (a). Note that $x^*$ is not necessarily unique, but $f^* = f(x^*)$ (the minimal value $f$ can take) is, because we are in the convex setting. Then we have for all $k \geq 1$:

$$\begin{aligned} 0 &\leq f(x^k) - f(x^*) \\ &\leq \frac{1}{2\gamma k}\left( \left\|x^0 - x^*\right\|^2 - \left\|x^k - x^*\right\|^2 \right) \\ &\leq \frac{1}{2\gamma k}\left\|x^0 - x^*\right\|^2 \end{aligned}$$

Taking the infimum on both sides, we get for all $k \geq 1$:

$$0 \leq f(x^k) - f^* \leq \frac{1}{2\gamma k} \inf_{x^* \in \arg\min f} \left\|x^0 - x^*\right\|^2 = \mathrm{dist}^2_{\arg\min f}(x^0).$$

(c) We have $f(x^k) \to f^*$ as $k \to \infty$, but what about the convergence of $x^k$ itself?

Take $x^* \in \arg\min f$ and set $x = x^*$ in (7.2). Then we have for all $k \geq 0$:

$$\begin{aligned} & 2\gamma(f(x^{k+1}) - f(x^*)) \leq \left\|x^k - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2 \\ \Rightarrow & \left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2 . \end{aligned}$$

Therefore, we have that $(\left\|x^k - x^*\right\|^2)_{k\geq 0}$ is non-increasing and bounded, and hence convergent. This means that there exist $l_{x^*} \in \mathbb{R}^+$ such that

$$l_{x^*} = \lim_{k\to\infty} \left\|x^k - x^*\right\|^2 .$$

Note that $l_{x^*}$ depends on $x^*$ and is not necessarily zero. This implies that $(\left\|x^k - x^*\right\|)_{k\leq 0}$ is bounded and, since $\left\|x^k\right\| \leq \left\|x^k - x^*\right\| + \|x^*\|$ for all $k \geq 0$, we know that $(x^k)_{k\geq 0}$ is also bounded (but we don't know yet whether it's converging). Boundedness implies that $(x^k)_{k\geq 0}$ has at least one accumulation point. We will prove that it has exactly one accumulation point, which implies that the whole sequence converges. Indeed, assume that there are two accumulation points $x'$ and $x''$. Then there exist subsequences $(x^{k_l})_{l\geq 0}$, $(x^{k_j})_{j\geq 0}$ such that

$$\begin{aligned} & x^{k_l} \to x' \ (l \to \infty) \\ & x^{k_j} \to x'' \ (j \to \infty). \end{aligned}$$

By continuity of $f$ and by part (b), we have that

$$\begin{aligned} & f(x^{k_l}) \to f(x') \ \Rightarrow \ f(x') = f^* \ \Rightarrow \ x' \in \arg\min f \\ & f(x^{k_j}) \to f(x'') \ \Rightarrow \ f(x'') = f^* \ \Rightarrow \ x'' \in \arg\min f. \end{aligned}$$

We have the following for all $k \geq 0$:

$$2(x^k)^T(x' - x'') = \left\|x^k - x''\right\|^2 - \left\|x^k - x'\right\|^2 - \|x''\|^2 + \|x'\|^2 .$$

Then we have for all $l \geq 0$ and for all $j \geq 0$:

$$\begin{aligned} 2(x^{k_l})^T(x' - x'') &= \left\|x^{k_l} - x''\right\|^2 - \left\|x^{k_l} - x'\right\|^2 - \|x''\|^2 - \|x'\|^2 \\ 2(x^{k_j})^T(x' - x'') &= \left\|x^{k_j} - x''\right\|^2 - \left\|x^{k_j} - x'\right\|^2 - \|x''\|^2 - \|x'\|^2 . \end{aligned}$$

Now let $l \to \infty$ and $j \to \infty$:

$$\begin{aligned} 2(x')^T(x' - x'') &= l_{x''} - l_{x'} - \|x''\|^2 - \|x'\|^2 \\ 2(x'')^T(x' - x'') &= l_{x''} - l_{x'} - \|x''\|^2 - \|x'\|^2 . \end{aligned}$$

Adding these two equations, we get

$$\begin{aligned} & (x')^T(x' - x'') = (x'')^T(x' - x'') \\ \Leftrightarrow\ & (x' - x'')^T(x' - x'') = 0 \\ \Leftrightarrow\ & \|x' - x''\|^2 = 0 \ \Leftrightarrow\ x' = x''. \end{aligned}$$

Thus, the whole sequence $(x^k)_{k\geq 0}$ is convergent to an element $\bar{x} \in \mathbb{R}^n$. Since $f(x^k) \to f^*$, we have:

$$\begin{aligned} & x^k \to \bar{x}\ (k \to \infty) \\ \Rightarrow\ & f(x^k) \to f(\bar{x})\ (k \to \infty) \\ \Rightarrow\ & f(\bar{x}) = f^* \ \Rightarrow\ \bar{x} \in \arg\min f. \end{aligned}$$

□

**Remark.** In Theorem (7.2)(c) we cannot say to which minimum we converge (insert picture).

In Theorem (7.2)(b), note that

$$\lim_{k\to\infty} \frac{\text{dist}^2_{\arg\min f}(x^0)}{2\gamma k} = 0.$$

Theorem (7.2)(c) implies Theorem (7.2)(b), but the converse is not true: there are methods for which we only know (b) but not (c).

Theorem (7.2)(b) tells us how many iterations we need to achieve a desired accuracy:

$$f(x^*) - f^* < \varepsilon \ \Leftrightarrow\ k > \frac{\text{dist}^2}{2\gamma\varepsilon}.$$

Note the role of $x^0$ in Theorem (7.2)(b): If we choose $x^0 \in \arg\min f$, we are immediately done (of course this never happens, but YKWIM).

Theorem (7.2)(b) implies sublinear (not as fast as linear) convergence: $f(x^k) \to f(x^*)$ with $\mathcal{O}(\frac{1}{k})$.

**Example.** We can use the gradient method to solve a linear system $Ax = b$. Indeed, rewrite

$$Ax = b \Leftrightarrow \min_{x \in \mathbb{R}^n} \|Ax - b\|^2 .$$

Define $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|Ax - b\|^2$. Then $f$ is convex, differentiable and its gradient $\nabla f(x) = 2A^*(Ax - b)$ is Lipschitz continuous with constant $L_{\nabla f} := 2\|A^*A\|$. Indeed, we have

$$\|\nabla f(x) - \nabla f(y)\| = 2\|A^*A(x - y)\| \leq 2\|A^*A\| \|x - y\| .$$

Therefore, if we choose $\gamma \in (0, \frac{1}{2\|A^*A\|})$, we can solve $Ax = b$ with the iteration $x^{k+1} := x^k - \gamma 2A^*(Ax^k - b)$.

**Remark 7.3.** The gradient method can be seen as an explicit time discretization of the following gradient flow system:

$$\begin{cases} \dot{x}(t) = -\nabla f(x(t)) \\ x(0) = x^0 \end{cases} \tag{7.5}$$

Indee, we have for all $k \geq 0$:

$$\frac{x(t_{k+1}) - x(t_k)}{\gamma_k} = -\nabla f(x(t_k)).$$

Setting $x(t_0) := x^0$, $x(t_k) := x^k$ and $\gamma_k = \gamma$ for all $k \geq 1$, we get

$$x^{k+1} - x^k = -\gamma \nabla f(x^k) \;\Leftrightarrow\; x^{k+1} = x^k - \gamma \nabla f(x^k).$$

This is the gradient method! For (7.5), we have:

1. like in discrete time, $f(x(t)) \to f^*$ with rate $\mathcal{O}(\frac{1}{t})$, meaning that

$$0 \leq f(x(t)) - f^* \leq \frac{c}{t} \;\; \forall\, t > 0.$$

2. $x(t)$ converges to an element in $\arg\min f$ as $t \to \infty$.

These statements can be proved using similar techniques as the ones we already used. Such a continuous interpretation exists for many algorithms!

