1. Introduction

Optimization problem

Let $W \subset \mathbb{R}^n$ be a nonempty set and $f\colon W \to \mathbb{R}$ a function. We consider the problem of finding minima of f in W, taking in particular

- $W = \mathbb{R}^n$ (unconstrained optimization),
- $W = \{x \in \mathbb{R}^n\colon g_1(x) = 0, \dots, g_m(x) = 0\}$, where $g_1, \dots, g_m$ are functions $\mathbb{R}^n \to \mathbb{R}^n$ (equality constraints),
- $W = \{x \in \mathbb{R}^n\colon g_1(x) \leqslant 0, \dots, g_m(x) \leqslant 0\}$, where $g_1, \dots, g_m$ are functions $\mathbb{R}^n \to \mathbb{R}^n$ (inequality constraints).

The set W is called a feasible set/region.

Definition 1 *A point $x_0 \in W$ is called a global minimum of f in W if*

$$f(x) \geqslant f(x_0) \quad \textit{for all } x \in W.$$

Definition 2 *A point $x_0 \in W$ is called a local minimum of f in W if there exists $\varepsilon > 0$ such that*

$$f(x) \geqslant f(x_0) \quad \textit{for all } x \in W \cap B(x_0, \varepsilon),$$

where $B(x_0, \varepsilon)$ is the ball whose centre is x_0 and the radius is ε.

Any global minimum is a local minimum. A minimum is called strict if in the definitions above there is $f(x) > f(x_0)$ for $x \neq x_0$. In a similar way we define global and local maxima. A point x_0 is a (global or local) extremum if it is a minimum or a maximum.

Minima need not exist, if no point x_0 fulfills the definitions. A global minimum does not exist if $\inf_{x \in W} f(x) = -\infty$ or $\inf_{x \in W} f(x) = c$ and $f(x) > c$ for all $x \in W$.

Example. Let $f(x) = x \cos x$. If $W = \mathbb{R}$ then $\inf_{x \in W} f(x) = -\infty$, and there is no global minimum and an infinite set of local minima. If $W = [a, b]$, where $a, b \in \mathbb{R}$,

then a global minimum exists. If $W = (a, b)$ then minima either exist or not, depending on the choice of a, b. In general, a continuous function is not guaranteed to have extrema if the feasible set is not compact, e.g. if it is open.

Existence of minima of a continuous function

Theorem 1 *If the set $W \in \mathbb{R}^n$ is compact and $f\colon W \to \mathbb{R}$ is a continuous function, then f reaches its infimum and supremum in W, i.e., there exist $x_0, y_0 \in W$ such that*

$$f(x_0) \leqslant f(x) \leqslant f(y_0) \quad \textit{for all } x \in W.$$

Definition 3 *A function $f\colon W \to \mathbb{R}$ is called coercive if $f(x) \to \infty$ for $\|x\| \to \infty$. Equivalently,*

$$\forall_{r>0} \exists_{s>0} \forall_{x \in W} \ \|x\| > s \ \Rightarrow \ f(x) > r.$$

If W is a bounded set, then any function $f\colon W \to \mathbb{R}$ is coercive.

Theorem 2 *If $W \subset \mathbb{R}^n$ is a closed set and $f\colon W \to \mathbb{R}$ is continuous and coercive, then there exists a minimum x_0 of f in W.*

Proof. For a point $y \in W$ we define the set $U_y = \{x \in W\colon f(x) \leqslant f(y)\}$. The set U_y is nonempty and closed, as the function f is continuous and the inequality in the definition of U_y is nonsharp and W is closed. This set is also bounded: for $r = f(y)$, from the coercivity of f there exists $s > 0$ such that if $\|x\| > s$, then $f(x) > r = f(y)$; hence, $x \notin U_y$ and $U_y \subset B(0, s)$. It follows that U_y is a closed and bounded set, i.e., it is compact. Therefore there exists a global minimum x_0 of f in U_y. Due to $f(x) > f(y) \geqslant f(x_0)$ for $x \notin U_y$, x_0 is also a global minimum of f in W. □

Theorem 3 *Let $W \subset \mathbb{R}^n$ be nonempty and let $f\colon W \to \mathbb{R}$ be a continuous function. If there exists $y \in W$ such that for any sequence $(x_n)_n \subset W$ such that*

$$x_n \to \operatorname{cl} W \setminus W \quad \textit{or} \quad \|x_n\| \to \infty$$

there is $\liminf_{n\to\infty} f(x_n) > f(y)$, then there exists a minimum x_0 of the function f.

Proof. The set U_y is defined as before. To show that it is closed, we take any sequence $(\boldsymbol{x}_n)_n \subset U_y$ which converges to $\bar{\boldsymbol{x}}$. It suffices to show that $\bar{\boldsymbol{x}} \in U_y$. From $\boldsymbol{x}_n \in U_y$ we have $f(\boldsymbol{x}_n) \leqslant f(\boldsymbol{y})$ and if $\bar{\boldsymbol{x}} \notin W$, then we have an inconsistency with the assumption. Hence, $\bar{\boldsymbol{x}} \in W$. As the function f is continuous in W, there is $f(\bar{\boldsymbol{x}}) \leqslant f(\boldsymbol{y})$, hence $\bar{\boldsymbol{x}} \in U_y$. The set U_y is also bounded, which follows from the assumed implication $\|\boldsymbol{x}_n\| \to \infty \;\Rightarrow\; \liminf_{n\to\infty} f(\boldsymbol{x}_n) > f(\boldsymbol{y})$. The proof is completed just like the proof of the previous theorem. □

Local minima of functions of one variable

Let $W \subset \mathbb{R}$ be an open set.

Theorem 4 *(necessary condition of the 1st order) If $x_0 \in W$ is a local minimum or maximum of f and $f'(x_0)$ exists, then $f'(x_0) = 0$.*

Proof. Let x_0 be a local minimum. For sufficiently small $h > 0$ there is $f(x_0 - h) \geqslant f(x_0) \leqslant f(x_0 + h)$ and then

$$\begin{aligned} \frac{f(x_0-h)-f(x_0)}{-h} \leqslant 0 \quad &\Rightarrow \quad \lim_{h\to 0}\frac{f(x_0-h)-f(x_0)}{-h} \leqslant 0 \quad \Rightarrow \quad f'(x_0) \leqslant 0,\\ \frac{f(x_0+h)-f(x_0)}{+h} \geqslant 0 \quad &\Rightarrow \quad \lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{+h} \geqslant 0 \quad \Rightarrow \quad f'(x_0) \geqslant 0, \end{aligned}$$

hence, $f'(x_0) = 0$. □

Theorem 5 *(necessary condition of the 2nd order) If $f\colon W \to \mathbb{R}$ is of class $C^2(W)$ and x_0 is a local minimum, then $f''(x_0) \geqslant 0$.*

If the set W is not open, then we cannot use the above theorems for $x_0 \in \partial W$. But the theorem below applies also in this case.

Theorem 6 *(sufficient condition of the 2nd order) If $f\colon W \to \mathbb{R}$ is of class $C^2(W)$ and $f'(x_0) = 0$, $f''(x_0) > 0$ at a point $x_0 \in W$, then f has a strict local minimum at x_0.*

Theorem 7 *If $W \subset \mathbb{R}$ is open, $f \in C^k(W)$ and $f'(x_0) = f''(x_0) = \cdots = f^{(k-1)}(x_0) = 0$, $f^{(k)}(x_0) \neq 0$ for $x_0 \in W$, then if k is odd, there is no extremum of f at x_0, and if k is even, then there is a local minimum if $f^{(k)}(x_0) > 0$ and a local maximum if $f^{(k)}(x_0) < 0$.*

Taylor's formulae

Theorem 8 *(Rolle's theorem) If a function $f\colon [a,b] \to \mathbb{R}$ is continuous in $[a,b]$, differentiable in (a,b) and $f(a) = f(b)$, then there exists a point $x_0 \in (a,b)$ such that $f'(x_0) = 0$.*

Proof. If f is constant, then the claim is obvious. Otherwise there exists an extremum x_0 of f in $[a,b]$ other than a and b: there is $f(x_0) = \sup_{x\in[a,b]} f(x) > f(a)$ or $f(x_0) = \inf_{x\in[a,b]} f(x) < f(a)$. Let x_0 be a maximum. Then $f(x) \leqslant f(x_0)$ for all $x \in [a,b]$ and

$$\frac{f(x)-f(x_0)}{x-x_0} \geqslant 0 \text{ if } x < x_0, \qquad \frac{f(x)-f(x_0)}{x-x_0} \leqslant 0 \text{ if } x > x_0.$$

Hence,

$$f'(x_0) = \lim_{x\nearrow x_0} \underbrace{\frac{f(x)-f(x_0)}{x-x_0}}_{\geqslant 0} = \lim_{x\searrow x_0} \underbrace{\frac{f(x)-f(x_0)}{x-x_0}}_{\leqslant 0},$$

therefore, $f'(x_0) = 0$. If x_0 is a minimum, the proof is similar. □

Theorem 9 *(mean value theorem) If a function $f\colon [a,b] \to \mathbb{R}$ is continuous in $[a,b]$ and differentiable in (a,b), then there exists a point $x_0 \in (a,b)$ such that*

$$f(b) - f(a) = f'(x_0)(b-a).$$

Proof. Let $g(x) \overset{\text{def}}{=} \big(f(b) - f(a)\big)x - (b-a)f(x)$. The function g is continuous in $[a,b]$ and differentiable in (a,b), moreover,

$$g(a) = f(b)a - f(a)b = g(b).$$

By Rolle's theorem, there exists $x_0 \in (a,b)$ such that $g'(x_0) = 0$. Hence,

$$0 = g'(x_0) = f(b) - f(a) - (b-a)f'(x_0).$$

The proof is completed by rearranging this formula. □

Theorem 10 *(Taylor's formula with the remainder in Peano form) Let* $f\colon [a,b] \to \mathbb{R}$ *be a function differentiable in* $[a,b]$ *and twice differentiable at some point* $x_0 \in (a,b)$*. Then for all* $x \in [a,b]$ *there is*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + o\bigl((x - x_0)^2\bigr).$$

Proof. Without loss of generality we assume $x_0 = 0$. Let

$$R(x) \overset{\text{def}}{=} f(x) - f(0) - f'(0)x - \frac{f''(0)}{2}x^2.$$

We need to show that $R(x) = o(x^2)$. From the continuity of f' we obtain

$$f(x) - f(0) = \int_0^x f'(y)\,dy.$$

The function f' is differentiable at 0. Hence, $f'(y) = f'(0) + f''(0)y + r(y)$, where $r(y) = o(y)$. This means that

$$\lim_{y=0} \frac{r(y)}{y} = 0,$$

i.e., for any $\varepsilon > 0$ there exists $\delta > 0$ such that $|y| < \delta \Rightarrow |r(y)| < \varepsilon|y|$.

Now we fix an $\varepsilon > 0$ and the related $\delta > 0$. For $|x| < \delta$ we integrate $f'(y)$:

$$f(x) - f(0) = \int_0^x \bigl(f'(0) + f''(0)y + r(y)\bigr)\,dy = f'(0)x + \frac{f''(0)}{2}x^2 + \int_0^x r(y)\,dy.$$

Hence, $R(x) = \int_0^x r(y)\,dy$. Using the estimate $|r(y)| < \varepsilon|y|$ for $|y| < \delta$, we obtain

$$|R(x)| \leqslant \int_0^x |r(y)|\,dy < \int_0^x \varepsilon|y|\,dy = \frac{\varepsilon x^2}{2}.$$

Hence,

$$\left|\frac{R(x)}{x^2}\right| < \frac{\varepsilon}{2}.$$

As $\varepsilon > 0$ may be arbitrary, $\lim_{x\to 0}\left|\frac{R(x)}{x^2}\right| = 0$, i.e., $R(x) = o(x^2)$. □

Just a little more effort is needed to prove the formula with more terms, applicable for functions having derivatives up to the order $k-1$ in (a,b) and the k-th order derivative at x_0:

$$f(x) = f(x_0) + \sum_{i=1}^{k} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + o\bigl((x - x_0)^k\bigr).$$

Theorem 11 *(Taylor's formula with the remainder in Lagrange form) Let* $f\colon [a,b] \to \mathbb{R}$ *be a function of class* $C^{k-1}[a,b]$ *and* k *times differentiable in* (a,b)*. For* $x_0 \in (a,b)$ *and* $x \in [a,b]$ *there is*

$$f(x) = f(x_0) + \sum_{i=1}^{k-1} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + \frac{f^{(k)}(\overline{x})}{k!}(x - x_0)^k,$$

where $\overline{x}$ *is a point between* x_0 *and* x*.*

Proof. The function $h(x) \overset{\text{def}}{=} f(x_0) + \sum_{i=1}^{k-1} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i$ is a polynomial of degree less than k. For $x \neq x_0$ let $g_x(y) \overset{\text{def}}{=} f(y) - h(y) - z_x(y - x_0)^k$, where $z_x = \frac{f(x)-h(x)}{(x-x_0)^k}$. It is easy to verify that $g_x(x_0) = g_x'(x_0) = \cdots = g_x^{(k-1)}(x_0) = g_x(x) = 0$. By Rolle's theorem, the derivative g_x' is equal to 0 at some point x_1 between x_0 and x; note that the point x_0 is a zero of multiplicity $k-1$ of g_x'. Using the induction and Rolle's theorem in the similar way, we show the existence of the sequence of points, $x_2, \ldots, x_k$ such that $g_x^{(i)}(x_i) = 0$ and each point x_i is between x_0 and x_{i-1}.

The point $\overline{x} = x_k$ is a zero of $g_x^{(k)}$ located between x_0 and x, i.e.,

$$0 = g_x^{(k)}(x_k) = f^{(k)}(x_k) - z_x k!.$$

Hence, $z_x = \frac{f^{(k)}(x_k)}{k!}$. By substitititing this expression and $y = x$ to the definition of g_x, due to $g_x(x) = 0$, we obtain the needed formula. □

Global extrema

Theorem 12 *Let* $I \subset \mathbb{R}$ *be an interval, open or closed at one or both ends, or even unbounded. Let* $f\colon I \to \mathbb{R}$ *be of class* $C^1(I)$ *and* $C^2(\operatorname{int} I)$*. Let* $x_0 \in I$ *and* $f'(x_0) = 0$*. If* $f''(x) \geqslant 0$ *for all* $x \in I$*, then* x_0 *is a global minimum of* f*. If* $f''(x) \leqslant 0$ *for all* $x \in I$*, then* x_0 *is a global maximum of* f*. If in addition* $f''(x_0) > 0$ *or respectively* $f''(x_0) < 0$*, then* x_0 *is a unique (strict) global minimum or maximum.*

Proof. By the Taylor's formula we have

$$f(x) = f(x_0) + \frac{1}{2}f''(\overline{x})(x - x_0)^2,$$

where $\overline{x}$ is a point between x_0 and x. Hence, the last term of the formula above determines the inequality between $f(x)$ and $f(x_0)$.

Assume that $f''(x) \geqslant 0$ for all $x \in I$ and $f''(x_0) > 0$. By $f'(x_0) = 0$ we obtain

$$f'(x) = f'(x) - f'(x_0) = \int_{x_0}^{x} f''(y)\,dy \geqslant 0$$

for $x > x_0$. Similarly we show that $f'(x) \leqslant 0$ for $x < x_0$. As $f''(x_0) > 0$ and f'' is continuous, it follows that f'' is positive in a neighbourhood of x_0. Hence, the integrals are positive, which implies $f'(x) > 0$ for $x > x_0$ and $f'(x) < 0$ for $x < x_0$. Thus f is decreasing for $x < x_0$ and increasing for $x > x_0$ and, therefore, x_0 is a strict minimum. The proof for the case of maximum is similar. □

2. Extrema of functions of two or more variables

Let $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is an open set. Points of $\mathbb{R}^n$, $\mathbf{x} = (x_1, \dots, x_n)$, are identified with column matrices, $[x_1, \dots, x_n]^T$, but it is convenient to write $f(\mathbf{x})$ and $f(x_1, \dots, x_n)$, which denotes the same thing. We use the Euclidean norm, $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2 + \cdots + x_n^2}$.

The <u>gradient</u> of f is a row matrix, $Df(\mathbf{x}) = \bigl[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\bigr]$.

The <u>Hessian</u> of f at $\mathbf{x} \in W$ is the $n \times n$ matrix,

$$D^2f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

<u>Definition 4</u> *The function f is <u>differentiable</u> at $\mathbf{x}_0 \in W$ if there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \boldsymbol{\alpha}^T(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|), \quad \mathbf{x} \in W.$$

The function f is <u>twice differentiable</u> at $\mathbf{x}_0 \in W$ if in addition there exists a matrix $H \in \mathbb{R}^{n\times n}$ such that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \boldsymbol{\alpha}^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T H(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2), \quad \mathbf{x} \in W.$$

If a function is twice differentiable, then there exists a symmetric matrix H mentioned in the definition above; if a nonsymmetric matrix H satisfies the formula in this definition, so does the symmetric matrix $\frac{1}{2}(H + H^T)$.

<u>Theorem 13</u> *I) If a function f is differentiable at $\mathbf{x}_0$, then the gradient $Df(\mathbf{x}_0)$ exists and is equal to $\boldsymbol{\alpha}^T$. Conversely, if $Df(\mathbf{x})$ exists in a neighbourhood of $\mathbf{x}_0$ and is continuous at $\mathbf{x}_0$, then f is differentiable at $\mathbf{x}_0$.*

II) if the Hessian $D^2f(\mathbf{x})$ exists in a neighbourhood of $\mathbf{x}_0$ and is continuous at $\mathbf{x}_0$, then f is twice differentiable at $\mathbf{x}_0$; the Hessian is then a symmetric matrix, $H = D^2f(\mathbf{x}_0)$.

<u>Remark.</u> If the function f is differentiable at a point $\mathbf{x}$, the (real) value of the product of matrices $Df(\mathbf{x})\boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^n$, is the directional derivative of the function f in the direction of the vector $\boldsymbol{v}$ at $\mathbf{x}$. If the function f is twice differentiable, then $\boldsymbol{v}^T D^2f(\mathbf{x})\boldsymbol{v}$ is equal to the second order directional derivative of f in the direction of $\boldsymbol{v}$.

<u>Remark.</u> To use second order derivatives *in practice* we need to assume the continuity of the Hessian.

<u>Remark.</u> A function f whose domain is an *open* set $W \subset \mathbb{R}^n$ is said to be of class C^1 (C^2) in W if it is continuous in W together with its first (and second) order derivatives. If the set W is not open, the function is said to be of class C^1 (C^2) if there exists an extension $\tilde{f}$ of class C^1 (C^2) of the function f to an open set $\tilde{W}$ such that $W \subset \tilde{W}$. Then we can consider the derivatives of f at the boundary points of W; if $W \subset \mathrm{cl}(\mathrm{int}\, W)$, then (due to their continuity) the derivatives are uniquely determined by the values of f in W.

<u>Lemma 1</u> *Let $W \subset \mathbb{R}^n$ be an open set. If a function $f\colon W \to \mathbb{R}$ is of class C^2 and a line segment $\overline{\mathbf{x}_0\mathbf{x}}$ is contained in W, then*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T D^2f(\overline{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0),$$

where $\overline{\mathbf{x}}$ is an interior point the line segment $\overline{\mathbf{x}_0\mathbf{x}}$.

<u>Proof.</u> Apply the Taylor's formula to the function $g(t) = f\bigl(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)\bigr)$, $t \in [0, 1]$. □

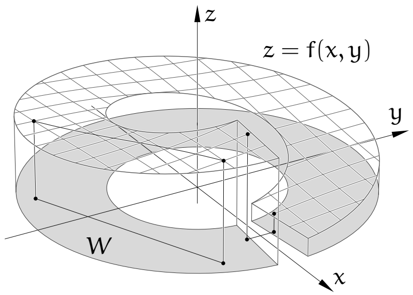

Figure 1: Applicability of the Taylor's formula for a multivariate function

Definition 5 *The set $W \subset \mathbb{R}^n$ is convex if*

$$\lambda x + (1-\lambda) y \in W$$

for all $x, y \in W$, $\lambda \in [0,1]$.

Corollary 1 *If $W \subset \mathbb{R}^n$ is open and convex and $f\colon W \to \mathbb{R}$ is of class $C^2(W)$ then for all $x_0, x \in W$ there is*

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T D^2 f(\overline{x})(x - x_0),$$

where $\overline{x} = (1-\lambda)x_0 + \lambda x$ for some $\lambda \in (0,1)$.

Proof. As the set W is convex, if $x_0, x \in W$, then $\overline{x_0 x} \subset W$; the claim follows from the lemma. □

The necessary first order condition

We consider a set $W \subset \mathbb{R}^n$ with a nonempty interior.

Theorem 14 *(necessary 1st order condition) If a function $f\colon W \to \mathbb{R}$ is differentiable at a point $x_0 \in \operatorname{int} W$ and x_0 is a local extremum of f, then $Df(x_0) = 0^T$.*

Proof. From $x_0 \in \operatorname{int} W$ it follows that the function $g_i(t) = f(x_0 + t e_i)$ (where $e_i = [0, \dots, 0, \underset{\substack{\uparrow \\ i}}{1}, 0, \dots, 0]^T$) is well defined. It has the local extremum at 0. By the necessary first order condition for functions of one variable there must be $g_i'(0) = 0$, which implies $\frac{\partial f}{\partial x_i} = 0$. As this holds for all $i = 1, \dots, n$, the gradient of f is the zero matrix $1 \times n$. □

Definition 6 *A point $x_0 \in \operatorname{int} W$ is called a critical point of the function $f\colon W \to \mathbb{R}$ if f is differentiable at x_0 and $Df(x_0) = 0^T$.*

Positive- and negative-definite matrices

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $A = [a_{ij}]$, $a_{ij} = a_{ji}$. It defines a quadratic form

$$F(x) = x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

Definition 7 *The matrix A or the quadratic form F is*

- *positive definite if $F(x) > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$ (we write $A > 0$),*
- *nonnegative definite if $F(x) \geqslant 0$ for all $x \in \mathbb{R}^n$ (we write $A \geqslant 0$),*
- *negative definite if $F(x) < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$ (we write $A < 0$),*
- *nonpositive definite if $F(x) \leqslant 0$ for all $x \in \mathbb{R}^n$ (we write $A \leqslant 0$),*
- *indefinite if there exist vectors $x, y \in \mathbb{R}^n$ such that $F(x) > 0$, $F(y) < 0$.*

At the first glance we can notice that if *not all* diagonal elements a_{ii} are positive (nonnegative) then the matrix A is not positive-definite (nonnegative-definite) and if *not all* diagonal coefficients are negative (nonpositive) then the matrix is not negative-definite (nonpositive-definite). The basic characteristic of positive-definite matrices is given by

Theorem 15 *(Sylvester's criterion) Let A_i be a matrix obtained from A by rejecting its last $n - i$ rows and columns (in particular $A_1 = [a_{11}]$, $A_n = A$).*

I) The matrix A is positive-definite if and only if $\det A_i > 0$ for $i = 1, \dots, n$,

II) The matrix A is nonnegative-definite if and only if $\det A_i \geqslant 0$ for $i = 1, \dots, n$.

A matrix A is negative-definite (nonpositive-definite) if $-A$ is positive-definite (nonnegative-definite). Another characteristic is related to the algebraic eigenproblem. From the linear algebra we know that all eigenvalues of a real symmetric matrix are real numbers; for any such a matrix there exists an orthogonal basis of $\mathbb{R}^n$ whose elements are eigenvectors of this matrix. A positive-, nonnegative-, negative- or nonpositive-definite symmetric matrix has respectively all eigenvalues positive, nonnegative, negative or nonpositive.

Conditions of the second order

Theorem 16 *(necessary 2nd order condition) If f is a function of class C^2 in an open set $W \subset \mathbb{R}^n$ and $x_0 \in W$ is a local minimum, then the matrix $D^2f(x_0)$ is nonnegative-definite. If x_0 is a local maximum, then $D^2f(x_0)$ is nonpositive-definite.*

Proof. Let x_0 be a local minimum. Let $\mathbf{h} \in \mathbb{R}^n \setminus \{0\}$ and $g(t) = f(x_0 + t\mathbf{h})$, where $t \in \mathbb{R}$ is chosen so as to obtain $x_0 + t\mathbf{h} \in W$. The function g has a local minimum at 0. As f is of class C^2, so is g. By the second order necessary condition for the univariate case, $g''(0) \geqslant 0$. The second order derivative of the composite function g is

$$g''(0) = \mathbf{h}^T D^2 f(x_0)\mathbf{h}.$$

As the vector $\mathbf{h}$ may be arbitrary, the matrix $D^2f(x_0)$ is nonnegative-definite. □

Theorem 17 *(sufficient 2nd order condition) If f is a function of class C^2 in an open set $W \subset \mathbb{R}^n$, $Df(x_0) = 0^T$ and the matrix $D^2f(x_0)$ is positive-definite (negative-definite), then x_0 is a local minimum (maximum) of f.*

Proof. Assume that $D^2f(x_0) > 0$. Let $\alpha\colon W \to \mathbb{R}$ be the function defined by

$$\alpha(x) = \inf_{\|\mathbf{h}\|=1} \mathbf{h}^T D^2 f(x)\mathbf{h}.$$

The function value $\alpha(x)$ is the minimal eigenvalue of the matrix $D^2f(x)$; the infimum is the minimum taken at the vector $\mathbf{h}$ which is a unit eigenvector corresponding to the minimal eigenvalue of the Hessian. Due to the continuity of the Hessian of f, the function α is continuous. Hence, there exists a ball $B(x_0, \varepsilon)$, $\varepsilon > 0$, such that $\alpha(x) > 0$ for all $x \in B(x_0, \varepsilon)$.

For a fixed $x \in B(x_0, \varepsilon)$, due to the Taylor's formula we have

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T D^2 f(\bar{x})(x - x_0),$$

where $\bar{x}$ is a point of the line segment $\overline{x_0x} \subset B(x_0, \varepsilon)$. The gradient of f vanishes at x_0 and

$$(x - x_0)^T D^2 f(\bar{x})(x - x_0) = \|x - x_0\|^2 \frac{(x - x_0)^T}{\|x - x_0\|} D^2 f(\bar{x}) \frac{(x - x_0)}{\|x - x_0\|} \geqslant \|x - x_0\|^2 \alpha(\bar{x}).$$

Hence,

$$f(x) - f(x_0) \geqslant \frac{1}{2}\|x - x_0\|^2 \alpha(\bar{x}) > 0.$$

It follows that x_0 is a strict local minimum. The proof for a maximum is similar. □

Global extrema

Let W be a convex set and $f\colon W \to \mathbb{R}$ a function of class $C^1(W)$ and $C^2(\operatorname{int} W)$.

Theorem 18 *If $x_0 \in \operatorname{int} W$ is a critical point of f, then*

I) If $D^2f(x) \geqslant 0$ for all $x \in \operatorname{int} W$, then x_0 is a global minimum,

II) If $D^2f(x) \leqslant 0$ for all $x \in \operatorname{int} W$, then x_0 is a global maximum.

If in addition $D^2f(x_0) > 0$ or $D^2f(x_0) < 0$ respectively, then x_0 is a strict minimum or maximum.

Proof. If $x \in W$, then by convexity of W the entire line segment $\overline{x_0x}$ is contained in W. By the Taylor's formula,

$$f(x) = f(x_0) + \frac{1}{2}(x - x_0)^T D^2 f(\bar{x})(x - x_0),$$

for a point $\bar{x} \in \overline{x_0x}$. From the inequality $D^2f(\bar{x}) \geqslant 0$ (or $D^2f(\bar{x}) \leqslant 0$) it follows that the last term above is nonnegative (or nonpositive), which proves that x_0 is a minimum (or a maximum).

If in addition to (I) we have $D^2f(x_0) > 0$, then we can consider the function $g(t) = f\bigl(x_0 + t(x - x_0)\bigr)$, $t \in [0, 1]$. Due to the convexity of W, $x_0 + t(x - x_0) \in W$, so the function g is well defined. From the assumptions it follows that $g'(0) = 0$, $g''(0) > 0$ and $g''(t) \geqslant 0$. Therefore g has a strict global minimum at 0, i.e. $f(x) > f(x_0)$. As the choice of $x \in W$ is arbitrary, x_0 is a strict global minimum of f.

The proof for the case of $D^2f(x_0) < 0$ holding in addition to (II) is similar. □

3. Convex sets and functions

Lemma 2 *The set $W \subset \mathbb{R}^n$ is convex if and only if for all $m \geqslant 2$ and for all points $x_1, \dots, x_m \in W$ and numbers $a_1, \dots, a_m \geqslant 0$, $a_1 + \cdots + a_m = 1$, there is*

$$a_1 x_1 + \cdots + a_m x_m \in W.$$

Lemma 3 *Let $W \subset \mathbb{R}^n$ be a convex set with a nonempty interior. Then*

I) For any $\boldsymbol{x} \in W$ and $\boldsymbol{x}_0 \in \operatorname{int} W$ the line segment $\overline{\boldsymbol{x}_0\boldsymbol{x}} \setminus \{\boldsymbol{x}\}$ is contained in the interior of W:

$$\lambda \boldsymbol{x}_0 + (1-\lambda)\boldsymbol{x} \in W \quad \textit{for all } \lambda \in (0, 1].$$

II) $W \subset \operatorname{cl}(\operatorname{int} W)$

Proof. Let the points $\boldsymbol{x}_0$ and $\boldsymbol{x}$ satisfy the assumptions. As $\operatorname{int} W$ is open, there exists a ball $B(\boldsymbol{x}_0, \varepsilon) \subset \operatorname{int} W$. The union of all line segments, whose one end point is $\boldsymbol{x}$ and the other end point is in this ball, is a “cone” with the vertex $\boldsymbol{x}$ and the base $B(\boldsymbol{x}_0, \varepsilon)$. This cone is a subset of W and its interior contains the line segment $\overline{\boldsymbol{x}_0\boldsymbol{x}} \setminus \{\boldsymbol{x}\}$. This completes the proof of (I). (II) follows immediately. □

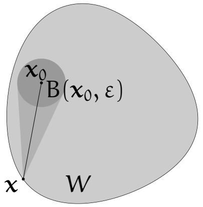

Figure 2: Illustration of Lemma 2

Theorem 19 *(weak separation theorem) Let $U, V \subset \mathbb{R}^n$ be nonempty convex sets such that $U \cap V = \emptyset$. There exists a hyperplane separating the sets U and V, i.e., there exists a nonzero vector $\boldsymbol{a} \in \mathbb{R}^n$ such that*

$$\boldsymbol{a}^T\boldsymbol{x} \leqslant \boldsymbol{a}^T\boldsymbol{y} \quad \textit{for all } \boldsymbol{x} \in U,\ \boldsymbol{y} \in V.$$

As the mapping $\boldsymbol{x} \to \boldsymbol{a}^T\boldsymbol{x}$ is continuous in $\mathbb{R}^n$, from the above we obtain

Corollary 2 *Let $U, V \subset \mathbb{R}^n$ be nonempty convex sets such that $\operatorname{int} U \neq \emptyset$ and $(\operatorname{int} U) \cap V = \emptyset$. Then there exists a hyperplane separating the sets U and V.*

Theorem 20 *(strong separation theorem) Let $U, V \subset \mathbb{R}^n$ be nonempty closed convex sets, let U be compact and let $U \cap V = \emptyset$. Then there exists a hyperplane strictly separating the sets U and V, i.e., there exists a nonzero vector $\boldsymbol{a} \in \mathbb{R}^n$ such that*

$$\sup_{x \in U} \boldsymbol{a}^T\boldsymbol{x} < \inf_{y \in V} \boldsymbol{a}^T\boldsymbol{y}.$$

There may be more than one hyperplane described by these theorems; one of them is the set defined as follows:

$$\{\boldsymbol{x} \in \mathbb{R}^n \colon \boldsymbol{a}^T\boldsymbol{x} = \alpha\}, \quad \alpha = \sup_{x \in U} \boldsymbol{a}^T\boldsymbol{x}.$$

Proof of the strong separation theorem. Let $d \colon U \times V \to \mathbb{R}$ be a function given by the formula $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$. As the set U is bounded, the function d is coercive; it may tend to infinity only by taking an appropriate sequence of points $\boldsymbol{y} \in V$. As the function d is continuous and coercive and its domain $U \times V$ is closed, it takes a minimum at a point $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in U \times V$. As $U \cap V = \emptyset$, there is $\boldsymbol{a} = \boldsymbol{y}_0 - \boldsymbol{x}_0 \neq \boldsymbol{0}$. Below we demonstrate that it is a vector satisfying the claim.

First we show that $\boldsymbol{a}^T\boldsymbol{y} \geqslant \boldsymbol{a}^T\boldsymbol{y}_0$ for all $\boldsymbol{y} \in V$. Let

$$g(t) \overset{\text{def}}{=} \Big(d\big(\boldsymbol{x}_0, \boldsymbol{y}_0 + t(\boldsymbol{y} - \boldsymbol{y}_0)\big)\Big)^2, \quad t \in \mathbb{R}.$$

There is

$$g(t) = \|\boldsymbol{y}_0 - \boldsymbol{x}_0\|^2 + 2t(\boldsymbol{y}_0 - \boldsymbol{x}_0)^T(\boldsymbol{y} - \boldsymbol{y}_0) + t^2(\boldsymbol{y} - \boldsymbol{y}_0)^T(\boldsymbol{y} - \boldsymbol{y}_0).$$

This function is differentiable for all $t \in \mathbb{R}$ and, as the set V is convex, $g(0) \leqslant g(t)$ for $t \in [0, 1]$. Hence, $g'(0) \geqslant 0$, i.e.,

$$(\boldsymbol{y}_0 - \boldsymbol{x}_0)^T(\boldsymbol{y} - \boldsymbol{y}_0) = \boldsymbol{a}^T(\boldsymbol{y} - \boldsymbol{y}_0) \geqslant 0.$$

In a similar way we can show that $\boldsymbol{a}^T\boldsymbol{x} \leqslant \boldsymbol{a}^T\boldsymbol{x}_0$ for all $\boldsymbol{x} \in U$. □

Proof of the weak separation theorem. Consider the set $C = V - U = \{\boldsymbol{y} - \boldsymbol{x} \colon \boldsymbol{x} \in U,\ \boldsymbol{y} \in V\}$. This set is convex and $\boldsymbol{0} \notin C$. It suffices to find a nonzero vector $\boldsymbol{a} \in \mathbb{R}^n$ such that $\boldsymbol{a}^T\boldsymbol{x} \geqslant 0$ for all $\boldsymbol{x} \in C$.

Let $A_x \overset{\text{def}}{=} \{\, a \in \mathbb{R}^n\colon \|a\| = 1,\ a^T x \geqslant 0 \,\}$. We are going to show that $\bigcap_{x\in C} A_x \neq \emptyset$. Suppose that $\bigcap_{x\in C} A_x = \emptyset$. Let $B_x = S \setminus A_x$, where S is the unit sphere in $\mathbb{R}^n$. The sets B_x are open subsets of S. If the intersection of all sets A_x, where $x \in C$, is empty, then the family $\{\, B_x\colon x \in C \,\}$ is an open coverage of S, which is a compact set. Hence, there exists a finite coverage $\{\, B_{x_1}, \dots, B_{x_k}\colon x_1, \dots, x_k \in C \,\}$ of S. Let

$$\hat{C} \overset{\text{def}}{=} \operatorname{conv}\{x_1, \dots, x_k\} = \Big\{\, \sum_{i=1}^{k} \lambda_i x_i\colon \lambda_1, \dots, \lambda_k \geqslant 0,\ \sum_{i=1}^{k} \lambda_i = 1 \,\Big\}.$$

The set $\hat{C}$ is convex and closed and it is a subset of C. Hence, $0 \notin \hat{C}$. By the strong separation theorem used to the sets $\{0\}$ and $\hat{C}$, there exists a nonzero vector a such that

$$a^T x > 0 \quad \text{for all } x \in \hat{C}.$$

In particular, $a^T x_i > 0$ i.e., $\frac{a}{\|a\|} \in A_{x_i}$ for $i = 1, \dots, k$, which contradicts the supposition that $\bigcap_{i=1}^{k} A_{x_i} = \emptyset$. □

Convex functions

Definition 8 *A function $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is convex, is called*

- *convex, if for all $x, y \in W$ and $\lambda \in (0, 1)$ there is*

$$f\big(\lambda x + (1-\lambda) y\big) \leqslant \lambda f(x) + (1-\lambda) f(y),$$

- *strictly convex, if for all $x, y \in W$ and $\lambda \in (0, 1)$ there is*

$$f\big(\lambda x + (1-\lambda) y\big) < \lambda f(x) + (1-\lambda) f(y),$$

A function f is (strictly) concave if $-f$ is (strictly) convex.

Theorem 21 *If a function $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is convex, is Lebesgue-measurable and such that*

$$f\Big(\frac{x+y}{2}\Big) \leqslant \frac{f(x)+f(y)}{2} \quad \text{for all } x, y \in W,$$

then f is a convex function.

We shall prove a simpler theorem:

Theorem 22 *If a function $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is convex, is continuous and such that*

$$f\Big(\frac{x+y}{2}\Big) \leqslant \frac{f(x)+f(y)}{2} \quad \text{for all } x, y \in W,$$

then f is a convex function.

Proof. Using induction with respect to k, we show that the inequality of the definition of convex functions holds for all $\lambda = \frac{p}{2^k}$, where $p = 0, 1, \dots, 2^k$. If $k = 1$, then this inequality is satisfied by assumption. Suppose that the inequality is satisfied for some $k \geqslant 1$. Let $p, q \in \mathbb{Z}$, $p, q \geqslant 0$ and $p + q = 2^{k+1}$. Suppose that $p \leqslant q$. Then $p \leqslant 2^k \leqslant q$ and we can write

$$z = \frac{p}{2^{k+1}} x + \frac{q}{2^{k+1}} y = \frac{1}{2}\Big(\frac{p}{2^k} x + \frac{q - 2^k}{2^k} y + y\Big).$$

Then,

$$\begin{aligned} f(z) &\leqslant \frac{1}{2} f\Big(\frac{p}{2^k} x + \frac{q-2^k}{2^k} y\Big) + \frac{1}{2} f(y) \\ &\leqslant \frac{1}{2}\frac{p}{2^k} f(x) + \frac{1}{2}\frac{q-2^k}{2^k} f(y) + \frac{1}{2} f(y) = \frac{p}{2^{k+1}} f(x) + \frac{q}{2^{k+1}} f(y). \end{aligned}$$

The first inequality follows from the assumption of the theorem and the second one from the inductive assumption. If $p > q$, then it suffices to exchange x and y.

The set of numbers $\frac{p}{2^k}$, $k = 1, 2, \dots$ and $p = 0, \dots, 2^k$, is dense in the interval $[0, 1]$. By the continuity of f we obtain the desired inequality for any $\lambda \in (0, 1)$. □

Properties of convex functions

Below we assume that $W \subset \mathbb{R}^n$ is convex.

Definition 9 *The epigraph of a function $f\colon W \to \mathbb{R}$ is the set*

$$\operatorname{epi}(f) = \{\, (x, z) \in W \times \mathbb{R}\colon z \geqslant f(x) \,\}.$$

Definition 10 *The sublevel set or the trench of a function $f\colon W \to \mathbb{R}$ is the set*

$$W_\alpha(f) = \{\, x \in W\colon f(x) \leqslant \alpha \,\}, \quad \alpha \in \mathbb{R}.$$

Theorem 23 *(epigraph theorem) A function* f *is convex if and only if its epigraph is a convex set.*

Theorem 24 *If a function* f *is convex, then its sublevel sets* $W_\alpha(f)$ *are convex for all* $\alpha \in \mathbb{R}$.

Remark. There exist nonconvex functions whose all sublevel sets are convex.

Theorem 25 *If a function* f *is convex, then it is also continuous in* $\operatorname{int} W$.

Theorem 26 *(supporting hyperplane theorem) If* f *is a convex function, then at each point* $\bar{x} \in \operatorname{int} W$ *there exists a supporting hyperplane, i.e., there exists* $\xi \in \mathbb{R}^n$ *such that*

$$f(x) \geqslant f(\bar{x}) + \xi^T(x - \bar{x}) \quad \textit{for all } x \in W.$$

Moreover, if f *is strictly convex, then*

$$f(x) > f(\bar{x}) + \xi^T(x - \bar{x}) \quad \textit{for all } x \in W \setminus \{\bar{x}\}.$$

If f *is differentiable at* $\bar{x}$*, then in both cases we can take* $\xi = \mathrm{D}f(\bar{x})^T$.

Proof. The set $\operatorname{epi}(f)$ is convex. We apply the weak separation theorem to the sets $U = \operatorname{int}\operatorname{epi}(f)$ and $V = \{(\bar{x}, f(\bar{x}))\}$. There exists a nonzero vector $\mathbf{a} = \{(\xi, \alpha)\} \in \mathbb{R}^{n+1}$ such that

$$\xi^T x + \alpha y \leqslant \xi^T \bar{x} + \alpha f(\bar{x}) \quad \text{for all } (x, y) \in \operatorname{epi}(f).$$

The inequality above holds for all $y \geqslant f(x)$. Hence, $\alpha \leqslant 0$. It turns out that $\alpha \neq 0$. To prove it, suppose that $\alpha = 0$. Then, for all $x \in W$ there is $\xi^T(x - \bar{x}) \leqslant 0$. As $\bar{x} \in \operatorname{int} W$, we know that there exists an $\varepsilon > 0$ such that $\bar{x} + \varepsilon\xi \in W$. Let $x = \bar{x} + \varepsilon\xi$. Then $0 \geqslant \xi^T(x - \bar{x}) = \varepsilon\xi^T\xi = \varepsilon\|\xi\|^2$; hence, $\xi = \mathbf{0}$. This contradicts the possibility $\mathbf{a} \neq \mathbf{0}$, and thus $\alpha < 0$.

We can rescale the vector $\mathbf{a}$ to obtain $\alpha = -1$. With that, for all $x \in W$ we obtain

$$\xi^T x - f(x) \leqslant \xi^T \bar{x} - f(\bar{x}),$$

which may be rewritten as

$$f(x) \geqslant f(\bar{x}) + \xi^T(x - \bar{x}),$$

which completes the proof of the first claim.

Suppose that f is strictly convex. Let $\bar{x} \in \operatorname{int} W$. By the first claim, there is $f(x) \geqslant f(\bar{x}) + \xi^T(x - \bar{x})$ for all $x \in W$. Suppose that there exists $x \in W \setminus \{\bar{x}\}$ such that $f(x) = f(\bar{x}) + \xi^T(x - \bar{x})$. By the strict convexity of f we obtain

$$f\Bigl(\frac{\bar{x} + x}{2}\Bigr) < \frac{1}{2}\bigl(f(x) + f(\bar{x})\bigr) = f(\bar{x}) + \frac{1}{2}\bigl(f(x) - f(\bar{x})\bigr) = f(\bar{x}) + \frac{1}{2}\xi^T(x - \bar{x}).$$

On the other hand, by the existence of the supporting hyperplane, we obtain

$$f\Bigl(\frac{\bar{x} + x}{2}\Bigr) \geqslant f(\bar{x}) + \xi^T\Bigl(\frac{x + \bar{x}}{2} - x\Bigr) = f(\bar{x}) + \xi^T\frac{x - \bar{x}}{2}.$$

The two inequalities are inconsistent. Hence, if f is a strictly convex function, there must be $f(x) > f(\bar{x}) + \xi^T(x - \bar{x})$ and the second claim is proved.

Suppose that f is differentiable at $\bar{x}$. For $x \in W \setminus \{\bar{x}\}$ and $\lambda \in (0, 1)$, by convexity of f we obtain

$$\begin{aligned} f(x) - f(\bar{x}) &= \frac{(1-\lambda)f(\bar{x}) + \lambda f(x) - f(\bar{x})}{\lambda} \\ &\geqslant \frac{f\bigl((1-\lambda)\bar{x} + \lambda x\bigr) - f(\bar{x})}{\lambda} = \frac{f\bigl(\bar{x} + \lambda(x - \bar{x})\bigr) - f(\bar{x})}{\lambda}. \end{aligned}$$

With this estimation of the divided difference we go to the limit

$$f(x) - f(\bar{x}) \geqslant \lim_{\lambda \searrow 0} \frac{f\bigl((1-\lambda)\bar{x} + \lambda x\bigr) - f(\bar{x})}{\lambda} = \mathrm{D}f(\bar{x})(x - \bar{x}).$$

The limit exists and is equal to $\mathrm{D}f(\bar{x})(x - \bar{x})$ due to the differentiability of f. If f is strictly convex, then we can repeat the proof of the second claim with ξ^T replaced by $\mathrm{D}f(\bar{x})$. Then we get the sharp inequality $f(x) - f(\bar{x}) > \mathrm{D}f(\bar{x})(x - \bar{x})$ for $x \neq \bar{x}$. □

Corollary 3 *If a function* f *is convex and differentiable at* $\bar{x} \in \operatorname{int} W$*, then* $\bar{x}$ *is a global minimum of* f *if and only if* $\mathrm{D}f(\bar{x}) = \mathbf{0}^T$.

Proof. The gradient of a differentiable function at a minimal point must be equal to $\mathbf{0}^T$; hence, $\mathrm{D}f(\bar{x}) = \mathbf{0}^T$ is a necessary condition. Suppose that it is satisfied. Then, for any $x \in W$ we have

$$f(x) \geqslant f(\bar{x}) + \mathrm{D}f(\bar{x})(x - \bar{x}) = f(\bar{x}),$$

which proves that $\overline{x}$ is a global minimum. □

Properties of convex functions

Theorem 27 *Let $W \subset \mathbb{R}^n$ be a convex set with a nonempty interior. If at each point $\overline{x} \in \operatorname{int} W$ there exists a vector $\xi \in \mathbb{R}^n$ such that*

$$f(x) \geqslant f(\overline{x}) + \xi^T(x - \overline{x}) \quad \textit{for all } x \in W,$$

then the function f is convex. If the inequality is sharp for $x \neq \overline{x}$, then f is strictly convex.

Proof. Let $x \in \operatorname{int} W$, $y \in W$ and $\lambda \in (0, 1)$. Denote $x_\lambda = \lambda x + (1 - \lambda)y$. We are going to prove that $f(x_\lambda) \leqslant \lambda f(x) + (1 - \lambda)f(y)$. By Lemma 3, $x_\lambda \in \operatorname{int} W$. By assumption, there exists $\xi \in \mathbb{R}^n$ such that

$$f(x) \geqslant f(x_\lambda) + \xi^T(x - x_\lambda), \quad f(y) \geqslant f(x_\lambda) + \xi^T(y - x_\lambda).$$

Hence,

$$\lambda f(x) + (1 - \lambda)f(y) \geqslant f(x_\lambda) + \xi^T\big[\lambda(x - x_\lambda) + (1 - \lambda)(y - x_\lambda)\big] = f(x_\lambda),$$

as the terms in the brackets cancel each other out. The convexity of f is proved. If the assumed inequalities are sharp, then also the inequalities in the calculation above are sharp and the function f is strictly convex. □

Theorem 28 *Let $W \subset \mathbb{R}^n$ be nonempty, open and convex and let $f\colon W \to \mathbb{R}$ be twice differentiable. Then,*

I) f is convex if and only if the Hessian $D^2f(x)$ is nonnegative-definite for all $x \in W$,

II) if the Hessian is positive-definite for all $x \in W$, then f is strictly convex (this is not a necessary condition).

Proof. Suppose that the Hessian is nonnegative-definite for all $x \in W$. Then, by Corollary 1, for all $\overline{x}, x \in W$ we have

$$f(x) = f(\overline{x}) + Df(\overline{x})(x - \overline{x}) + \frac{1}{2}(x - \overline{x})^T D^2 f(\tilde{x})(x - \overline{x}),$$

where $\tilde{x}$ is a point of the line segment $\overline{\overline{x}x}$. As the Hessian is assumed to be nonnegative-definite, the last term above is nonnegative. Hence,

$$f(x) \geqslant f(\overline{x}) + Df(\overline{x})(x - \overline{x}).$$

This inequality holds for all $\overline{x}, x \in W$, the function f is convex by Theorem 27.

If the Hessian is positive-definite in W, then for $x \neq \overline{x}$ the last inequality is sharp, and the function f is strictly convex.

Now we prove that the convexity of f implies that the Hessian is nonnegative-definite. Assume that f is convex. Let $\overline{x} \in W$ and $h \in \mathbb{R}^n \setminus \{0\}$ be fixed. As the set W is open, there exists $\delta > 0$ such that $\overline{x} + th \in W$ for all $t \in (-\delta, \delta)$. Let $g(t) \overset{\text{def}}{=} f(\overline{x} + th)$. It is a convex and twice differentiable function of one variable. By Theorem 26,

$$g(t) \geqslant g(0) + g'(0)t, \quad t \in (-\delta, \delta).$$

Using the Taylor's formula with the remainder in Peano form, we obtain

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + o(t^2), \quad t \in (-\delta, \delta).$$

The last inequality and the Taylor's formula give us the following estimation:

$$\frac{1}{2}g''(0)t^2 + o(t^2) \geqslant 0.$$

After dividing both sides by t^2 we get

$$\frac{1}{2}g''(0) + \frac{o(t^2)}{t^2} \geqslant 0.$$

With t tending to 0, the second term vanishes, which leaves us $g''(0) \geqslant 0$. With this we return to the function f:

$$g'(t) = Df(\overline{x} + th)h, \quad g''(t) = h^T D^2 f(\overline{x} + th)h.$$

Hence, $g''(0) = h^T D^2 f(\overline{x})h$. As the vector h may be arbitrarily chosen, the Hessian at $\overline{x}$ is nonnegative-definite. □

Subdifferential

We are going to generalise the notion of derivative to non-differentiable convex functions. Let $W \subset \mathbb{R}^n$ be a convex set and $f\colon W \to \mathbb{R}$ be a convex function.

Definition 11 *A vector $\xi \in \mathbb{R}^n$ is called the subgradient of the function f at a point $x_0 \in W$, if*

$$f(x) \geqslant f(x_0) + \xi^T(x - x_0) \quad \textit{for all } x \in W.$$

The set of all subgradients of f at x_0 is called the subdifferential and it is denoted by $\partial f(x_0)$.

Corollary 4 *If $W \subset \mathbb{R}^n$ is a convex set with the nonempty interior, then $f\colon W \to \mathbb{R}$ is a convex function if and only if at each point $x \in \operatorname{int} W$ there exists a subgradient, i.e.,*

$$\partial f(x) \neq \emptyset.$$

Proof. By the supporting hyperplane theorem, the convexity implies the existence of the gradient at each point of W. By Theorem 27, it is a sufficient condition. □

Lemma 4 *Let $W \subset \mathbb{R}^n$ be convex and let $f\colon W \to \mathbb{R}$ be a convex function. Then the subdifferential $\partial f(x)$ is a convex and closed set. If $x \in \operatorname{int} W$, then the subdifferential is also bounded, and thus it is a compact set.*

Proof. The proof of convexity and closedness is an exercise. Let $\overline{x} \in \operatorname{int} W$ be fixed. There exists $\varepsilon > 0$ such that the ball $B(\overline{x}, \varepsilon) \subset \operatorname{int} W$. For any $\xi \in \partial f(\overline{x})$ we have

$$f(x) \geqslant f(\overline{x}) + \xi^T(x - \overline{x}) \quad \text{for all } x \in W.$$

Therefore,

$$\sup_{x \in B(\overline{x},\varepsilon)} f(x) \geqslant f(\overline{x}) + \sup_{x \in B(\overline{x},\varepsilon)} \xi^T(x - \overline{x}).$$

The left hand side does not depend on ξ, and, by continuity of f in $\operatorname{int} W$, it is finite. The supremum on the right hand side is attended for $x = \overline{x} + \varepsilon\xi/\|\xi\|$, and it is equal to $\varepsilon\|\xi\|$. Hence,

$$\varepsilon\|\xi\| \leqslant \sup_{x \in B(\overline{x},\varepsilon)} f(x) - f(\overline{x}),$$

which proves that the set $\partial f(\overline{x})$ is bounded. □

Definition 12 *The directional derivative of a function f at a point $\overline{x}$ in the direction $\mathbf{d}$ is the limit*

$$f'(\overline{x}; \mathbf{d}) = \lim_{\lambda \searrow 0} \frac{f(\overline{x} + \lambda\mathbf{d}) - f(\overline{x})}{\lambda}.$$

Definition 13 *A divided difference of order 0 of a function $f\colon I \subset \mathbb{R} \to \mathbb{R}$ at a point t_i is the number $f[t_i] = f(t_i)$. A divided difference of order $k > 0$ at different points $t_i, \ldots, t_{i+k}$ is given by the recursive formula*

$$f[t_i, \ldots, t_{i+k}] = \frac{f[t_i, \ldots, t_{i+k-1}] - f[t_{i+1}, \ldots, t_{i+k}]}{t_i - t_{i+k}}.$$

Lemma 5 *I) Divided differences of any order are symmetric functions of the arguments $t_i, \ldots, t_{i+k}$, i.e., $f[t_i, \ldots, t_{i+k}] = f[t_{\sigma(i)}, \ldots, t_{\sigma(i+k)}]$ for any permutation σ of the set $\{i, \ldots, i+k\}$.*

II) If f is a convex function, then the first order divided difference $f[t_i, t_{i+1}]$ is a monotone (nondecreasing) function of the arguments t_i, t_{i+1}.

Proof. The proof of (I) is an exercise. To prove (II) we notice that

$$f[x_0, x_1, x_2] = c_0 f(x_0) + c_1 f(x_1) + c_2 f(x_2),$$
$$c_0 = \frac{1}{(x_1 - x_0)(x_2 - x_0)}, \quad c_1 = \frac{1}{(x_2 - x_1)(x_0 - x_1)}, \quad c_2 = \frac{1}{(x_0 - x_2)(x_1 - x_2)}.$$

Assume that $x_0 < x_1 < x_2$; then, $c_0, c_2 > 0$, $c_1 < 0$. Then,

$$\lambda = \frac{x_1 - x_0}{x_2 - x_0} \in (0, 1), \quad (1 - \lambda) = \frac{x_2 - x_1}{x_2 - x_0}.$$

We can check that if $f(x_1) = (1 - \lambda)f(x_0) + \lambda f(x_2)$, then $f[x_0, x_1, x_2] = 0$; as the value of a convex function at x_1 is less than or equal to this expression, due to $c_1 < 0$ there is $f[x_0, x_1, x_2] \geqslant 0$.

Now we use the symmetry of the divided differences; we choose the numbers t_0, t_1, t_2 such that $t_1 < t_2$. Then,

$$0 \leqslant f[t_2, t_0, t_1] = \frac{f[t_2, t_0] - f[t_0, t_1]}{t_2 - t_1} = \frac{f[t_2, t_0] - f[t_1, t_0]}{t_2 - t_1} = \frac{f[t_0, t_2] - f[t_0, t_1]}{t_2 - t_1}.$$

Hence, if f is convex and $t_2 > t_1$, then $f[t_2, t_0] \geqslant f[t_1, t_0]$ and $f[t_0, t_2] \geqslant f[t_0, t_1]$. □

Lemma 6 *Let $W \subset \mathbb{R}^n$ be a convex open set and let $f\colon W \to \mathbb{R}$ be a convex function. Then, for all $\mathbf{d} \in \mathbb{R}^n$ and $x \in W$*

I) there exists the directional derivative $f'(\bar{x}; \mathbf{d})$,

II) $f'(\bar{x}; \mathbf{d}) = \inf_{\lambda>0} \frac{f(\bar{x}+\lambda\mathbf{d})-f(\bar{x})}{\lambda}$,

III) $f'(\bar{x}; \mathbf{d}) \geqslant -f'(\bar{x}; -\mathbf{d})$.

Proof. Let $g(t) \overset{\text{def}}{=} f(x + t\mathbf{d})$ for t such that $x + t\mathbf{d} \in W$. As W is open, the function g is defined in an interval $(-\delta, \delta)$. This function is convex. By Lemma 5, its divided difference is monotone, i.e., for $t_1, t_2 \in (-\delta, \delta) \setminus \{0\}$, $t_1 < t_2$, we have

$$\frac{g(t_1) - g(0)}{t_1} \leqslant \frac{g(t_2) - g(0)}{t_2}. \qquad (*)$$

By the monotonicity of the divided difference, there exists the left-side derivative, $g'(0^-)$, and the right-side derivative, $g'(0^+)$, such that $g'(0^-) \leqslant g'(0^+)$, and

$$g'(0^-) = \sup_{t<0} \frac{g(t) - g(0)}{t}, \quad g'(0^+) = \inf_{t>0} \frac{g(t) - g(0)}{t}.$$

It suffices to notice that $f'(x; \mathbf{d}) = g'(0^+)$ and $f'(x; -\mathbf{d}) = -g'(0^-)$. □

Lemma 7 *Let $W \subset \mathbb{R}^n$ be a convex open set and let $f\colon W \to \mathbb{R}$ be a convex function. Then a vector ξ is a subgradient if and only if*

$$f'(\bar{x}; \mathbf{d}) \geqslant \xi^T\mathbf{d} \quad \textit{for all } \mathbf{d} \in \mathbb{R}^n.$$

Proof. Let $\bar{x} \in W$ and $\xi \in \partial f(\bar{x})$. Then, for $\lambda > 0$ and $\mathbf{d} \in \mathbb{R}^n$ such that $\bar{x} + \lambda\mathbf{d} \in W$, there is

$$f(\bar{x} + \lambda\mathbf{d}) \geqslant f(\bar{x}) + \lambda\xi^T\mathbf{d}.$$

Hence,

$$\frac{f(\bar{x} + \lambda\mathbf{d}) - f(\bar{x})}{\lambda} \geqslant \xi^T\mathbf{d},$$

i.e., $f'(\bar{x}; \mathbf{d}) \geqslant \xi^T\mathbf{d}$.

Now, let $\xi \in \mathbb{R}^n$ be a vector such that $f'(\bar{x}; \mathbf{d}) \geqslant \xi^T\mathbf{d}$ for all $\mathbf{d} \in \mathbb{R}^n$. By Lemma 6(II), for $\lambda > 0$ we obtain

$$f'(\bar{x}; \mathbf{d}) \leqslant \frac{f(\bar{x} + \lambda\mathbf{d}) - f(\bar{x})}{\lambda}.$$

Hence,

$$f(\bar{x} + \lambda\mathbf{d}) \geqslant f(\bar{x}) + \lambda\xi^T\mathbf{d}.$$

As λ and $\mathbf{d}$ may be arbitrary (such that $\bar{x} + \lambda\mathbf{d} \in W$), the vector ξ is a subgradient. □

Theorem 29 *Let $f\colon W \to \mathbb{R}$ be a convex function in an open convex set $W \subset \mathbb{R}^n$. For each point $\bar{x} \in W$ and vector $\mathbf{d} \in \mathbb{R}^n$ there is*

$$f'(\bar{x}; \mathbf{d}) = \max_{\xi\in\partial f(\bar{x})} \xi^T\mathbf{d}.$$

Moreover, the function f is differentiable at $\bar{x}$ if and only if the subdifferential $\partial f(\bar{x})$ has only one element. This element is $Df(\bar{x})^T$.

Proof. By Lemma 7, $f'(\bar{x}; \mathbf{d}) \geqslant \xi^T\mathbf{d}$ for all $\xi \in \partial f(\bar{x})$. Hence,

$$f'(\bar{x}; \mathbf{d}) \geqslant \max_{\xi\in\partial f(\bar{x})} \xi^T\mathbf{d}.$$

The opposite inequality may be proved using the weak separation theorem. Let

$$\begin{aligned} C_1 &= \{\, (x, z) \in W \times \mathbb{R}\colon z > f(x) \,\} \\ C_2 &= \{\, (x, z) \in W \times \mathbb{R}\colon x = \bar{x} + \lambda\mathbf{d},\ z = f(\bar{x}) + \lambda f'(\bar{x}; \mathbf{d}),\ \lambda \geqslant 0 \,\}. \end{aligned}$$

Note that C_1 is the interior of the epigraph of f; hence, C_1 is a convex set. The set C_2 is a halfline with the origin at $(\bar{x}, f(\bar{x}))$ and the direction $(\mathbf{d}, f'(\bar{x}; \mathbf{d}))$, which is also a convex set. This halfline is the graph of a linear approximation of f along the line segment $\{\, \bar{x} + \lambda\mathbf{d}\colon \lambda \geqslant 0 \,\} \cap W$.

By Lemma 6, $f'(x; \mathbf{d}) \leqslant \frac{f(\bar{x}+\lambda\mathbf{d})-f(\bar{x})}{\lambda}$, i.e.,

$$f(\bar{x} + \lambda\mathbf{d}) \geqslant f(\bar{x}) + \lambda f'(x; \mathbf{d}).$$

Hence, the sets C_1 and C_2 are disjoint. By the weak separation theorem, there exists a nonzero vector $(\mu, \gamma) \in \mathbb{R}^{n+1}$ such that

$$\mu^T x + \gamma z \geqslant \mu^T(\bar{x} + \lambda\mathbf{d}) + \gamma\bigl(f(\bar{x}) + \lambda f'(\bar{x}; \mathbf{d})\bigr), \quad \text{for all } (x, z) \in C_1,\ \lambda \in [0, L),$$

where $L = \sup\{\,\lambda > 0 \colon \bar{x} + \lambda\mathbf{d} \in W\,\}$. The number γ cannot be negative, as the left hand side might be arbitrarily small (after choosing a large z). Also, γ cannot be zero, as in that case the inequality $\mu^T(x - \bar{x}) \geqslant \lambda\mu^T\mathbf{d}$ would have to hold for all $x \in W$, and this is possible only with $\mu = 0$. This inconsistency with $(\mu, \gamma) \neq 0$ proves that $\gamma > 0$.

By rescaling the vector (μ, γ), we can assume $\gamma = 1$. Then,

$$\mu^T x + z \geqslant \mu^T(\bar{x} + \lambda\mathbf{d}) + \big(f(\bar{x}) + \lambda f'(\bar{x}; \mathbf{d})\big), \quad \text{for all } (x, z) \in C_1,\ \lambda \in [0, L),$$

With z tending to $f(x)$ we obtain the following inequality, which holds for all $x \in W$ and $\lambda \in [0, L)$:

$$\mu^T x + f(x) \geqslant \mu^T(\bar{x} + \lambda\mathbf{d}) + \big(f(\bar{x}) + \lambda f'(\bar{x}; \mathbf{d})\big) \tag{*}$$

With $\lambda = 0$, we obtain

$$\mu^T(x - \bar{x}) + f(x) \geqslant f(\bar{x}),$$

i.e.,

$$f(x) \geqslant f(\bar{x}) - \mu^T(x - \bar{x}),$$

Hence, $-\mu \in \partial f(\bar{x})$. Now, substituting $\lambda > 0$ and $x = \bar{x}$ in (*), we obtain

$$-\mu^T(\lambda\mathbf{d}) \geqslant \lambda f'(\bar{x}; \mathbf{d}),$$

i.e.,

$$\sup_{\xi \in \partial f(\bar{x})} \xi^T\mathbf{d} \geqslant f(\bar{x}; \mathbf{d}).$$

The proof of the first claim is complete.

To prove the second claim, we notice that the function f is differentiable at $\bar{x}$ if and only if there exists $\alpha \in \mathbb{R}^n$ such that $f'(\bar{x}; \mathbf{d}) = \alpha^T\mathbf{d}$ for all $\mathbf{d} \in \mathbb{R}^n$ (then $\alpha = Df(\bar{x})^T$). Thus, if the set $\partial f(\bar{x})$ has only one element, then f is differentiable at $\bar{x}$.

Suppose that f is diferentiable at $\bar{x}$. Then, for sufficiently small $\lambda > 0$ and $\mathbf{d} \in \mathbb{R}^n$ (without loss of generality we assume that $\|\mathbf{d}\| = 1$), we have

$$f(\bar{x} + \lambda\mathbf{d}) = f(\bar{x}) + \lambda Df(\bar{x})\mathbf{d} + o(\lambda).$$

By definition of the subgradient, we have

$$f(\bar{x} + \lambda\mathbf{d}) \geqslant f(\bar{x}) + \lambda\xi^T\mathbf{d},$$

where ξ is a subgradient. By subtracting the sides of the above, we obtain

$$\lambda\big(\xi^T - Df(\bar{x})\big)\mathbf{d} - o(\lambda) \leqslant 0.$$

After dividing both sides of this inequality by λ and passing with λ to 0, we obtain

$$\big(\xi^T - Df(\bar{x})\big)\mathbf{d} \leqslant 0.$$

The substitution $\mathbf{d} = \pm\frac{\xi - Df(\bar{x})^T}{\|\xi - Df(\bar{x})^T\|}$ yields the equality

$$\xi^T = Df(\bar{x}),$$

which means that the subdifferential consists of one element. □

Theorem 30 *Let $W \subset \mathbb{R}^n$ be an open convex set and $f_1, f_2 \colon W \to \mathbb{R}$ be convex functions.*

I) Let $f = f_1 + f_2$. Then, $\partial f(x) = \partial f_1(x) + \partial f_2(x)$, i.e.,

$$\partial f(x) = \partial f_1(x) + \partial f_2(x) = \{\, \xi_1 + \xi_2 \colon \xi_1 \in \partial f_1(x),\ \xi_2 \in \partial f_2(x) \,\}.$$

II) Let $f = \max(f_1, f_2)$. Then,

$$\partial f(x) = \begin{cases} \partial f_1(x) & \text{if } f_1(x) > f_2(x), \\ \operatorname{conv}\big(\partial f_1(x) \cup \partial f_2(x)\big) & \text{if } f_1(x) = f_2(x), \\ \partial f_2(x) & \text{if } f_1(x) < f_2(x), \end{cases}$$

where $\operatorname{conv}\big(\partial f_1(x) \cup \partial f_2(x)\big)$ is the convex hull of the union $\partial f_1(x) \cup \partial f_2(x)$, i.e., the set of all convex combinations of the subgradients in both subdifferentials.

Proof. (I): Let $\bar{x} \in W$. Let $\xi_1 \in \partial f_1(\bar{x})$ and $\xi_2 \in \partial f_2(\bar{x})$. Then, for all $x \in W$ we have

$$f_1(x) \geqslant f_1(\bar{x}) + \xi_1^T(x - \bar{x}),$$
$$f_2(x) \geqslant f_2(\bar{x}) + \xi_2^T(x - \bar{x}).$$

By adding the above inequalities side by side we obtain

$$f(x) \geqslant f(\bar{x}) + (\xi_1 + \xi_2)^T(x - \bar{x}),$$

i.e., $\xi_1 + \xi_2 \in \partial f(\bar{x})$. Hence, $\partial f_1(\bar{x}) + \partial f_2(\bar{x}) \subset \partial f(\bar{x})$. Suppose that there exists $\xi \in \partial f(\bar{x})$ such that $\xi \notin \partial f_1(\bar{x}) + \partial f_2(\bar{x})$. By Lemma 4, the subdifferentials $\partial f_1(\bar{x})$

and $\partial f_2(\overline{x})$ are compact convex sets. Their algebraic sum is, therefore, also a compact convex set. By the strong separation theorem, applied to the sets $\{\xi\}$ and $\partial f_1(\overline{x}) + \partial f_2(\overline{x})$, there exists $\mu \in \mathbb{R}^n$, such that

$$\mu^T\xi_1 + \mu^T\xi_2 < \mu^T\xi, \quad \text{for all } \xi_1 \in \partial f_1(\overline{x}) \text{ and } \xi_2 \in \partial f_2(\overline{x}).$$

We take ξ_1, ξ_2 to maximise the left hand side. By Theorem 29,

$$f_1'(\overline{x};\mu) + f_2'(\overline{x};\mu) < \xi^T\mu \leqslant f'(\overline{x};\mu).$$

On the other hand, by the properties of directional derivatives,

$$f_1'(\overline{x};\mu) + f_2'(\overline{x};\mu) = f'(\overline{x};\mu).$$

This is an inconsistency; a vector ξ with assumed properties cannot exist, which completes the proof of (I).