**Remark 7.4.** (a) Can we make the gradient method faster than $\mathcal{O}(\frac{1}{k})$? Yes! Nesterov introduced in 1983 the so called *accelerated* or *fast gradient method* for solving (7.1). For $x^0 = x^1 \in \mathbb{R}^n$, $\gamma \in (0, \frac{1}{L_{\nabla f}}]$, $t_1 = 1$ and $t_{k+1} = \frac{1+\sqrt{4t_k^2+1}}{2}$ (note that $t_{k+1}$ is the solution of $t_{k+1}^2 - t_{k+1} - t_k^2 = 0$) for all $k \geq 1$, let

$$\begin{cases} y^k := x^k + \dfrac{t_k - 1}{t_{k-1}}(x^k - x^{k-1}) \\ x^{k+1} := y^k - \gamma \nabla f(y^k), \end{cases} \tag{7.6}$$

where $(y^k)_{k\geq 1}$ is called the *momentum sequence*. It holds for all $k \geq 2$:

$$0 \leq f(x^k) - f^* \leq \frac{2}{\gamma(k+1)^2} \text{dist}^2_{\arg\min f}(x^0).$$

This is faster than the gradient method! However, it is not known whether the sequence $(x^k)_{k\geq 1}$ converges.

(b) The continuous counterpart of (7.6) was proposed by Su-Boyd-Candes in 2015 and has the following formulation for $t \geq t_0 > 0$ and $\alpha > 0$:

$$\begin{cases} \ddot{x}(t) + \dfrac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0 \\ x(t_0) = x^0, \dot{x}(t_0) = u^0. \end{cases} \tag{7.7}$$

The following holds (Attouch, 2015):

- If $\alpha \geq 3$, then $f(x(t)) \to f^*$ with rate $\mathcal{O}(\frac{1}{t^2})$ as $t \to \infty$, as in discrete time.
- If $\alpha > 3$, then $x(t)$ converges to an element of $\arg\min f$ and $f(x(t)) \to f^*$ with rate $o(\frac{1}{t^2})$ as $t \to \infty$, meaning $t^2(f(x(t) - f^*) \to 0$ as $t \to \infty$.

(c) Explicit time discretization of (7.7) leads to the following accelerated/fast gradient method (Chambolle-Dossal, 2015). For $x^0 = x^1 \in \mathbb{R}^n$, $\gamma \in (0, \frac{1}{L_{\nabla f}}]$, $t_k = \frac{k+\alpha-2}{\alpha-1}$, $\alpha \geq 3$ and for all $k \geq 1$, let

$$\begin{cases} y^k := x^k + \dfrac{t_k - 1}{t_{k-1}}(x^k - x^{k-1}) \\ x^{k+1} := y^k - \gamma \nabla f(y^k), \end{cases} \tag{7.8}$$

For $\alpha \geq 3$, we get the benefit that $f(x^k) \to f^*$ with rate $o(\frac{1}{t^2})$ and $(x^k)_{k\geq 0}$ converges to an element in $\arg\min f$ as $k \to \infty$. This is all we want! This is the best method so far for the convex case.

To conclude this chapter, we will discuss the convergence properties of the (classical) gradient method when minimizing strongly convex functions. Recall the following definition from Exercise 23.

**Definition.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *strongly convex with modulus $\mu > 0$* if

$$f(\lambda x + (1-\lambda)y) + \mu\lambda(1-\lambda)\left\|x-y\right\|^2 \leq \lambda f(x) + (1-\lambda)f(y)$$

for all $x, y \in \mathbb{R}^n$ and all $\lambda \in [0,1]$. Equivalently, $f$ is strongly convex with modulus $\mu$ if

$$g : \mathbb{R}^n \to \mathbb{R},\, g(x) = f(x) - \mu\left\|x\right\|^2$$

is convex.

**Theorem 7.5.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex with modulus $\mu > 0$ and differentiable with $L_{\nabla f}$-Lipschitz continuous gradient. Let $x^0 \in \mathbb{R}^n$ and $\gamma \in (0, \frac{4\mu}{L^2_{\nabla f}})$ and $c := 1 - \gamma(4\mu - \gamma L^2_{\nabla f} \in (0,1)$. For the sequence $x^{k+1} := x^k - \gamma\nabla f(x^k)$, it holds*

*(a) for all $k \geq 0$, $\left\|x^{k+1} - x^*\right\| \leq \sqrt{c}\left\|x^k - x^*\right\|$, where $x^*$ is the unique global minimum of $f$.*

*(b) $\left\|x^k - x^*\right\| \to 0$ with rate $\mathcal{O}(\sqrt{c}^k)$ as $k \to \infty$.*

*(c) $f(x^k) \to f^*$ with rate $\mathcal{O}(c^k)$ as $k \to \infty$.*

*Proof.* (a) Let $T : \mathbb{R}^n \to \mathbb{R}^n$, $T(x) = x - \gamma\nabla f(x)$. We will prove that $T$ is a contraction. The function $f$ is strongly convex, therefore $f - \mu\left\|\cdot\right\|^2$ is convex and we get from Exercise 11 for all $x, y \in \mathbb{R}^n$:

$$\begin{aligned} &\langle \nabla f(y) - 2\mu y - \nabla f(x) - 2\mu x, y - x\rangle \geq 0 \\ \Leftrightarrow\ &\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq 2\mu\langle y - x, y - x\rangle = 2\mu\left\|y - x\right\|^2, \end{aligned}$$

where $\langle x, y\rangle = x^T y$ denotes the standard inner product. Then we have for all $x, y \in \mathbb{R}^n$ :

$$\begin{aligned} \left\|T(x) - T(y)\right\|^2 &= \left\|x - y - \gamma(\nabla f(x) - \nabla f(y))\right\|^2 \\ &= \left\|x-y\right\|^2 - 2\gamma\langle\nabla f(x) - \nabla f(y), x - y\rangle + \gamma^2\left\|\nabla f(x) - \nabla f(y)\right\|^2. \end{aligned}$$

Since $\langle\nabla f(x) - \nabla f(y), x - y\rangle \geq 2\mu\left\|y - x\right\|^2$ and $\left\|\nabla f(x) - \nabla f(y)\right\|^2 \leq L^2_{\nabla f}\left\|x - y\right\|^2$, we get

$$\left\|T(x) - T(y)\right\| \leq \left\|x - y\right\|(1 - 4\gamma\mu + \gamma^2 L^2_{\nabla f}) = c\left\|x - y\right\|,$$

and thus, $T$ is a contraction because $c \in (0, 1)$. According to the Contraction Mapping Theorem, the sequence $x^{k+1} := T(x^k)$ for all $k \geq 0$ converges to the unique fixpoint of $x^*$ (which exists). Then we have

$$x^* = T(x^*) = x^* - \gamma \nabla f(x^*) \;\Leftrightarrow\; \nabla f(x^*) = 0,$$

which is equivalent to $x^*$ being global minimum of $f$. Furthermore, we get for all $k \geq 0$

$$\left\|x^{k+1} - x^*\right\| \leq \sqrt{c} \left\|x^k - x^*\right\|.$$

(b) Follows directly from (a) by applying the inequality $k$ times.

(c) By inequality (7.2), we get

$$\begin{aligned} 0 \leq f(x^{k+1}) - f^* &= f(x^{k+1}) - f(x^*) \\ &\leq \frac{1}{2\gamma} \left\|x^k - x^*\right\|^2 \\ &\leq \frac{1}{2\gamma} c^k \left\|x^0 - x^*\right\|^2. \end{aligned}$$

This is equivalent to

$$f(x^{k+1}) - f^* \leq \frac{\left\|x^0 - x^*\right\|^2}{2\gamma c} \cdot c^{k+1},$$

which proves the claim.

□

**Remark.** Under the assumptions of Theorem (7.5), we have that $\mu \leq \frac{L_{\nabla f}}{2}$, which implies

$$\gamma \in \left(0, \frac{4\mu}{L_{\nabla f}^2}\right) \subseteq \left(0, \frac{2}{L_{\nabla f}}\right).$$

Q: Don't we want $\gamma \in (0, \frac{1}{L_{\nabla f}})$? Indeed, we have for all $x, y \in \mathbb{R}^n$ by Exercises 11 and 23 and Lemma (7.1):

$$\mu \left\|x - y\right\|^2 + \nabla f(x)^T (y - x) + f(x) \leq f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L_{\nabla f}}{2} \left\|y - x\right\|^2.$$

Therefore, we have $\mu \leq \frac{L_{\nabla f}}{2}$.

# 8 Conjugate Gradient (CG) algorithms

## 8.1 The CG algorithm for linear systems of equations (standard)

Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $b \in \mathbb{R}^n$. Consider the linear system

$$Ax = b. \tag{8.1}$$

We attach to (8.1) the convex optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{2}x^T Ax - b^T x \tag{8.2}$$

**Remark.** We could apply the gradient method we know to solve (8.2), but this would provide us only with an approximate solution. We will see that the CG method gives us the *exact* solution after $n$ iterations.

Note that $\nabla f(x) = Ax - b$ and $f$ is convex, since $\nabla^2 f(x) = A$ is positive definite. Then we get that solving (8.1) is equivalent to solving (8.2), indeed:

$$\begin{aligned} & x^* \text{ is the unique solution of (8.1)} \\ \Leftrightarrow\ & \nabla f(x^*) = 0 \\ \Leftrightarrow\ & x^* \text{ is the unique global minimum of (8.2).} \end{aligned}$$

**Definition.** Let $A$ be symmetric positive definite. Then

$$\langle \cdot, \cdot, \rangle_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ \langle u, v, \rangle_A = u^T Av$$

defines a scalar product.

**Lemma 8.1.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite, $b \in \mathbb{R}^n$ and $x^0 \in \mathbb{R}^n$. Let $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n \setminus \{0\}$ be such that*

$$(d^i)^T Ad^j = 0 \ \ \forall\, i, j = 0, \ldots, n-1, j \neq i. \tag{8.3}$$

*Then, the algorithm of successive minimization along the directions $d^0, \ldots, d^{n-1}$, i.e., the calculation of $(x^k)_{k\geq 0}$ via*

$$x^{k+1} = x^k + t_k d^k \text{ with } f(x^k + t_k d^k) = \min_{t\in\mathbb{R}} f(x^k + td^k) \tag{8.4}$$

*for $k = 0, \dots, n-1$, provides (at latest) after $n$ steps, with $x^n$, the unique global minimum of $f$. In addition, for $k = 0, \dots, n-1$ and $g^k := Ax^k - b = \nabla f(x^k)$, it holds:*

$$t_k := \frac{-(g^k)^T d^k}{(d^k)^T A d^k} \text{ and } (g^{k+1})^T d^j = 0 \text{ for } j = 0, \dots, k. \tag{8.5}$$

*Proof.* For $k = 0, \dots, n-1$, let (compare with Exercise ??)

$$\begin{aligned} g(t) &= f(x^k + td^k) = \frac{1}{2}t^2 (d^k)^T A d^k + t(Ax^k - b)^T d^k + \frac{1}{2}(x^k)^T A x^k - b^T x^k \\ g'(t) &= t(d^k)^T A d^k + (g^k)^T d^k \overset{!}{=} 0 \;\Leftrightarrow t = \frac{-(g^k)^T d^k}{(d^k)^T A d^k} \\ g''(t) &= (d^k)^T A d^k > 0 \;\; \forall t \in \mathbb{R}. \end{aligned}$$

Therefore,

$$t_k := \frac{-(g^k)^T d^k}{(d^k)^T A d^k} \tag{*}$$

is the minimizer of $g$. Since $x^{k+1} = x^k + t_k d^k$, we have

$$\begin{aligned} (g^{k+1})^T d^k &= (Ax^{k+1} - b)^T d^k \\ &= (Ax^k - b + t_k A d^k)^T d^k \\ &= (g^k)^T d^k + t_k (d^k)^T A d^k = 0, \end{aligned} \tag{8.6}$$

where the last equality follows by (*). Moreover, we have for all $i \neq j$:

$$\begin{aligned} (g^{i+1} - g^i)^T d^j &= (Ax^{i+1} - Ax^i)^T d^j \\ &= (x^{i+1} - x^i)^T A d^j \\ &= t_i (d^i)^T A d^j = 0, \end{aligned} \tag{**}$$

where we used the symmetry of $A$ and the assumptions on $d^i$ and $t_i$. Next, observe that for all $k = 0, \dots, n-1$ and for all $j = 0, \dots, k$, we have by (8.6) and (**):

$$(g^{k+1})^T d^j = (g^{j+1})^T d^j + \sum_{i=j+1}^{k} (g^{i+1} - g^i)^T d^j = 0. \tag{8.7}$$

By (8.3), $d^0, \dots, d^{n-1}$ are orthogonal with respect to $\langle \cdot, \cdot \rangle_A$. We claim that they are linearly independent. Indeed, we have rest of the proof missing. □

**Remark.** Compare equation (8.5) with exercise 20(?). Furthermore, the fact that $(g^{k+1})^T d^j = 0$ for $j = 0, \dots, k$ is very important and is the main idea of CG: the new gradient is perpendicular to all previous directions.

The vectors $d^0, \dots d^{n-1}$ are called $A$-orthogonal or $A$-conjugate. How can these be created in a clever way, without having to store a lot of information when $n$ is very large?

**Method 1.** Gram-Schmidt orthogonalization approach with respect to $\langle \cdot, \cdot \rangle_A$. Take $v_1, \dots, v_n$ linearly independent. Then define

$$\begin{aligned} u_1 &:= v_1 \\ u_2 &:= v_2 - \frac{\langle v_2, u_1 \rangle_A}{\langle u_1, u_1 \rangle_A} \\ u_3 &:= v_3 - \frac{\langle v_3, u_1 \rangle_A}{\langle u_1, u_1 \rangle_A} - \frac{\langle v_3, u_3 \rangle_A}{\langle u_2, u_2 \rangle_A} \\ &\dots \end{aligned}$$

**Method 2.** Alternatively, $d^0, \dots d^{n-1}$ can be successively generated when running the algorithm. In addition, $d^k$ can be chosen as a descent direction of $f$ at $x^k$.

Let us choose

$$\begin{aligned} d^0 &:= -\nabla f(x^0) = -(Ax^0 - b) = -g^0 \\ \Rightarrow (g^0)^T d^0 &= -\left\| g^0 \right\|^2 . \end{aligned}$$

Assume that $d^0, \dots, d^l \in \mathbb{R}^n \setminus \{0\}$ with $(d^i)^T A d^j = 0$ for all $i \neq j$, $i, j = 0, \dots, l$ have been constructed (so far, we're not specifying *how* they have been constructed), where $l \in \{0, 1, \dots, n-2\}$.

For $k \in \{0, \dots, l\}$, we have (see Lemma (8.1))

$$t_k = \frac{-(g^k)^T d^k}{(d^k)^T A d^k} \text{ and } (g^{k+1})^T d^j = 0 \text{ for } j = 0, \dots, k.$$

Recall that $x^{l+1} = x^l + t_l d^l$. Assume that $g^{l+1} := Ax^{l+1} - b \neq 0$. (In other words: that $x^{l+1}$ is not a solution of (8.1)). Let us make use of an ansatz for $d^{l+1}$:

$$d^{l+1} := -g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i. \tag{8.8}$$

We must have

$$\begin{aligned}
&\forall\, j = 0, \ldots, l: \ (d^{l+1})^T A d^j = 0 \\
\Leftrightarrow\, &\forall\, j = 0, \ldots, l: \ (g^{l+1})^T A d^j = \beta_j^l (d^j)^T A d^j \\
\Leftrightarrow\, &\forall\, j = 0, \ldots, l: \ \beta_j^l = \frac{(g^{l+1})^T A d^j}{(d^j)^T A d^j}.
\end{aligned} \tag{8.9}$$

This gives an expression for $d^{l+1}$.

This construction has two advantages:

- By (8.8) and (8.5), we have

$$\nabla f(x^{l+1})^T d^{l+1} = (g^{l+1})^T d^{l+1} = -(g^{l+1})^T g^{l+1} = -\left\|g^{l+1}\right\|^2 < 0.$$

  This implies, on the one hand, that $d^{l+1} \neq 0$ and, on the other hand, that $d^{l+1}$ is a descent direction of $f$ at $x^{l+1}$.

- Furthermore, we have

$$t_{l+1} = -\frac{(g^{l+1})^T d^{l+1}}{(d^{l+1})^T A d^{l+1}} > 0.$$

Now, assume that $d^0, \ldots, d^l \in \mathbb{R}^n \setminus \{0\}$ with $(d^i)^T A d^j = 0$ for all $i \neq j$, $i, j = 0, \ldots, l$ have been constructed with the approach we just described, where $l \in \{0, 1, \ldots, n-2\}$.