Now we prove (II). The form of the subdifferential ∂f in the sets W_1 and W_2 defined as $W_i = \{x \in W\colon f_i(x) > f_{3-i}(x)\}$ is obvious, which leaves the set $W_0 = \{x \in W\colon f_1(x) = f_2(x)\}$ to investigate. Let $\overline{x} \in W$ and $f_1(\overline{x}) = f_2(\overline{x})$. Denote $A = \operatorname{conv}\big(\partial f_1(x) \cup \partial f_2(x)\big)$. For $i = 1, 2$ and $x \in W$ we have

$$f(x) - f(\overline{x}) \geqslant f_i(x) - f(\overline{x}) = f_i(x) - f_i(\overline{x}) \geqslant \xi_i^T(x - \overline{x}), \quad \text{for all } \xi_i \in \partial f_i(\overline{x}).$$

From the above we obtain $\partial f_1(\overline{x}) \cup \partial f_2(\overline{x})$. By convexity of the subdifferential, $A \subset \partial f(\overline{x})$. Suppose that there exists $\xi \in \partial f(\overline{x}) \setminus A$. The set A is convex and compact. By the strong separation theorem applied to the sets $\{\xi\}$ and A, there exists a vector $\mu \in \mathbb{R}^n$ and a constant b such that

$$\mu^T\tilde{\xi} < b < \mu^T\xi \quad \text{for all } \tilde{\xi} \in A.$$

In particular, $\mu^T\xi_i < b$ for $\xi_i \in \partial f_i(\overline{x})$, $i = 1, 2$. By Theorem 29,

$$\max\big\{f_1'(\overline{x};\mu), f_2'(\overline{x};\mu)\big\} \leqslant b.$$

Similarly, $b < \xi^T\mu \leqslant f'(\overline{x};\mu)$; hence,

$$\max\big\{f_1'(\overline{x};\mu), f_2'(\overline{x};\mu)\big\} < f'(\overline{x};\mu). \tag{*}$$

On the other hand, by definition of the directional derivative, due to $f(\overline{x}) = f_1(\overline{x}) = f_2(\overline{x})$, we obtain the equality

$$\frac{f(\overline{x}+\lambda\mathbf{d}) - f(\overline{x})}{\lambda} = \max\left\{\frac{f_1(\overline{x}+\lambda\mathbf{d}) - f(\overline{x})}{\lambda}, \frac{f_2(\overline{x}+\lambda\mathbf{d}) - f(\overline{x})}{\lambda}\right\}, \quad \lambda > 0.$$

Passing with λ to 0, we obtain

$$f'(\overline{x};\mathbf{d}) = \max\big\{f_1'(\overline{x};\mathbf{d}), f_2'(\overline{x};\mathbf{d})\big\}.$$

With $\mathbf{d} = \mu$ we obtain an inconsistency with (*). Hence, the set $\partial f(\overline{x}) \setminus A$ is empty. □

Theorem 31 *Let $W \subset \mathbb{R}^n$ be an open and convex set, let $f\colon W \to \mathbb{R}$ be a convex function and let A be an $n \times m$ matrix. If $\tilde{W} = \{x \in \mathbb{R}^m\colon Ax \in W\}$, then $\tilde{W}$ is an open convex set and the function $F\colon \tilde{W} \to \mathbb{R}$ given by the formula $F(x) = f(Ax)$ at any point $x \in \tilde{W}$ has the subdifferential given by*

$$\partial F(x) = A^T\partial f(Ax).$$

Proof. Let $\overline{x} \in \tilde{W}$ and let $\xi \in \partial f(A\overline{x})$. Then,

$$f(Ax) \geqslant f(A\overline{x}) + \xi^T(Ax - A\overline{x}) = f(A\overline{x}) + (A^T\xi)^T(x - \overline{x}),$$

i.e., $A^T\xi \in \partial F(\overline{x})$. Hence, $A^T\partial f(A\overline{x}) \subset \partial F(A\overline{x})$. Suppose that there exists $\xi \in \partial F(A\overline{x}) \setminus A^T\partial f(A\overline{x})$. The set $A^T\partial f(A\overline{x})$ is convex and closed, as the image of a closed and convex set in a linear transformation. We apply the strong separation theorem to this set and $\{\xi\}$. There exists $\mu \in \mathbb{R}^m$ and $b \in \mathbb{R}$ such that

$$\mu^TA^T\tilde{\xi} < b < \mu^T\xi \quad \text{for all } \tilde{\xi} \in \partial f(A\overline{x}).$$

By taking the supremum over $\tilde{\xi} \in \partial f(A\overline{x})$, and using Theorem 29, we obtain $f'(A\overline{x};A\mu) \leqslant b$. The right hand side may be estimated by the directional derivative: $\mu^T\xi \leqslant F'(\overline{x};\mu)$. Hence,

$$f'(A\overline{x};A\mu) \leqslant b < F'(\overline{x};\mu).$$

But the directional derivatives satisfy the equality $F'(\overline{x};\mathbf{d}) = f'(A\overline{x};A\mathbf{d})$ for all $\mathbf{d} \in \mathbb{R}^m$. Thus, we have an inconsistency, which proves that $\partial F(A\overline{x}) \setminus A^T\partial f(A\overline{x})$ is the empty set. □

4. Extrema of convex functions with constraints

We consider a convex function $f\colon W \to \mathbb{R}$ defined in a convex set $W \subset \mathbb{R}^n$ and the following problem:

$$\begin{cases} f(x) \to \min, \\ x \in W. \end{cases}$$

A <u>global solution</u> is a feasible point $\bar{x}$ such that $f(\bar{x}) \leqslant f(x)$ for all $x \in W$.
A <u>local solution</u> is a point $\bar{x} \in W$ such that there exists $\varepsilon > 0$ such that $f(\bar{x}) \leqslant f(x)$ for all $x \in W \cap B(\bar{x}, \varepsilon)$, i.e., if the point $\bar{x}$ is a minimum of f in its neighbourhood. The local solution is <u>strict</u> if $f(\bar{x}) < f(x)$ for $x \neq \bar{x}$.

<u>Theorem 32</u> *Let the set $W \subset \mathbb{R}^n$ be convex, and the function $f\colon W \to \mathbb{R}^n$ be convex. If $\bar{x} \in W$ is a local solution of the problem above, then*

I) $\bar{x}$ is a global solution,

II) the set of global solutions is convex,

III) if f is strictly convex, then $\bar{x}$ is a strict local solution,

IV) if $\bar{x}$ is a strict local solution, then it is the unique global solution.

We do not assume the differentiability of f.

<u>Proof.</u> (I) is proved by contradiction. Suppose that there exists $x^* \in W$ such that $f(x^*) < f(\bar{x})$. As $\bar{x}$ is a local solution, $f(\bar{x}) \leqslant f(x)$ for all $x \in W \cap B(\bar{x}, \varepsilon)$, for some $\varepsilon > 0$. By convexity of W, this set contains the line segment $\overline{\bar{x}x^*}$. This line segment has a nonempty intersection with the ball $B(\bar{x}, \varepsilon)$; there exists $\lambda \in (0, 1)$ such that $\lambda\bar{x} + (1-\lambda)x^* \in B(\bar{x}, \varepsilon)$. By convexity of f, we obtain

$$f\bigl(\lambda\bar{x} + (1-\lambda)x^*\bigr) \leqslant \lambda f(\bar{x}) + (1-\lambda)f(x^*) < f(\bar{x}),$$

which contradicts $\bar{x}$ being a local minimum.

The proofs of (II), (III) and (IV) are left as exercises. □

So far, we have shown that a necessary and sufficient condition for a minimum of a differentiable convex function in an *open* set is the zero of the derivative or the gradient. This result may be generalised to arbitrary convex sets.

<u>Theorem 33</u> *Let the set $W \subset \mathbb{R}^n$ be convex and the function $f\colon W \to \mathbb{R}$ be convex. If f is differentiable at $\bar{x} \in W$, then there is the following equivalence: $\bar{x}$ is a minimum if and only if $Df(\bar{x})(x - \bar{x}) \geqslant 0$ for all $x \in W$.*

<u>Remark.</u> To speak of differentiability of f at a point $\bar{x}$, this function must be defined in a neighbourhood of this point, i.e., in a ball $B(\bar{x}, \varepsilon)$, $\varepsilon > 0$. If $\bar{x}$ is at the boundary of W, then we assume that f is defined in $W \cup B(\bar{x}, \varepsilon)$, though we omit it in the theorem's assumptions.

<u>Remark.</u> If $\bar{x} \in \operatorname{int} W$, then the condition in the theorem is equivalent to $Df(\bar{x}) = 0^T$.

<u>Proof.</u> Let $Df(\bar{x})(x - \bar{x}) \geqslant 0$ for all $x \in W$. Suppose that there is no minimum at $\bar{x}$. Then, there exists a point $x' \in W$ such that $f(x') < f(\bar{x})$. We construct a sequence $x_k = (1 - \frac{1}{k})\bar{x} + \frac{1}{k}x'$. By convexity of W, this sequence is contained in W. We consider the directional derivative of f in the direction of the vector $x' - \bar{x}$:

$$\begin{aligned} f'(\bar{x}; x' - \bar{x}) &= \lim_{k\to\infty} \frac{f\bigl(\bar{x} + \frac{1}{k}(x' - \bar{x})\bigr) - f(\bar{x})}{1/k} = \lim_{k\to\infty} \frac{f(x_k) - f(\bar{x})}{1/k} \\ &\leqslant \lim_{k\to\infty} \frac{(1 - \frac{1}{k})f(\bar{x}) + \frac{1}{k}f(x') - f(\bar{x})}{1/k} = f(x') - f(\bar{x}) < 0. \end{aligned}$$

By the assumption, we have

$$f'(\bar{x}; x' - \bar{x}) = Df(\bar{x})(x' - \bar{x}) \geqslant 0$$

This inconsistency proves that $\bar{x}$ is a minimum of f in W.

Now suppose that $\bar{x}$ is a solution. Let $x \in W$. The convexity of W implies that $\bar{x} + \lambda(x - \bar{x}) = (1-\lambda)\bar{x} + \lambda x \in W$ for all $\lambda \in [0, 1]$. By definition of the derivative,

$$Df(\bar{x})(x - \bar{x}) = \lim_{\lambda\searrow 0, \lambda<1} \frac{f\bigl(\bar{x} + \lambda(x - \bar{x})\bigr) - f(\bar{x})}{\lambda}.$$

As $\bar{x}$ is a minimum, $f\bigl(\bar{x} + \lambda(x - \bar{x})\bigr) \geqslant f(\bar{x})$. Hence, $Df(\bar{x})(x - \bar{x}) \geqslant 0$. □

<u>Corollary 5</u> *If $\bar{x} \in W$, where $W \subset \mathbb{R}^n$ is convex, is a local minimum of a (not necessarily convex) function $f\colon W \to \mathbb{R}$, differentiable at $\bar{x}$, then $Df(\bar{x})(x - \bar{x}) \geqslant 0$ for all $x \in W$.*

Theorem 34 *Let $\mathbb{X} \subset \mathbb{R}^n$ be a convex and open set and $f\colon \mathbb{X} \to \mathbb{R}$ be a convex function. Suppose that the feasible set W is a subset of $\mathbb{X}$. Then $\bar{\mathbf{x}} \in W$ is a minimum if and only if there exists $\boldsymbol{\xi} \in \partial f(\bar{\mathbf{x}})$, such that $\boldsymbol{\xi}^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ for all $\mathbf{x} \in W$.*

Corollary 6 *If $\bar{\mathbf{x}} \in \operatorname{int} W$, then f has a global minimum at $\bar{\mathbf{x}}$ if and only if $\mathbf{0} \in \partial f(\bar{\mathbf{x}})$.*

Proof. Suppose that there exists $\boldsymbol{\xi} \in \partial f(\bar{\mathbf{x}})$ such that $\boldsymbol{\xi}^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ for all $\mathbf{x} \in W$. As $\boldsymbol{\xi}$ is a subgradient, it follows that

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^T(\mathbf{x} - \bar{\mathbf{x}}), \quad \mathbf{x} \in W.$$

Now it suffices to use the assumption to notice that $f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}})$, i.e., $\bar{\mathbf{x}}$ is a minimum.

Now suppose that $\bar{\mathbf{x}} \in W$ is a minimum. We define two sets:

$$C_1 = \{\, (\mathbf{x}, z) \in \mathbb{R}^{n+1} \colon \mathbf{x} \in \mathbb{X},\ z > f(\mathbf{x}) - f(\bar{\mathbf{x}}) \,\},$$
$$C_2 = \{\, (\mathbf{x}, z) \in \mathbb{R}^{n+1} \colon \mathbf{x} \in W,\ z \leqslant 0 \,\}.$$

Both sets are convex and the interior of C_1 is nonempty (the interior of C_2 may be empty if the interior of W is empty). From $\bar{\mathbf{x}}$ being a solution it follows that $C_1 \cap C_2 = \emptyset$. We use the weak separation theorem: there exists a nonzero vector $(\boldsymbol{\mu}, \gamma) \in \mathbb{R}^{n+1}$ and a constant b such that

$$\boldsymbol{\mu}^T\mathbf{x} + \gamma z \leqslant b, \quad \text{for all } \mathbf{x} \in \mathbb{X},\ z > f(\mathbf{x}) - f(\bar{\mathbf{x}}),$$
$$\boldsymbol{\mu}^T\mathbf{x} + \gamma z \geqslant b, \quad \text{for all } \mathbf{x} \in W,\ z \leqslant 0.$$

Before we proceed, let's take a look at Figure 3. As we can see, the two sets "touch" each other at the point $(\bar{\mathbf{x}}, 0)$. The separating hyperplane must therefore contain this point. It is tangent to the graph of the function $\mathbf{x} \to f(\mathbf{x}) - f(\bar{\mathbf{x}})$ and in fact it determines a subgradient of this function, which is also a subgradient of the function f at $\bar{\mathbf{x}}$.

Now we prove it analytically. We can write

$$\boldsymbol{\mu}^T(\mathbf{x} - \bar{\mathbf{x}}) + \gamma z \leqslant \tilde{b}, \quad \text{for all } \mathbf{x} \in \mathbb{X},\ z > f(\mathbf{x}) - f(\bar{\mathbf{x}}), \tag{*}$$
$$\boldsymbol{\mu}^T(\mathbf{x} - \bar{\mathbf{x}}) + \gamma z \geqslant \tilde{b}, \quad \text{for all } \mathbf{x} \in W,\ z \leqslant 0, \tag{**}$$

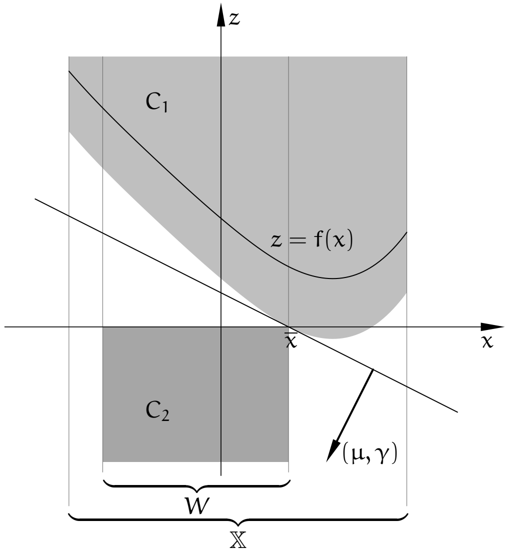

Figure 3: The sets C_1 and C_2

where $\tilde{b} = b - \boldsymbol{\mu}^T\bar{\mathbf{x}}$. We notice that γ cannot be positive, as we can take an arbitrarily small z in (**), which leads to inconsistency.

Taking $\mathbf{x} = \bar{\mathbf{x}}$ and $z = 0$ in (**), we obtain $\tilde{b} \leqslant 0$. On the other hand, with $\mathbf{x} = \bar{\mathbf{x}}$, the inequality (*) turns into $\gamma z \leqslant \tilde{b}$ for $z > 0$; hence, $\tilde{b} \geqslant 0$. Therefore, $\tilde{b} = 0$. Using this fact we show that γ cannot be zero. From (*), due to $\mathbb{X}$ being open, we would then obtain $\boldsymbol{\mu} = \mathbf{0}$ which contradicts $(\boldsymbol{\mu}, \gamma) \neq \mathbf{0}$. Thus we proved that $\gamma < 0$. Taking $z = f(\mathbf{x}) - f(\bar{\mathbf{x}})$ in (*), we obtain

$$\boldsymbol{\mu}^T(\mathbf{x} - \bar{\mathbf{x}}) + \gamma\bigl(f(\mathbf{x}) - f(\bar{\mathbf{x}})\bigr) \leqslant 0.$$

After dividing the sides by γ, which is negative, we obtain

$$\frac{\boldsymbol{\mu}^T}{\gamma}(\mathbf{x} - \bar{\mathbf{x}}) + f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geqslant 0,$$

which proves that $\boldsymbol{\xi} = -\frac{\boldsymbol{\mu}}{\gamma} \in \partial f(\bar{\mathbf{x}})$. Taking $z = 0$ in (**), we obtain $\boldsymbol{\mu}^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant \tilde{b} = 0$. After dividing the sides by $-\gamma$, we obtain $\boldsymbol{\xi}^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$, and the proof is complete. □

Pseudoconvex functions

We introduce a family of functions such that

$$\mathrm{D}f(\overline{x}) = 0^{\mathsf{T}} \quad \Leftrightarrow \quad \overline{x} \text{ is a global minimum of } f.$$

Definition 14 *Let $W \subset \mathbb{R}^n$ be convex, open and nonempty and let $f\colon W \to \mathbb{R}$. The function f is pseudoconvex in W if it is differentiable in W and*

$$\mathrm{D}f(x)(y - x) \geqslant 0 \quad \Rightarrow \quad f(y) \geqslant f(x) \qquad \text{for all } x, y \in W.$$

A function f is strictly pseudoconvex in W if

$$\mathrm{D}f(x)(y - x) \geqslant 0 \quad \Rightarrow \quad f(y) > f(x) \qquad \text{for all } x, y \in W,\ x \neq y.$$

A function f is pseudoconvex at a point $\overline{x} \in W$ if it is differentiable at $\overline{x}$ and

$$\mathrm{D}f(\overline{x})(y - \overline{x}) \geqslant 0 \quad \Rightarrow \quad f(y) \geqslant f(\overline{x}) \qquad \text{for all } y \in W.$$

Similarly is defined a function strictly pseudoconvex at a point $\overline{x} \in W$.

A function f is (strictly) pseudoconcave if $-f$ is (strictly) pseudoconvex.

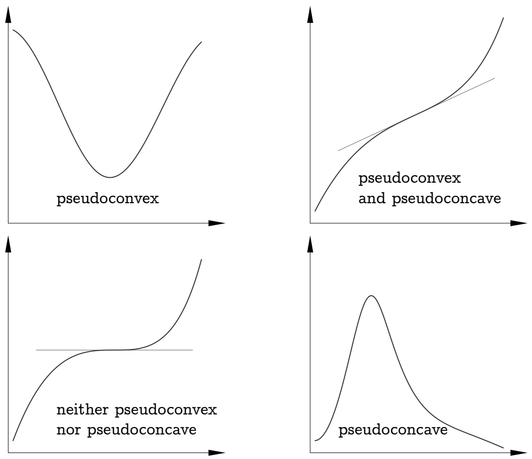

Figure 4: Examples explaining the notion of pseudoconvexity

Remark. Pseudoconvexity at a point is a property of a function related with the entire set W, even if the differentiability is needed at that point only.

Remark. A condition equivalent to that in the definition is the following:

$$f(y) < f(x) \quad \Rightarrow \quad \mathrm{D}f(x)(y - x) < 0.$$

Lemma 8 *Let $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is nonempty, open and convex. If f is (strictly) convex and differentiable in W, then f is (strictly) pseudoconvex.*

Proof. Suppose that f is convex. By the supporting hyperplane theorem, for any $\overline{x}, x \in W$ we have

$$f(x) \geqslant f(\overline{x}) + \mathrm{D}f(\overline{x})(x - \overline{x}).$$

Thus, if $\mathrm{D}f(\overline{x})(x - \overline{x}) \geqslant 0$, then $f(x) \geqslant f(\overline{x})$ and f is indeed pseudoconvex. The proof of strict pseudoconvexity of a strictly convex function is similar. □

Lemma 9 *Let $f\colon W \to \mathbb{R}$, where $W \subset \mathbb{R}^n$ is nonempty, open and convex. If a function f is pseudoconvex at $\overline{x} \in W$, then $\overline{x}$ is a global minimum of f if and only if $\mathrm{D}f(\overline{x}) = 0^{\mathsf{T}}$.*

Proof. Identical as the proof of Corollary 3. □

Lemma 10 *Let $W \subset \mathbb{R}^n$ be convex and let $f\colon W \to \mathbb{R}$ be pseudoconvex. Then, $\overline{x}$ is a solution of the minimization problem if and only if $\mathrm{D}f(\overline{x})(x - \overline{x}) \geqslant 0$ for all $x \in W$.*

Proof. Identical as the proof of Theorem 33. □

Finding maxima of convex functions

Definition 15 *An extremal point of a convex set $W \subset \mathbb{R}^n$ is such a point $\overline{x} \in W$, which is not an internal point of any line segment contained in W, i.e., if $\overline{x} = \lambda x_1 + (1 - \lambda)x_2$, where $\lambda \in (0, 1)$ and $x_1, x_2 \in W$, then $x_1 = x_1 = \overline{x}$.*

Definition 16 *The convex hull of points $A = \{x_i\colon i \in I\}$ is the set of points being convex combinations of all finite subsets of the set A.*

Equivalently, the convex hull may be defined as the smallest convex set containing the set A.

Definition 17 *The dimension of a convex set $U \in \mathbb{R}^n$ is the dimension of the smallest affine subspace $A \in \mathbb{R}^n$ containing U (affine hull of U), i.e., the set*

$$\operatorname{aff} U = \Big\{ \sum_{i=1}^{k} \lambda_i x_i \colon x_1, \dots, x_k \in U \Big\}.$$

Remark. Any affine subspace of $\mathbb{R}^n$, subject to a translation, may become a linear subspace.

Remark. A set U whose dimension is m may be seen as a subset of $\mathbb{R}^m$.

Remark. A convex set $U \subset \mathbb{R}^n$ has a nonempty interior if and only if its dimension is n.

Theorem 35 *(Krein–Milman, finite-dimensional case) Let $U \subset \mathbb{R}^n$ be convex and compact. It is then the convex hull of the set of its extremal points.*

Lemma 11 *Let $U \subset \mathbb{R}^n$ be a convex set with a nonempty interior and let $\bar{x} \in \partial U$. The point $\bar{x}$ is an element of a hyperplane such that U is contained in one of two halfspaces separated by this hyperplane, which we call a supporting hyperplane.*

Proof. By the weak separation theorem applied to $\operatorname{int} U$ and $V = \{\bar{x}\}$ there exists $a \in \mathbb{R}^n \setminus \{0\}$ such that $a^T x \leqslant a^T \bar{x}$ for all $x \in U$. The hyperplane sought is

$$H = \{ x \in \mathbb{R}^n \colon a^T x = a^T \bar{x} \}.$$

The set U is contained in the halfspace $\{ x \in \mathbb{R}^n \colon a^T x \leqslant a^T \bar{x} \}$. □

First proof of the Krein–Milman theorem. We use induction with respect to the dimension m of the compact and convex set U. The cases $m = 0$ (U consists of a single point) and $m = 1$ (U is a line segment) are obvious. The induction step is the following: assume that each convex and compact set of dimension not greater than m is the convex hull of the set of its extremal points. Let U be a convex and compact set of dimension $m+1$. We look at U as a subset of $\mathbb{R}^{m+1}$. Its interior is then nonempty. Let $\bar{x} \in U$.

First, let $\bar{x} \in \partial U$. By Lemma 11, there exists a supporting hyperplane H of the set U. The set $U_{\bar{x}} = U \cap H$ is convex and compact and its dimension is at most m. By the inductive assumption, $\bar{x}$ is a convex combination of the extremal points of $U_{\bar{x}}$. It has to be shown that these points are also extremal points of U. But this is a consequence of the fact that no extremal point of $U_{\bar{x}}$ is in the interior of a line segment whose both end points are in U.

Let $\bar{x} \in \operatorname{int} U$. We can take an arbitrary line passing through $\bar{x}$; its intersection with U is a line segment whose end points x_1, x_2 are located on the boundary of U. Both these points are convex combinations of extremal points of U, and so is the point $\bar{x}$. □

Second proof of the Krein–Milman theorem. If U is a subset of $\mathbb{R}^1$, then the claim is trivial. Assume that any convex and compact subset of $\mathbb{R}^m$ is the convex hull of the set of its extremal points. We shall prove the theorem for the subsets of $\mathbb{R}^{m+1}$. Let W be the convex hull of the set of extremal points of U. Obviously, $W \subset U$. Suppose that there exists $\bar{x} \in U \setminus W$. Then, there exists a ball $B(\bar{x}, \varepsilon)$ disjoint with W. By the strong separation theorem, there exists a nonzero vector $a \in \mathbb{R}^{m+1}$ and a number α such that $a^T x \leqslant \alpha$ for $x \in W$ and $a^T \bar{x} > \alpha$. Let $\beta = \sup_{x \in U} a^T x$. As U is compact, β is finite. The hyperplane $P = \{ x \in \mathbb{R}^{m+1} \colon a^T x = \beta \}$ does not intersect W, but it has a common point with U; indeed, $P_U \overset{\mathrm{def}}{=} P \cap U$ is nonempty, as U is compact; hence, the supremum β of $a^T x$ is taken at some point $x \in U$. We are going to show that in the set P_U there is an extremal point of U, which is inconsistent with the definition of W. The set P_U is a nonempty, compact and convex set of dimension m. It may be seen as a subset of $\mathbb{R}^m$; by the inductive assumption it is the convex hull of the set of its extremal points. Let $\bar{y}$ be one of extremal points of P_U; suppose that $\bar{y}$ is a convex combination of some points of U: $\bar{y} = \lambda y_1 + (1-\lambda) y_2$, $y_1, y_2 \in U$, $\lambda \in (0, 1)$. Then, $\beta = a^T \bar{y} = \lambda a^T y_1 + (1-\lambda) a^T y_2$. By definition of β, both terms, $a^T y_1$ and $a^T y_2$ must be equal to β; hence, $y_1, y_2 \in P_U$. But $\bar{y}$ is an extremal point of P_U, therefore, $\bar{y} = y_1 = y_2$ and thus $\bar{y}$ is an extremal point of U. □

Theorem 36 *Let $f \colon W \to \mathbb{R}$ be a convex and continuous function defined in a convex and compact set $W \subset \mathbb{R}^n$. Then at least one of global solutions of* *the problem*

$$\begin{cases} f(x) \to \max, \\ x \in W. \end{cases}$$

is an extremal point of the set W.

Proof. A continuous function in a compact set achievess its extrema. Therefore, the problem formulated above has a solution $\overline{x} \in W$. By Theorem 35, the point $\overline{x}$ is a convex combination of a finite number of extremal points of the set W, i.e.,

$$\overline{x} = a_1 x_1 + \cdots + a_m x_m,$$

where $x_1, \ldots, x_m$ are extremal points, $a_1, \ldots, a_m > 0$ and $a_1 + \cdots + a_m = 1$.
By convexity of f we obtain

$$f(\overline{x}) \leqslant a_1 f(x_1) + \cdots + a_m f(x_m).$$

As $\overline{x}$ is a maximum of f in W, there must be $f(x_1) = \cdots = f(x_m)$. □

Definition 18 *A set $W \subset \mathbb{R}^n$ is called a polyhedral set if it is the intersection of a finite number of halfspaces, i.e.,*

$$W = \{x \in \mathbb{R}^n \colon \mathbf{p}_i^T x \leqslant \alpha_i,\ i = 1, \ldots, m\},$$

where $\mathbf{p}_i \in \mathbb{R}^n \setminus \{0\}$, $\alpha_i \in \mathbb{R}$.

Lemma 12 *A polyhedral set is convex and closed.*

Proof. Obvious. □

5. Necessary condition of the first order

We consider the optimization problem

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{x} \in W, \end{cases} \tag{*}$$

where $W \subset \mathbb{R}^n$ and $f\colon W \to \mathbb{R}$. Let $\overline{\mathbf{x}}$ be a local solution. We are going to connect the local geometry of the set W at $\overline{\mathbf{x}}$ with the behaviour of the function f, i.e., with directions of descent of its values. By the local geometry we understand the set of directions at which we can move without leaving the set W.

Definition 19 *The cone of tangents $T(\overline{\mathbf{x}})$ to the set W at $\overline{\mathbf{x}} \in \operatorname{cl} W$ is the set of vectors $\mathbf{d} \in \mathbb{R}^n$ such that*

$$\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \overline{\mathbf{x}})$$

for some numbers $\lambda_k > 0$ and points $\mathbf{x}_k \in W$ such that $\mathbf{x}_k \to \overline{\mathbf{x}}$.

According to the definition the vector $\mathbf{d}$ is an element of the cone of tangents $T(\overline{\mathbf{x}})$ if it is the limit of a sequence of vectors determined by a sequence of feasible points $(\mathbf{x}_k)_k$ tending to $\overline{\mathbf{x}}$. It may formally be described as follows

$$T(\overline{\mathbf{x}}) = \Big\{ \mathbf{d} \in \mathbb{R}^n\colon \mathbf{d} = \lambda \lim_{k\to\infty} \frac{\mathbf{x}_k - \overline{\mathbf{x}}}{\|\mathbf{x}_k - \overline{\mathbf{x}}\|} \\ \text{for some } (\mathbf{x}_k)_k \subset W,\ \mathbf{x}_k \to \overline{\mathbf{x}},\ \mathbf{x}_k \neq \overline{\mathbf{x}},\ \lambda \geqslant 0 \Big\},$$

which is to be proved as an exercise, as well as the lemma below.

Lemma 13 *I) The set $T(\overline{\mathbf{x}})$ is a cone, i.e., $\lambda\mathbf{d} \in T(\overline{\mathbf{x}})$ for all $\mathbf{d} \in T(\overline{\mathbf{x}})$ and $\lambda \geqslant 0$. In particular, $0 \in T(\overline{\mathbf{x}})$.*

II) If $\overline{\mathbf{x}} \in \operatorname{int} W$, then $T(\overline{\mathbf{x}}) = \mathbb{R}^n$.

III) The cone $T(\overline{\mathbf{x}})$ is closed.

Definition 20 *Let $f\colon \mathbb{X} \to \mathbb{R}$ be differentiable at $\overline{\mathbf{x}} \in \mathbb{X}$. The set of descent directions(or improving directions) of f at $\overline{\mathbf{x}}$ is the set*

$$D(\overline{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^n\colon Df(\overline{\mathbf{x}})\mathbf{d} < 0 \,\}.$$

Theorem 37 *Let $\overline{\mathbf{x}}$ be a local solution of the problem (*). If f is differentiable at $\overline{\mathbf{x}}$, then*

$$D(\overline{\mathbf{x}}) \cap T(\overline{\mathbf{x}}) = \emptyset.$$

Proof. Let $\mathbf{d} \in T(\overline{\mathbf{x}})$. Then, $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \overline{\mathbf{x}})$ for some sequence of points $(\mathbf{x}_k)_k$ converging to $\overline{\mathbf{x}}$ and a sequence $(\lambda_k)_k \subset (0,\infty)$. As the function f is differentiable at $\overline{\mathbf{x}}$, there is

$$f(\mathbf{x}_k) = f(\overline{\mathbf{x}}) + Df(\overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}}) + o(\|\mathbf{x}_k - \overline{\mathbf{x}}\|).$$

As $\overline{\mathbf{x}}$ is a local solution, $f(\mathbf{x}_k) \geqslant f(\overline{\mathbf{x}})$ for k sufficiently large. Together with the formula above, we obtain the following estimation:

$$0 \leqslant f(\mathbf{x}_k) - f(\overline{\mathbf{x}}) = Df(\overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}}) + o(\|\mathbf{x}_k - \overline{\mathbf{x}}\|)$$

By multiplying the sides by λ_k, we obtain

$$0 \leqslant Df(\overline{\mathbf{x}})\big(\lambda_k(\mathbf{x}_k - \overline{\mathbf{x}})\big) + \lambda_k o(\|\mathbf{x}_k - \overline{\mathbf{x}}\|).$$

A little trick allows us to deal with the remainder; here we go with k to infinity:

$$0 \leqslant Df(\overline{\mathbf{x}}) \underbrace{\big(\lambda_k(\mathbf{x}_k - \overline{\mathbf{x}})\big)}_{\to \mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k - \overline{\mathbf{x}}\|}_{\to\|\mathbf{d}\|} \underbrace{\frac{o(\|\mathbf{x}_k - \overline{\mathbf{x}}\|)}{\|\mathbf{x}_k - \overline{\mathbf{x}}\|}}_{\to 0}.$$

We proved that $Df(\overline{\mathbf{x}})\mathbf{d} \geqslant 0$, i.e., $\mathbf{d} \notin D(\overline{\mathbf{x}})$. □

Example. Consider the following problem:

$$\begin{cases} x_1^2 + x_2^2 \to \min, \\ x_1 + x_2 \geqslant 1. \end{cases}$$

We denote $f(x_1, x_2) = x_1^2 + x_2^2$, $W = \{\, \mathbf{x} \in \mathbb{R}^2\colon x_1 + x_2 \geqslant 1 \,\}$.

We investigate the sets $T(\overline{\mathbf{x}})$ and $D(\overline{\mathbf{x}})$ at three points: $(1,1)$, $(1,0)$, $(\frac{1}{2}, \frac{1}{2})$.

The point $(1,1)$ is located in the interior of W, i.e., $T(\overline{\mathbf{x}}) = \mathbb{R}^2$. The set of descent directions is

$$\begin{aligned} D(\overline{\mathbf{x}}) &= \{\, \mathbf{d} \in \mathbb{R}^2\colon Df(\overline{\mathbf{x}})\mathbf{d} < 0 \,\} = \{\, \mathbf{d} \in \mathbb{R}^2\colon [2,2]\mathbf{d} < 0 \,\} \\ &= \{\, (d_1, d_2) \in \mathbb{R}^2\colon d_1 + d_2 < 0 \,\}. \end{aligned}$$

Obviously, the two sets have a nonempty intersection; hence, there is no minimum at $(1,1)$.

The point $(1,0)$ is located at the boundary of W, and we have

$$T(\bar{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^2 \colon d_1 + d_2 \geqslant 0 \,\}, \quad D(\bar{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^2 \colon d_1 < 0 \,\}.$$

The intersection of the two sets is nonempty and thus there is no minimum at $(1,0)$.

For the point $(\frac{1}{2}, \frac{1}{2})$ we find the sets

$$\begin{aligned} T(\bar{\mathbf{x}}) &= \{\, \mathbf{d} \in \mathbb{R}^2 \colon d_1 + d_2 \geqslant 0 \,\}, \\ D(\bar{\mathbf{x}}) &= \{\, \mathbf{d} \in \mathbb{R}^2 \colon [1,1]\mathbf{d} < 0 \,\} = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_1 + d_2 < 0 \,\}. \end{aligned}$$

Their intersection is empty and thus *it is possible* (it is still to be verified) that there is a minimum at $(\frac{1}{2}, \frac{1}{2})$.

Inequality constraints

Now we consider the problem given in the following form:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \dots, m, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{**}$$

where $\mathbb{X} \subset \mathbb{R}^n$ is an open set and $f, g_1, \dots, g_m \colon \mathbb{X} \to \mathbb{R}$. Here the feasible set is

$$W = \{\, \mathbf{x} \in \mathbb{X} \colon g_1(\mathbf{x}) \leqslant 0, \dots, g_m(\mathbf{x}) \leqslant 0 \,\}.$$

The functions $g_1, \dots, g_m$ are called inequality constraints and the problem (**) is called the optimization problem with inequality constraints.

We assume that the functions g_i are continuous; then the motion around a point $\bar{\mathbf{x}}$ is restricted by only those functions equal to 0 at $\bar{\mathbf{x}}$. The others, due to their continuity, are less than zero in a neighbourhood of $\bar{\mathbf{x}}$.

Definition 21 *The set of active (or binding or tight) constraints at a point $\bar{\mathbf{x}}$ is the set*

$$I(\bar{\mathbf{x}}) = \big\{\, i \in \{1, \dots, m\} \colon g_i(\bar{\mathbf{x}}) = 0 \,\big\}.$$

We are going to connect the properties of active constraints at a point $\bar{\mathbf{x}} \in W$ with the local geometry of the set W around $\bar{\mathbf{x}}$. To do this we introduce the following definition:

Definition 22 *Let $\bar{\mathbf{x}} \in W$ and let the functions g_i which describe the constraints active at $\bar{\mathbf{x}}$ be differentiable at $\bar{\mathbf{x}}$. The cone of tangents for active (binding) constraints is the set*

$$T_{\mathrm{lin}}(\bar{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^n \colon Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0 \text{ for all } i \in I(\bar{\mathbf{x}}) \,\}.$$

The cone of tangents for active constraints is a polyhedral set, i.e., it is convex and closed.

Lemma 14 *If $\bar{\mathbf{x}} \in W$, then $T(\bar{\mathbf{x}}) \subset T_{\mathrm{lin}}(\bar{\mathbf{x}})$.*

Proof. The proof is similar to that of Theorem 37. Let $\mathbf{d} \in T(\bar{\mathbf{x}})$. Then, $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$ for some sequence of points $(\mathbf{x}_k)_k \subset W$ converging to $\bar{\mathbf{x}}$ and a sequence of positive numbers λ_k. Let $i \in I(\bar{\mathbf{x}})$. As g_i is differentiable at $\bar{\mathbf{x}}$, we have

$$g_i(\mathbf{x}_k) = g_i(\bar{\mathbf{x}}) + Dg_i(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

The i-th constraint is active at $\bar{\mathbf{x}}$; hence, $g_i(\bar{\mathbf{x}}) = 0$. Obviously, $g_i(\mathbf{x}_k) \leqslant 0$ due to $\mathbf{x}_k \in W$. From the above we have the estimation

$$0 \geqslant g_i(\mathbf{x}_k) - g_i(\bar{\mathbf{x}}) = Dg_i(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

We multiply the sides by λ_k to obtain

$$0 \geqslant Dg_i(\bar{\mathbf{x}})\big(\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})\big) + \lambda_k o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

The same trick as before gives us

$$0 \geqslant Dg_i(\bar{\mathbf{x}}) \underbrace{\big(\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})\big)}_{\to \mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k - \bar{\mathbf{x}}\|}_{\to \|\mathbf{d}\|} \underbrace{\frac{o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|)}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}}_{\to 0}.$$

Here we proved that $Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0$. This holds for all $i \in I(\bar{\mathbf{x}})$; hence, $\mathbf{d} \in T_{\mathrm{lin}}(\bar{\mathbf{x}})$. □

Example. Let $W = \{\, (x_1, x_2) \in \mathbb{R}^2 \colon x_1^2 + x_2^2 \leqslant 1,\ x_2 \geqslant 0) \,\}$. In the canonical form we have

$$W = \{\, (x_1, x_2) \in \mathbb{X} = \mathbb{R}^2 \colon x_1^2 + x_2^2 - 1 \leqslant 0, -x_2 \leqslant 0) \,\},$$

and $g_1(x_1, x_2) = x_1^2 + x_2^2 - 1$, $g_2 = -x_2$. We look at three points of the set W: $(\frac{1}{2}, \frac{1}{2})$, $(0,1)$, $(1,0)$.

No constraint is active at the point $(\frac{1}{2}, \frac{1}{2})$; there is $I\big((\frac{1}{2}, \frac{1}{2})\big) = \emptyset$ and $T\big((\frac{1}{2}, \frac{1}{2})\big) = T_{\text{lin}}\big((\frac{1}{2}, \frac{1}{2})\big) = \mathbb{R}^2$.

At the point $(0, 1)$ we have $I\big((0,1)\big) = \{1\}$, $T\big((0,1)\big) = \{\, (d_1, d_2) \in \mathbb{R}^2\colon d_2 \leqslant 0 \,\}$ and

$$T_{\text{lin}}\big((0,1)\big) = \{\, \mathbf{d} \in \mathbb{R}^2\colon Dg_1(0,1)\mathbf{d} \leqslant 0 \,\} = \{\, \mathbf{d} \in \mathbb{R}^2\colon [0,2]\mathbf{d} \leqslant 0 \,\} = T\big((0,1)\big).$$

At the point $(1, 0)$ there is $I\big((1,0)\big) = \{1, 2\}$,
$T\big((1,0)\big) = \{\, (d_1, d_2) \in \mathbb{R}^2\colon d_1 \leqslant 0,\ d_2 \geqslant 0 \,\}$ and

$$\begin{aligned} T_{\text{lin}}\big((1,0)\big) &= \{\, \mathbf{d} \in \mathbb{R}^2\colon Dg_1(1,0)\mathbf{d} \leqslant 0,\ Dg_2(1,0)\mathbf{d} \leqslant 0 \,\} \\ &= \{\, \mathbf{d} \in \mathbb{R}^2\colon [2,0]\mathbf{d} \leqslant 0,\ [0,-1]\mathbf{d} \leqslant 0 \,\} = T\big((1,0)\big). \end{aligned}$$

Example. The same set W may have another description:
$W = \{\, (x_1, x_2) \in \mathbb{R}^2\colon x_1^2 + x_2^2 - 1 \leqslant 0,\ -x_2^3 \leqslant 0 \,\}$. The second constraint is now described by the function $g_2(x_1, x_2) = -x_2^3$. The cone of tangent directions at $(1, 0)$ is unchanged, but

$$\begin{aligned} T_{\text{lin}}\big((1,0)\big) &= \{\, \mathbf{d} \in \mathbb{R}^2\colon Dg_1(1,0)\mathbf{d} \leqslant 0,\ Dg_2(1,0)\mathbf{d} \leqslant 0 \,\} \\ &= \{\, \mathbf{d} \in \mathbb{R}^2\colon [2,0]\mathbf{d} \leqslant 0,\ [0,0](1,0)\mathbf{d} \leqslant 0 \,\} \\ &= \{\, (d_1, d_2) \in \mathbb{R}^2\colon d_1 \leqslant 0 \,\} \end{aligned}$$

and in this case $T(\bar{\mathbf{x}}) \neq T_{\text{lin}}(\bar{\mathbf{x}})$.

Necessary Kuhn–Tucker conditions

By Lemma 14, $T(\bar{\mathbf{x}}) \subset T_{\text{lin}}(\bar{\mathbf{x}})$. Often, but not always, the two sets are equal. It is an important property, being the starting point to the entire theory of nonlinear optimization by Kuhn and Tucker. We begin studying it with the lemma:

Lemma 15 *(G. Farkas, 1901) Let A be an $m \times n$ real matrix and let $\mathbf{d} \in \mathbb{R}^n$. Then exactly one of the two following systems of equations and inequalities has a solution:*

$$(1)\left\{\begin{array}{l} A\mathbf{x} \leqslant \mathbf{0}, \\ \mathbf{d}^T\mathbf{x} > 0, \\ \mathbf{x} \in \mathbb{R}^n \end{array}\right. \qquad (2)\left\{\begin{array}{l} A^T\mathbf{y} = \mathbf{d}, \\ \mathbf{y} \geqslant \mathbf{0}, \\ \mathbf{y} \in \mathbb{R}^m. \end{array}\right.$$

Proof. First we show that if (2) has a solution, then (1) does not. Let $\mathbf{y}$ satisfy (2). Then, $\mathbf{d} = A^T\mathbf{y}$. After substituting this to (1) we obtain

$$\left\{\begin{array}{l} A\mathbf{x} \leqslant \mathbf{0}, \\ \mathbf{y}^TA\mathbf{x} > 0. \end{array}\right.$$

The first inequality means that each coordinate of the vector $A\mathbf{x}$ is nonpositive. As the coordinates of $\mathbf{y}$ are nonnegative, the scalar product $\mathbf{y}^TA\mathbf{x}$ is not positive. This is inconsistent with the second inequality above; hence, (1) does not have a solution.

Now suppose that (2) does not have a solution. Let

$$V \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^n\colon \mathbf{x} = A^T\mathbf{y},\ \mathbf{y} \in \mathbb{R}^m,\ \mathbf{y} \geqslant \mathbf{0}\}.$$

The set V is a polyhedral set; hence, it is convex and closed. As (2) does not have a solution, $\mathbf{d} \notin V$. By the strong separation theorem applied to the sets V and $U = \{\mathbf{d}\}$, there exists a vector $\mathbf{a} \in \mathbb{R}^n$ such that

$$\mathbf{a}^T\mathbf{d} > \sup_{x \in V} \mathbf{a}^T\mathbf{x}.$$

Below we show that $\bar{\mathbf{x}} = \mathbf{a}$ is a solution of (1). Let $\alpha = \sup_{x \in V} \bar{\mathbf{x}}^T\mathbf{x}$. From $\mathbf{0} \in V$ it follows that $\alpha \geqslant 0$; hence, $\mathbf{a}^T\mathbf{d} = \mathbf{d}^T\bar{\mathbf{x}} > 0$. It remains to be proved that $A\bar{\mathbf{x}} \leqslant \mathbf{0}$. Suppose that the i-th coordinate of $A\bar{\mathbf{x}}$ is positive. By definition of V, for any $\mathbf{y} \geqslant \mathbf{0}$ there is $\alpha \geqslant \bar{\mathbf{x}}^TA^T\mathbf{y} = \mathbf{y}^TA\bar{\mathbf{x}}$. Let $\mathbf{y}_k = k\mathbf{e}_i$ (i.e., the i-th coordinate of the vector $\mathbf{y}_k$ is k and all other coordinates are zero). If the i-th coordinate $(A\bar{\mathbf{x}})_i$ of $A\bar{\mathbf{x}}$ is positive, then,

$$\lim_{k\to\infty} \mathbf{y}_k^TA\bar{\mathbf{x}} = \lim_{k\to\infty} k(A\bar{\mathbf{x}})_i = \infty,$$

which is inconsistent with $\mathbf{y}^TA\bar{\mathbf{x}} \leqslant \alpha$ for all $\mathbf{y} \geqslant \mathbf{0}$. Hence, $A\bar{\mathbf{x}} \leqslant \mathbf{0}$. □

Remark. If a vector $\mathbf{x}$ is a solution of System (1), then for all $a > 0$ the vector $a\mathbf{x}$ is also a solution; hence, the set of solutions of (1) is either empty or infinite. On the other hand, if the rows of the matrix A are linearly independent and System (2) has a solution $\mathbf{y}$, then this solution is unique.

An example is shown in Figure 5. The matrix A is 2×2, its rows are the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ (they are row matrices 1×2). For System (1) from the lemma, the condition $A\mathbf{x} \leqslant \mathbf{0}$ is equivalent to two inequalities, $\mathbf{a}_1\mathbf{x} \leqslant 0$ and $\mathbf{a}_2\mathbf{x} \leqslant 0$. The sets of vectors $\mathbf{x}$ satisfying each of the two inequalities are hatched with dashed lines; their intersection is the set of vectors $\mathbf{x}$ such that $A\mathbf{x} \leqslant \mathbf{0}$. The set of vectors $\mathbf{x}$

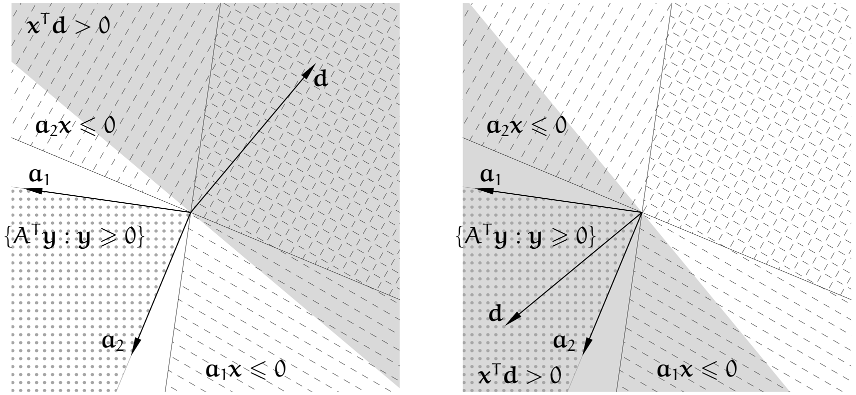
$\mathbf{x}^T\mathbf{d} > 0$
$\mathbf{d}$
$\mathbf{a}_2\mathbf{x} \leqslant 0$
$\mathbf{a}_1$
$\{A^T\mathbf{y} : \mathbf{y} \geqslant 0\}$
$\mathbf{a}_2$
$\mathbf{a}_1\mathbf{x} \leqslant 0$
$\mathbf{a}_2\mathbf{x} \leqslant 0$
$\mathbf{a}_1$
$\{A^T\mathbf{y} : \mathbf{y} \geqslant 0\}$
$\mathbf{d}$
$\mathbf{a}_2$
$\mathbf{x}^T\mathbf{d} > 0$
$\mathbf{a}_1\mathbf{x} \leqslant 0$

Figure 5: Farkas' lemma: on the left side (1) has a solution and on the right side (2) has a solution

such that $\mathbf{d}^T\mathbf{x} > 0$, where $\mathbf{d}$ is given, is grey. Clearly, (1) has solutions if and only if the grey area and the doubly hatched area intersect. This is the case in the picture on the left side. On the right side, System (2) has no solution. To find a geometric interpretation of (2), we need to see how the set $V = \{A^T\mathbf{y} \colon \mathbf{y} \geqslant \mathbf{0}\}$ looks like. Any vector $A^T\mathbf{y}$ may be represented as a linear combination $y_1\mathbf{a}_1^T + y_2\mathbf{a}_2^T$, where $\mathbf{y} = (y_1, y_1)$. One can easily notice that the set V is a cone spanned by the vectors $\mathbf{a}_1^T$ and $\mathbf{a}_2^T$; it is dotted on the picture. Equation $A^T\mathbf{y} = \mathbf{d}$ has a solution when the vector $\mathbf{d}$ is inside the dotted part of the cone. This is the case in the picture on the right side. On the left side the vector $\mathbf{d}$ is outside of the dotted area, and System (2) does not have a solution.

Theorem 38 *(Kuhn–Tucker theorem) Let $\bar{\mathbf{x}}$ be a local solution of the problem (**). If the functions f and g_i, where $i \in I(\bar{\mathbf{x}})$, are differentiable at $\bar{\mathbf{x}}$ and $T(\bar{\mathbf{x}}) = T_{\text{lin}}(\bar{\mathbf{x}})$, then there exists $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m) \in [0,\infty)^m$ such that*

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum\limits_{i\in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1, 2, \ldots, m. \end{cases}$$

Remark. The second condition is called the complementary slackness condition. *Whenever the constraint $g_i \leqslant 0$ is slack (meaning $g_i < 0$), the constraint $\mu_i \geqslant 0$ must not be slack (meaning $\mu_i = 0$) and reverse.*

Remark. Often the system of Kuhn–Tucker equations is written with the sum over all indices $i = 1, \ldots, m$:

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum\limits_{i=1}^{m} \mu_i Dg_i(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1, 2, \ldots, m. \end{cases}$$

This notation is an abuse, as the functions g_i which describe inactive constraints need not be differentiable at $\bar{\mathbf{x}}$. On the other hand, the derivatives are multiplied by μ_i, equal to 0 for inactive constraints, which is a sort of justification for this notation.

Remark. The assumptions of the Kuhn–Tucker theorem are obviously satisfied if $\bar{\mathbf{x}} \in \operatorname{int} W$, i.e., $I(\bar{\mathbf{x}})$ is the empty set. Then we have $Df(\bar{\mathbf{x}}) = \mathbf{0}^T$, $\boldsymbol{\mu} = \mathbf{0}$.

Proof of the Kuhn–Tucker theorem. By Theorem 37 we have $D(\bar{\mathbf{x}}) \cap T(\bar{\mathbf{x}}) = \emptyset$. By assumption, $D(\bar{\mathbf{x}}) \cap T_{\text{lin}}(\bar{\mathbf{x}}) = \emptyset$, which means that the system

$$\begin{cases} Df(\bar{\mathbf{x}})\mathbf{z} < 0, \\ Dg_i(\bar{\mathbf{x}})\mathbf{z} \leqslant 0, \quad i \in I(\bar{\mathbf{x}}), \end{cases}$$

has no solution $\mathbf{z} \in \mathbb{R}^n$. Let $\mathbf{d} = -\big(Df(\bar{\mathbf{x}})\big)^T$, let $k = |I(\bar{\mathbf{x}})|$ and let A be the $k \times n$ matrix whose rows are gradients of the active constraints $Dg_i(\bar{\mathbf{x}})$. By the Farkas' lemma, there exists a vector $\mathbf{y} \in [0,\infty)^k$ such that $\mathbf{y}^TA = \mathbf{d}$, i.e.,

$$Df(\bar{\mathbf{x}}) + \mathbf{y}^T\mathbf{A} = \mathbf{0}^T.$$

Let $\boldsymbol{\mu} \in [0,\infty)^m$ be defined as follows: $(\mu_i)_{i\in I(\bar{\mathbf{x}})} = \mathbf{y}$ and $(\mu_i)_{i\notin I(\bar{\mathbf{x}})} = \mathbf{0}$. Then, the equality above is equivalent to

$$Df(\bar{\mathbf{x}}) + \sum_{i\in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) = \mathbf{0}^T.$$

By definition of $\boldsymbol{\mu}$ it is obvious that $\mu_i g_i(\bar{\mathbf{x}}) = 0$ for all $i = 1, \ldots, m$. □

The assumptions of the Kuhn–Tucker theorem are called the necessary conditions of the first order. Due to the importance of the vector $\boldsymbol{\mu}$ in what follows, we give it a name:

Definition 23 *The vector $\boldsymbol{\mu}$ which appears in the necessary first-order condition is called the vector of Lagrange multipliers.*

6. Regularity conditions and examples

Constraints qualifications

Three conditions defined below are sufficient for the equality $T(\bar{x}) = T_{\text{lin}}(\bar{x})$; we shall prove it.

Definition 24 *At a point $\bar{x} \in W \subset \mathbb{X}$*

- *the linear independence condition is satisfied if the functions g_i are continuous at $\bar{x}$ for $i \notin I(\bar{x})$ and the vectors $Dg_i(\bar{x})$ are linearly independent for $i \in I(\bar{x})$,*
- *the affine function condition is satisfied if the functions g_i are continuous at $\bar{x}$ for $i \notin I(\bar{x})$ and the functions g_i are affine for $i \in I(\bar{x})$,*
- *the Slater condition is satisfied if the functions g_i are continuous at $\bar{x}$ for $i \notin I(\bar{x})$, the functions g_i are pseudoconvex at $\bar{x}$ for $i \in I(\bar{x})$ (i.e., $Df(\bar{x})(y - \bar{x}) \geqslant 0 \Rightarrow f(y) \geqslant f(\bar{x})$) and there exists a point $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for $i \in I(\bar{x})$.*

Note that the point x considered in the Slater condition needs not satisfy the inactive constraints, i.e., it is not required that $x \in W$.

Theorem 39 *If the affine function condition is satisfied at $\bar{x} \in W$, then $T(\bar{x}) = T_{\text{lin}}(\bar{x})$.*

Proof. The set inclusion $T(\bar{x}) \subset T_{\text{lin}}(\bar{x})$ follows from Lemma 14. We need to prove the opposite inclusion. Let $\mathbf{d} \in T_{\text{lin}}(\bar{x})$. We are going to prove the existence of $\lambda^* > 0$ such that the entire line segment $\bar{x} + \lambda\mathbf{d}$, $\lambda \in [0, \lambda^*]$, is a subset of W.

We can notice that if $i \notin I(\bar{x})$, then $g_i(\bar{x}) < 0$. As these functions g_i are continuous, there exists $\lambda^* > 0$ such that $g_i(\bar{x} + \lambda\mathbf{d}) \leqslant 0$ for all $\lambda \in [0, \lambda^*]$. It remains to be proved that this inequality is satisfied also for the active constraints. Let $i \in I(\bar{x})$. By definition of $\mathbf{d}$, $Dg_i(x)\mathbf{d} \leqslant 0$. The active constraint g_i is an affine function equal to 0 at $\bar{x}$, i.e., it has the form $g_i(x) = \mathbf{a}_i^T(x - \bar{x})$ for some nonzero vector $\mathbf{a}_i \in \mathbb{R}^n$. There is $Dg_i(\bar{x})\mathbf{d} = \mathbf{a}_i^T\mathbf{d}$. Therefore, for any $\lambda \geqslant 0$ we have

$$0 \geqslant \lambda\mathbf{a}_i^T\mathbf{d} = g_i(\bar{x} + \lambda\mathbf{d});$$

hence,

$$\{\, \bar{x} + \lambda\mathbf{d} \colon \lambda \in [0, \lambda^*] \,\} \subset W.$$

It remains to construct appropriate sequences $(x_k)_k \subset W$, and $(\lambda_k)_k \subset (0, \infty)$. Let

$$x_k = \bar{x} + \frac{\lambda^*}{k}\mathbf{d}, \quad \lambda_k = \frac{k}{\lambda^*}.$$

Then, $x_k \in W$, $x_k \to \bar{x}$ and $\lambda_k(x_k - \bar{x}) = \mathbf{d}$ for all k, and thus $\lim_{k\to\infty} \lambda_k(x_k - \bar{x}) = \mathbf{d}$. □

The proofs of the other two regularity conditions refer to the set

$$T_{\text{int}}(\bar{x}) \overset{\text{def}}{=} \{\, \mathbf{d} \in \mathbb{R}^n \colon Dg_i(\bar{x})\mathbf{d} < 0 \text{ for all } i \in I(\bar{x}) \,\}.$$

Note that $T_{\text{int}}(\bar{x}) = \emptyset$ if $g_i(\bar{x}) = 0$ and $Dg_i(\bar{x}) = 0^T$ for some $i \in \{1, \ldots, m\}$.

Lemma 16 *If the functions g_i, $i \in I(\bar{x})$ are differentiable at $\bar{x} \in W$ and the other constraints g_i are continuous, then from $\mathbf{d} \in T_{\text{int}}(\bar{x})$ it follows that $\bar{x} + \lambda\mathbf{d} \in \operatorname{int} W$ for sufficiently small $\lambda > 0$.*

Proof. If $i \notin I(\bar{x})$, then $g_i(\bar{x}) < 0$. By the continuity of g_i, there is $g_i(\bar{x} + \lambda\mathbf{d}) < 0$ for all λ sufficiently small, say, $\lambda \in (0, \lambda^*]$. If $i \in I(\bar{x})$, then, by differentiability of g_i at $\bar{x}$, we have

$$\lim_{\lambda \searrow 0} \frac{g_i(\bar{x} + \lambda\mathbf{d}) - g_i(\bar{x})}{\lambda} = Dg_i(\bar{x})\mathbf{d} < 0,$$

because $\mathbf{d} \in T_{\text{int}}(\bar{x})$. This inequality holds also if $\lambda > 0$ is arbirarily small. Therefore,

$$\frac{g_i(\bar{x} + \lambda\mathbf{d}) - g_i(\bar{x})}{\lambda} < 0,$$

and $g_i(\bar{x} + \lambda\mathbf{d}) - g_i(\bar{x}) = g_i(\bar{x} + \lambda\mathbf{d}) < 0$. □

Lemma 17 *Let $\bar{x} \in W$, let the functions g_i, where $i \in I(\bar{x})$, be differentiable at $\bar{x}$ and let the functions g_i, where $i \notin I(\bar{x})$ be continuous at $\bar{x}$. Then,*

I) $T_{\text{int}} \subset T(\bar{x})$,

II) If $T_{\text{int}}(\bar{x}) \neq \emptyset$, then $\operatorname{cl}\big(T_{\text{int}}(\bar{x})\big) = T_{\text{lin}}(\bar{x})$.

Proof. (I) follows directly from Lemma 16. To prove (II) we notice that $T_{\text{int}}(\overline{x})$ is the interior of the set $T_{\text{lin}}(\overline{x})$. As $T_{\text{lin}}(\overline{x})$ is a polyhedral set, it is convex and closed. The claim follows from Lemma 3. □

Theorem 40 *If the Slater condition is satisfied at $\overline{x} \in W$, then $T(\overline{x}) = T_{\text{lin}}(\overline{x})$.*

Proof. First we show that $T_{\text{int}}(\overline{x}) \neq \emptyset$. Let $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for $i \in I(\overline{x})$. By pseudoconvexity of g_i at $\overline{x}$ we have $Dg_i(\overline{x})(x^* - \overline{x}) < 0$ for $i \in I(\overline{x})$, i.e., $x^* - \overline{x} \in T_{\text{int}}(\overline{x})$.

By Lemma 17, $\text{cl}\,\big(T_{\text{int}}(\overline{x})\big) = T_{\text{lin}}(\overline{x})$. We have also proved that $T_{\text{int}}(\overline{x}) \subset T(\overline{x}) \subset T_{\text{lin}}(\overline{x})$ and that the set $T(\overline{x})$ is closed. Hence, $T(\overline{x}) = T_{\text{lin}}(\overline{x})$. □

Lemma 18 *(P. Gordan, 1873) Let A be an $m \times n$ matrix. Then exactly one of the following two systems,*

$$\text{(1)}\begin{cases} Ax < 0, \\ x \in \mathbb{R}^n, \end{cases} \qquad \text{(2)}\begin{cases} A^T y = 0, \\ y \geqslant 0,\ y \neq 0, \\ y \in \mathbb{R}^m, \end{cases}$$

has a solution.

Proof. First we prove that the two systems cannot have solutions at the same time. Suppose that they have; let x and y satisfy (1) and (2), respectively. As y satisfies (2), we have $y^T A x = 0$. On the other hand, x satisfies (1), i.e., $(Ax)_i < 0$ for all $i = 1, \dots, m$. Knowing that $y_i \geqslant 0$ for all i and $y \neq 0$, we have $y^T A x < 0$. This inconsistency proves that at least one of the systems does not have a solution.

Now we prove that at least one of the systems has a solution, by proving that if (1) has no solution, then (2) has at least one. We define the convex sets

$$U \overset{\text{def}}{=} (-\infty, 0)^m, \quad V \overset{\text{def}}{=} \{\, z \in \mathbb{R}^m \colon z = Ax \text{ for some } x \in \mathbb{R}^n \,\}.$$

The set U is the interior of a cone and V is a linear subspace. If (1) has no solution, then the two sets are disjoint. By the weak separation theorem, there exists a nonzero vector $a \in \mathbb{R}^m$ such that

$$\sup_{z \in U} a^T z \leqslant \inf_{z \in V} a^T z.$$

We show that $a \geqslant 0$. On the contrary, suppose that one of the coordinates of a, say, $a_i < 0$. Consider the sequence of vectors $z_k = -\big(k e_i + \frac{1}{k}\sum_{j \neq i} e_j\big)$. Then, $z_k \in U$ and $\lim_{k\to\infty} a^T z_k = \infty$, which is an inconsistency, as $\sup_{z \in U} a^T z$ is finite (we have $0 \in V$; hence, $\inf_{z \in V} a^T z \leqslant 0$).

We proved that all coordinates of a are nonnegative. Hence, $\sup_{z \in U} a^T z = 0$. Let $z = A(-A^T a)$. Then, $z \in V$, i.e., $a^T z \geqslant 0$. Therefore,

$$0 \leqslant a^T z = -a^T A A^T a = -\|A^T a\|^2,$$

which implies that $\|A^T a\| = 0$, i.e., $A^T a = 0$. The vector $y = a$ is, therefore, a solution of (2). □

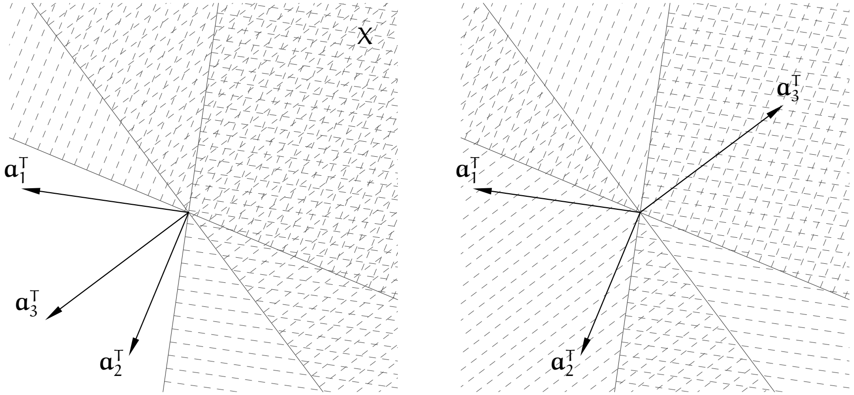

Figure 6: Gordan's lemma

Figure 6 illustrates the Gordan's lemma. The matrix A is 3×2, its transposed rows are the vectors in $\mathbb{R}^2$. Obviously, in this example they are linearly dependent. The sets of solutions of the inequalities $a_i x < 0$ are the hatched halfplanes (without the boundaries).

On the left side there exist vectors $v \in \mathbb{R}^2$ (e.g., a_3^T) such that for all i the number $a_i v$ is positive; hence, the coefficients of the linear combination $a_1^T y_1 + a_2^T y_2 + a_3^T y_3 = 0$, if not all equal to 0, must have both signs, i.e., at least one of them is positive and at least one is negative. Thus, the system (2) has no solution. The cone X being the intersection of all hatched halfplanes is nonempty; the vectors x satisfying (1) are elements of this cone.

On the right side we have vectors a_1, a_2, a_3, such that there exists a linear combination with positive coefficients, which is the zero vector. Therefore, there exists a vector $\mathbf{y} \in \mathbb{R}^3$ satisfying (2). The set X—the intersection of the appropriate halfplanes—is empty.

Theorem 41 *If the linear independence condition is satisfied at a point* $\overline{x} \in W$, *then* $T(\overline{x}) = T_{\text{lin}}(\overline{x})$.

Proof. Just as in the case of the Slater condition, it suffices to prove that $T_{\text{int}}(\overline{x}) \neq \emptyset$. Let A be the matrix whose rows are gradients of active constraints. By their linear independence, a nonzero vector $\boldsymbol{\mu} \in \mathbb{R}^{|I(\overline{x})|}$ such that $A^T\boldsymbol{\mu} = \mathbf{0}$ does not exist. In other words, System (2) of Lemma 18 has no solution. Hence, System (1) has a solution, i.e., there exists $\mathbf{d} \in \mathbb{R}^n$ such that $A\mathbf{d} < \mathbf{0}$, which means that

$$Dg_i(\overline{x})\mathbf{d} < 0 \quad \text{for all } i \in I(\overline{x}).$$

Hence, $\mathbf{d} \in T_{\text{int}}(\overline{x})$. □

Examples

Example. Consider an optimization problem in the set

$$W = \{\, \mathbf{x} \in \mathbb{R}^2 \colon x_1^2 + x_2^2 \leqslant 1,\ x_1 + 2x_2 \leqslant 1,\ x_1 - 3x_2 \leqslant 1 \,\}.$$

There is $\mathbb{X} = \mathbb{R}^2$. The constraints are described by the functions

$$g_1(x_1, x_2) = x_1^2 + x_2^2 - 1, \quad g_2(x_1, x_2) = x_1 + 2x_2 - 1, \quad g_3(x_1, x_2) = x_1 - 3x_2 - 1.$$

At the point $\overline{x} = (1, 0)$ all three constraints are active and

$$Dg_1(\overline{x}) = [2, 0], \quad Dg_2(\overline{x}) = [1, 2], \quad Dg_3(\overline{x}) = [1, -3].$$

The linear independence condition is not satisfied at $\overline{x}$. Also, not all active constraints are described by affine functions. But all these functions are convex and differentiable, i.e., they are pseudoconvex. At $\mathbf{x} = (0, 0)$ we have $g_i(\mathbf{x}) = -1 < 0$ for $i = 1, 2, 3$. Thus, the Slater condition is satisfied.

Example (Kuhn, Tucker, 1951). Consider the optimization problem

$$\begin{cases} x_1 \to \min, \\ x_2 \leqslant x_1^3, \\ x_2 \geqslant 0. \end{cases}$$

The constraints are described by the functions

$$g_1(x_1, x_2) = -x_1^3 + x_2, \quad g_2(x_1, x_2) = -x_2.$$

At each feasible point except $(0, 0)$ the linear independence condition is satisfied. However, the solution of this problem is $\overline{x} = (0, 0)$, but

$$T(\overline{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_1 \geqslant 0,\ d_2 = 0 \,\} \neq T_{\text{lin}}(\overline{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_2 = 0 \,\},$$

as $Dg_1(\overline{x}) = [0, 1]$, $Dg_2(\overline{x}) = [0, -1]$, which implies $Dg_1(\overline{x})\mathbf{d} \leqslant 0 \Rightarrow d_2 \leqslant 0$ and $Dg_2(\overline{x})\mathbf{d} \leqslant 0 \Rightarrow d_2 \geqslant 0$. On the other hand,

$$D(\overline{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_1 < 0 \,\},$$

because $Df(\overline{x}) = [1, 0]$; hence, $Df(\overline{x})\mathbf{d} < 0 \Rightarrow d_1 < 0$.

For the optimization problem

$$\begin{cases} x_2 \to \min, \\ x_2 \leqslant x_1^3, \\ x_2 \geqslant 0, \end{cases}$$

we still have $T(\overline{x}) \neq T_{\text{lin}}(\overline{x})$. But for this new problem $D(\overline{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_2 < 0 \,\}$. It follows that at $\overline{x} = (0, 0)$

$$T(\overline{x}) \cap D(\overline{x}) = T_{\text{lin}}(\overline{x}) \cap D(\overline{x}) = \emptyset.$$

The above condition is sufficient for the claim of the Kuhn–Tucker theorem to hold.

Example. Let A be a symmetric $n \times n$ matrix. Consider the optimization problem

$$\begin{cases} \mathbf{x}^T A\mathbf{x} \to \max, \\ \|\mathbf{x}\| \leqslant 1, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

First, we rewrite the problem in the canonical form; note that the constraint is equivalent to $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} \leqslant 1$:

$$\begin{cases} -\mathbf{x}^T A\mathbf{x} \to \max, \\ \mathbf{x}^T\mathbf{x} - 1 \leqslant 0, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

By W we denote the feasible set. At each point the linear independence condition and the Slater condition are satisfied. Therefore, the solution of the problem satisfies the necessary first order conditions (i.e., the Kuhn–Tucker conditions). We are looking for all points “under suspicion”.

The Kuhn–Tucker conditions have the form

$$\begin{cases} -2x^T A + 2\mu x^T = 0^T, \\ \mu(x^T x - 1) = 0, \\ \mu \geqslant 0, \ x \in \mathbb{R}^n. \end{cases}$$

First we check the case $x^T x - 1 < 0$. Then (due to the second equation), $\mu = 0$ and the first equation is equivalent to $x^T A = 0^T$ or $Ax = 0$. This equation is satisfied by all vectors $x \in \ker A$ such that $\|x\| < 1$. In particular, it has at least one solution, $x = 0$.

Then we check $x^T x - 1 = 0$; from the second equation nothing may be said about μ. However, the first equation takes the form $\mu x^T = x^T A$, or $Ax = \mu x$. The solutions are the unit eigenvectors corresponding to nonnegative eigenvalues μ of the matrix A. The set of solutions may be empty if all the eigenvalues are negative. To conclude, the set of points “under suspicion”, i.e., satisfying the necessary first order condition, is

$$\begin{aligned} \{\, x \in \mathbb{R}^n \colon & (x \in \ker A \text{ and } \|x\| < 1) \text{ or} \\ & (\|x\| = 1 \text{ and } x \text{ is an eigenvector of } A \\ & \ \text{associated with a nonnegative eigenvalue})\,\}. \end{aligned}$$

At this point we do not have any technique of finding solutions, except for using the common sense. If $x \in \ker A$, then the target function $x^T A x$ is equal to 0. For any eigenvector x the value of the target function is the associated eigenvalue. If the greatest eigenvalue is positive, then each of its unit eigenvectors is a global solution. If the matrix A has no positive eigenvalue, then any vector from $\ker A$ whose norm is not greater than 1 is a solution.