Then we have for all $k = 0, \ldots, l+1$:

$$(g^k)^T d^k = -\left\|g^k\right\|^2 < 0 \text{ and } t_k > 0. \tag{8.10}$$

In addition, by (8.5) and (8.8), we have the following:

$$\begin{aligned}
j = 0 &: (g^{l+1})^T g^j = (g^{l+1})^T(-d^0) = 0 \\
\forall\, j = 1, \ldots, l &: (g^{l+1})^T g^j = (g^{l+1})^T(-d^j + \sum_{i=0}^{j-1} \beta_i^{j-1} d^i) = 0.
\end{aligned}$$

So, we have $(g^{l+1})^T g^j = 0$ for all $j = 0, \ldots, l$. This implies

$$g^{j+1} - g^j = Ax^{j+1} - Ax^j = t_j A d^j \ \Rightarrow\ A d^j = \frac{1}{t_j}(g^{j+1} - g^j).$$

Now let's finally look at the following:

$$(g^{l+1})^T A d^j = \frac{1}{t_j}[(g^{l+1})^T g^{j+1} - (g^{l+1})^T g^j] = 0 \ \ \forall\, j = 0, \dots, l-1.$$

Then, (8.9) implies that $\beta_j^l = 0$ for all $j = 0, \dots, l-1$. Then, recalling (8.8), we have that $d^{l+1} = -g^{l+1} + \beta_l^l d^l$. Lastly, let's simplify the expression for $\beta_l^l$ using (8.8):

$$\beta_l^l = \frac{(g^{l+1})^T A d^l}{d^l A d^l} = \frac{1}{t_l}\frac{(g^{l+1})^T (g^{l+1} - g^l)}{(d^l)^T A d^l} = \frac{\left\|g^{l+1}\right\|^2}{(d^l)^T A d^l} \cdot \frac{(d^l)^T A d^l}{(-g^l)^T d^l} = \frac{\left\|g^{l+1}\right\|^2}{\left\|g^l\right\|^2}.$$

In summary:

**Algorithm 3** CG algorithm for systems of linear equations

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, set $g^0 := Ax^0 - b$, $d^0 := -g^0$, $\varepsilon \geq 0$ and set $k := 0$
2: If $\left\|g^k\right\| \leq \varepsilon$, stop
3: Set
       $t_k := \frac{\left\|g^k\right\|^2}{(d^k)^T A d^k}$
4: Set
       $x^{k+1} := x^k + t_k d^k$, $g^{k+1} := g^k + t_k A d^k$, $\beta_k := \frac{\left\|g^{k+1}\right\|^2}{\left\|g^k\right\|^2}$, $d^{k+1} := -g^{k+1} + \beta_k d^k$
5: Set $k := k + 1$ and go to Step 2
```

For the convergence analysis, we will assume $\varepsilon = 0$.

**Example 8.2.** This example is temporarily here to fix the numbering of the theorems (the CG method for linear systems should be Algorithm 8.2).

**Remark.** Note that in Step 4 of Algorithm 3, we could have set $g^{k+1} := Ax^{k+1} - b$. This however would require a new matrix-vector multiplication $Ax^{k+1}$. We already have $Ad^k$ available from Step 3, so it is more efficient to set $g^{k+1} := g^k + t_k A d^k$.

**Theorem 8.3** (Convergence theorem for CG algorithm for linear systems)**.** *Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite, $b \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = \frac{1}{2}x^T A x - b^T x$. Then Algorithm 3 provides after at most $n$ steps the minimum $x^*$ of $f$.*

*Proof.* If $Ax^m - b = g^m = 0$ for $m < n$, the algorithm stops for $x^* = x^m$. Otherwise, we get $d^0, \dots, d^{m-1} \in \mathbb{R}^n \setminus \{0\}$ $A$-conjugate directions, and according to Lemma (8.1), we have $x^* = x^n$. □

**Remark 8.4.** We do not need to store $A$ as a matrix, we only need a formula for the map $d \mapsto Ad$. This makes a big difference!

## 8.2 The Fletcher-Reeves algorithm

We "extend" the algorithm 8.2 to the solving of optimization problems of the form (4.1):

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a a continuously differentiable function, by taking $g^k := \nabla f(x^k)$ and by using the strong Wolf-Powell strategy as a step size rule.

**Algorithm 4** Fletcher-Reeves algorithm

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, $0 < \sigma < \rho < \frac{1}{2}$, $d^0 := -\nabla f(x^0)$, $\varepsilon \geq 0$ and set $k := 0$
2: If $\|\nabla f(x^k)\| \leq \varepsilon$, stop
3: Choose $t_k > 0$ such that
       $f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k \nabla f(x^k)^T d^k$ and $|\nabla f(x^k + t_k d^k)^T d^k| \leq -\rho \nabla f(x^k)^T d^k$
4: Set
       $x^{k+1} := x^k + t_k d^k$, $\beta_k^{FR} := \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}$, $d^{k+1} := -\nabla f(x^{k+1}) + \beta_k^{FR} d^k$
5: Set $k := k + 1$ and go to Step 2
```

**Example 8.5.** This example is temporarily here to fix the numbering of the theorems (the Fletcher-Reeves algorithm should be Algorithm 8.5).

**Remark 8.6.** Different from the general formulation of the strong Wolfe-Powell strategy, we choose $\rho \in (\sigma, \frac{1}{2})$. This more restrictive assumption will be essential in the convergence analysis.

For the convergence analysis, we assume $\varepsilon = 0$.

**Theorem 8.7.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and bounded from below. Then the Fletcher-Reeves algorithm is well defined.*

*Proof.* We will show that in every iteration, the step size $t_k > 0$ can be chosen according to the strong Wolfe-Powell strategy. According to Theorem (5.2) (a), it is enough to prove that

$$\nabla f(x^k)^T d^k < 0 \;\; \forall\, k \geq 0.$$

We will prove by induction that for all $k \geq 0$, we have

$$-\sum_{j=0}^{k} \rho^j \leq \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} \leq -2 + \sum_{j=0}^{k} \rho^j. \tag{8.11}$$

Indeed, we have for $k = 0$:

$$-1 \leq \frac{\nabla f(x^0)^T d^0}{\|\nabla f(x^0)\|^2} = -1 \leq -2 + 1 = -1,$$

where the first equality holds by the choice of $d^0$. Now, let (8.11) be fulfilled for a $k \geq 0$. Then we have for $k+1$, by Step 3 of Algorithm 4:

$$\begin{aligned} & \rho \nabla f(x^k)^T d^k \leq \nabla f(x^{k+1})^T d^k \leq -\rho \nabla f(x^k)^T d^k \\ \Leftrightarrow \; & -1 + \rho \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} \leq -1 + \frac{\nabla f(x^{k+1})^T d^k}{\|\nabla f(x^k)\|^2} \leq -1 - \rho \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2}, \end{aligned} \tag{8.12}$$

where we used that $x^{k+1} = x^k + t_k d^k$. Moreover, we have by Step 4 of Algorithm 4:

$$\frac{\nabla f(x^{k+1})^T d^k}{\|\nabla f(x^{k+1})\|^2} = \frac{\nabla f(x^{k+1})^T (-\nabla f(x^{k+1} + \beta_k^{FR} d^k}{\|\nabla f(x^{k+1})\|^2} = -1 + \frac{\nabla f(x^{k+1})^T d^k}{\|\nabla f(x^k)\|^2}, \tag{8.13}$$

where the second equality follows by definition of $\beta_k^{FR}$. This implies:

$$\begin{aligned}
-\sum_{j=0}^{k+1}\rho^j &= -1-\rho\sum_{j=0}^{k}\rho^j \\
&\leq -1+\rho\frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} && \text{By (8.11)} \\
&\leq -1+\frac{\nabla f(x^{k+1})^T d^k}{\|\nabla f(x^k)\|^2} && \text{By (8.12)} \\
&= \frac{\nabla f(x^{k+1})^T d^{k+1}}{\|\nabla f(x^{k+1})\|^2} && \text{By (8.13)} \\
&\leq -1-\rho\frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} && \text{By (8.12)} \\
&\leq -1+\rho\sum_{j=0}^{k}\rho^j && \text{By (8.11)} \\
&= -2+\sum_{j=0}^{k+1}\rho^j,
\end{aligned}$$

which proves that (8.11) holds for all $k \geq 0$. Then we have for all $k \geq 0$:

$$\frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} < -2+\sum_{j=0}^{\infty}\rho^j = -2+\frac{1}{1-\rho} = \frac{2\rho-1}{1-\rho} < 0.$$

□

Next, we want to analyze the convergence properties of the sequence $(x^k)_{k\geq 0}$ generated by Algorithm (4). Ideally, we would like to show that the angle condition holds, but this is very difficult to prove. So we use another technique.