Example. Consider the problem

$$\begin{cases} x_1 + x_2 \to \min, \\ x_2 \geqslant x_1^2, \\ x_2 \leqslant 0. \end{cases}$$

It is easy to notice that the solution is $\bar{x} = (0, 0)$, which is the only feasible point. On the other hand, $Df = [1, 1]$, $Dg_1 = [2x_1, -1]$ and $Dg_2 = [0, 1]$. Therefore, for any numbers μ_1 and μ_2, there is

$$Df(\bar{x}) + \mu_1 Dg_1(\bar{x}) + \mu_2 Dg_2(\bar{x}) \neq 0,$$

i.e., the Kuhn–Tucker conditions are not satisfied. But $T(\bar{x}) = \{0\} \neq T_{\text{lin}}(\bar{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_2 = 0 \,\}$. Also, $\emptyset = T(\bar{x}) \cap D(\bar{x}) \neq T_{\text{lin}}(\bar{x}) \cap D(\bar{x})$, as $D(\bar{x}) = \{\, (d_1, d_2) \in \mathbb{R}^2 \colon d_1 + d_2 < 0 \,\}$. As we can see, the assumptions of the Kuhn–Tucker theorem are not satisfied at $\bar{x}$.

7. Quasi-convex functions and sufficient conditions

Quasi-convexity

Below we extend the class of functions whose maxima are located at the extremal points of their domains.

Definition 25 *Let $W \subset \mathbb{R}^n$ be a convex set and let $f\colon W \to \mathbb{R}$. The function f is quasi-convex if for all $\mathbf{x}, \mathbf{y} \in W$ and $\lambda \in [0,1]$ there is*

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \max\{f(\mathbf{x}), f(\mathbf{y})\}.$$

The function f is quasi-concave if $-f$ is quasi-convex, i.e., if

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \geqslant \min\{f(\mathbf{x}), f(\mathbf{y})\} \quad \textit{for all } \mathbf{x}, \mathbf{y} \in W,\ \lambda \in [0,1].$$

The function f is quasi-linear if it is both quasi-convex and quasi-concave.

Theorem 42 *Let $W \subset \mathbb{R}^n$ be a convex set and let $f\colon W \to \mathbb{R}$. The function f is quasi-linear if and only if each sublevel set of this function is convex.*

Proof. Suppose that the function f is quasi-convex and let $\alpha \in \mathbb{R}$ be fixed. Let $\mathbf{x}, \mathbf{y} \in W_\alpha(f)$. Then, $f(\mathbf{x}) \leqslant \alpha$ and $f(\mathbf{y}) \leqslant \alpha$. For any $\lambda \in (0,1)$ we obtain

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \max\{f(\mathbf{x}), f(\mathbf{y})\} \leqslant \alpha.$$

Hence, $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in W_\alpha(f)$, i.e., $W_\alpha(f)$ is a convex set.

Now suppose that $W_\alpha(f)$ is convex for all $\alpha \in \mathbb{R}$. We fix $\mathbf{x}, \mathbf{y} \in W$ and $\lambda \in (0,1)$. By assumption, the set $W_\alpha(f)$, where $\alpha = \max\{f(\mathbf{x}), f(\mathbf{y})\}$, is convex. It follows that $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in W_\alpha(f)$; hence,

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \alpha = \max\{f(\mathbf{x}), f(\mathbf{y})\}$$

and the proof is complete. □

Corollary 7 *Any convex function is a quasi-convex function.*

Example. The function $f(x) = -e^x$ is quasi-convex even if it is strictly concave. For $\alpha \geqslant 0$ the set $W_\alpha(f) = \mathbb{R}$ and if $\alpha < 0$, then $W_\alpha(f) = [\ln(-\alpha), \infty)$. All these sets are convex; hence, by Theorem 42, the function f is quasi-convex. It is also quasi-concave, i.e., it is quasi-linear.

Example. The function $f(x) = x^2$ is quasi-convex (as it is convex), but it is not quasi-concave. This may be checked by looking at the sublevel sets of $-f$: for $\alpha < 0$ we have $W_\alpha(f) = (-\infty, -\sqrt{-\alpha}] \cup [\sqrt{-\alpha}, \infty)$, which is not a convex set.

Lemma 19 *If a set $W \subset \mathbb{R}^n$ is convex, then a function $f\colon W \to \mathbb{R}$ is quasi-linear if and only if its restriction to any interval is a monotone function.*

Proof is left as an exercise.

Example. The function $f(x) = -e^{-x^2}$ has the following sublevel sets:

$$W_\alpha(f) = \begin{cases} \emptyset & \text{if } \alpha \leqslant -1, \\ \Big[-\sqrt{-\ln(-\alpha)}, \sqrt{-\ln(-\alpha)}\Big] & \text{if } \alpha \in (-1, 0], \\ \mathbb{R} & \text{if } \alpha > 0. \end{cases}$$

All these sets are convex; hence, the function f is quasi-convex.

Example. The function $f\colon [0,\infty)^2 \to \mathbb{R}$ given by the formula $f(x_1, x_2) = -x_1x_2$ is quasi-convex. Its sublevel sets for $\alpha \geqslant 0$ are trivial, $W_\alpha(f) = [0,\infty)^2$, and if $\alpha < 0$, then the sublevel sets have a boundary being one branch of a hyperbola. This function is neither convex nor concave, as its Hessian has the eigenvalues -1 and 1.

Example. Let $\mathbf{a}, \mathbf{c} \in \mathbb{R}^n$ and $b, d \in \mathbb{R}$. Let $D = \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{c}^T\mathbf{x} + d > 0\}$. Then, the rational function $f\colon D \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \frac{\mathbf{a}^T\mathbf{x} + b}{\mathbf{c}^T\mathbf{x} + d}$$

is quasi-linear. The proof is left as an exercise.

Theorem 43 *Let $W \subset \mathbb{R}^n$ be a convex set and let $f\colon W \to \mathbb{R}$.*

I) If the function f is quasi-convex and differentiable at a point $\mathbf{y} \in W$, then

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) \Rightarrow Df(\mathbf{y})(\mathbf{x} - \mathbf{y}) \leqslant 0 \quad \textit{for all } \mathbf{x} \in W.$$

II) Suppose that the function f is differentiable in the entire set W. Then f is quasi-convex if and only if

$$f(x) \leqslant f(y) \Rightarrow Df(y)(x-y) \leqslant 0 \text{ for all } x, y \in W.$$

Remark. The implication $f(x) \leqslant f(y) \Rightarrow Df(y)(x-y) \leqslant 0$ is equivalent to

$$Df(y)(x-y) > 0 \Rightarrow f(x) > f(y).$$

If the function f is quasi-linear and $f(x) = f(y)$, then $Df(y)(x-y) = 0$.

Proof. (I) We fix $x, y \in W$ such that $f(x) \leqslant f(y)$. For any $\lambda \in (0,1)$ we have

$$f\big(y + \lambda(x-y)\big) = f\big(\lambda x + (1-\lambda)y\big) \leqslant \max\{f(x), f(y)\} = f(y).$$

Therefore,

$$\frac{f\big(y + \lambda(x-y)\big) - f(y)}{\lambda} \leqslant 0.$$

By definition of the directional derivative, we obtain

$$Df(y)(x-y) = \lim_{\lambda \searrow 0} \frac{f\big(y + \lambda(x-y)\big) - f(y)}{\lambda} \leqslant 0$$

The proof of (II) is left as a (not very easy) exercise. □

Theorem 44 *If a function $f\colon W \to \mathbb{R}$ defined in a convex set $W \subset \mathbb{R}^n$ is pseudoconvex, then it is quasi-convex.*

Proof. By assuming that the function f is not quasi-convex we shall get an inconsistency with its pseudoconvexity. Let $x, y \in W$ and $\lambda \in (0,1)$ such that

$$f\big(\lambda x + (1-\lambda)y\big) > \max\{f(x), f(y)\}.$$

Let $z = \lambda x + (1-\lambda)y$. By pseudoconvexity, we have

$$f(x) < f(z) \Rightarrow Df(z)(x-z) < 0,$$
$$f(y) < f(z) \Rightarrow Df(z)(y-z) < 0.$$

The vectors $x - z$ and $y - z$ have the same direction, but opposite orientations. The directional derivatives of f at z cannot be negative in both directions. □

Theorem 45 *Let a function $f\colon W \to \mathbb{R}$ defined in a convex open set $W \subset \mathbb{R}^n$ be quasi-convex and continuous. If the function f is differentiable at a point $\bar{x} \in W$ and $Df(\bar{x}) \neq 0^T$, then f is pseudoconvex at $\bar{x}$.*

Proof. We need to prove that for all $y \in W$ the condition $Df(\bar{x})(y - \bar{x}) \geqslant 0$ implies that $f(y) \geqslant f(\bar{x})$. Let A denote the affine space perpendicular to $Df(\bar{x})$ passing through $\bar{x}$. As $Df(\bar{x}) \neq 0^T$, the dimension of A is $n-1$, i.e., A is a hyperplane in $\mathbb{R}^n$.

We notice that if $y \in W \setminus A$ and $Df(\bar{x})(y - \bar{x}) \geqslant 0$, then the directional derivative is (strictly) positive: $Df(\bar{x})(y - x) > 0$. It follows (see the remark above) that $f(y) > f(\bar{x})$, which is what was to be proved. Now we fix $y \in W \cap A$. As W is open and A is a hyperplane, there exists a sequence of points $(y_k)_k \subset W \setminus A$ converging to y and such that $Df(\bar{x})(y_k - \bar{x}) > 0$, i.e., $f(y_k) > f(\bar{x})$. By continuity of f we obtain $f(y) \geqslant f(\bar{x})$. □

Finding maxima of quasi-convex functions

Theorem 46 *Let $f\colon W \to \mathbb{R}$ be a quasi-convex and continuous function defined in a convex and compact set $W \subset \mathbb{R}^n$. Then at least one of the global solutions of the problem*

$$\begin{cases} f(x) \to \max, \\ x \in W \end{cases}$$

is an extremal point of the set W.

Proof. A convex function reaches its extremal values in a compact set. Therefore, a solution $\bar{x}$ exists. By Theorem 35, the point $\bar{x}$ is either an extremal point or a convex combination of a finite set of extremal points of W, i.e.,

$$\bar{x} = a_1 x_1 + \cdots + a_m x_m,$$

where $a_1, \ldots, a_m > 0$, $a_1 + \cdots + a_m = 1$. By quasi-convexity of f,

$$f(\bar{x}) \leqslant \max\{f(x_1), \ldots, f(x_m)\}.$$

As the function f at $\bar{x}$ takes its maximal value in W, the equality $f(\bar{x}) = f(x_i)$ has to hold for some $i \in \{1, \ldots, m\}$. □

Sufficient conditions

Below we consider the optimization problem with both non-equality and equality constraints:

$$\begin{cases} f(x) \to \min, \\ g_i(x) \leqslant 0, \; i = 1, \dots, m, \\ h_j(x) = 0, \; j = 1, \dots, l, \\ x \in \mathbb{X}, \end{cases} \tag{*}$$

where $\mathbb{X} \subset \mathbb{R}^n$ is an open set and $f, g_1, \dots, g_m, h_1, \dots, h_l \colon \mathbb{X} \to \mathbb{R}$. Thus, the feasible set is

$$W = \{x \in \mathbb{X} \colon g_1(x) \leqslant 0, \; \dots, \; g_m(x) \leqslant 0, \; h_1(x) = \dots = h_l(x) = 0\}.$$

The functions g_i are called non-equality constraints, the functions h_j are called equality constraints and the problem above is called an optimization problem with mixed constraints.

The theorem below describes sufficient conditions for a point satisfying the first order necessary conditions to be a global solution of (*).

Theorem 47 *Let $\bar{x} \in W$. Assume that*

- *the functions g_i, where $i \notin I(\bar{x})$, are continuous at $\bar{x}$, the functions g_i, where $i \in I(\bar{x})$, are differentiable at $\bar{x}$ and quasi-convex,*
- *the functions h_j, where $j \in \{1, \dots, l\}$, are quasi-linear and differentiable at $\bar{x}$,*
- *the function f is pseudoconvex at $\bar{x}$.*

If there exists $\mu \in [0, \infty)^m$ and $\lambda \in \mathbb{R}^l$ which satisfy the following first order condition:

$$\begin{cases} Df(\bar{x}) + \sum_{i \in I(\bar{x})} \mu_i Dg_i(\bar{x}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{x}) = 0^T, \\ \mu_i g_i(\bar{x}) = 0, \; i = 1, \dots, m, \end{cases}$$

then the point $\bar{x}$ is a global solution.

Proof. Let $x \in W$. We multiply the sides of the first equation by $x - \bar{x}$:

$$Df(\bar{x})(x - \bar{x}) + \sum_{i \in I(\bar{x})} \mu_i Dg_i(\bar{x})(x - \bar{x}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{x})(x - \bar{x}) = 0$$

By Theorem 43, $Dh_j(\bar{x})(x - \bar{x}) = 0$ for all j, as $h_j(x) = h_j(\bar{x}) = 0$. By the same theorem, also $Dg_i(\bar{x})(x - \bar{x}) \leqslant 0$ for $i \in I(\bar{x})$, as $0 = g_i(\bar{x}) \geqslant g_i(x)$. From the above we conclude that

$$Df(\bar{x})(x - \bar{x}) \geqslant 0.$$

By the definition of a pseudoconvex function, $f(x) \geqslant f(\bar{x})$. As the choice of $x \in W$ was arbitrary, $\bar{x}$ is a global solution. □

Remark. If the assumptions of Theorem 47 are satisfied locally, in a neighbourhood of $\bar{x}$, then $\bar{x}$ is a local solution.

Remark. By Theorem 45, instead of the pseudoconvexity of f at $\bar{x}$, we can assume its continuity in $\mathbb{X}$, quasi-convexity and the condition $Df(\bar{x}) \neq 0^T$. It is one of necessary conditions, given by Arrow and Enthoven in 1961.

8. A necessary condition for mixed constraints

Below we are going to derive a necessary first-order condition for the optimization problem given as follows:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0,\ i = 1, \ldots, m \\ h_j(\mathbf{x}) = 0,\ j = 1, \ldots, l, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{*}$$

where $\mathbb{X} \subset \mathbb{R}^n$ is an open set and $f, g_1, \ldots, g_m, h_1, \ldots, h_l \colon \mathbb{X} \to \mathbb{R}$. The set of feasible points is thus

$$W = \{\mathbf{x} \in \mathbb{X}\colon g_1(\mathbf{x}) \leqslant 0, \ldots, g_m(\mathbf{x}) \leqslant 0, h_1(\mathbf{x}) = \cdots = h_l(\mathbf{x}) = 0\}.$$

Example. Consider the following problem:

$$\begin{cases} f(\mathbf{x}) \to \min \\ \mathbf{a}^T\mathbf{x} + b = 0, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

for some $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The equality constraint may be replaced by two inequality constraints:

$$\begin{cases} f(\mathbf{x}) \to \min \\ \mathbf{a}^T\mathbf{x} + b \leqslant 0, \\ -\mathbf{a}^T\mathbf{x} - b \leqslant 0, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

The constraints are affine functions; hence, the affine function condition is satisfied at each point. If $\overline{\mathbf{x}}$ is a local solution, then there exists a vector of Lagrange multipliers $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ such that the Kuhn–Tucker conditions are satisfied:

$$\begin{cases} Df(\overline{\mathbf{x}}) + \mu_1\mathbf{a}^T + \mu_2(-\mathbf{a}^T) = \mathbf{0}^T, \\ \mu_1(\mathbf{a}^T\mathbf{x} + b) = 0, \\ \mu_2(-\mathbf{a}^T\mathbf{x} - b) = 0, \\ \mu_1, \mu_2 \geqslant 0. \end{cases}$$

The point $\overline{\mathbf{x}}$, being a solution, is feasible, i.e., it satisfies the constraints: $\mathbf{a}^T\overline{\mathbf{x}} + b = 0$. Therefore, the second and third equalities are trivially satisfied. The conditions above are thus equivalent to

$$\begin{cases} Df(\overline{\mathbf{x}}) + (\mu_1 - \mu_2)\mathbf{a}^T = \mathbf{0}^T, \\ \mu_1, \mu_2 \geqslant 0. \end{cases}$$

Denote $\lambda = \mu_1 - \mu_2$. As these two numbers must only be nonnegative, $\lambda \in \mathbb{R}$. As a result we obtain

$$Df(\overline{\mathbf{x}}) + \lambda\mathbf{a}^T = \mathbf{0}^T, \quad \lambda \in \mathbb{R},$$

which is the Kuhn–Tucker condition for equality constraints.

The example above suggests that the theory for inequality constraints, developed earlier, is also suitable for dealing with equality constraints. Unfortunately, it is not the case. Affine constraints are very special. If any constraint is not affine and we split it to a pair of inequality constraints, as above, then neither the linear independence condition nor the Slater condition is satisfied at any point of the set W.

Necessary first-order condition

The theory described below is a direct extension of what we have done for the optimization problems with inequality contraints. We begin with extending the definition of the set $T_{\text{lin}}(\overline{\mathbf{x}})$:

Definition 26 *Let $\overline{\mathbf{x}} \in W$, let g_i, where $i \in I(\overline{\mathbf{x}})$ be functions differentiable at $\overline{\mathbf{x}}$ and let the functions $h_1, \ldots, h_l$ be differentiable at $\overline{\mathbf{x}}$. The set*

$$T_{\text{lin}}(\overline{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^n\colon Dg_i(\overline{\mathbf{x}})\mathbf{d} \leqslant 0 \textit{ for } i \in I(\overline{\mathbf{x}}), Dh_j(\overline{\mathbf{x}})\mathbf{d} = 0 \textit{ for } i = 1, \ldots, l\}$$

is called the cone of tangents for active (binding) constraints.

Just as in the case of inequality constraints, the cone of tangents for binding constraints is a polyhedral set, i.e., it is a convex and closed set. However, with at least one equality constraint the interior of this set is empty.

The necessary condition for the existence of a local solution of the optimization problem with mixed constraints given below assumes the equality of the cone of tangents to the set W and the cone of tangents for binding constraints. Later we show generalised regularity conditions which imply this equality.

Theorem 48 *(Kuhn–Tucker theorem) Let $\overline{x}$ be a local solution of the problem (*). If the functions $f, g_i, i \in I(\overline{x}), h_j, j = 1, \dots, l$ are differentiable at $\overline{x}$ and $T(\overline{x}) = T_{\text{lin}}(\overline{x})$, then there exist vectors $\boldsymbol{\mu} \in [0,\infty)^m$ and $\boldsymbol{\lambda} \in \mathbb{R}^l$ such that*

$$\begin{cases} Df(\overline{x}) + \sum\limits_{i \in I(\overline{x})} \mu_i Dg_i(\overline{x}) + \sum\limits_{j=1}^{l} \lambda_j h_j(\overline{x}) = 0^T, \\ \mu_i g_i(\overline{x}) = 0, \ i = 1, \dots, m. \end{cases}$$

Proof. By Theorem 37, $D(\overline{x}) \cap T(\overline{x}) = \emptyset$. Then, by assumption, $D(\overline{x}) \cap T_{\text{lin}}(\overline{x}) = \emptyset$, which means that there is no solution z of the system

$$\begin{cases} Df(\overline{x})z < 0, \\ Dg_i(\overline{x})z \leqslant 0, \ i \in I(\overline{x}), \\ Dh_j(\overline{x})z \leqslant 0, \ j = 1, \dots, l, \\ -Dh_j(\overline{x})z \leqslant 0, \ j = 1, \dots, l. \end{cases}$$

We use the Farkas' lemma with $\mathbf{d} = -Df(\overline{x})$ and with the following matrix A:

$$A = \begin{bmatrix} Dh_j(\overline{x}), & j = 1, \dots, l \\ -Dh_j(\overline{x}), & j = 1, \dots, l \\ Dg_i(\overline{x}), & i \in I(\overline{x}) \end{bmatrix}.$$

Hence, there exists $\mathbf{y} \in [0,\infty)^{|I(\overline{x})|+2l}$ such that $\mathbf{y}^T A = -Df(\overline{x})$, or

$$Df(\overline{x}) + \mathbf{y}^T A = 0^T. \qquad (\diamond)$$

Let $\lambda_j = y_j - y_{l+j}$ for $j = 1, \dots, l$. Let the coordinates of the vector $\boldsymbol{\mu} \in \mathbb{R}^m$ corresponding to the active constraints be equal to the last $|I(\overline{x})|$ coordinates of $\mathbf{y}$ and let the other coordinates of $\boldsymbol{\mu}$ be equal to 0. Then, $(\diamond)$ is equivalent to

$$Df(\overline{x}) + \sum_{i \in I(\overline{x})} \mu_i Dg_i(\overline{x}) + \sum_{j=1}^{l} \lambda_j h_j(\overline{x}) = 0^T.$$

By definition of $\boldsymbol{\mu}$ it is obvious that $\mu_i g_i(\overline{x}) = 0$ for all i. □

Regularity constraints

Below three sufficient conditions for the equality $T(\overline{x}) = T_{\text{lin}}(\overline{x})$, called the regularity conditions, are defined.

Definition 27 *At a point $\overline{x} \in W$*

- *the linear independence condition is satisfied if the functions g_i, where $i \notin I(\overline{x})$, are continuous at $\overline{x}$ and all the other inequality and equality constraints are continuously differentiable in a neighbourhood of $\overline{x}$ and the vectors $Dg_i(\overline{x})$ for $i \in I(\overline{x})$ and Dh_j for $j = 1, \dots, l$ are linearly independent,*
- *the affine function condition is satisfied if the functions $g_i(\overline{x})$ for $i \in I(\overline{x})$ and h_j for $j = 1, \dots, l$ are affine, while g_i, for $i \notin I(\overline{x})$, are continuous at $\overline{x}$,*
- *the Slater condition is satisfied if*
  —the functions g_i, $i \in I(\overline{x})$ are pseudoconvex at $\overline{x}$,
  —the functions g_i, $i \notin I(\overline{x})$ are continuous at $\overline{x}$,
  —the functions h_j, $j = 1, \dots, l$, are affine,
  —there exists $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for $i \in I(\overline{x})$ and $h_j(x^*) = 0$ for $l = 1, \dots, l$.

Theorem 49 *If the affine function condition is satisfied at a point $\overline{x} \in W$, then $T(\overline{x}) = T_{\text{lin}}(\overline{x})$.*

Proof. Just as in the last example, we change the affine equality constraints into pairs of affine inequality constraints. The claim follows from Theorem 39. □

Theorem 50 *If the Slater condition is satisfied at a point $\overline{x} \in W$, then $T(\overline{x}) = T_{\text{lin}}(\overline{x})$.*

Proof. The functions h_j, for $j = 1, \dots, l$, have the form

$$h_j(\mathbf{y}) = \mathbf{a}_j^T \mathbf{y} + b_j, \quad \mathbf{a}_j \in \mathbb{R}^n, \ b_j \in \mathbb{R}.$$

We introduce the generalisation of set $T_{\text{int}}(\bar{\mathbf{x}})$ for mixed constraints:

$$T_{\text{int}}(\bar{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^n \colon Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0 \text{ for } i \in I(\bar{\mathbf{x}}), Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0 \text{ for } j = 1, \dots, l \}.$$

(1) We prove that $T_{\text{int}}(\bar{\mathbf{x}}) \neq \emptyset$. We take $\mathbf{x}$ from the Slater condition. By pseudoconvexity of g_i, we have

$$Dg_i(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) < 0, \quad i \in I(\bar{\mathbf{x}}).$$

For all j we also have

$$\mathbf{a}_j^T(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{a}_j^T\mathbf{x} + b_j - \mathbf{a}_j^T\bar{\mathbf{x}} - b_j = h_j(\mathbf{x}) - h_j(\bar{\mathbf{x}}) = 0.$$

Hence, $(\mathbf{x} - \bar{\mathbf{x}}) \in T_{\text{int}}(\bar{\mathbf{x}})$.

(2) We prove that $T_{\text{int}}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}})$. We choose any $\mathbf{d} \in T_{\text{int}}(\bar{\mathbf{x}})$. It suffices to show that there exists a line segment contained in W whose one end is $\bar{\mathbf{x}}$ and whose direction is $\mathbf{d}$. Let $\mathbf{y}(\lambda) = \bar{\mathbf{x}} + \lambda\mathbf{d}$. As the inactive constraints are continuous functions, there exists $\varepsilon > 0$ such that $g_i\big(\mathbf{y}(\lambda)\big) \leqslant 0$ for $\lambda \in [0, \varepsilon]$ and $i \notin I(\bar{\mathbf{x}})$. From $\mathbf{d} \in T_{\text{int}}(\bar{\mathbf{x}})$ it follows also that $h_j\big(\mathbf{y}(\lambda)\big) = 0$ for $j = 1, \dots, l$ (because h_j are affine functions). And as the functions g_i, where $i \in I(\bar{\mathbf{x}})$, are differentiable at $\bar{\mathbf{x}}$ and $\mathbf{d} \in T_{\text{int}}(\bar{\mathbf{x}})$, it follows that

$$\lim_{\lambda \searrow 0} \frac{g_i\big(\mathbf{y}(\lambda)\big) - g_i(\bar{\mathbf{x}})}{\lambda} = Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0.$$

Hence, $g_i\big(\mathbf{y}(\lambda)\big) - g_i(\bar{\mathbf{x}}) < 0$ for λ small enough.

(3) We prove that $\operatorname{cl} T_{\text{int}}(\bar{\mathbf{x}}) = T_{\text{lin}}(\bar{\mathbf{x}})$. The sets $T_{\text{int}}(\bar{\mathbf{x}})$ and $T_{\text{lin}}(\bar{\mathbf{x}})$ are contained in the affine subspace H determined by the affine equality constraints. Therefore, there exists an affine mapping P, such that the image of the subspace H is $\mathbb{R}^{n'}$, where $n' = \dim H$ (if the vectors $\mathbf{a}_j$ are linearly independent, then $n' = n - l$). The mapping P restricted to H is a one-to-one function. The topologies of H and $\mathbb{R}^{n'}$ are, therefore, identical, which means that it suffices to prove this claim for the images $T'_{\text{int}}(\bar{\mathbf{x}})$ and $T'_{\text{lin}}(\bar{\mathbf{x}})$ of the sets $T_{\text{int}}(\bar{\mathbf{x}})$ and $T_{\text{lin}}(\bar{\mathbf{x}})$ under the mapping P. We notice that $T'_{\text{int}}(\bar{\mathbf{x}})$ is open in $\mathbb{R}^{n'}$ and nonempty. It is also the interior of $T'_{\text{lin}}(\bar{\mathbf{x}})$. By Lemma 17 we have $T'_{\text{int}}(\bar{\mathbf{x}}) = T'_{\text{lin}}(\bar{\mathbf{x}})$.

(4) The proof that $T(\bar{\mathbf{x}}) \subset T_{\text{lin}}(\bar{\mathbf{x}})$ is identical to the proof of Lemma 14.

(5) It suffices to recall that $T(\bar{\mathbf{x}})$ is a closed set. Therefore,

$$\operatorname{cl} T_{\text{int}}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}}) \subset T_{\text{lin}}(\bar{\mathbf{x}}) = \operatorname{cl} T_{\text{int}}(\bar{\mathbf{x}}). \quad \square$$

Before tackling the third regularity condition, we recall the implicit function theorem, which we need to describe the cone of tangents to a surface determined by equality constraints.

Theorem 51 *(implicit function theorem) Let $\mathbf{f} \colon \mathbb{X} \to \mathbb{R}^n$, where $\mathbb{X} \subset \mathbb{R}^{n+m}$ is an open set, be a function of class C^k. Assume that $\mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \mathbf{0}$, where $\bar{\mathbf{x}} \in \mathbb{R}^n$, $\bar{\mathbf{y}} \in \mathbb{R}^m$, $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathbb{X}$. By A_x we denote the matrix of partial derivatives of $\mathbf{f}$ with respect to the first n variables at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$: $A_x \in \mathbb{R}^{n \times n}$, $(A_x)_{ij} = \frac{\partial f_i}{\partial u_j}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.*

If the matrix A_x is nonsingular, then there exists an open set $W \subset \mathbb{R}^m$, whose element is $\bar{\mathbf{y}}$, and a function $\mathbf{g} \colon W \to \mathbb{R}^n$ of class C^k, such that $\mathbf{g}(\bar{\mathbf{y}}) = \bar{\mathbf{x}}$ and $\big(\mathbf{g}(\mathbf{y}), \mathbf{y}\big) \in \mathbb{X}$ and $\mathbf{f}\big(\mathbf{g}(\mathbf{y}), \mathbf{y}\big) = \mathbf{0}$ for all $\mathbf{y} \in W$. Moreover, $D\mathbf{g}(\bar{\mathbf{y}}) = -(A_x)^{-1}A_y$, where the matrix $A_y \in \mathbb{R}^{n \times m}$ consists of the derivatives of $\mathbf{f}$ at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ with respect to the last m variables: $(A_y)_{ij} = \frac{\partial f_i}{\partial u_{n+j}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.

We consider the surface S described by a system of m^* equations:

$$S = \{\mathbf{x} \in \mathbb{X} \colon c_i(\mathbf{x}) = 0,\ i = 1, \dots, m^*\},$$

where $\mathbb{X}$ is an open set. By $T^S(\bar{\mathbf{x}})$ we denote the cone of tangents to S at a point $\bar{\mathbf{x}} \in S$.

Theorem 52 *Let $k \geqslant 1$. Assume that the functions $c_1, \dots, c_{m^*}$, are of class C^k in a neighbourhood of $\bar{\mathbf{x}}$ and assume that their gradients $Dc_i(\bar{\mathbf{x}})$ for $i = 1, \dots, m^*$ are linearly independent. Then,*

$$T^S(\bar{\mathbf{x}}) = T^S_{\text{lin}}(\bar{\mathbf{x}}) \overset{\text{def}}{=} \{\, \mathbf{d} \in \mathbb{R}^n \colon Dc_i(\bar{\mathbf{x}})\mathbf{d} = 0,\ i = 1, \dots, m^* \}.$$

Moreover, for any $\mathbf{d} \in T^S(\bar{\mathbf{x}})$ there exists an $\varepsilon > 0$ and a parametric curve $\mathbf{y} \colon (-\varepsilon, \varepsilon) \to S$ of class C^k such that $\mathbf{y}(0) = \bar{\mathbf{x}}$ and $\mathbf{y}'(0) = \mathbf{d}$.

Proof. First we show that $T^S(\bar{\mathbf{x}}) \subset T^S_{\text{lin}}(\bar{\mathbf{x}})$. Let $\mathbf{d} \in T^S(\bar{\mathbf{x}})$. Then, $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$ for some sequences $(\mathbf{x}_k)_k \subset S$, $\mathbf{x}_k \neq \bar{\mathbf{x}}$, and $(\lambda_k)_k \subset (0, \infty)$. By definition of the directional derivative, for each $i = 1, \dots, m^*$ we obtain

$$\underbrace{c_i(\mathbf{x}_k)}_{=0} = \underbrace{c_i(\bar{\mathbf{x}})}_{=0} + Dc_i(\bar{\mathbf{x}})\underbrace{\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})}_{\to\mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k - \bar{\mathbf{x}}\|}_{\to\|\mathbf{d}\|}\underbrace{\frac{o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|)}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}}_{\to 0},$$

i.e., $Dc_i(\bar{\mathbf{x}})\mathbf{d} = 0$. Hence, $\mathbf{d} \in T^S_{\text{lin}}(\bar{\mathbf{x}})$.

It remains to be proved that $T^S_{\text{lin}}(\bar{\mathbf{x}}) \subset T^S(\bar{\mathbf{x}})$, which is more difficult. We fix $\mathbf{d} \in T^S_{\text{lin}}(\bar{\mathbf{x}})$. We are going to construct a curve located on S and passing through $\bar{\mathbf{x}}$, whose derivative at $\bar{\mathbf{x}}$ is $\mathbf{d}$. We denote $\mathbf{c}(\mathbf{x}) = \big(c_1(\mathbf{x}), \dots, c_{m^*}(\mathbf{x})\big)$ and we define the function $\mathbf{\Phi}\colon \mathbb{R}^{m^*} \times \mathbb{R} \to \mathbb{R}^{m^*}$ by the formula

$$\mathbf{\Phi}(\mathbf{u}, t) = \mathbf{c}\big(\bar{\mathbf{x}} + t\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}\big).$$

We can see that $\mathbf{\Phi}(\mathbf{0}, 0) = \mathbf{0}$. By $D_{\mathbf{u}}\mathbf{\Phi}$ we denote the matrix of partial derivatives with respect to the coordinates of the vector $\mathbf{u}$: $D_{\mathbf{u}}\mathbf{\Phi} = \big(\frac{\partial \Phi_i}{\partial u_j}\big)_{i,j=1,\dots,m^*}$. At the point $(\mathbf{0}, 0)$ we have $D_{\mathbf{u}}\mathbf{\Phi}(\mathbf{0}, 0) = D\mathbf{c}(\bar{\mathbf{x}})\big(D\mathbf{c}(\bar{\mathbf{x}})\big)^T$. Recall that due to the assumption, the matrix $D\mathbf{c}(\bar{\mathbf{x}})$ is of full rank (equal to m^*), i.e., $D_{\mathbf{u}}\mathbf{\Phi}(\mathbf{0}, 0)$ is nonsingular. By the implicit function theorem, there exists $\varepsilon > 0$ and a function $\mathbf{u}\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^{m^*}$ of class C^k such that $\mathbf{\Phi}\big(\mathbf{u}(t), t\big) = \mathbf{0}$ and $\mathbf{u}(0) = \mathbf{0}$. Let

$$\mathbf{y}(t) = \bar{\mathbf{x}} + t\mathbf{d} + \big(D\mathbf{c}(\bar{\mathbf{x}})\big)^T\mathbf{u}(t).$$

This curve, according to the construction, is located on the surface S, as $\mathbf{c}\big(\mathbf{y}(t)\big) = \mathbf{\Phi}\big(\mathbf{u}(t), t\big) = \mathbf{0}$ for $t \in (-\varepsilon, \varepsilon)$ and $\mathbf{y}(0) = \bar{\mathbf{x}}$. The derivative of the composition of functions $\mathbf{c} \circ \mathbf{y}$ is

$$\frac{d}{dt}\mathbf{c}\big(\mathbf{y}(t)\big) = D\mathbf{c}\big(\mathbf{y}(t)\big)\big(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(t)\big),$$

and at $t = 0$ we have

$$\frac{d}{dt}\mathbf{c}\big(\mathbf{y}(t)\big)\Big|_{t=0} = D\mathbf{c}\big(\mathbf{y}(0)\big)\big(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(0)\big).$$

On the other hand, we know that $\mathbf{c}\big(\mathbf{y}(t)\big) = \mathbf{0}$, i.e., the derivative above is equal to 0: $D\mathbf{c}\big(\mathbf{y}(0)\big)\big(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(0)\big) = 0$. But we have chosen $\mathbf{d} \in T^S_{\text{lin}}(\bar{\mathbf{x}})$, which in our notation means that $D\mathbf{c}(\bar{\mathbf{x}})\mathbf{d} = \mathbf{0}$. Hence, $D\mathbf{c}(\bar{\mathbf{x}})\big(D\mathbf{c}(\bar{\mathbf{x}})\big)^T\mathbf{u}'(0) = \mathbf{0}$. As the rank of the matrix $D\mathbf{c}(\bar{\mathbf{x}})$ is equal to m^*, we obtain $\mathbf{u}'(0) = \mathbf{0}$. To complete the proof we compute the derivative of $\mathbf{y}$:

$$\mathbf{y}'(t) = \mathbf{d} + \big(Dx(\bar{\mathbf{x}})\big)^T\mathbf{u}'(t),$$

and the above, for $t = 0$, gives us $\mathbf{y}'(0) = \mathbf{d}$. This equality allows us to conclude that $\mathbf{d} \in T^S(\bar{\mathbf{x}})$. □

Remark. The theorem above establishes a well known fact about spaces tangent to manifolds. From its assumptions it follows that S locally (in a neighbourhood of $\bar{\mathbf{x}}$) is a differential manifold of class C^k. The tangent space at $\bar{\mathbf{x}}$ is defined as the set of vectors being derivatives (at $\bar{\mathbf{x}}$) of curves contained in this manifold and passing through $\bar{\mathbf{x}}$ (which is equivalent to the definition of $T(\bar{\mathbf{x}})$). The equality $T_{\text{lin}}(\bar{\mathbf{x}}) = T(\bar{\mathbf{x}})$ means that the tangent space is the kernel of the linear mapping $D\mathbf{c}(\bar{\mathbf{x}})$.

Theorem 52 will often be used in what follows; it is the main tool used in the proof of the sufficient second-order condition. Using this theorem, we can easily prove the equality $T(\bar{\mathbf{x}}) = T_{\text{lin}}(\bar{\mathbf{x}})$ assuming that the linear independence condition is satisfied:

Theorem 53 *If the linear independence condition is satisfied at a point $\bar{\mathbf{x}} \in W$, then $T(\bar{\mathbf{x}}) = T_{\text{lin}}(\bar{\mathbf{x}})$.*

Proof. Let $\mathbf{d} \in T_{\text{lin}}(\bar{\mathbf{x}})$. Let $\hat{I}(\bar{\mathbf{x}}) = \{\, i \in I(\bar{\mathbf{x}})\colon Dg_i(\bar{\mathbf{x}})\mathbf{d} = 0 \,\}$. We define the surface

$$S = \{\mathbf{x} \in \mathbb{X}\colon c_k(\mathbf{x}) = 0,\ \text{where } c_k = g_i \text{ if } i \in I(\bar{\mathbf{x}}) \text{ or } c_k = h_j,\ j = 1, \dots, l\}.$$

Then,

$$T^S(\bar{\mathbf{x}}) = \{\, \mathbf{d} \in \mathbb{R}^n\colon Dg_i(\bar{\mathbf{x}})\mathbf{d} = 0,\ i \in \hat{I}(\bar{\mathbf{x}}),\ Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0,\ j = 1, \dots, l \,\}.$$

By Theorem 52, there exists a curve $\mathbf{y}\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $\mathbf{y}(0) = \bar{\mathbf{x}}$, $\mathbf{y}'(0) = \mathbf{d}$ and $g_i\big(\mathbf{y}(t)\big) = 0$, $i \in \hat{I}(\bar{\mathbf{x}})$, $h_j\big(\mathbf{y}(t)\big) = 0$, $j = 1, \dots, l$. For $i \in I(\bar{\mathbf{x}}) \setminus \hat{I}(\bar{\mathbf{x}})$ let $\hat{g}_i(t) = g_i\big(\mathbf{y}(t)\big)$, $t \in (-\varepsilon, \varepsilon)$. Then, $\hat{g}_i'(0) = Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0$, i.e., there exists $\varepsilon_i > 0$ such that $\hat{g}_i(t) < 0$ for $t \in [0, \varepsilon_i)$. For $i \notin I(\bar{\mathbf{x}})$, by continuity of g_i, there is $g_i\big(\mathbf{y}(t)\big) < 0$ in a neighbourhood of 0. Therefore, there exists $\bar{\varepsilon} > 0$ such that $\mathbf{y}(t) \in W$ for $t \in [0, \bar{\varepsilon})$. Hence, trivially, $\mathbf{d} \in T(\bar{\mathbf{x}})$.

The proof of the set inclusion $T(\bar{\mathbf{x}}) \subset T_{\text{lin}}(\bar{\mathbf{x}})$ is identical to the proof of Lemma 14.
□

9. Second-order conditions

We consider the optimization problem with mixed constraints. We assume that at a point $\bar{x} \in W$ there is $T(\bar{x}) = T_{\text{lin}}(\bar{x})$ and the first-order necessary condition is satisfied, i.e., there exist vectors $\boldsymbol{\mu} \in [0, \infty)^m$ and $\boldsymbol{\lambda} \in \mathbb{R}^l$ such that

$$\begin{cases} Df(\bar{x}) + \sum_{i \in I(\bar{x})} \mu_i Dg_i(\bar{x}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{x}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{x}) = 0, \quad i = 1, \dots, m. \end{cases}$$

Definition 28 *The Lagrange function is the function given by the formula*

$$L(x, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(x) + \sum_{i \in I(\bar{x})} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x)$$

The first-order condition may be written in a shorter form

$$\begin{cases} D_x L(x, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{x}) = 0, \quad i = 1, \dots, m, \end{cases}$$

where D_x denotes the derivative with respect to x.

Definition 29 *The set*

$$I^*(\bar{x}) = \{ i \in I(\bar{x}) \colon \mu_i > 0 \}$$

is called the set of strongly binding inequality constraints. The set

$$I^0(\bar{x}) = I(\bar{x}) \setminus I^*(\bar{x})$$

is called the set of weakly binding inequality constraints.

Theorem 54 *(necessary second-order condition) Suppose that $\bar{x}$ is a local solution of the problem with mixed constraints and the linear independence condition is satisfied at $\bar{x}$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be vectors of Lagrange multipliers of the first-order condition. If the functions f, g_i for $i \in I(\bar{x})$ and $h_1, \dots, h_l$ are twice differentiable in a neighbourhood of $\bar{x}$, then*

$$\mathbf{d}^T D_x^2 L(x, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathbf{d} \geqslant 0$$

for all $\mathbf{d} \in \mathbb{R}^n$ such that

$$\begin{aligned} &Dg_i(\bar{x})\mathbf{d} = 0, \quad i \in I(\bar{x}), \\ &Dh_j(\bar{x})\mathbf{d} = 0, \quad j = 1, \dots, l. \end{aligned}$$

Proof. Let $\mathbf{d} \in \mathbb{R}^n$ be a vector satisfying the conditions given above. By Theorem 52 there exists $\varepsilon > 0$ and a parametric curve $\mathbf{y} \colon (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ of class C^2 having the following properties: $\mathbf{y}(0) = \bar{x}$, $\mathbf{y}'(0) = \mathbf{d}$ and for $t \in (-\varepsilon, \varepsilon)$ there is $h_j(\mathbf{y}(t)) = 0$, $j = 1, \dots, l$ and $g_i(\mathbf{y}(t)) = 0$, $i \in I(\bar{x})$. Hence, the function $F(t) \overset{\text{def}}{=} L(\mathbf{y}(t), \boldsymbol{\mu}, \boldsymbol{\lambda})$ is equal to $f(\mathbf{y}(t))$ for $t \in (-\varepsilon, \varepsilon)$. The continuity of the inactive constraints allows us to conclude that $\mathbf{y}(t) \in W$ for t in a neighbourhood of 0. Hence, F has a local minimum at 0, as $\mathbf{y}(0)$ is a local minimum of f in W. From the assumptions it follows that F is of class C^2. The existence of its minimum at 0 implies that $F''(0) \geqslant 0$, i.e.,

$$0 \leqslant \mathbf{d}^T D_x^2 L(\bar{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathbf{d} + D_x L(\bar{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathbf{y}''(0).$$

The above completes the proof, as at the point $\bar{x}$ the necessary conditions are satisfied, in particular

$$D_x L(\bar{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathbf{0}^T. \quad \square$$

Theorem 54 may be generalised as follows:

Theorem 55 *If the assumptions of Theorem 54 are satisfied, then the inequality*

$$\mathbf{d}^T D_x^2 L(x, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathbf{d} \geqslant 0$$

is satisfied for all $\mathbf{d} \in \mathbb{R}^n$ such that

$$\begin{aligned} &Dg_i(\bar{x})\mathbf{d} = 0, \quad i \in I^*(\bar{x}), \\ &Dg_i(\bar{x})\mathbf{d} \leqslant 0, \quad i \in I^0(\bar{x}), \\ &Dh_j(\bar{x})\mathbf{d} = 0, \quad j = 1, \dots, l. \end{aligned}$$

The next theorem describes a sufficient condition for a local solution. Note that if this condition is satisfied then the solution is strict. This leaves us a “grey zone”, where the necessary condition is satisfied, but the sufficient condition is not, just like in the case of unconstrained optimization.

Theorem 56 *(sufficient second-order condition) Suppose that the first-order necessary condition is satisfied at a point $\bar{x} \in W$ and the functions g_i for $i \in I^*(\bar{x})$ and $h_1 \dots, h_l$ are twice differentiable at $\bar{x}$. If*

$$\mathbf{d}^T D_x^2 L(x, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathbf{d} > 0$$

for all $\mathbf{d} \in \mathbb{R}^n \setminus \{0\}$ *such that*

$$\begin{aligned} &Dg_i(\overline{x})\mathbf{d} = 0, \quad i \in I^*(\overline{x}), \\ &Dg_i(\overline{x})\mathbf{d} \leqslant 0, \quad i \in I^0(\overline{x}), \qquad (\otimes) \\ &Dh_j(\overline{x})\mathbf{d} = 0, \quad j = 1, \ldots, l, \end{aligned}$$

then the point $\overline{x}$ *is a strict local solution.*

Note that no regularity condition is assumed in this theorem.

Proof. The proof is done by contradiction. Suppose that $\overline{x}$ is not a strict local minimum. Then, there exists a sequence of feasible points, $(x_k)_k \in W$, convergent to $\overline{x}$ and such that $x_k \neq \overline{x}$ and $f(x_k) \leqslant f(\overline{x})$ for all k. Let

$$\mathbf{d}_k = \frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} \quad \text{and} \quad s_k = \|x_k - \overline{x}\|.$$

Then, $x_k = \overline{x} + s_k \mathbf{d}_k$. As $\|\mathbf{d}_k\| = 1$ for all k, there exists a subsequence $(\mathbf{d}_{k_p})_p$ convergent to a unit vector $\mathbf{d}$. To simplify the notation, assume that the original sequence $(\mathbf{d}_k)_k$ converges to $\mathbf{d}$. It is obvious that $\lim_{k\to\infty} s_k = 0$.

In what follows we are going to prove two properties of the vector $\mathbf{d}$:
(a) $\mathbf{d}^T D_x^2 L(\overline{x}, \mu, \lambda)\mathbf{d} \leqslant 0$ and (b), the vector $\mathbf{d}$ satisfies the conditions $(\otimes)$. Which contradicts the assumptions.

(a) By definition of the second order derivative,

$$L(x, \mu, \lambda) = L(\overline{x}, \mu, \lambda) + D_x L(\overline{x}, \mu, \lambda)(x - \overline{x}) + \frac{1}{2}(x - \overline{x})^T D_x^2 L(\overline{x}, \mu, \lambda)(x - \overline{x}) + o(\|x - \overline{x}\|)$$

From $f(x_k) \leqslant f(\overline{x})$, $g_i(x_k) \leqslant g_i(\overline{x})$ for $i \in I(\overline{x})$ and $h(x_k) = h(\overline{x}) = 0$ it follows that $L(x_k, \mu, \lambda) \leqslant L(\overline{x}, \mu, \lambda)$. From the first-order condition it follows that $D_x L(\overline{x}, \mu, \lambda) = \mathbf{0}^T$. Hence,

$$(x - \overline{x})^T D_x^2 L(\overline{x}, \mu, \lambda)(x - \overline{x}) + o(\|x - \overline{x}\|) \leqslant 0.$$

To the above we substitute $x_k = \overline{x} + s_k \mathbf{d}_k$ and we obtain

$$s_k^2 \mathbf{d}_k^T D_x^2 L(\overline{x}, \mu, \lambda)\mathbf{d}_k + o(s_k^2 \|\mathbf{d}_k\|^2) \leqslant 0.$$

After dividing both sides of the above by s_k^2 and recalling that $\|\mathbf{d}_k\| = 1$ we obtain

$$\mathbf{d}_k^T D_x^2 L(\overline{x}, \mu, \lambda)\mathbf{d}_k + \frac{o(s_k^2)}{s_k^2} \leqslant 0.$$

With k tending to ∞ the second term above tends to 0 and $\mathbf{d}_k$ tends to $\mathbf{d}$; hence,

$$\mathbf{d}^T D_x^2 L(\overline{x}, \mu, \lambda)\mathbf{d} \leqslant 0.$$

(b) is proved in a similar way. For a function f differentiable at $\overline{x}$ we have

$$f(x_k) = f(\overline{x}) + Df(\overline{x})(x_k - \overline{x}) + o(\|x_k - \overline{x}\|).$$

With $f(x_k) \leqslant f(\overline{x})$ it follows that

$$Df(\overline{x})(x_k - \overline{x}) + o(\|x_k - \overline{x}\|) \leqslant 0.$$

Using again $x_k = \overline{x} + s_k \mathbf{d}_k$ we obtain

$$Df(\overline{x})\mathbf{d}_k + \frac{o(s_k)}{s_k} \leqslant 0.$$

At the limit for $k \to \infty$ we obtain $Df(\overline{x})\mathbf{d} \leqslant 0$. Taking into account that $g_i(x_k) \leqslant g_i(\overline{x}) = 0$ for $i \in I(\overline{x})$ and proceeding as previously, we obtain $Dg_i(\overline{x})\mathbf{d} \leqslant 0$, $i \in I(\overline{x})$. Similarly we prove that $Dh_j(\overline{x})\mathbf{d} = 0$ for $j = 1, \ldots, l$.

We multiply the first equation of the necessary condition by $\mathbf{d}$:

$$Df(\overline{x})\mathbf{d} + \sum_{i \in I(\overline{x})} \mu_i Dg_i(\overline{x})\mathbf{d} + \sum_{j=1}^{l} \lambda_j Dh_j(\overline{x})\mathbf{d} = 0.$$

All terms of the sum above are less than or equal to 0. As their sun is 0, all of them must be zero. In particular,

$$Df(\overline{x})\mathbf{d} = 0, \quad \text{and} \quad Dg_i(\overline{x})\mathbf{d} = 0, \; i \in I^*(\overline{x}).$$

Thus we proved that the vector $\mathbf{d}$ satisfies the conditions $(\otimes)$. □

A general procedure for optimization problems with mixed constraints is the following:

Step 1. We are looking for candidates for solutions, which form two sets:

$$\begin{aligned} A_1 &= \{x \in \mathbb{X}\colon \textit{no regularity condition} \text{ is satisfied at } x\}, \\ A_2 &= \{x \in \mathbb{X}\colon \text{a regularity condition} \\ &\qquad \text{and the necessary first-order condition are satisfied at } x\}. \end{aligned}$$

Step 2. We check, whether the assumptions of Theorem 47, i.e., sufficient first-order conditions, are satisfied at each point of the sets A_1 and A_2. If they are, then we obtain global solutions.

The steps that follow should be done if we have not found any global solution or we need to find *all* global solutions or we need to find *all* local solutions. We remove all global solutions from the sets A_1 and A_2; the sets of remaining points we denote by A_1' and A_2'.

Step 3. From A_2' we exclude all points not satisfying the necessary condition of the second order, thus obtaining the set A_2''.

Step 4. At each point of $A_1' \cup A_2''$ we check the sufficient condition of the second order. The points satisfying it are local solutions.

Step 5. The analysis of the other points of $A_1' \cup A_2''$ must be done using other methods.

Example. Below is a detailed description of the analysis, step by step, of the following problem:

$$\begin{cases} (x_1 - 1)^2 + x_2^2 \to \min, \\ 2kx_1 - x_2^2 \leqslant 0, \\ \mathbf{x} = (x_1, x_2) \in \mathbb{R}^2, \end{cases}$$

where $k > 0$ is a parameter.

We notice that at each point, where the constraint is active, the linear independence condition is satisfied: the gradient of the constraint function is $[2k, 2x_2] \neq 0^T$. Hence, $A_1 = \emptyset$.

The Lagrange function for this problem is

$$L(x_1, x_2; \mu) = (x_1 - 1)^2 + x_2^2 + \mu(2kx_1 - x_2^2).$$

The first-order necessary condition is thus

$$\begin{cases} [2(x_1 - 1), 2x_2] + \mu[2k, -2x_2] = [0, 0], \\ \mu(2kx_1 - x_2^2) = 0, \\ \mu \geqslant 0. \end{cases}$$

Is it possible that $\mu = 0$? By the first equation we obtain $2(x_1 - 1) = 0$, $2x_2 = 0$, i.e., $x_1 = 1$, $x_2 = 0$. This is not a feasible point for any parameter $k > 0$. Therefore μ must be positive and the necessary first-order condition may be satisfied only at the boundary of the feasible set, where $2kx_1 - x_2^2 = 0$.

Again, by the first equation,

$$\begin{cases} x_1 - 1 + \mu k = 0, \\ \quad x_2 - \mu x_2 = 0. \end{cases}$$

There is either $\mu_2 = 1$ or $x_2 = 0$. If $x_2 = 0$, then due to the constraint we have $x_1 = 0$. The point $(0, 0)$ together with the Lagrange multiplier $\mu = 1/k$ satisfies the first-order necessary constraint.

Now we consider $\mu = 1$. Then, from $x_1 - 1 - \mu k = 0$ we obtain $x_1 = 1 - k$. If $k > 1$, then $x_1 < 0$ and no point with such a first coordinate is feasible. If $k = 1$, then we get the point $(0, 0)$ and for $k \in (0, 1)$ we have two points:

$$x_1 = 1 - k, \quad x_2 = \pm\sqrt{2k(1-k)}.$$

We conclude that $A_1 = \emptyset$ and

$$A_2 = \{(0, 0)\} \quad \text{if } k \geqslant 1,$$
$$A_2 = \Big\{(0, 0),\ \big(1 - k, \sqrt{2k(1-k)}\big), \big(1 - k, -\sqrt{2k(k-1)}\big)\Big\} \quad \text{if } 0 < k < 1.$$

The function $g_1(x_1, x_2) = 2kx_1 - x_2^2$ is not quasi-convex, so we cannot use Theorem 47 describing a sufficient first-order condition. Thus we have $A_2' = A_2$.

Now we pass to Step 3. We check the sufficient second-order condition for points of A_2'. We take a look at the point $(0, 0)$; with this point we have the Lagrange multiplier $\mu = 1/k$. The gradient and the Hessian of the Lagrange function are

$$D_x L(x_1, x_2; 1/k) = [2(x_1 - 1) + 2, 2x_2(1 - 1/k)]$$

$$D_x^2 L(x_1, x_2; 1/k) = \begin{bmatrix} 2 & 0 \\ 0 & 2(1 - 1/k) \end{bmatrix}$$

As the theorem states, it suffices to check if

$$\mathbf{d}^T \begin{bmatrix} 2 & 0 \\ 0 & 2(1 - 1/k) \end{bmatrix} \mathbf{d} \geqslant 0$$

for all vectors $\mathbf{d} \in \mathbb{R}^2 \setminus \{0\}$ such that $Dg_1(0, 0)\mathbf{d} = [2k, 0]\mathbf{d} = 0$, i.e., the vectors with the coordinate $d_1 = 0$. The last inequality is $(1 - 1/k)d_2^2 \geqslant 0$; it is satisfied if $k \geqslant 1$. Thus, if $k \geqslant 1$ it *may be* a local solution. If $k < 1$, then the inequality is not satisfied and there is no solution at $(0, 0)$.

Now we assume that $0 < k < 1$ and we consider the other two points. In both cases $\mu = 1$. Then we have

$$D_x L(x_1, x_2; 1/k) = [2(x_1 - 1) + 2k, 0]$$
$$D_x^2 L(x_1, x_2; 1/k) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

The Hessian is nonnegative-definite; hence, the necessary second-order condition is satisfied at both points. Thus, we have

$$A_2'' = \{(0,0)\} \quad \text{if } k \geqslant 1,$$
$$A_2'' = \left\{\left(1 - k, \sqrt{2k(1-k)}\right), \left(1 - k, -\sqrt{2k(k-1)}\right)\right\} \quad \text{if } 0 < k < 1.$$

Now we check the sufficient second-order condition. If $k > 1$ the Hessian is positive-definite and by Theorem 56 there is a strict local minimum at $(0,0)$. We cannot make this conclusion if $k = 1$.

We assume $0 < k < 1$ and we analyse the point $\overline{x} = (1 - k, \sqrt{2k(k-1)})$. We need to check the assumptions of Theorem 56:

$$\mathbf{d}^T \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{d} = 2d_1^2 > 0$$

for $\mathbf{d} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$ such that $[2k, -2\sqrt{2k(1-k)}]\mathbf{d} = 0$, i.e., $d_1 = \frac{1}{k}\sqrt{2k(1-k)}d_2$. If $\mathbf{d} \neq \mathbf{0}$, then $d_1 \neq 0$ and at the point $\overline{x}$ the sufficient condition of the second order is satisfied: this point is a strict local solution. Similarly we prove that the point $(1 - k, -\sqrt{2k(k-1)})$ is also a strict local solution.

It remains to analyse the case $k = 1$. Though the point $(0,0)$ is a global minimum, we have not managed to prove it using the Kuhn–Tucker theory. On the other hand, it is easy to do it in an elementary way.

10. Dual problems

Described below are so-called dual problems, i.e., another view at optimization problems with inequality constraints. In contrast to the Kuhn–Tucker approach, the dual setting does not assume differentiability of the target function f nor of the constraint functions g_i. Moreover, with suitable assumptions, the solution of the original problem may easily be obtained from the solution of the dual problem. Solving the dual problem is finding the maximum of a concave target function in the nonnegative octant. As we shall see, the concavity may lead to a good convergence of numerical optimization algorithms. The simplicity of the feasible set is another feature contributing to the simplicity of implementation and to speed of numerical algorithms. Sometimes the dual problem is easier to solve analytically; examples will be given.

Sufficient condition

Definition 30 *Let A, B be arbitrary sets and let $h\colon A \times B \to \mathbb{R}$ be a function. The point $(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}}) \in A \times B$ is called a saddle point of the function h if*

$$h(\bar{\mathbf{x}}, \boldsymbol{\mu}) \leqslant h(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}}) \leqslant h(\mathbf{x}, \bar{\boldsymbol{\mu}}) \quad \textit{for all } \mathbf{x} \in A,\ \boldsymbol{\mu} \in B.$$

Example. The simplest example is the “centre of the saddle”, $A = B = \mathbb{R}$, $h(x, \mu) = x^2 - \mu^2$. The function h has the saddle point at $(0, 0)$, which is the minimum with respect to x and the maximum with respect to μ.

It turns out that the global solution of the following problem:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0,\ i = 1, \dots, m, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{*}$$

is related to the saddle point of the Lagrange function defined for this problem. Recall that W denotes the feasible set, i.e.,

$$W = \{\mathbf{x} \in \mathbb{X} \colon g_i(\mathbf{x}) \leqslant 0, i = 1, \dots, m\}.$$

Theorem 57 *If $(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}}) \in W \times [0, \infty)^m$ is a saddle point of the Lagrange function*

$$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=0}^{m} \mu_i g_i(\mathbf{x})$$

in the set $W \times [0, \infty)^m$, i.e.,

$$L(\bar{\mathbf{x}}, \boldsymbol{\mu}) \leqslant L(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}}) \leqslant L(\mathbf{x}, \bar{\boldsymbol{\mu}}) \quad \textit{for all } \mathbf{x} \in W,\ \boldsymbol{\mu} \in [0, \infty)^m,$$

then $\bar{\mathbf{x}}$ is a global solution of the problem (*) and $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ for $i = 1, \dots, m$.

Proof. First we prove that $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ for $i = 1, \dots, m$. The inequality $L(\bar{\mathbf{x}}, \boldsymbol{\mu}) \leqslant L(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}})$ may be expanded as follows:

$$f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \mu_i g_i(\bar{\mathbf{x}}) \leqslant f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}).$$

Hence, for all $\boldsymbol{\mu} \in [0, \infty)^m$ we have

$$\sum_{i=1}^{m} \mu_i g_i(\bar{\mathbf{x}}) \leqslant \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}).$$

By substituting $\boldsymbol{\mu} = \frac{1}{2}\bar{\boldsymbol{\mu}}$, we obtain

$$\sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) \geqslant 0.$$

The point $\bar{\mathbf{x}}$ is feasible, i.e., $g_i(\bar{\mathbf{x}}) \leqslant 0$ for all i. Taking into acount that all coordinates of $\bar{\boldsymbol{\mu}}$ are nonnegative, we conclude that $\sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ and each term is nonnegative. From the above it follows directly that $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ for all i.

Now we use the second inequality, $L(\bar{\mathbf{x}}, \bar{\boldsymbol{\mu}}) \leqslant L(\mathbf{x}, \bar{\boldsymbol{\mu}})$ for all $\mathbf{x} \in W$, to prove that $\bar{\mathbf{x}}$ is a global solution. We expand the inequality:

$$f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) \leqslant f(\mathbf{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\mathbf{x}).$$

Earlier we have proved that $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ for all i. Due to $\mathbf{x} \in W$, there is $\bar{\mu}_i g_i(\mathbf{x}) \leqslant 0$. Hence,

$$f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x}) \quad \text{for all } x \in W. \ \square$$

Remarks.

- We do not assume that $\mathbb{X}$ is open, also the functions f and g_i do not have to be continuous.
- The feasible set does not have to be convex.

- No regularity conditions are necessary.
- The theorem does not offer any methods of searching the saddle point. It might be found using the necessary first-order conditions, and then Theorem 57 may be used as the sufficient condition.
- Theorem 57 is the base for the dual approach and despite the previous remark it is useful in the development of numerical optimization algorithms.

Necessary condition for convex programming

Now we assume that the set $\mathbb{X} \subset \mathbb{R}^n$ is convex and the functions f and g_i, $i = 1, \ldots, m$, are convex. For such an optimization problem the saddle point of the Lagrange function is a necessary condition for the global solution. We begin the analysis with the simpler case, where all functions are differentiable, and later we prove the theorem which does not assume differentiability. As it was mentioned before, the lack of the requirement of differentiability distinguishes the saddle point method from the Kuhn–Tucker method.

Lemma 20 *Suppose that the set $\mathbb{X}$ in the convex programming problem is open and the functions $f, g_1, \ldots, g_m$ are differentiable at a point $\bar{x}$. If $\bar{x}$ is a local solution of the problem (*) and one of the regularity conditions: linear independence, affine function or Slater condition, is satisfied, then there exists $\bar{\mu} \in [0, \infty)^m$, such that $(\bar{x}, \bar{\mu})$ is a saddle point of the Lagrange function in the set $\mathbb{X} \times [0, \infty)^m$.*

Proof. By the Kuhn–Tucker theorem (Theorem 38), there exists a vector of Lagrange multipliers $\bar{\mu} \in [0, \infty)^m$ such that the first-order condition is satisfied (the assumptions of this theorem are satisfied due to regularity of $\bar{x}$ and Theorems 39–41). The function

$$L(x, \mu) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x),$$

being the linear combination of convex functions with nonnegative coefficients, is convex. Therefore,

$$L(x, \bar{\mu}) \geqslant L(\bar{x}, \bar{\mu}) + D_x L(\bar{x}, \bar{\mu})(x - \bar{x}).$$

By the Kuhn–Tucker theorem,

$$D_x L(\bar{x}, \bar{\mu}) = Df(\bar{x}) + \sum_{i=1}^{m} \bar{\mu}_i Dg_i(\bar{x}) = 0^T,$$

i.e., $D_x L(\bar{x}, \bar{\mu})(x - \bar{x}) = 0$. Hence, $L(x, \bar{\mu}) \geqslant L(\bar{x}, \bar{\mu})$.

To prove the inequality $L(\bar{x}, \bar{\mu}) \geqslant L(\bar{x}, \mu)$, we notice that

$$\sum_{i=1}^{m} \mu_i g_i(\bar{x}) \leqslant 0 = \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{x}),$$

because $\mu_i \geqslant 0$ and $g_i(\bar{x}) \leqslant 0$. The last equality is the claim of Theorem 38. □

Remark. By Theorem 47, each point satisfying the first-order condition is a global solution of the convex programming problem. Therefore, we do not need to distinguish global and local solutions.

Theorem 58 *Let $\bar{x} \in \mathbb{X}$ be a global solution of the convex programming problem (*) and let there exist a point $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for all $i = 1, \ldots, m$. Then, there exists $\bar{\mu} \in [0, \infty)^m$ such that $(\bar{x}, \bar{\mu})$ is a saddle point of the Lagrange function in the space $\mathbb{X} \times [0, \infty)^m$, i.e.,*

$$L(\bar{x}, \mu) \leqslant L(\bar{x}, \bar{\mu}) \leqslant L(x, \bar{\mu}) \quad \text{for all } x \in \mathbb{X} \text{ and } \mu \in [0, \infty)^m.$$

Moreover, $\bar{\mu}_i g_i(\bar{x}) = 0$ for $i = 1, \ldots, m$.

Remark. In Theorems 57 and 58 the saddle point of the Lagrange function is considered in different spaces. In the second theorem the space is wider, as the first variable goes through the entire set $\mathbb{X}$, not just the feasible set W. Now we obtain the equivalence of the existence of a solution at the saddle point of the Lagrange function and the existence of the global solution of the convex programming problem.

Proof. Like in the proof of the necessary first-order condition (Theorem 38), the crucial role is played by the separation theorem. It will indicate the Lagrange multiplier vector $\bar{\mu}$.

Denote $g(x) = \big(g_1(x), \ldots, g_m(x)\big)$. We define the following subsets of $\mathbb{R}^{m+1}$:

$$\begin{aligned}
A &= \{\, \bar{y} = (y_0, y) \in \mathbb{R} \times \mathbb{R}^n \colon y_0 \geqslant f(x),\ y \geqslant g(x),\ x \in \mathbb{X} \,\},\\
B &= \{\, \bar{y} = (y_0, y) \in \mathbb{R} \times \mathbb{R}^n \colon y_0 = f(x),\ y = g(x),\ x \in \mathbb{X} \,\},\\
C &= \{\, \bar{y} = (y_0, y) \in \mathbb{R} \times \mathbb{R}^n \colon y_0 < f(\bar{x}),\ y < 0 \,\}.
\end{aligned}$$

The "inequality between vectors" notation is to be understood as the inequality between their corresponding coordinates.

The set C is the Cartesian product of the interval $(-\infty, f(\overline{\boldsymbol{x}})]$ and the cone $\{\boldsymbol{y} < \boldsymbol{0}\}$. Obviously, this set is convex. As $\overline{\boldsymbol{x}}$ is the minimum of f, there is $B \cap C = \emptyset$. The set B is not convex; therefore we cannot use the separation theorem for the sets B and C. A remedy is to take the set A, whose subset is B. Suppose that there exists $\overline{\boldsymbol{y}} = (y_0, \boldsymbol{y}) \in A \cap C$. It follows that there exists $\boldsymbol{x}' \in \mathbb{X}$ such that

$$y_0 \geqslant f(\boldsymbol{x}'),\ \boldsymbol{y} \geqslant \boldsymbol{g}(\boldsymbol{x}'),\ y_0 < f(\overline{\boldsymbol{x}}),\ \boldsymbol{y} < \boldsymbol{0}.$$

From the inequalities above we conclude that $f(\boldsymbol{x}') < f(\overline{\boldsymbol{x}})$ and $\boldsymbol{g}(\boldsymbol{x}') < \boldsymbol{0}$. Thus, $\boldsymbol{x}'$ is a feasible point at which the value of f is less than $f(\overline{\boldsymbol{x}})$. It contradicts $\overline{\boldsymbol{x}}$ being a solution.

Example. Before proceeding with the proof, we can take a look at the sets A, B, C for the following problem:

$$\begin{cases} -x \to \min, \\ x^2 - 1 \leqslant 0, \\ x \in \mathbb{X} = \mathbb{R}. \end{cases}$$

Its solution is $\overline{x} = 1$. There is only one constraint, and thus the sets are subsets of $\mathbb{R}^2$. As we can see in Figure 7, the part of the set B for $y_0 \leqslant 0$ is a part of the boundary of A, and the rest of it (for $y_0 > 0$) is located in the interior of A.

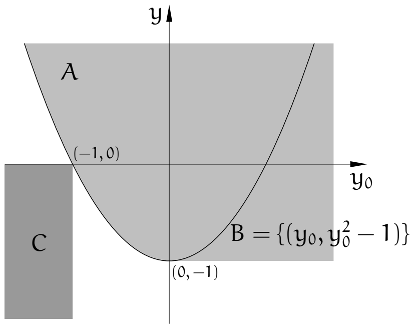

Figure 7: The sets A, B, C in the example

Back to the proof. The convexity of C is already established. The convexity of A may be proved directly. Let $\overline{\boldsymbol{y}}', \overline{\boldsymbol{y}}'' \in A$ and $\lambda \in (0,1)$. Then, there exist points $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{X}$ such that

$$y_0' \geqslant f(\boldsymbol{x}'),\ \boldsymbol{y}' \geqslant \boldsymbol{g}(\boldsymbol{x}'), \qquad y_0'' \geqslant f(\boldsymbol{x}''),\ \boldsymbol{y}'' \geqslant \boldsymbol{g}(\boldsymbol{x}'').$$

Let $\boldsymbol{x} = \lambda\boldsymbol{x}' + (1-\lambda)\boldsymbol{x}''$. By convexity of $\mathbb{X}$, $\boldsymbol{x} \in \mathbb{X}$. We also have

$$\lambda y_0' + (1-\lambda)y_0'' \geqslant \lambda f(\boldsymbol{x}') + (1-\lambda)f(\boldsymbol{x}'') \geqslant f\big(\lambda\boldsymbol{x}' + (1-\lambda)\boldsymbol{x}''\big) = f(\boldsymbol{x}).$$

The first inequality above results from the assumed properties of $\boldsymbol{x}'$, $\boldsymbol{x}''$, and the second inequality is a consequence of the convexity of f. In a similar way, using the convexity of the components of $\boldsymbol{g}$, we show that

$$\lambda\boldsymbol{y}' + (1-\lambda)\boldsymbol{y}'' \geqslant \boldsymbol{g}(\boldsymbol{x}).$$

Hence, $\overline{\boldsymbol{y}} = \lambda\overline{\boldsymbol{y}}' + (1-\lambda)\overline{\boldsymbol{y}}'' \in A$, as with the point $\boldsymbol{x}$ defined above,

$$y_0 \geqslant f(\boldsymbol{x}),\ \boldsymbol{y} \geqslant \boldsymbol{g}(\boldsymbol{x}).$$

By the weak separation theorem, there exists a nonzero vector $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^{m+1}$ such that

$$\tilde{\boldsymbol{\mu}}^T\overline{\boldsymbol{y}} \geqslant \tilde{\boldsymbol{\mu}}^T\overline{\boldsymbol{z}}, \quad \text{for all } \overline{\boldsymbol{y}} \in A,\ \overline{\boldsymbol{z}} \in C.$$

From $\sup_{\overline{\boldsymbol{z}}\in C} \tilde{\boldsymbol{\mu}}^T\overline{\boldsymbol{z}} < \infty$ it follows that $\tilde{\boldsymbol{\mu}} \geqslant \boldsymbol{0}$. Due to the continuity of linear functions, we can take $\overline{\boldsymbol{z}}$ from the closure of C:

$$\tilde{\boldsymbol{\mu}}^T\overline{\boldsymbol{y}} \geqslant \tilde{\boldsymbol{\mu}}^T\overline{\boldsymbol{z}}, \quad \text{for all } \overline{\boldsymbol{y}} \in A,\ \overline{\boldsymbol{z}} \in \operatorname{cl} C.$$

Hence, for $\overline{\boldsymbol{z}} = \big(f(\overline{\boldsymbol{x}}), \boldsymbol{0}\big)$ we have

$$\tilde{\mu}_0 y_0 + \sum_{i=1}^{m} \tilde{\mu}_i y_i \geqslant \tilde{\mu}_0 f(\overline{\boldsymbol{x}}) \quad \text{for all } (y_0, \boldsymbol{y}) \in A. \tag{$\otimes$}$$

In particular, this inequality holds for $y_0 = f(\boldsymbol{x})$ and $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{X}$:

$$\tilde{\mu}_0 f(\boldsymbol{x}) + \sum_{i=1}^{m} \tilde{\mu}_i g_i(\boldsymbol{x}) \geqslant \tilde{\mu}_0 f(\overline{\boldsymbol{x}}).$$

Now we prove that $\tilde{\mu}_0 \neq 0$, which together with the observation $\tilde{\boldsymbol{\mu}} \geqslant \boldsymbol{0}$ implies $\tilde{\mu}_0 > 0$. The proof is done by contradiction. Suppose that $\tilde{\mu}_0 = 0$. Then, by the last inequality, we have

$$\sum_{i=1}^{m} \tilde{\mu}_i g_i(\boldsymbol{x}) \geqslant 0 \quad \text{for all } (y_0, \boldsymbol{y}) \in A.$$

In particular, the above holds for the point $\boldsymbol{x}^*$ from the assumptions of the theorem. However, at this point we have $g_i(\boldsymbol{x}^*) < 0$ for $i = 1, \ldots, m$. Together with the fact that $\tilde{\boldsymbol{\mu}} \geqslant \boldsymbol{0}$ this implies $\tilde{\mu}_1 = \cdots = \tilde{\mu}_m = 0$. It follows that $\tilde{\boldsymbol{\mu}} = \boldsymbol{0}$, and this is inconsistent with the choice of $\tilde{\boldsymbol{\mu}}$ based on the separation theorem.