**Theorem 8.8.** *Let $x^0 \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, bounded from below and let $\nabla f$ be Lipschitz continuous on $\mathcal{L}(x^0)$. Then, for every sequence $(x^k)_{k\geq 0}$ generated by Algorithm 4, it holds*

$$\liminf_{k\to\infty} \left\|\nabla f(x^k)\right\| = 0$$

*(i.e., there exist a subsequence $(x^{k_l})_{l\geq 0} \subseteq \mathcal{L}(x^0)$? such that $\nabla f(x^{k_l}) \to 0$ as $l \to \infty$).*

*Proof.* Brace yourselves: this is a beautiful proof. Let $(x^k)_{k\geq 0}$ be a sequence generated by Algorithm 4. By (8.11), we have for all $k \geq 0$:

$$
\begin{aligned}
-\frac{1}{1-\rho} = -\sum_{j=0}^{\infty} \rho^j \leq -\sum_{j=0}^{j} \rho^j \leq \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} \\
\leq -2 + \sum_{j=0}^{k} \rho^j \leq -2 + \sum_{j=0}^{\infty} \rho^j = \frac{2\rho - 1}{1-\rho} < 0. \qquad (8.14)
\end{aligned}
$$

Assume that

$$\liminf_{k\to\infty} \left\|\nabla f(x^k)\right\| \neq 0.$$

Since $\left\|\nabla f(x^k)\right\| \geq 0$ for all $k \geq 0$, this means that

$$\liminf_{k\to\infty} \left\|\nabla f(x^k)\right\| > 0.$$

This means that

$$\sup_{n\geq 0} \inf_{k\geq n} \{\left\|\nabla f(x^k)\right\|\} > 0,$$

which means that

$$\exists\, \delta > 0 \text{ s.t. } \left\|\nabla f(x^k)\right\| > \delta \;\; \forall\, k \geq 0. \qquad (*)$$

Note that all of this is possible because we assume that $\varepsilon = 0$ (the algorithm doesn't stop). We know from Theorem (5.2) (b) that there exists $\theta > 0$ such that for all $k \geq 0$, we have

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

By (8.14), this is equivalent to the following for all $k \geq 0$:

$$
\begin{aligned}
f(x^k) - f(x^{k+1}) &\geq \theta \left( \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|^2} \right)^2 \cdot \frac{\left\|\nabla f(x^k)\right\|^4}{\|d^k\|^2} \\
&\geq \theta \left( \frac{2\rho - 1}{1-\rho} \right)^2 \cdot \frac{\left\|\nabla f(x^k)\right\|^4}{\|d^k\|^2}. \qquad (8.15)
\end{aligned}
$$

Define

$$\frac{1}{\gamma_k} := \frac{\left\|\nabla f(x^k)\right\|^4}{\left\|d^k\right\|^2},$$

then we have

$$\begin{aligned}
\gamma_k &= \frac{(d^k)^T d^k}{\|\nabla f(x^k)\|^4} \\
&= \frac{(-\nabla f(x^k) + \beta_{k-1}^{FR} d^{k-1})^T(-\nabla f(x^k) + \beta_{k-1}^{FR} d^{k-1})}{\|\nabla f(x^k)\|^4} \\
&= \frac{1}{\|\nabla f(x^k)\|^2} - 2\frac{\left\|\nabla f(x^k)\right\|^2}{\|\nabla f(x^{k+1})\|^2} \cdot \frac{\nabla f(x^k)^T d^{k-1}}{\|\nabla f(x^k)\|^4} + \frac{\left\|\nabla f(x^k)\right\|^4}{\|\nabla f(x^{k+1})\|^4} \cdot \frac{\left\|d^{k-1}\right\|^2}{\|\nabla f(x^k)\|^4} \\
&\leq \frac{1}{\|\nabla f(x^k)\|^2} - 2\rho \cdot \frac{\nabla f(x^{k-1})^T d^{k-1}}{\|\nabla f(x^{k-1})\|^2} \cdot \frac{1}{\|\nabla f(x^k)\|^2} + \frac{\left\|d^{k-1}\right\|^2}{\|\nabla f(x^{k-1})\|^4} \\
&\leq \frac{1}{\|\nabla f(x^k)\|^2} + \frac{2\rho}{1-\rho} \cdot \frac{1}{\|\nabla f(x^k)\|^2} + \gamma_{k-1} \\
&= \frac{1+\rho}{1-\rho} \cdot \frac{1}{\|\nabla f(x^k)\|^2} + \gamma_{k-1},
\end{aligned}$$

where we used the definitions of $\beta_k^{FR}$, $\gamma_k$, of the Strong Wolfe-Powell strategy (first inequality) and (8.14) (second inequality). Telescoping leads to the following for all $k \geq 1$:

$$\begin{aligned}
\gamma_k &\leq \frac{1+\rho}{1-\rho} \sum_{j=1}^{k} \frac{1}{\|\nabla f(x^j)\|^2} + \gamma_0 \\
&= \frac{1+\rho}{1-\rho} \sum_{j=1}^{k} \frac{1}{\|\nabla f(x^j)\|^2} + \frac{1}{\|\nabla f(x^0)\|^2} \\
&\leq \frac{1+\rho}{1-\rho} \sum_{j=0}^{k} \frac{1}{\|\nabla f(x^j)\|^2} \\
&\leq (k+1)\frac{1+\rho}{1-\rho} \cdot \frac{1}{\delta^2},
\end{aligned}$$

where the last inequality follows from (*). Then, (8.15) implies that for all $k \geq 0$, we have

$$f(x^k) - f(x^{k+1}) \geq \theta\left(\frac{2\rho - 1}{1-\rho}\right)^2 \cdot \delta^2 \cdot \frac{1-\rho}{1+\rho} \cdot \frac{1}{k+1} = c \cdot \frac{1}{k+1}.$$

Again, telescoping leads to the following for all $k \geq 1$:

$$f(x^0) - f(x^{k+1}) \geq c \sum_{j=0}^{k} \frac{1}{j+1}.$$

Since the LHS of this inequality is a real number and the RHS goes to infinity, we have a contradiction. □

**Remark 8.9.** Theorem (8.8) is weaker than Theorem (6.3).

Indeed, let $(x^{k_l})_{l\geq 0}$ be a subsequence of $(x^k)_{k\geq 0}$ such that

$$\lim_{l\to\infty} \nabla f(x^{k_l}) = 0.$$

Assume that $(x^{k_l})_{l\geq 0}$ has an accumulation point $x^*$. This means that there exist a subsequence $(x^{k_{l_t}})_{t\geq 0}$ of $(x^{k_l})_{l\geq 0}$ such that

$$\lim_{t\to\infty} x^{k_{l_t}} = x^*.$$

By continuity of $\nabla f$, we get

$$\lim_{t\to\infty} \nabla f(x^{k_{l_t}}) = \nabla f(x^*).$$

On the other hand, by Theorem (8.8), we know that $\lim_{t\to\infty} \nabla f(x^{k_{l_t}}) = 0$, and therefore we have $\nabla f(x^*) = 0$ which implies that $x^*$ is a critical point of $f$.

The subsequence $(x^{k_l})_{l\geq 0}$ has an accumulation point if, for instance, $\mathcal{L}(x^0)$ is bounded. This is the case if, for instance, $f$ is coercive, which means that

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

**Theorem 8.10.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, $x^0 \in \mathbb{R}^n$, $\mathcal{L}(x^0)$ convex and $f$ strongly convex on $\mathcal{L}(x^0)$. Then the whole sequence $(x^k)_{k\geq 0}$ generated by Algorithm 4 converges to the unique minimizer $x^*$ of $f$.*

*Proof.* By Exercise 24, we know that $\mathcal{L}(x^0)$ is compact. Furthermore, by Exercise 19, we know that $\nabla f$ is Lipschitz continuous on $\mathcal{L}(x^0)$. We claim that $f$ is bounded from below. Indeed, by Exercise 24 (c), we know the following

$$\begin{aligned}
&\forall\, x \in \mathcal{L}(x^0) : f(x) \geq f(x^*) + \mu \left\|x - x^*\right\|^2 \geq f(x^*) \\
&\forall\, x \notin \mathcal{L}(x^0) : f(x) > f(x^0),
\end{aligned}$$

so $f$ is bounded from below. By Theorem (8.8), we know that there exists a subsequence $(x^{k_j})_{j\geq 0} \subseteq \mathcal{L}(x^0)$ of $(x^k)_{k\geq 0}$ such that $\nabla f(x^{k_j}) \to 0$ as $j \to \infty$. Since $\mathcal{L}(x^0)$ is bounded, rest of the proof missing. □

**Remark.** Under the assumptions of Theorem (8.10), if $x^*$ is the unique minimizer of $f$ on $\mathcal{L}(x^0)$, then $x^*$ is the unique minimizer of $f$ on $\mathbb{R}^n$. Indeed, assume that is not the case. Then there would exist $x' \in \mathbb{R}^n$ with $f(x') < f(x^*) \leq f(x^0)$, which would mean that $x' \in \mathcal{L}(x^0)$, which is a contradiction.

# 9 The Newton algorithm

Unlike all the methods we have considered so far, Newton's method uses second order information. In our setting, this is a matrix (the Hessian), and usually one does not want to have to store matrices. This is the reason why Newton's method is not really used in practice. So what makes this method great? It converges very fast!

## 9.1 Convergence rates

**Definition 9.1.** Let $(x^k)_{k\geq 0} \subseteq \mathbb{R}^n$ be a given sequence. (p-convergence)

(a) The sequence $(x^k)_{k\geq 0}$ is said to *converge linearly* to $x^* \in \mathbb{R}^n$ if there exist $q \in (0,1)$ such that

$$\left\|x^{k+1} - x^*\right\| \leq q \left\|x^k - x^*\right\| \quad \forall\, k \geq k_0.$$

(b) The sequence $(x^k)_{k\geq 0}$ is said to *converge superlinearly* to $x^* \in \mathbb{R}^n$ is there exist a sequence $(\varepsilon_k)_{k\geq 0} \searrow 0$ such that

$$\left\|x^{k+1} - x^*\right\| \leq \varepsilon_k \left\|x^k - x^*\right\| \quad \forall\, k \geq 0.$$

(c) If $x^k \to x^*$ as $k \to \infty$, then $(x^k)_{k\geq 0}$ is said to *converge quadratically* to $x^* \in \mathbb{R}^n$ is there exists $Q > 0$ such that

$$\left\|x^{k+1} - x^*\right\| \leq Q \left\|x^k - x^*\right\|^2 \quad \forall\, k \geq 0.$$

**Remark 9.2.** Superlinear convergence implies linear convergence, and linear convergence implies convergence, because

$$\left\|x^{k+1} - x^*\right\| \leq q^{k-k_0+1} \left\|x^0 - x^k\right\| \to 0 \ (k \to \infty).$$

However, the last condition alone in Definition (9.1) (c) does not imply convergence, reason why we assume it.

## 9.2 The Newton algorithm for nonlinear equations

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable mapping. Note: there is an extension of this algorithm for $F : \mathbb{R}^n \to \mathbb{R}^m$ but we won't consider it. We want to find $x^* \in \mathbb{R}^n$ that solves the nonlinear equation

$$F(x) = 0. \tag{9.1}$$

Assume that $x^k$ is an approximation of $x^*$. In order to find $x^{k+1}$, we consider the linearization of $F$ at $x^k$:

$$F_k : \mathbb{R}^n \to \mathbb{R}^n, \quad F_k(x) = F(x^k) + \nabla F(x^k)(x - x^k), \tag{9.2}$$

where $\nabla F$ denotes the Jacobian of $F$. The new iterate $x^{k+1}$ is chosen as a solution of the linear system

$$F_k(x) = 0. \tag{9.3}$$

This was the idea of the method! Now let's make some things more precise. If $\nabla F(x^k)^{-1}$ exists, then

$$x^{k+1} = x^k - \nabla F(x^k)^{-1} F(x^k).$$

In general, we don't want to calculate the inverse of a matrix explicitly, because it's very costly. We actually need only a solution $d^k \in \mathbb{R}^n$ of the*Newton equation*

$$\nabla F(x^k)d = -F(x^k)$$

(this equation is linear, so a solution can be found using Gram-Schmidt or, if $\nabla F(x^k)$ is SPD, we can also use CG) and to set afterwards

$$x^{k+1} = x^k + d^k.$$

Insert picture!

**Algorithm 5** Newton algorithm for nonlinear equations

```
1: Choose starting point $x^0 \in \mathbb{R}^n$ and set $k := 0$
2: If $F(x^k) = 0$, stop
3: Find $d^k \in \mathbb{R}^n$ as a solution of the Newton equation
       $$\nabla F(x^k)d = -F(x^k)$$
4: Set $x^{k+1} := x^k + d^k$, $k := k + 1$ and go to Step 2
```

**Example 9.3.** This example is temporarily here to fix the numbering of the theorems (the Newton algorithm should be Algorithm 9.3).

The biggest drawback of the Newton algorithm so far is that it is a local method (i.e., $x^0$ must be chosen in a neighbourhood of $x^*$).

In the following, the symbol $\|\cdot\|$ denotes the operator norm induced by the Euclidean norm.

**Lemma 9.4** (Banach Lemma). *(a) Let $M \in \mathbb{R}^{n\times n}$ be a matrix with $\|M\| < 1$. Then $\mathbb{I} - M$ is regular (and therefore invertible) and it holds*

$$\left\|(\mathbb{I} - M)^{-1}\right\| \leq \frac{1}{1 - \|M\|}.$$

*(b) Let $A, B \in \mathbb{R}^{n\times n}$ with $\|\mathbb{I} - BA\| < 1$, then $A$ and $B$ are regular and it holds*

$$\left\|B^{-1}\right\| \leq \frac{\|A\|}{1 - \|\mathbb{I} - BA\|}.$$

(Proof: Exercise)

**Lemma 9.5.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable, $x^* \in \mathbb{R}^n$ and $\nabla F(x^*)$ regular. Then there exists $\varepsilon > 0$ and $c > 0$ such that for all $x \in U_\varepsilon(x^*)$, the matrix $\nabla F(x)$ is regular and $\|\nabla F(x)^{-1}\| \leq c$.*

*Proof.* By (definition of) continuity of $\nabla F$, there exists $\varepsilon > 0$ such that for all $x \in U_\varepsilon(x^*)$, it holds

$$\|\nabla F(x) - \nabla F(x^*)\| \leq \frac{1}{2\left\|\nabla F(x^*)^{-1}\right\|}.$$

Note that the expression on the LHS is well-defined, because $\nabla F(x^*)$ is regular. Recall the following facts for $A, B \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$:

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \;\Rightarrow\; \|Ax\| \leq \|A\|\,\|x\| \;\Rightarrow\; \|A\|\,\|B\|.$$

Then we have for all $x \in U_\varepsilon(x^*)$:

$$\begin{aligned}\left\|\mathbb{I} - \nabla F(x^*)^{-1}\nabla F(x)\right\| &= \left\|\nabla F(x^*)^{-1}(\nabla F(x^*) - \nabla F(x))\right\| \\ &\leq \left\|\nabla F(x^*)^{-1}\right\|\|\nabla F(x^*) - \nabla F(x)\| \leq \frac{1}{2} < 1.\end{aligned}$$

By Lemma (9.4) (b), we get that $\nabla F(x)$ is regular and

$$\left\|\nabla F(x^*)^{-1}\right\| \leq \frac{\|\nabla F(x^*)^{-1}\|}{1 - \|\mathbb{I} - \nabla F(x^*)^{-1}\nabla F(x)\|} \leq 2\left\|\nabla F(x^*)^{-1}\right\| =: c.$$

□

**Lemma 9.6.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be an operator, $x^* \in \mathbb{R}^n$ and $(x^k)_{k\geq 0} \to x^*$ as $k \to \infty$.*

*(a) If $F$ is continuously differentiable, then there exists $(\varepsilon_k)_{k\geq 0} \searrow 0$ such that for all $k \geq 0$, we have* (Lin. convergence)

$$\left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \leq \varepsilon_k \left\|x^k - x^*\right\|.$$

*(b) If $F$ is continuously differentiable and $\nabla F$ is locally Lipschitz continuous at $x^*$, then there exists $C > 0$ such that for all $k \geq 0$, we have* (quadratic)

$$\left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \leq C \left\|x^k - x^*\right\|^2.$$