As we now know, $\tilde{\mu}_0 > 0$. We define

$$\overline{\mu} = \Bigl(\frac{\tilde{\mu}_1}{\tilde{\mu}_0}, \dots, \frac{\tilde{\mu}_m}{\tilde{\mu}_0}\Bigr).$$

Obviously, $\overline{\mu} \in [0, \infty)^m$. As $\overline{x}$, being a solution, is a feasible point, there is $g_i(\overline{x}) \leqslant 0$ for $i = 1, \dots, m$ and $\sum_{i=1}^m \overline{\mu}_i g_i(\overline{x}) \leqslant 0$. We add this sum to the right-hand side of the inequality $(\otimes)$ divided by $\tilde{\mu}_0$:

$$f(x) + \sum_{i=1}^m \overline{\mu}_i g_i(x) \geqslant f(\overline{x}) + \sum_{i=1}^m \overline{\mu}_i g_i(\overline{x}) \quad \text{for all } x \in \mathbb{X}.$$

In other words,

$$L(x, \overline{\mu}) \geqslant L(\overline{x}, \overline{\mu}) \quad \text{for all } x \in \mathbb{X}.$$

It remains to prove the other inequality of the saddle point. By taking $x = \overline{x}$ and dividing both sides of $(\otimes)$ by $\tilde{\mu}_0$, we obtain $\sum_{i=1}^m \overline{\mu}_i g_i(\overline{x}) \geqslant 0$. On the other hand, the point $\overline{x}$ is feasible, i.e., $g_i(\overline{x}) \leqslant 0$. Recalling that $\overline{\mu} \geqslant \mathbf{0}$, we conclude that each term of this sum is nonpositive. Hence, we obtain

$$\overline{\mu}_i g_i(\overline{x}) = 0, \quad i = 1, \dots, m.$$

For any other μ we have $\sum_{i=1}^m \mu_i g_i(\overline{x}) = 0$, i.e.,

$$\sum_{i=1}^m \mu_i g_i(\overline{x}) \leqslant \sum_{i=1}^m \overline{\mu}_i g_i(\overline{x}) \quad \text{for all } \mu \in [0, \infty)^m.$$

This is equivalent to

$$L(\overline{x}, \mu) \leqslant L(\overline{x}, \overline{\mu}) \quad \text{for all } \mu \in [0, \infty)^m. \quad \square$$

Primal and dual problems

The theory of saddle points leads to formulating primal and dual problems. Consider the optimization problem (*) and its Lagrange function $L(x, \mu)$. We define the function $L_P \colon W \to (-\infty, \infty]$:

$$L_P(x) = \sup_{\mu \in [0,\infty)^m} L(x, \mu).$$

As we can notice,

$$L_P(x) = \begin{cases} f(x) & \text{if } g(x) \leqslant \mathbf{0}, \\ \infty & \text{else.} \end{cases}$$

Therefore the problem (*) may be rewritten in what seems to be a simpler form

$$L_P(x) \to \min, \quad x \in \mathbb{X}.$$

Alas, the problem above reduces to the original problem, so it does not give any “added value”, but soon it will. Before we reveal this value, we define another function, $L_D(\mu) \colon [0, \infty)^m \to [-\infty, \infty)$:

$$L_D(\mu) = \inf_{x \in \mathbb{X}} L(x, \mu).$$

Remarks. (1) for any $x \in \mathbb{X}$ and $\mu \in [0, \infty)^m$ there is $L_P(x) \geqslant L(x, \mu) \geqslant L_D(\mu)$.
(2) If $(\overline{x}, \overline{\mu})$ is a saddle point of the Lagrange function in $\mathbb{X} \times [0, \infty)^m$, then $L_P(\overline{x}) = L_D(\overline{\mu})$.

If $(\overline{x}, \overline{\mu})$ is a saddle point, then $L(\overline{x}, \overline{\mu}) \leqslant L(x, \overline{\mu})$. In view of the first remark above it gives us $L(\overline{x}, \overline{\mu}) = L_D(\overline{\mu})$. Similarly we prove the equality $L(\overline{x}, \overline{\mu}) = L_P(\overline{x})$. These observations lead us in the right direction. We are going to use the functions L_P and L_D in the search of saddle points.

Definition 31 *The primal problem is the optimization problem*

$$L_P(x) \to \min, \quad x \in \mathbb{X}.$$

Its dual problem is the optimization problem

$$L_D(\mu) \to \max, \quad \mu \in [0, \infty)^m.$$

By the properties mentioned in the last remark, the value of the target function of the primal problem at the solution is greater than or equal to the value of the target function of the dual problem at the solution:

$$\inf_{x \in \mathbb{X}} L_P(x) \geqslant \sup_{\mu \in [0,\infty)^m} L_D(\mu).$$

Moreover, the solution of the dual problem gives us an estimation from below of the function f.

Lemma 21 *(weak duality theorem) For any feasible point $x \in W$ and any vector $\mu \in [0, \infty)^m$ there is*

$$f(x) \geqslant L_D(\mu).$$

Hence,

$$f(x) \geqslant \sup_{\mu \in [0,\infty)^m} L_D(\mu).$$

Proof. We have

$$f(x) \geqslant L(x, \mu) \geqslant L_D(\mu).$$

The first inequality holds for $x \in W$ because $g_i(x) \leqslant 0$ for all i. The second inequality is a consequence of Remark (1) above. □

Definition 32 *The duality gap is the difference between the values of target functions at the solutions of the primal and dual problem:*

$$\inf_{x \in \mathbb{X}} L_P(x) - \sup_{\mu \in [0,\infty)^m} L_D(\mu).$$

The saddle point condition written in terms of the primal and dual functions is the following: $(\overline{x}, \overline{\mu})$ is a saddle point if

$$L_P(\overline{x}) = L(\overline{x}, \overline{\mu}) = L_D(\overline{\mu}).$$

In other words, if the Lagrange function has a saddle point, then the duality gap is zero. This is the case if, for example, the assumptions of Theorem 58 are satisfied.

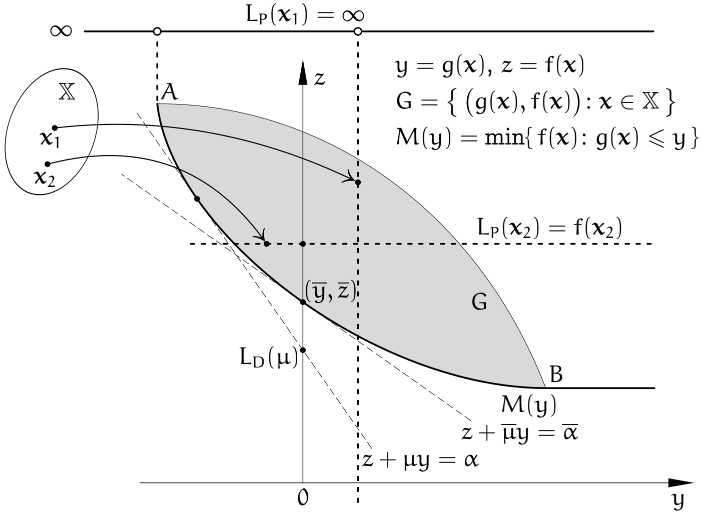
∞
$L_P(x_1) = \infty$
z
$y = g(x)$, $z = f(x)$
$G = \{ (g(x), f(x)) : x \in \mathbb{X} \}$
$M(y) = \min\{ f(x) : g(x) \leqslant y \}$
$\mathbb{X}$
A
x_1
x_2
$L_P(x_2) = f(x_2)$
$(\overline{y}, \overline{z})$
G
$L_D(\mu)$
B
$M(y)$
$z + \overline{\mu}y = \overline{\alpha}$
$z + \mu y = \alpha$
0
y

Figure 8: A scheme of solutions of a primal and dual problem

Figure 8 shows the solution of a primal and dual problem of an optimization problem with only one inequality constraint; $m = 1$. The set G is the set of pairs

of values $\big(f(x), g(x)\big)$ for $x \in \mathbb{X}$. The lines $z + \mu y = \alpha$ show the values of the Lagrange function $L(x, \mu) = f(x) + \mu g(x)$. From the definition of the primal function L_P it follows that $L_P(x) = f(x)$ if $g(x) \leqslant 0$, as $\sup_{\mu \geqslant 0}\{ \alpha : \alpha = z + \mu y \}$ is taken for $\mu = 0$. This is the case of the point x_2. If $g(x) > 0$, then $L_P(x) = +\infty$. This is the case of the point x_1. The dual function $L_D(\mu)$ may be found by considering the lines $z + \mu y = \alpha$, looking for $\inf_{x \in \mathbb{X}} \alpha$ with a fixed μ. It may be seen in the picture that the minimum is taken for the line tangent to the boundary of the area G and the value of this function is the intersection point of the line with the line $y = 0$. One can also see that the line which yields the greatest value of L_D is the line $z + \overline{\mu}y = \overline{\alpha}$ whose inclination (tangent of the angle between the y axis and the line) is $-\overline{\mu}$, tangent to the boundary of G at the point $(\overline{y}, \overline{z})$. This point coresponds to the solution of the primal problem, because $\overline{z} = \inf_{x \in \mathbb{X}, g(x) \leqslant 0} f(x)$.

Theorem 59 *(strong duality theorem) Let $\mathbb{X}$ be a nonempty convex subset of $\mathbb{R}^n$ and let the functions f and g_i, $i = 1, \dots, m$, be convex in $\mathbb{X}$. Assume in addition that there exists a point $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for all i. Then,*

$$\inf_{x \in \mathbb{X}} L_P(x) = \sup_{\mu \in [0,\infty)^m} L_D(\mu).$$

If $\inf_{x \in \mathbb{X}} L_P(x)$ is finite, then $\sup_{\mu \in [0,\infty)^m} L_D(\mu)$ is taken at a point $\overline{\mu}$ such that $\overline{\mu} \geqslant 0$. If $\inf_{x \in \mathbb{X}} L_P(x)$ is taken at a point $\overline{x}$, then $\overline{\mu}_i g_i(\overline{x}) = 0$ for $i = 1, \dots, m$.

Proof. Let $\gamma = \inf_{x \in \mathbb{X}} L_P(x)$. If $\gamma = -\infty$, then (by Lemma 21) $\sup_{\mu \in [0,\infty)^m} L_D(\mu) = -\infty$ and the claim is true. Now assume that $\gamma > -\infty$.

From the proof of Theorem 58 applied to the sets

$$A = \{ \overline{\boldsymbol{y}} = (y_0, \boldsymbol{y}) \in \mathbb{R} \times \mathbb{R}^n : y_0 \geqslant f(x),\ \boldsymbol{y} \geqslant \boldsymbol{g}(x),\ x \in \mathbb{X} \},$$
$$C = \{ \overline{\boldsymbol{y}} = (y_0, \boldsymbol{y}) \in \mathbb{R} \times \mathbb{R}^n : y_0 < f(\overline{x}),\ \boldsymbol{y} < \boldsymbol{0} \},$$

we conclude that

$$f(x) + \sum_{i=1}^{m} \overline{\mu}_i g_i(x) \geqslant \gamma \quad \text{for all } x \in \mathbb{X}. \tag{$\oplus$}$$

As a consequence, we obtain

$$L_D(\overline{\mu}) = \inf_{x \in \mathbb{X}} \Big(f(x) + \sum_{i=1}^{m} \overline{\mu}_i g_i(x) \Big) \geqslant \gamma.$$

From the remark made before Definition 31 it follows that

$$\gamma = \inf_{x \in \mathbb{X}} L_P(x) \geqslant \sup_{\mu \in [0,\infty)^m} L_D(\mu) = L_D(\overline{\mu}) \geqslant \gamma.$$

Hence, $L_D(\overline{\mu}) = \gamma$ and $\overline{\mu}$ is a solution of the dual problem.

If $\inf_{x \in \mathbb{X}} L_P(x)$ is taken at a point $\overline{x}$, then due to the definition of L_P we have $L_P(\overline{x}) = f(\overline{x})$. The point $\overline{x}$ is a solution of the primal problem; there is $g_i(\overline{x}) \leqslant 0$ for $i = 1, \ldots, m$ and $f(\overline{x}) = \gamma$. By substituting $x = \overline{x}$ in the inequality ($\oplus$), we get $\sum_i \overline{\mu}_i g_i(\overline{x}) \geqslant 0$. Because $\mu_i \geqslant 0$ and $g_i(\overline{x}) \leqslant 0$, the equality $\overline{\mu}_i g_i(\overline{x}) = 0$ for $i = 1, \ldots, m$ follows. □

The following algorithm of solving the optimization problem (*) using dual methods may be proposed:

1. Solve the dual problem. The solution gives a lower bound for the solution of the primal problem, due to Lemma 21.
2. Suppose that there exists a finite solution $\overline{\mu} \in [0, \infty)^m$ of the dual problem and a point $\overline{x} \in \mathbb{X}$ such that $L_D(\overline{\mu}) = L_P(\overline{x})$. If $\overline{x}$ is a feasible point and $f(\overline{x}) = L_D(\overline{\mu})$, then $(\overline{x}, \overline{\mu})$ is a saddle point of the Lagrange function and by Theorem 57 the point $\overline{x}$ is a solution of the problem (*).

Let's explain the conditions in the second step. From $L_D(\overline{\mu}) = L(\overline{x}, \overline{\mu})$ it follows that $L(\overline{x}, \overline{\mu}) \leqslant L(x, \overline{\mu})$ for all $x \in \mathbb{X}$. This is the second inequality of the saddle point. To verify the first inequality, recall that $L_P(x) = f(x)$ for any feasible point x and $\inf_{x \in \mathbb{X}} L_P(x) \geqslant L_D(\mu)$ for all $\mu \in [0, \infty)^m$. In the second step we assume that $f(\overline{x}) = L_D(\overline{\mu})$, which implies that

$$L_P(\overline{x}) = f(\overline{x}) = L_D(\overline{\mu})$$

and, thus, $(\overline{x}, \overline{\mu})$ is indeed a saddle point.

11. Sensitivity theory

So far, the Lagrange multipliers seemed to be just a technical trick useful in finding solutions of optimization problems with constraints. Below we show that they represent costs of changing the constraints. The equality and inequality constraints are dealt with separately.

Equality constraints

Consider a problem with equality constraints:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ h_j(\mathbf{x}) = 0, \ j = 1, \dots, l, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{*}$$

where $\mathbb{X} \subset \mathbb{R}^n$ is an open set and $f, h_1, \dots, h_l \colon \mathbb{X} \to \mathbb{R}$. To simplify the notation, we denote $\mathbf{h}(\mathbf{x}) = \big(h_1(\mathbf{x}), \dots, h_l(\mathbf{x})\big)$. Now we introduce a perturbation:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{h}(\mathbf{x}) = \mathbf{z}, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{**}$$

where $\mathbf{z} \in \mathbb{R}^l$.

<u>Theorem 60</u> *Let $\overline{\mathbf{x}}$ be a solution of the problem (*) and let $\overline{\boldsymbol{\lambda}}$ be the corresponding vector of Lagrange multipliers. Assume that the functions $f, h_1, \dots, h_l$ are of class C^2 in a neighbourhood of $\overline{\mathbf{x}}$, the gradients of the constraint functions are linearly independent and*

$$\mathbf{d}^T D_x^2 L(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) \mathbf{d} > 0 \tag{$\odot$}$$

for all nonzero vectors $\mathbf{d} \in \mathbb{R}^n$ such that $Dh_j(\overline{\mathbf{x}})\mathbf{d} = 0$, $j = 1, \dots, l$. Then, there exists a neighbourhood $\tilde{O}$ of the point $\mathbf{0} \in \mathbb{R}^l$ and a function $\mathbf{x} \colon \tilde{O} \to \mathbb{X}$ of class C^1 such that $\mathbf{x}(\mathbf{0}) = \overline{\mathbf{x}}$ and $\mathbf{x}(\mathbf{z})$ is a strict local solution of the modified problem (**). Moreover,

$$D(f \circ \mathbf{x})(\mathbf{0}) = -\overline{\boldsymbol{\lambda}}^T.$$

<u>Proof.</u> By Theorem 48, the point $\overline{\mathbf{x}}$ is a solution of the system of equations

$$\begin{cases} D_x L(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = \mathbf{0}^T, \\ \mathbf{h}(\overline{\mathbf{x}}) = \mathbf{0}, \end{cases}$$

where

$$D_x L(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) = Df(\overline{\mathbf{x}}) + \overline{\boldsymbol{\lambda}}^T D\mathbf{h}(\overline{\mathbf{x}}) = Df(\overline{\mathbf{x}}) + \sum_{j=1}^{l} \overline{\lambda}_j Dh_j(\overline{\mathbf{x}}).$$

After adding the perturbation $\mathbf{z}$ to the right-hand side of the second equality we are going to show that there exists a solution being a function of class C^1 of the perturbation. Consider the system

$$\begin{cases} D_x L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}^T, \\ \mathbf{h}(\mathbf{x}) = \mathbf{z}, \end{cases}$$

with unknown $\boldsymbol{\lambda}$ and $\mathbf{x}$. We define the function $\mathbf{G} \colon \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^n \times \mathbb{R}^l$ by the formula

$$\mathbf{G}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{z}) = \begin{bmatrix} \big(D_x L(\mathbf{x}, \boldsymbol{\lambda})\big)^T \\ \mathbf{h}(\mathbf{x}) - \mathbf{z} \end{bmatrix}.$$

The modified system may be rewritten as $\mathbf{G}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{z}) = \mathbf{0}$.

We know that $\mathbf{G}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}, \mathbf{0}) = \mathbf{0}$. Using the implicit function theorem, we weave the first two variables as functions of the third one. To do that, we consider the matrix of derivatives of $\mathbf{G}$ (the blocks 0 and $-I$ are $l \times l$):

$$D\mathbf{G}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}, \mathbf{0}) = \begin{bmatrix} D^2 L(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) & \big(D\mathbf{h}(\overline{\mathbf{x}})\big)^T & 0 \\ D\mathbf{h}(\overline{\mathbf{x}}) & 0 & -I \end{bmatrix}.$$

The linear independence of the gradients of the constraints implies that the submatrix

$$D\mathbf{G}(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}, \mathbf{0}) = \begin{bmatrix} D^2 L(\overline{\mathbf{x}}, \overline{\boldsymbol{\lambda}}) & \big(D\mathbf{h}(\overline{\mathbf{x}})\big)^T \\ D\mathbf{h}(\overline{\mathbf{x}}) & 0 \end{bmatrix}.$$

is nonsingular (proof is an exercise). The assumptions of the implicit function theorem (Theorem 51) are, therefore, satisfied and there exists a neighbourhood O of the point $\mathbf{0} \in \mathbb{R}^l$ and functions $\mathbf{x} \colon O \to \mathbb{X}$ and $\boldsymbol{\lambda} \colon O \to \mathbb{R}^l$ of class C^1 such that for all $\mathbf{z} \in O$ there is $\mathbf{G}\big(\mathbf{x}(\mathbf{z}), \boldsymbol{\lambda}(\mathbf{z})\big) = \mathbf{0}$, i.e.,

$$D_x L\big(\mathbf{x}(\mathbf{z}), \boldsymbol{\lambda}(\mathbf{z})\big) = \mathbf{0}^T, \quad \mathbf{h}\big(\mathbf{x}(\mathbf{z})\big) = \mathbf{z}.$$

Using the fact that the functions $D_x^2 L$, $D\mathbf{h}$, $\mathbf{x}$, $\boldsymbol{\lambda}$ are continuous and the inequality ($\odot$) is satisfied for the original problem, we conclude that there exists a (possibly smaller) neighbourhood $\tilde{O}$ of $\mathbf{0} \in \mathbb{R}^l$ such that

$$\mathbf{d}^T D_x^2 L\big(\mathbf{x}(\mathbf{z}), \boldsymbol{\lambda}(\mathbf{z})\big) \mathbf{d} > 0$$

for all $z \in \tilde{O}$ and nonzero vectors $\mathbf{d} \in \mathbb{R}^n$ such that $Dh_j\big(\mathbf{x}(z)\big)\mathbf{d} = 0$, $j = 1, \dots, l$. The key to this result is that $(\odot)$ is a sharp inequality. By Theorem 56 the point $\mathbf{x}(z)$ is therefore a strict solution of the modified problem (**). Recall that $\mathbf{x}$ is a function of class C^1. Hence, we can define the derivative of the composition

$$D(f \circ \mathbf{x})(0) = Df(\overline{\mathbf{x}})D\mathbf{x}(0).$$

To complete the proof we need two observations. First, after multiplying the sides of the necessary condition for the original problem,

$$Df(\overline{\mathbf{x}}) + \overline{\boldsymbol{\lambda}}^T D\mathbf{h}(\overline{\mathbf{x}}) = \mathbf{0}^T$$

by $D\mathbf{x}(0)$ we obtain

$$Df(\overline{\mathbf{x}})D\mathbf{x}(0) + \overline{\boldsymbol{\lambda}}^T D\mathbf{h}(\overline{\mathbf{x}})D\mathbf{x}(0) = \mathbf{0}^T.$$

Second, by differentiating $\mathbf{h}\big(\mathbf{x}(z)\big)$ with respect to z, at $z = 0$ we obtain the following derivative: $D(\mathbf{h} \circ z)(0) = D\mathbf{h}(\overline{\mathbf{x}})D\mathbf{x}(0) = I$. The equation above may, therefore, be simplified to the following:

$$Df(\overline{\mathbf{x}})D\mathbf{x}(0) + \overline{\boldsymbol{\lambda}}^T = \mathbf{0}^T.$$

The claim follows immediately. □

Theorem 60 may be understood as follows: a small change of the j-th constraint from 0 to ε causes the change of the local minimum of f by $-\overline{\lambda}_j\varepsilon + o(\varepsilon) \approx -\overline{\lambda}_j\varepsilon$.

Inequality constraints

We use a different approach for inequality constraints. We focus on a convex optimization problem:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \dots, m, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \qquad (\overset{**}{*})$$

where $\mathbb{X} \subset \mathbb{R}^n$ is a convex set and the functions $f, g_1, \dots, g_n \colon \mathbb{X} \to \mathbb{R}$ are convex. To simplify the notation we denote $\mathbf{g}(\mathbf{x}) = \big(g_1(\mathbf{x}), \dots, g_m(\mathbf{x})\big)$. The problem may be rewritten as

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \\ \mathbf{x} \in \mathbb{X}, \end{cases}$$

Consider the modified problem

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{g}(\mathbf{x}) \leqslant z, \\ \mathbf{x} \in \mathbb{X}, \end{cases}$$

Definition 33 *Let D_M denote the set of vectors $z \in \mathbb{R}^m$ such that the feasible set for the modified problem, $W_z = \{\mathbf{x} \in \mathbb{X} \colon \mathbf{g}(z) \leqslant z\}$, is nonempty. The function*

$$M(z) = \inf_{\mathbf{x} \in \mathbb{X},\, \mathbf{g}(\mathbf{x}) \leqslant z} f(\mathbf{x}),$$

defined for all $z \in D_M$, is called the perturbation function and the set D_M is called the perturbation function domain.

Note that $M(z) < \infty$ for all $z \in D_M$, but it is possible that $M(z) = -\infty$.

The graph of the perturbation function in Figure 8 is the curve $M(y)$; as we can see it is a convex function. Note that this function is well defined between the points $A = (y_A, z_A)$ and $B = (y_B, z_B)$, because there exists $\mathbf{x} \in \mathbb{X}$ such that $\big(g(\mathbf{x}), f(\mathbf{x})\big) \in G$, and then $y = g(\mathbf{x})$ is an element of the domain of the function $M(y)$. If $y < y_A$, then the feasible set is empty and such points y are outside the domain of the perturbation function. For $y > y_B$ the perturbation function is a constant.

Theorem 61 *1. The set D_M is convex.*

2. The function $M \colon D_M \to \mathbb{R} \cup \{-\infty\}$ is convex.

3. If there exists a point $\mathbf{x}^* \in \mathbb{X}$ such that $\mathbf{g}(\mathbf{x}^*) < \mathbf{0}$, then $\operatorname{int} D_M \neq \emptyset$ and $\mathbf{0} \in \operatorname{int} D_M$.

Proof. From the convexity of each component of $\mathbf{g}$ it follows that

$$\begin{aligned} &\mathbf{g}(\mathbf{x}_1) \leqslant z_1,\ \mathbf{g}(\mathbf{x}_2) \leqslant z_2 \quad \Rightarrow \\ &\qquad \mathbf{g}\big(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\big) \leqslant \lambda z_1 + (1-\lambda)z_2 \quad \text{for all } \lambda \in [0, 1] \end{aligned} \qquad (\otimes)$$

(the inequalities are to be understood componentwise). This observation will be used a number of times in the proof below.

(1) Let $z_1, z_2 \in D_M$ and let $\lambda \in (0,1)$. There exist $x_1, x_2 \in \mathbb{X}$ such that $g(x_1) \leqslant z_1$ and $g(x_2) \leqslant z_2$. From $(\otimes)$ we obtain $g\big(\lambda x_1 + (1-\lambda)x_2\big) \leqslant \lambda z_1 + (1-\lambda)z_2$; hence, $\lambda z_1 + (1-\lambda)z_2 \in D_M$.

(2) Let $z_1, z_2 \in D_M$ and let $\lambda \in (0,1)$. Then,

$$\begin{aligned}
\lambda M(z_1) + (1-\lambda)M(z_2) &= \inf_{x_1 \in \mathbb{X},\, g(x_1) \leqslant z_1} \big(\lambda f(x_1)\big) + \inf_{x_2 \in \mathbb{X},\, g(x_2) \leqslant z_2} \big((1-\lambda) f(x_1)\big) \\
&= \inf_{\substack{x_1 \in \mathbb{X},\, g(x_1) \leqslant z_1 \\ x_2 \in \mathbb{X},\, g(x_2) \leqslant z_2}} \big(\lambda f(x_1) + (1-\lambda) f(x_2)\big) \\
&\geqslant \inf_{\substack{x_1 \in \mathbb{X},\, g(x_1) \leqslant z_1 \\ x_2 \in \mathbb{X},\, g(x_2) \leqslant z_2}} f\big(\lambda x_1 + (1-\lambda)x_2\big) \\
&\geqslant \inf_{x \in \mathbb{X},\, g(x) \leqslant \lambda z_1 + (1-\lambda) z_2} f(x).
\end{aligned}$$

The first inequality above is a consequence of the convexity of f and the second is obtained from $(\otimes)$, which implies the following:

$$\begin{aligned}
&\{\lambda x_1 + (1-\lambda)x_2 \colon x_1, x_2 \in \mathbb{X},\ g(x_1) \leqslant z_1,\ g(x_2) \leqslant z_2\} \subset \\
&\qquad \{x \in \mathbb{X} \colon g(x) \leqslant \lambda z_1 + (1-\lambda) z_2\}.
\end{aligned}$$

(3) We need to prove that the feasible set W_z is nonempty for any z from some neighbourhood of $0 \in \mathbb{R}^m$. By assumption, there exists $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for $i = 1, \ldots, m$. Let $a = \min\{-g_1(x^*), \ldots, -g_m(x^*)\}$. Then, for all $z \in [-a, a]^m$ we have $x^* \in W_z$; hence, $[-a, a]^m \subset D_M$. □

Remark. (1) If $M_{\overline{z}} = -\infty$ for some $\overline{z} \in D_M$, then by convexity of M for any $z \in D_M$ and $\lambda \in (0,1)$ we have $M\big(\lambda \overline{z} + (1-\lambda)z\big) = -\infty$.
(2) As a consequence, if $M(\overline{z}) = -\infty$ for some $\overline{z} \in D_M$, then $M(z) = -\infty$ for all $z \in \operatorname{int} D_M$.
(3) Moreover, if there exists $\overline{z} \in \operatorname{int} D_M$ such that $M(\overline{z}) > -\infty$, then $M(z) > -\infty$ for all $z \in D_M$.

Theorem 62 *If in a problem of convex optimization there exists a point $x^* \in \mathbb{X}$ such that $g_i(x^*) < 0$ for $i = 1, \ldots, m$ and $M(0) > -\infty$, then $M(z) > -\infty$ for all $z \in D_M$ and there exists a vector $\mu \in [0, \infty)^m$ which determines a supporting hyperplane of M:*

$$M(z) \geqslant M(0) - \mu^T z, \quad z \in D_M.$$

Proof. By Theorem 61, M is a convex function and $0 \in \operatorname{int} D_M$. Hence, by Remark (3) above, we have $M(z) > -\infty$ for all $z \in D_M$. The existence of the supporting plane is a consequence of the supporting plane theorem (Theorem 26):

$$M(z) \geqslant M(0) - \mu^T z, \quad z \in D_M,$$

for some $\mu \in \mathbb{R}^m$. We have to prove that all coordinates of μ are nonnegative. On the contrary, suppose that $\mu_i < 0$ for some $i \in \{1, \ldots, m\}$. As $0 \in \operatorname{int} D_M$, if a number a is sufficiently small, the point $\overline{z} = a e_i$ is an element of D_M. Due to μ_i being negative, we obtain

$$M(\overline{z}) \geqslant M(0) - \mu_i a > M(0).$$

On the other hand, $W_0 \subset W_{\overline{z}}$ (due to $\overline{z} \geqslant 0$, i.e., $M(\overline{z}) \leqslant M(0)$. This inconsistency proves that $\mu \in [0, \infty)^m$. □

The vector μ is called the sensitivity vector for the problem $(\overset{**}{*})$. By Theorem 26, if the function M is differentiable at 0, then $\mu = -\big(DM(0)\big)^T$. Therefore, μ denotes the speed and the direction of changes of the minimal value of f caused by a perturbation of constraints, just like in the case of equality constraints discussed earlier.

Now we take a look at the relation of the sensitivity vector with the saddle point and the first-order condition. Note that the connection of the saddle point and the sensitivity vector does not assume the convexity of the optimization problem.

Theorem 63 *1. If $(\overline{x}, \overline{\mu})$ is a saddle point of the Lagrange function in the set $\mathbb{X} \times [0, \infty)^m$, then $\overline{\mu}$ is a sensitivity vector (i.e., it determines a supporting plane). This claim does not assume convexity of the problem.*

2. Suppose that the functions $f, g_1, \ldots, g_m$ are differentiable at $\overline{x}$ and convex. If the first-order condition is satisfied at $\overline{x}$ with a Lagrange multipliers vector $\overline{\mu} \in [0, \infty)^m$, then $\overline{\mu}$ is the sensitivity vector.

Proof. (1) Let $L_z(x, \mu)$ denote the Lagrange function for a modified problem. Then,

$$L_z(x, \mu) = f(x) + \sum_{i=1}^m \mu_i \big(g_i(x) - z_i\big) = L(x, \mu) - \mu^T z.$$

The point $(\overline{x}, \overline{\mu})$ is a saddle point; hence,

$$M(0) = L(\overline{x}, \overline{\mu}) = \inf_{x \in \mathbb{X}} L(x, \overline{\mu}).$$

Therefore,

$$M(0) = \inf_{x \in \mathbb{X}} L(x, \overline{\mu}) = \inf_{x \in \mathbb{X}} \left(L_z(x, \overline{\mu}) + \overline{\mu}^\mathsf{T} z \right) = \inf_{x \in \mathbb{X}} L_z(x, \overline{\mu}) + \overline{\mu}^\mathsf{T} z. \tag{$\otimes$}$$

Note that for any $x \in W_z$ and $\mu \in [0, \infty)^m$ we have $f(x) \geqslant L_z(x, \mu)$; in particular,

$$M(z) = \inf_{x \in W_z} f(x) \geqslant \inf_{x \in W_z} L_z(x, \overline{\mu}) \geqslant \inf_{x \in \mathbb{X}} L_z(x, \overline{\mu}).$$

By substituting the above to $(\otimes)$, we obtain

$$M(0) \leqslant M(z) + \overline{\mu}^\mathsf{T} z.$$

(2) By Lemma 20, the point $(\overline{x}, \overline{\mu})$ is a saddle point of the Lagrange function. The claim may thus be obtained from the first claim. □

12. Introduction to numerical methods of optimization

So far, we studied the theory of optimization problems; more precisely, we proved theorems which *enable* finding candidates for solutions and recognising solutions among the candidates. Unfortunately, in most cases we need to solve systems of nonlinear equations. In practice, these systems may not have solutions possible to describe by algebraic formulae (and also the systems may be too large to be dealt with analytically). What may be done in such cases, is using numerical methods, which can find approximations of solutions with given accuracy. Some of those methods are described below. Many algorithms do not look for points satisfying the first-order necessary conditions; instead, they construct, step by step, sequences of points convergent to solutions. It does not mean that Lagrange multipliers and dual methods are useless in numerical computations. On the contrary, the classical approach with equality constraints uses extensively Lagrange multipliers, even if below we survey other methods.

Definition 34 *An iterative process is a four-tuple (Q, I, Ω, h), where Q is a set, $I \subset Q$, $\Omega \subset Q$ and $h\colon Q \to Q$ is a mapping in the set Q, which is an identity mapping of the subset Ω. This four-tuple represents a computation process; I is the set of initial data, Ω is the set of solutions and the function h describes the computation. Given an initial point $x \in I$, the process generates the sequence*

$$x_0 = x, \qquad x_{k+1} = h(x_k), \quad k = 0, 1, \ldots$$

The iterative process terminates after n steps if $x_n \in \Omega$ (in accordance to the definition of h, as in that case $x_{n+1} = x_n \in \Omega$). An algorithm is an iterative process which terminates after a finite number of steps.