*Proof.* Recall that (Fréchet) differentiabilty of $F$ at $x^*$ means

$$\lim_{x\to x^*} \frac{\|F(x) - F(x^*) - \nabla F(x^*)(x - x^*)\|}{\|x - x^*\|} = 0. \tag{*}$$

(a) By the triangle inequality, we have for all $k \geq 0$:

$$\begin{aligned}&\left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \leq \\ &\leq \left\|F(x^k) - F(x^*) - \nabla F(x^*)(x^k - x^*)\right\| + \left\|(\nabla F(x^k) - \nabla F(x^*))(x^k - x^*)\right\|.\end{aligned}$$

Let's look at the two summands separately. First, define

$$\varepsilon_k' := \begin{cases} \frac{\left\|F(x^k)-F(x^*)-\nabla F(x^*)(x^k-x^*)\right\|}{\left\|x^k-x^*\right\|}, & \text{if } x^k \neq x^* \\ 0, & \text{otherwise} \end{cases}.$$

By (*), we have that $\varepsilon_k' \searrow 0$ as $x^k \to x^*$. Next, define

$$\varepsilon_k'' := \left\|\nabla F(x^k) - \nabla F(x^*)\right\|.$$

By continuity of $\nabla F$, we have that also $\varepsilon_k'' \searrow 0$ as $x^k \to x^*$. Define $\varepsilon_k := \varepsilon_k' + \varepsilon_k''$. Then, we get that for all $k \geq 0$:

$$\left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \leq \varepsilon_k \left\|x^k - x^*\right\|.$$

(b) Let $U_\delta(x^*)$ be the neighbourhood of $x^*$ where $\nabla F$ is L-Lipschitz continuous. Let $k_0 \geq 0$ be such that $x^k \in U_\delta(x^*)$ for all $k \geq k_0$. Then, by the Mean Value

Theorem in integral form, the following holds for all $k \geq k_0$:

$$F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*) =$$
$$= \left[ \int_0^1 \nabla F(x^* + t(x^k - x^*))(x^k - x^*)\, dt \right] - \nabla F(x^k)(x^k - x^*)$$
$$= \int_0^1 [\nabla F(x^* + t(x^k - x^*)) - \nabla F(x^k)](x^k - x^*)\, dt.$$

Taking the norm on both sides and invoking the $L$-Lipschitz continuity of $\nabla F$ (note that $x^* + t(x^k - x^* \in U_\delta(x^*))$, we get the following:

$$\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\| \leq$$
$$\leq \int_0^1 \|\nabla F(x^* + t(x^k - x^*)) - \nabla F(x^k)\| \left\|x^k - x^*\right\|\, dt$$
$$\leq \int_0^1 L(1-t) \left\|x^k - x^*\right\|^2\, dt = \frac{L}{2} \left\|x^k - x^*\right\|^2.$$

Define $C := L/2$ and this proves the claim for $k \geq k_0$. However, extending the proof for $k \geq 0$ is an easy exercise, or at least that is what the Prof. said.

□

**Theorem 9.7** (Local convergence of the Newton algorithm)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable, $x^*$ a solution of (9.1) and $\nabla F(x^*)$ a regular matrix. Then there exist $\varepsilon > 0$ such that for all $x^0 \in U_\varepsilon(x^*)$, the following statements hold:*

*(a) Algorithm 5 is well defined and generates a sequence which converges to $x^*$.*

*(b) The convergence of Algorithm 5 is superlinear.*

*(c) If $\nabla F$ is locally Lipschitz continuous at $x^*$, then the convergence is quadratic.*

*Proof.* (a) By Lemma (9.5), we know that there exist $\varepsilon_1, c > 0$ such that for all $x \in U_{\varepsilon_1}(x^*)$, the matrix $\nabla F(x)$ is regular and $\|\nabla F(x)^{-1}\| \leq c$. Moreover, since $F$ is differentiable and $\nabla F$ is continuous at $x^*$, we know that there exist $\varepsilon_2 > 0$ such that for all $x \in U_{\varepsilon_2}(x^*)$, it holds

$$\|F(x) - F(x^*) - \nabla F(x)(x - x^*)\| \leq$$
$$\leq \|F(x) - F(x^*) - \nabla F(x^*)(x - x^*)\| + \|\nabla F(x) - \nabla F(x^*)\| \|x - x^*\|$$
$$\leq \frac{1}{4c} \|x - x^*\| + \frac{1}{4c} \|x - x^*\| = \frac{1}{2c} \|x - x^*\|.$$

Let $\varepsilon := \min\{\varepsilon_1, \varepsilon_2\}$. For $x^0 \in U_\varepsilon(x^*)$, we know that $\nabla F(x^0)$ is regular and therefore $x^1 := x^0 - \nabla F(x^0)^{-1} F(x^0)$ is well defined. Furthermore, we have the following (don't forget that $F(x^*) = 0$):

$$\begin{aligned}
\left\|x^1 - x^*\right\| &= \left\|x^0 - \nabla F(x^0)^{-1} F(x^0) - x^0\right\| \\
&= \left\|\nabla F(x^0)^{-1}(F(x^*) - F(x^0) + \nabla F(x^0)(x^* - x^0))\right\| \\
&\leq \left\|\nabla F(x^0)^{-1}\right\| \left\|-F(x^*) + F(x^0) - \nabla F(x^0)(x^0 - x^*)\right\| \\
&\leq c \cdot \frac{1}{2c} \left\|x^0 - x^*\right\| = \frac{1}{2} \left\|x^0 - x^*\right\|,
\end{aligned}$$

which implies that $x^1 \in U_\varepsilon(x^*)$ and therefore that $x^2$ is well defined. Repeating this argument for all $k \geq 0$, we get

$$\left\|x^{k+1} - x^*\right\| \leq \frac{1}{2} \left\|x^k - x^*\right\| \leq \left(\frac{1}{2}\right)^k \left\|x^0 - x^*\right\|, \tag{*}$$

which implies that $(x^k)_{k\geq 0} \subseteq U_\varepsilon(x^*)$ and therefore that the method is well defined. Lastly, (*) shows that $x^k \to x^*$ as $k \to \infty$.

(b) We have for all $k \geq 0$:

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\| &= \left\|x^k - \nabla F(x^k)^{-1} F(x^k) - x^*\right\| \\
&= \left\|\nabla F(x^k)^{-1}(-F(x^k) + \nabla F(x^k)(x^k - x^*))\right\| \\
&= \left\|\nabla F(x^k)^{-1}(F(x^k) - \nabla F(x^k)(x^k - x^*))\right\| \\
&= \left\|\nabla F(x^k)^{-1}(F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*))\right\| \\
&\leq \left\|\nabla F(x^k)^{-1}\right\| \left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \\
&\leq c \left\|F(x^k) - F(x^*) - \nabla F(x^k)(x^k - x^*)\right\| \qquad (**)
\end{aligned}$$

By lemma (9.6) (a), we have that for some $\varepsilon_k \searrow 0$:

$$(**) \leq c \cdot \varepsilon_k \cdot \left\|x^k - x^*\right\|,$$

which shows (b).

(c) By lemma (9.6) (b), we have that for some $C > 0$:

$$(**) \leq c \cdot C \cdot \left\|x^k - x^*\right\|^2,$$

which shows (c).

□

Let's now emphasize the importance of the assumptions in Theorem (9.7).

**Example 9.8.** (a) The method may fail when $F$ is not continuously differentiable in a neighbourhood of $x^*$. For instance, take

$$\begin{aligned}&F : \mathbb{R} \to \mathbb{R}\\ &F(x) = x^{1/3}\\ &F'(x) = \frac{1}{3}x^{-2/3} \text{ for } x \neq 0.\end{aligned}$$

At $0$, the function $F$ is not differentiable. Furthermore, the unique zero is $x^* = 0$. For $x^0 \in \mathbb{R} \setminus \{0\}$, we have for all $k \geq 0$:

$$x^{k+1} = x^k - \frac{(x^k)^{1/3}}{\frac{1}{3}(x^k)^{-2/3}} = -2x^k = 4x^{k-1} = \ldots = (-2)^{k+1}x^0 \not\to 0 = x^*.$$

(b) The method may fail when the starting point is not close enough to $x^*$. For instance, take

$$\begin{aligned}&F : \mathbb{R} \to \mathbb{R}\\ &F(x) = x^3 - 2x + 2\\ &F'(x) = 3x^2 - 2,\end{aligned}$$

Insert picture. For the unique zero $x^*$, it holds $x^* \in (-2, -1)$. We have for all $k \geq 0$:

$$x^{k+1} = x^k - \frac{(x^k)^3 - 2x^k + 2}{3(x^k)^2 - 2}.$$

Then, for $x^0 = 0$, we have $x^1 = 1$ and $x^2 = 0$, so the method alternates between $0$ and $1$. In particular, it does not converge to $x^*$. This happens because $x^0$ is too far away from $x^*$.

(c) The convergence is not always quadratic. For instance, take

$$\begin{aligned}&F : \mathbb{R} \to \mathbb{R}\\ &F(x) = x + x^{4/3}\\ &F'(x) = 1 + \frac{4}{3}x^{1/3}.\end{aligned}$$

In particular, $F'(0) = 1$ is invertible, and thus, by Theorem (9.7), there exists $\varepsilon > 0$ such that for all $x \in U_\varepsilon(x^*)$:

$$x^{k+1} = x^k - \frac{x^k + (x^k)^{4/3}}{1 + \frac{4}{3}(x^k)^{1/3}} = \frac{\frac{1}{3}(x^k)^{4/3}}{1 + \frac{4}{3}(x^k)^{1/3}} \to 0 \text{ as } k \to \infty$$

superlinearly. However, if we check the definition of quadratic convergence, we see the following:

$$\frac{|x^{k+1} - 0|}{|x^k - 0|^2} = \frac{1}{3}\left|\frac{(x^k)^{4/3}}{(x^k)^2 + \frac{4}{3}(x^k)^{7/3}}\right| = \frac{1}{3}\left|\frac{1}{(x^k)^{2/3} + \frac{4}{3}(x^k)}\right| \to \infty \text{ as } k \to \infty,$$

since $x^k \to 0$. Therefore, the convergence is not quadratic. This happens because $F'$ is not locally Lipschitz continuous at $0$. Indeed, there is no $L > 0$ and $\varepsilon > 0$ such that for all $x, y \in U_\varepsilon(0)$:

$$|F'(x) - F'(y)|= \frac{4}{3}|x^{1/3} - y^{1/3}|\leq L|x - y|.$$

## 9.3 The Newton algorithm for optimization problems

We consider again the optimization problem (4.1):

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

but assume this time that $f : \mathbb{R}^n \to \mathbb{R}$ is *twice* continuously differentiable. The Newton algorithm for solving this problem is nothing else than the Newton algorithm for solving the nonlinear equation

$$\boxed{\nabla f(x) = 0,} \iff \min_{x\in\mathbb{R}^n} f(x) \qquad (9.4)$$

where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ denotes the gradient of $f$. The update rule reads for all $k \geq 0$:

$$x^{k+1} := x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k),$$

where $\nabla^2 f$ denotes the Hessian of $f$. In order to avoid the calculation of $(\nabla^2 f(x^k))^{-1}$, one can find $d^k \in \mathbb{R}^n$ fulfilling the Newton equation

$$\nabla^2 f(x^k) d = -\nabla f(x^k) \tag{9.5}$$

and make the update

$$x^{k+1} := x^k + d^k. \tag{9.6}$$

**Algorithm 6** Newton algorithm for optimization problems

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, $\varepsilon \geq 0$ and set $k := 0$
2: If $\|\nabla f(x^k)\| \leq \varepsilon$, stop
3: Find $d^k \in \mathbb{R}^n$ as a solution of the Newton equation
       $\nabla^2 f(x^k) d = -\nabla f(x^k)$
4: Set $x^{k+1} := x^k + d^k$, $k := k + 1$ and go to Step 2
```

**Example 9.9.** This example is temporarily here to fix the numbering of the theorems (the Newton algorithm should be Algorithm 9.9).

The local convergence theorem follows as a special case of Theorem (9.7). To this end, we set $\varepsilon = 0$ and assume that Algorithm 6 does not terminate after finitely many steps.

**Theorem 9.10** (Local convergence of Newton algorithm for optimization problems)**.** *Let $x^* \in \mathbb{R}^n$ be a critical point of (4.1), which means that $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*)$ be regular. Then there exist $\varepsilon > 0$ such that for all $x^0 \in U_\varepsilon(x^*)$ the following statements are true:*

*(a) The sequence generated by Algorithm 6 is well defined and $x^k \to x^*$ as $k \to \infty$.*

*(b) The convergence rate is superlinear.*

*(c) If $\nabla^2$ is locally Lipschitz continuous at $x^*$, the convergence is quadratic.*

**Remark 9.11.** (a) To guarantee the convergence to the critical point $x^*$, one has to start in the *unknown* neighbourhood $U_\varepsilon(x^*)$.

(b) The sequence $(x^k)_{k\geq 0}$ might convergence to a local maximum of $f$, as local maxima also satisfy $\nabla f(x^*) = 0$.

(i),(ii)

In the following, we will try to solve these issues. One can combine the fast convergence properties of the Newton method with the "global convergence features" (regarding the choice of $x^0$) of the gradient method. This is done by taking gradient steps first, in order to get in an appropriate neighbourhood of $x^*$, and then taking Newton steps and converging fast. (Newton + Gradient Method)

**Algorithm 7** Globalized Newton algorithm for optimization problems

```
1: Choose starting point $x^0 \in \mathbb{R}^n$, $\rho > 0$, $p > 2$, $\beta \in (0,1)$, $\sigma \in (0, \frac{1}{2})$, $\varepsilon \geq 0$ and set
   $k := 0$
2: If $\|\nabla f(x^k)\| \leq \varepsilon$, stop
3: Find $d^k \in \mathbb{R}^n$ as a solution of the Newton equation
       $\nabla^2 f(x^k)d = -\nabla f(x^k).$
   If the Newton equation has no solution or if
       $\nabla f(x^k)^T d^k > -\rho \|d^k\|^p,$
   take $d^k := -\nabla f(x^k)$.
4: Find $t_k := \max\{\beta^l : l = 0, 1, \ldots\}$ such that
       $f(x^k + t_k d^k) \leq f(x^k) + \sigma t_k (\nabla f(x^k))^T d^k$
5: Set $x^{k+1} := x^k + t_k d^k$, $k := k + 1$ and go to Step 2
```

**Example 9.12.** This example is temporarily here to fix the numbering of the theorems (the globalized Newton algorithm should be Algorithm 9.12).

**Remark.** In Step 3 of Algorithm 7, the existence of the solution is guaranteed if $x^k \in U_\varepsilon(x^*)$. This need not be the case. Therefore, if the Newton equation has no solution or if the solution is "bad", i.e. the equation

$$\nabla f(x^k)^T d^k \leq -\rho \|d^k\|^p \qquad (*)$$

is not fulfilled, we take a gradient step instead of a Newton step. Note that (*) not being fulfilled implies that $d^k$ is not a descent direction.

This way we combined the Newton algorithm with the gradient algorithm! We now set $\varepsilon = 0$ and assume that Algorithm 7 does not terminate after finitely many steps.

**Definition.** Let $(x^k)_{k\geq 0} \subseteq \mathbb{R}^n$ be a sequence and $x^*$ be an accumulation point of $(x^k)_{k\geq 0}$. We say that $x^*$ is an *isolated accumulation point* if there exists $\varepsilon > 0$ such that $U_\varepsilon(x^*)$ contains no further accumulation points of $(x^k)_{k\geq 0}$.

We will not prove the next theorem because it would take too much time.

**Theorem 9.13** (Convergence theorem of the globalized Newton method)**.** *Let $(x^k)_{k\geq 0}$ be a sequence generated by Algorithm 7 and $x^*$ be an accumulation point of $(x^k)_{k\geq 0}$. Then*

*(a) The element $x^*$ is a critical point of $f$.*

*(b) If $x^*$ is an isolated accumulation point, then $x^k \to x^*$ as $k \to \infty$.*

*(c) If $\nabla^2 f(x^*)$ is positive definite, then $x^*$ is a strict local minimum and $x^k \to x^*$ superlinearly as $k \to \infty$. If, in addition, $\nabla^2 f$ is locally Lipschitz continuous at $x^*$, then $x^k \to x^*$ quadratically as $k \to \infty$.*

**Remark.** Note that the whole sequence $(x^k)_{k\geq 0}$ converges! This is because of the influence of the Newton method.