In optimization we are interested in using algorithms to solve problems which have solutions. The algorithms ought to have the following properties:

1. Correctness of the algorithm, i.e., the property that for any admissible initial point $x \in I$ we get the correct result. For our purposes we assume that this property ensures the convergence of the sequence to a solution.
2. Stop property, which involves a condition for the iterative process to terminate. This condition is satisfied when an element of the sequence of points is in the set Ω, which means that it is an accurate enough approximation of the solution.

3. Effectiveness, related with the rate of convergence of the sequence to the solution.

General properties of optimization algorithms

Before discussing the algorithms we formulate the problem:

$$\begin{cases} f(x) \to \min, \\ x \in \mathbb{X} \subset \mathbb{R}^n. \end{cases}$$

Remark. For the algorithm to work it is necessary to provide procedures computing at given points x_k the function values $f(x_k)$ and sometimes its derivatives, $Df(x_k)$, $D^2f(x_k)$ etc. The evaluation of the function f and its derivatives is not a part of the algorithm; it is assumed that those procedures compute their results with some required accuracy. Sometimes they are seen as an "oracle" yielding the function values.

Definition 35 *Let x^* be a solution of an optimization problem and let $f^* = f(x^*)$. The stop criteria using an absolute tolerance at a level $\varepsilon > 0$ may be the following:*

1. $|f(x_k) - f^*| \leqslant \varepsilon$,
2. $\|x_k - x^*\| \leqslant \varepsilon$,
3. $\|Df(x_k)\| \leqslant \varepsilon$,
4. $|f(x_{k+1}) - f(x_k)| \leqslant \varepsilon$,
5. $\|x_{k+1} - x_k\| \leqslant \varepsilon$.

We can also use the stop criteria with a relative tolerance at a level $\varepsilon > 0$:

1. $|f(x_k) - f^*|/|f^*| \leqslant \varepsilon$,
2. $\|x_k - x^*\|/\|x^*\| \leqslant \varepsilon$,
3. $|f(x_{k+1}) - f(x_k)|/|f(x_k)| \leqslant \varepsilon$,
4. $\|x_{k+1} - x_k\|/\|x_k\| \leqslant \varepsilon$.

In both cases the first two criteria are the most natural, but they are the least practical, as we do not know the point $\boldsymbol{x}^*$ nor the minimal function value f^*. In practice we use the other criteria, even if they do not guarantee the termination of the process close to the actual solution of the problem.

Definition 36 *The greatest number* p *such that the inequality*

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leqslant c\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^p$$

is satisfied for all $k > K$ *(with some* $K \in \mathbb{N}$*), where* $c > 0$ *is a constant, is called the exponent of convergence of the algorithm.*

Remark. This definition is not easy to use in practice, because we do not know $\boldsymbol{x}^*$ (but it is useful in theoretical analysis of algorithms). The exponent of convergence may be measured *a posteriori* using the approximate criterion; p is the greatest number such that

$$\limsup_{k\to\infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^p} < \infty.$$

Note that a numerical computation does not produce the infinite sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, but only its finite initial subsequence.

Optimization of strictly quasi-convex functions

Definition 37 *Let* $W \subset \mathbb{R}^n$ *be a convex set. A function* $f\colon W \to \mathbb{R}$ *is strictly quasi-convex if for any* $\boldsymbol{x}, \boldsymbol{y} \in W$*,* $\boldsymbol{x} \neq \boldsymbol{y}$*, and* $\lambda \in (0,1)$ *there is*

$$f\bigl(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}\bigr) < \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}.$$

Lemma 22 1. *A strictly quasi-convex function has at most one minimum (local and global).*

2. *A strictly convex function is strictly quasi-convex.*

3. *A function defined on a line or on an* open *interval is strictly quasi-convex if it is strictly increasing, strictly decreasing, or there exists a point* $\bar{x}$ *in its domain such that this function is strictly decreasing for* $x < \bar{x}$ *and strictly increasing for* $x > \bar{x}$*.*

Proof is left as an exercise.

The strict quasi-convexity is a property making it possible to find a minimum in a closed interval without using derivatives. The main observation is made in the following lemma:

Lemma 23 *Let* $f\colon [a,b] \to \mathbb{R}$ *be a strictly quasi-convex function and let* $a \leqslant x \leqslant y \leqslant b$*.*

1. *If* $f(x) \geqslant f(y)$*, then* $f(z) > f(y)$ *for all* $z \in [a,x)$*.*
2. *If* $f(x) \leqslant f(y)$*, then* $f(z) > f(x)$ *for all* $z \in (y,b]$*.*

Proof. We prove the claim (1) by contradiction. Suppose that there exists $z \in [a,x)$ such that $f(z) \leqslant f(y)$. Then, by quasi-convexity of f, it follows that $f(x) < \max\{f(z), f(y)\} = f(y)$, which is inconsistent with the assumption that $f(x) \geqslant f(y)$. The claim (2) is proved in a similar way. □

Based on Lemma 23, we can construct many algorithms finding minima of strictly quasi-convex functions of one variable. First we take a look at the dichotomic subdivision algorithm. Its idea is quite simple: to find the minimum of a strictly quasi-convex function $f\colon [a,b] \to \mathbb{R}$, we choose two points, $\lambda < \mu$, in the interior of the interval $[a,b]$. By Lemma 23, we notice that if $f(\lambda) < f(\mu)$, then the interval with the minimum of f may be restricted to $[a,\mu]$ and if $f(\lambda) > f(\mu)$, then the interval may be restricted fo $[\lambda,b]$. The best strategy of choosing the two points is to obtain the next interval as short as possible. We do not know *a priori*, at which of the two points the function f takes a greater value. Therefore, to obtain the fastest convergence we should take into account the worst case and minimise the greater of the two numbers $\mu - a$, $b - \lambda$. This minimum is obtained with $\mu = \lambda = (a+b)/2$. As this solution does not yield two *different* points, we choose a small $\varepsilon_k > 0$ and we take $\lambda = (a+b)/2 - \varepsilon_k$ and $\mu = (a+b)/2 + \varepsilon_k$. The algorithm is the following:

Preparation: choose $\varepsilon_1 \in \bigl(0, (a+b)/2\bigr)$. Take $x_1 = a$, $y_1 = b$.

k-th step, repeated in a loop:

1. Compute $\lambda_k = (x_k + y_k)/2 - \varepsilon_k$ and $\mu_k = (x_k + y_k)/2 + \varepsilon_k$.

2. If $f(\lambda_k) < f(\mu_k)$, then take $x_{k+1} = x_k$ and $y_{k+1} = \mu_k$.
3. If $f(\lambda) \geqslant f(\mu_k)$, then take $x_{k+1} = \lambda_k$ and $y_{k+1} = y_k$.
4. Take $\varepsilon_{k+1} = \varepsilon_k/2$.

Stop condition: $y_{k+1} - x_{k+1} \leqslant 2\varepsilon$, where $\varepsilon > 0$ is the required accuracy (we can take the solution $\tilde{x} = (x_{k+1} + y_{k+1})/2$).

If a minimum $\bar{x}$ of f exists in $[a, b]$ (which *may not* be the case if f is not continuous), then, by Lemma 23, this minimum is located in the interval $[x_k, y_k]$ for all k. With $k \to \infty$ the lengths of these intervals tend to 0; note that $0 < \varepsilon_k < (y_{k-1} - x_{k-1})/2$ for all k. Hence, both sequences, $(x_k)_k$ and $(y_k)_k$, converge to the solution $\bar{x}$. This proves that the algorithm is correct.

If the iterations are terminated after the k-th step, then the approximation error of the exact solution by the midpoint $\tilde{x}$ of the interval $[x_{k+1}, y_{k+1}]$ is bounded by the half of length of this interval, which justifies the choice of the stop criterion.

We can prove that the lengths of the intervals $[x_k, y_k]$ are bounded by elements of a geometric sequence; there is

$$y_{k+1} - x_{k+1} \leqslant (b-a)c^k,$$

where $c = \frac{1}{2} + \frac{\varepsilon_1}{b-a}$. The exponent of convergence is $p = 1$; the convergence with such an exponent is called linear.

The dichotomic subdivision algorithm has to compute two function values in each step. The golden ratio algorithm computes just one function value in each step. The choice of points λ_k, μ_k in the interval $[x_k, y_k]$ is done as follows:

$$y_k - \lambda_k = \mu_k - x_k; \quad \text{hence, } \lambda_k = \tau x_k + (1-\tau)y_k,\ \ \mu_k = (1-\tau)x_k + \tau y_k,$$

where $\tau \in (0, 1)$ is such that

$$\lambda_k = (1-\tau)x_k + \tau\mu_k \quad \text{and} \quad \mu_k = \tau\lambda_k + (1-\tau)y_k.$$

The above determines the number τ; there is

$$\begin{aligned}\lambda_k &= \tau x_k + (1-\tau)y_k = (1-\tau)x_k + \tau\mu_k = (1-\tau)x_k + \tau\big((1-\tau)x_k + \tau y_k\big)\\ &= \tau x_k + y_k - \tau y_k = x_k - \tau^2 x_k + \tau^2 y_k.\end{aligned}$$

The last equality must hold for any x_k, y_k; hence,

$$\tau x_k = x_k - \tau^2 x_k, \quad y_k - \tau y_k = \tau^2 y_k,$$
$$\text{i.e., } (\tau^2 + \tau - 1)x_k = (\tau^2 + \tau - 1)y_k = 0.$$

The positive zero of the polynomial $p(\tau) = \tau^2 + \tau - 1$ is $\tau = (\sqrt{5} - 1)/2 \approx 0.618$; this number describes the golden ratio proportion. The golden ratio algorithm is the following:

Preparation: Take $x_1 = a$, $y_1 = b$, $\lambda_1 = \tau x_1 + (1-\tau)y_1$, $\mu_1 = (1-\tau)x_1 + \tau y_1$ and compute $f(\lambda_1)$ and $f(\mu_1)$.

k-th step, repeated in a loop:

1. If $f(\lambda_k) < f(\mu_k)$, then take $x_{k+1} = x_k$, $y_{k+1} = \mu_k$, $\lambda_{k+1} = \tau x_k + (1-\tau)y_k$, $\mu_{k+1} = \lambda_k$ and compute $f(\lambda_{k+1})$.
2. If $f(\lambda_k) \geqslant f(\mu_k)$, then take $y_{k+1} = y_k$, $x_{k+1} = \lambda_k$, $\mu_{k+1} = (1-\tau)x_k + \tau y_k$, $\lambda_{k+1} = \mu_k$ and compute $f(\mu_{k+1})$.

Stop condition: $y_{k+1} - x_{k+1} \leqslant 2\varepsilon$, where $\varepsilon > 0$ is the required accuracy.

The convergence of the sequences $(x_k)_k$, $(y_k)_k$ to the minimum $\bar{x}$ is slower than that of the sequences from the dichotomic subdivision algorithm; there is

$$y_{k+1} - x_{k+1} = (b-a)\tau^k.$$

However, due to the twice lower computational cost of one iteration, the golden ratio algorithm achieves the same accuracy of the result in a shorter time.

The slightly more complicated Fibonacci algorithm guarantees finding the minimum with the assumed accuracy after evaluating the function at the minimal number of points sufficient to minimise *any* strictly quasi-convex function.

Remark. (1) One can prove that the golden ratio algorithm may need to evaluate the function at most at one point more than the Fibonacci algorithm.
(2) There exist algorithms working faster for functions f having some properties in addition to the strict quasi-convexity, like differentiablility. For example, having a procedure of computing the derivative, we can use the secant method to find the zero of f'; if f'' satisfies the Lipschitz condition, then the exponent of convergence is $p = 1 + \tau \approx 1.618$.

The Fibonacci sequence is defined as follows:

$$F_0 = F_1 = 1,$$
$$F_{k+1} = F_k + F_{k-1}, \quad k = 1, 2, 3, \ldots$$

To use this algorithm we fix *a priori* the number of iterations, which is related with the desired accuracy in the way explained later. The numbers λ_k and μ_k are obtained from the formulae

$$\begin{cases} \lambda_k = x_k + \dfrac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k), \\ \mu_k = x_k + \dfrac{F_{n-k}}{F_{n-k+1}}(y_k - x_k), \end{cases} \qquad k = 1, \dots, n-1.$$

If $f(\lambda_k) \geqslant f(\mu_k)$, then we take $x_{k+1} = \lambda_k$ and $y_{k+1} = y_k$. The rules written above produce

$$\begin{aligned} \lambda_{k+1} &= x_{k+1} + \frac{F_{n-k-2}}{F_{n-k}}(y_{k+1} - x_{k+1}) \\ &= x_k + \frac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k) + \frac{F_{n-k-2}}{F_{n-k}}\left(y_k - x_k - \frac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k)\right) \\ &= x_k + \left(\frac{F_{n-k-1}}{F_{n-k+1}} + \frac{F_{n-k-2}}{F_{n-k}} - \frac{F_{n-k-2}}{F_{n-k}}\frac{F_{n-k-1}}{F_{n-k+1}}\right)(y_k - x_k) \\ &= x_k + \frac{F_{n-k-1}F_{n-k} + F_{n-k-2}F_{n-k+1} - F_{n-k-2}F_{n-k-1}}{F_{n-k+1}F_{n-k}}(y_k - x_k) \\ &= x_k + \frac{F_{n-k-1}F_{n-k} + F_{n-k-2}F_{n-k}}{F_{n-k+1}F_{n-k}}(y_k - x_k) \\ &= x_k + \frac{F_{n-k}}{F_{n-k+1}}(y_k - x_k) = \mu_k. \end{aligned}$$

If $f(\lambda_k) < f(\mu_k)$, then we take $x_{k+1} = x_k$ and $y_{k+1} = \mu_k$. A similar calculation proves that in this case

$$\mu_{k+1} = \lambda_k,$$

which means that in both cases we choose only one point in the interval $[x_k, y_k]$ at which the function value has not yet been computed.

Suppose that we need to find the minimum in the interval $[a, b]$ with the accuracy not worse than some $\varepsilon > 0$. The number of steps is $n - 1$, where n is the smallest number such that $F_n \geqslant (b - a)/\varepsilon$. Note that $\lambda_{n-1} = \mu_{n-1} = \frac{1}{2}(x_{n-1} + y_{n-1})$; the point obtained in last step is the midpoint of the interval $[x_{n-1}, y_{n-1}]$, whose length is less than or equal to $F_2\varepsilon$.

The Fibonacci algorithm is the following:

Preparation: Find the smallest n such that $F_n \geqslant (b - a)/\varepsilon$;
take $x_1 = a$, $y_1 = b$, $\lambda_1 = \big(F_{n-1}x_1 + (F_n - F_{n-1})y_1\big)/F_n$,
$\mu_1 = \big((F_n - F_{n-1})x_1 + F_{n-1}y_1\big)/F_n$ and compute $f(\lambda_1)$ and $f(\mu_1)$.

k-th step, repeated in a loop for $k = 1, \dots, n-2$:

1. If $f(\lambda_k) < f(\mu_k)$, then take $x_{k+1} = x_k$, $y_{k+1} = \mu_k$, $\mu_{k+1} = \lambda_k$, $\lambda_{k+1} = \big(F_{n-k}x_k + (F_{n-k+1} - F_{n-k})y_k\big)/F_{n-k+1}$ and compute $f(\lambda_{k+1})$.
2. If $f(\lambda_k) \geqslant f(\mu_k)$, then take $y_{k+1} = y_k$, $x_{k+1} = \lambda_k$, $\lambda_{k+1} = \mu_k$, $\mu_{k+1} = \big((F_{n-k+1} - F_{n-k})x_k + F_{n-k}y_k\big)/F_{n-k+1}$ and compute $f(\mu_{k+1})$.

The last step: If $f(\lambda_{n-1}) < f(\mu_{n-1})$, then take $\tilde{x} = \lambda_{n-1}$, else take $\tilde{x} = \mu_{n-1}$. The point $\tilde{x}$ is an approximation of the minimum with the required accuracy.

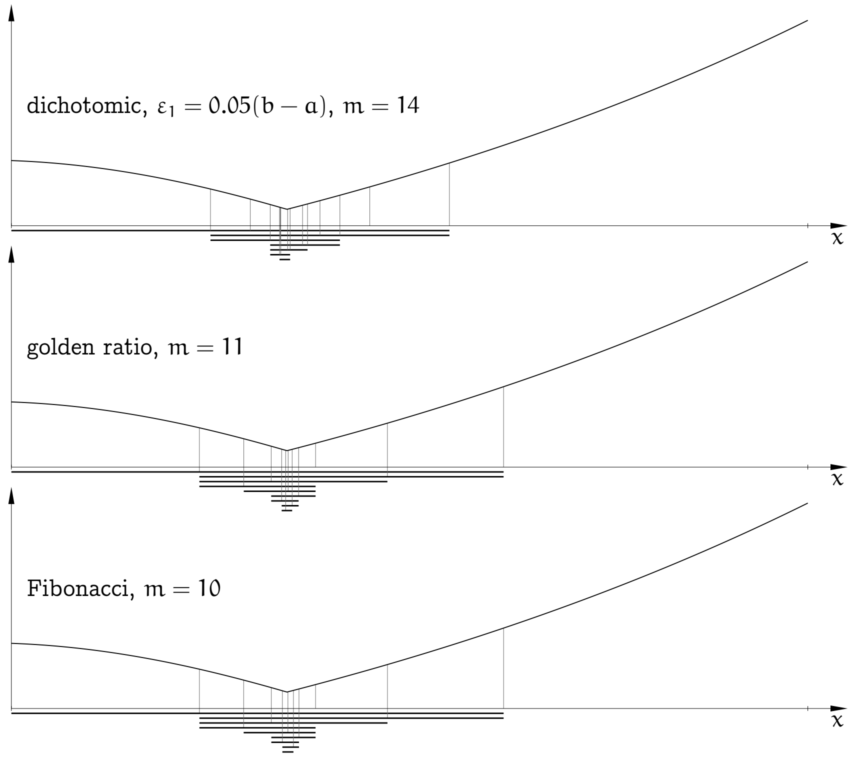

Figure 9: Three algorithms in action

An example is shown in Figure 9; the interval $[a, b] = [0, 1]$ and the assumed accuracy is 0.01. Below the x axis of each graph the subsequent intervals $[x_k, y_k]$ are drawn; m is the number of points at which the function value was computed.

13. Algorithms for unconstrained optimization

Now we take a look at multidimensional optimization methods for problems without constraints. We assume that the target function $f\colon \mathbb{R}^n \to \mathbb{R}$ is of class C^1 and its minimum is to be found. All methods considered here are based on the following scheme: beginning at a point $\mathbf{x}_1$, which we believe to be close enough to the solution, we construct a sequence of points $\mathbf{x}_1, \mathbf{x}_2, \ldots$ so as to obtain $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ for all k. We expect to find a minimum of f in this way. However, it may turn out that the concentration points of the sequence $(\mathbf{x}_k)_k$ are not solutions. We survey various methods of constructing sequences $(\mathbf{x}_k)_k$, paying attention to their convergence. All results described below apply also to functions defined in open subsets of $\mathbb{R}^n$. Descent methods will play first fiddle in optimization problems with constraints, considered later.

We consider the problem

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Descent methods are algorithms constructing consecutive points according to the formula

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

where $\alpha_k > 0$ and the vector $\mathbf{d}_k$ has a descent direction, i.e.,

$$\begin{aligned} Df(\mathbf{x}_k)\mathbf{d}_k < 0 &\quad \text{if } Df(\mathbf{x}_k) \neq \mathbf{0}^T, \\ \mathbf{d}_k = \mathbf{0} &\quad \text{if } Df(\mathbf{x}_k) = \mathbf{0}^T. \end{aligned}$$

With $\mathbf{d}_k \neq \mathbf{0}$ and α_k sufficiently small we have $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k$. We might expect the sequence $(\mathbf{x}_k)_k$ to converge to a minimum. But without additional information about the problem we have no guarantee of finding a global minimum. We are going to look for methods of making the sequence converge to a local minimum.

Steepest descent methods

The steepest descent methods take the vectors $\mathbf{d}_k$ parallel to $\big(Df(\mathbf{x}_k)\big)^T$. The idea is simple:

Preparation: Choose an initial point $\mathbf{x}_1$.

k-th step:

1. Choose α_k,
2. Take $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k\big(Df(\mathbf{x}_k)\big)^T$.

Stop condition: $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

It remains to select rules of choosing α_k. The rules below will also be used with any vectors $\mathbf{d}_k$ having descent directions, not necessarily $-\big(Df(\mathbf{x}_k)\big)^T$.

- Exact minimization rule: choose α_k such that

  $$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = \min_{\alpha \geqslant 0} f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$

- Limited minimization rule: with a fixed $A > 0$ choose α_k such that

  $$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = \min_{\alpha \in [0,A]} f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$

- Armijo rule: with fixed $s > 0$, $\beta, \sigma \in (0,1)$ we take $\alpha_k = \beta^{m_k} s$, where m_k is the smallest integer m such that

  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta^m s \mathbf{d}_k) \geqslant -\sigma\beta^m s Df(\mathbf{x}_k)\mathbf{d}_k.$$

  According to this rule, the following inequality holds:

  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta^{m-1} s \mathbf{d}_k) < -\sigma\beta^{m-1} s Df(\mathbf{x}_k)\mathbf{d}_k.$$

  The constant s is called a stride, β controls the rate of decreasing or increasing the stride (the smaller it is, the faster the stride changes) and σ influences the choice of α_k as follows: the smaller it is, the smaller m_k satisfies the conditions above, resulting in the greater value of α_k.

  In the steepest descent methods, where $\mathbf{d}_k = -\big(Df(\mathbf{x}_k)\big)^T$, the choice of m_k is slightly simpler due to $Df(\mathbf{x}_k)\mathbf{d}_k = -\|Df(\mathbf{x}_k)\|^2$: m_k is the smallest integer m such that

  $$f(\mathbf{x}_k) - f\big(\mathbf{x}_k - \beta^m s\big(Df(\mathbf{x}_k)\big)^T\big) \geqslant \sigma\beta^m s\|Df(\mathbf{x}_k)\|^2. \qquad (*)$$

Figure 10 shows the idea of the Armijo rule, which is easy to implement. We assume that $Df(\mathbf{x}_k) \neq \mathbf{0}^T$. Starting from the point $\mathbf{y} = \mathbf{x}_k - s\big(Df(\mathbf{x}_k)\big)^T$ we check the condition (*). If it does not hold, then we check the points $\mathbf{x}_k - s\beta\big(Df(\mathbf{x}_k)\big)^T$,

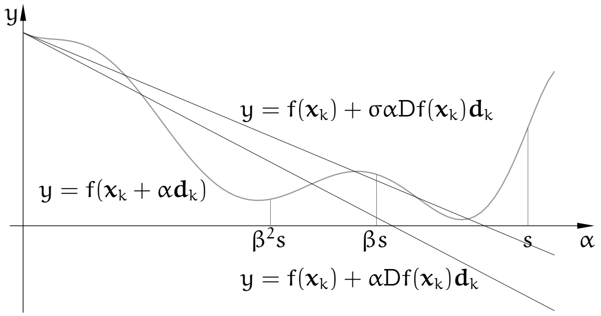

Figure 10: The Armijo rule in action; $\sigma = 0.8$, $\beta = 0.7$, $\mathbf{x}_{k+1} = \mathbf{x}_k + \beta^2 s\mathbf{d}_k$

$\mathbf{x}_k - s\beta^2\big(Df(\mathbf{x}_k)\big)^T$ etc. If (*) is satisfied with $m = 0$, corresponding to the point $\mathbf{y}$, we try the points $\mathbf{x}_k - s\beta^{-1}\big(Df(\mathbf{x}_k)\big)^T$, $\mathbf{x}_k - s\beta^{-2}\big(Df(\mathbf{x}_k)\big)^T$ etc. After a finite number of steps we find the number m satisfying (*) such that $m - 1$ does not satisfy this condition.

The exact minimization rule is well defined if the minimum on the right-hand side exists. The interval in which the minimum is to be found is unbounded; hence, the exact minimization rule does not have to be well defined. Later we show assumptions that guarantee its correctness. Restricting the interval to $[0, A]$ has two advantages: first, the problem always has a solution, as we minimise a continuous function in a compact set. Second, we can use faster methods of searching minima.

Both rules, exact and limited, find steps such that $f(\mathbf{x}_{k+1}) \leqslant f(\mathbf{x}_k)$; the proof of this fact is an exercise. By Inequality (*), also the Armijo rule produces such steps. There are, however, two questions. Is the stop criterion correct? Does the sequence $(\mathbf{x}_k)_k$ converge to a minimum?

Below we prove that any concentration point of the sequence $(\mathbf{x}_k)$ obtained by the steepest descent method is a critical point, i.e., it is a zero of the gradient of f. Such a point does not have to be a local minimum. However, if the function f is pseudoconvex, then we can be sure that it is a minimum, moreover, a global minimum.

Theorem 64 *Let $(\mathbf{x}_k)_k$ be a sequence of points obtained using the steepest descent method, with any of the three rules described above. Then, any concentration point of this sequence is a critical point.*

Proof. Let $\overline{\mathbf{x}}$ be a concentration point and let $(\mathbf{x}_{k_n})_n$ be a subsequence convergent to $\overline{\mathbf{x}}$. The proof is done by contradiction; we suppose that $Df(\overline{\mathbf{x}}) \neq 0^T$.

The main idea for the first two rules (exact and limited) is to show the existence of a constant $\gamma > 0$ such that with n big enough, i.e., with $\mathbf{x}_{k_n}$ close enough to $\overline{\mathbf{x}}$, there is

$$f(\mathbf{x}_{k_n+1}) \leqslant f(\mathbf{x}_{k_n}) - \gamma,$$

which means that it is possible to decrease the value of f in the step k_n at least by γ. Bearing in mind that the sequence $\big(f(\mathbf{x}_k)\big)_k$ is decreasing, we have

$$f(\mathbf{x}_{k_{n+1}}) \leqslant f(\mathbf{x}_{k_n+1}) \leqslant f(\mathbf{x}_{k_n}) - \gamma.$$

By passing with n to ∞ and using the fact that $\overline{\mathbf{x}}$ is a concentration point and f is continuous, we obtain the inconsistency with the inequality $\gamma > 0$.

Therefore we need to prove the existence of $\gamma > 0$. The derivative of f is continuous by assumption; hence, there exists a neighbourhood V of $\overline{\mathbf{x}}$ such that

$$\frac{\|Df(\mathbf{x}) - Df(\mathbf{y})\|}{\|Df(\mathbf{x})\|} \leqslant \frac{1}{2} \quad \text{for all } \mathbf{x}, \mathbf{y} \in V. \tag{**}$$

The main observation to justify the inequality above is that in a neighbourhood of $\overline{\mathbf{x}}$ the norm of gradient of f is strictly separated from 0.

Let $\delta > 0$ be such that $B(\overline{\mathbf{x}}, 2\delta) \subset V$ and $\delta \leqslant A \inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\|$, where A is the constant of the definition of the limited minimization rule (from (**), we conclude that $\inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\| > 0$). For a point $\mathbf{x} \in B(\overline{\mathbf{x}}, \delta)$, the point $\mathbf{x} - \frac{\delta}{\|Df(\mathbf{x})\|}\big(Df(\mathbf{x})\big)^T$ is an element of $B(\overline{\mathbf{x}}, 2\delta)$ and, therefore, also of V. Moreover, with the limited minimization rule we have

$$\frac{\delta}{\|Df(\mathbf{x})\|} \leqslant \frac{A \inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\|}{\|Df(\mathbf{x})\|} \leqslant A,$$

i.e., $\frac{\delta}{\|Df(\mathbf{x})\|} \in [0, A]$. By the mean value theorem,

$$\begin{aligned} f(\mathbf{x}) - f\Big(\mathbf{x} - \frac{\delta}{\|Df(\mathbf{x})\|}\big(Df(\mathbf{x})\big)^T\Big) &= Df(\boldsymbol{\theta})\Big(\frac{\delta}{\|Df(\mathbf{x})\|}\big(Df(\mathbf{x})\big)^T\Big) \\ &= \frac{\delta}{\|Df(\mathbf{x})\|} Df(\boldsymbol{\theta})\big(Df(\mathbf{x})\big)^T, \end{aligned}$$

where $\boldsymbol{\theta}$ is an intermediate point; hence, it is an element of V. Now we focus on the product of the gradients:

$$\begin{aligned} Df(\boldsymbol{\theta})\big(Df(\mathbf{x})\big)^T &= \big(Df(\mathbf{x}) + Df(\boldsymbol{\theta}) - Df(\mathbf{x})\big)\big(Df(\mathbf{x})\big)^T \\ &= \|Df(\mathbf{x})\|^2 + \big(Df(\boldsymbol{\theta}) - Df(\mathbf{x})\big)\big(Df(\mathbf{x})\big)^T \\ &\geqslant \|Df(\mathbf{x})\|^2 - \|Df(\boldsymbol{\theta}) - Df(\mathbf{x})\|\|Df(\mathbf{x})\|. \end{aligned}$$

We substitute this estimation to the previous formula:

$$f(\boldsymbol{x}) - f\Big(\boldsymbol{x} - \frac{\delta}{\|Df(\boldsymbol{x})\|}\big(Df(\boldsymbol{x})\big)^T\Big) \geqslant \delta\|Df(\boldsymbol{x})\| - \delta\|Df(\boldsymbol{\theta}) - Df(\boldsymbol{x})\|$$
$$= \delta\|Df(\boldsymbol{x})\|\Big(1 - \frac{\|Df(\boldsymbol{\theta}) - Df(\boldsymbol{x})\|}{\|Df(\boldsymbol{x})\|}\Big) \geqslant \frac{\delta}{2}\|Df(\boldsymbol{x})\|;$$

the last inequality follows from (**). The estimation above is true for all $\boldsymbol{x} \in B(\overline{\boldsymbol{x}}, \delta)$; in particular for all $\boldsymbol{x}_{k_n}$ with n big enough we have

$$f(\boldsymbol{x}_{k_n}) - f(\boldsymbol{x}_{k_n+1}) \geqslant f(\boldsymbol{x}_{k_n}) - f\Big(\boldsymbol{x}_{k_n} - \frac{\delta}{\|Df(\boldsymbol{x}_{k_n})\|}\big(Df(\boldsymbol{x}_{k_n})\big)^T\Big) \geqslant \frac{\delta}{2}\|Df(\boldsymbol{x}_{k_n})\|.$$

Therefore, we can take

$$\gamma = \frac{\delta}{2}\inf_{x \in B(\overline{x},\delta)} \|Df(\boldsymbol{x})\|.$$

For the steepest descent method with the Armijo rule we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geqslant \sigma\alpha_k\|Df(\boldsymbol{x}_k)\|^2,$$

which implies that the sequence $\big(f(\boldsymbol{x}_k)\big)_k$ is monotonically decreasing, i.e., it is either convergent, or it diverges to $-\infty$. Due to $f(\boldsymbol{x}_{k_n}) \to f(\overline{\boldsymbol{x}})$ there is $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \to 0$. It follows that also $\alpha_k\|Df(\boldsymbol{x}_k)\|^2 \to 0$. Due to $Df(\boldsymbol{x}_{k_n}) \to Df(\overline{\boldsymbol{x}}) \neq \mathbf{0}^T$, there must be $\alpha_k \to 0$.

On the other hand, the numbers α_k in the Armijo rule are chosen in the optimal way, i.e.,

$$f(\boldsymbol{x}_k) - f\big(\boldsymbol{x}_k - \alpha_k/\beta\big(Df(\boldsymbol{x}_k)\big)^T\big) < \sigma\alpha_k/\beta\|Df(\boldsymbol{x}_k)\|^2.$$

By applying the intermediate value theorem to the left-hand side of this inequality, we get

$$f(\boldsymbol{x}_k) - f\big(\boldsymbol{x}_k - \alpha_k/\beta\big(Df(\boldsymbol{x}_k)\big)^T\big) = Df\big(\boldsymbol{x}_k - \tilde{\alpha}_k/\beta\big(Df(\boldsymbol{x}_{k_n})\big)^T\big)\alpha_k/\beta\big(Df(\boldsymbol{x}_{k_n})\big)^T.$$

Therefore, the previous inequality takes the form

$$Df\big(\boldsymbol{x}_k - \tilde{\alpha}_k/\beta\big(Df(\boldsymbol{x}_k)\big)^T\big)\big(Df(\boldsymbol{x}_k)\big)^T < \sigma\|Df(\boldsymbol{x}_k)\|^2$$

Taking the subsequence $(\boldsymbol{x}_{k_n})_n$, we obtain

$$Df\big(\boldsymbol{x}_{k_n} - \tilde{\alpha}_{k_n}/\beta\big(Df(\boldsymbol{x}_{k_n})\big)^T\big)\big(Df(\boldsymbol{x}_{k_n})\big)^T < \sigma\|Df(\boldsymbol{x}_{k_n})\|^2$$

If we pass with this inequality to the limit, then we see that $\tilde{\alpha}_{k_n} \in [0, \alpha_{k_n}]$; $\tilde{\alpha}_{k_n} \to 0$ because $\alpha_{k_n} \to 0$. The last inequality at the limit gives us $\|Df(\overline{\boldsymbol{x}})\|^2 \leqslant \sigma\|Df(\overline{\boldsymbol{x}})\|^2$, i.e., $(1-\sigma)\|Df(\overline{\boldsymbol{x}})\|^2 \leqslant 0$. But $1 - \sigma \in (0, 1)$, which gives us an inconsistency with the supposition that $Df(\overline{\boldsymbol{x}}) \neq \mathbf{0}^T$. □

It turns out that slightly stronger assumptions ensure that $Df(\boldsymbol{x}_k) \to \mathbf{0}^T$ for the entire sequence obtained by the steepest descent method, not just a subsequence convergent to the concentration point.

Lemma 24 *Let f be a function of class C^1 bounded from below. Let $(\boldsymbol{x}_k)_k$ be a sequence of points obtained by a steepert descent method. If there exists a constant $c > 0$ (independent of k) such that*

$$f(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k) < f(\boldsymbol{x}_k) - c\|Df(\boldsymbol{x}_k)\|^2, \quad k = 1, 2, \ldots, \qquad (\substack{**\\*})$$

then either there exists a number K such that $Df(\boldsymbol{x}_K) = \mathbf{0}^T$, or the sequence $\big(Df(\boldsymbol{x}_k)\big)_k$ converges to $\mathbf{0}^T$.

Proof. Suppose that there exists an infinite sequence $(\boldsymbol{x}_k)_k$ obtained by a steepest descent method (i.e., a number K as described, does not exist). By $(\substack{**\\*})$, the sequence $\big(f(\boldsymbol{x}_k)\big)_k$ is monotonically decreasing. Being bounded from below, this sequence is convergent; hence, $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \to 0$. Due to $(\substack{**\\*})$, we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) > c\|Df(\boldsymbol{x}_k)\|^2,$$

therefore, $\|Df(\boldsymbol{x}_k)\| \to 0$. □

Theorem 65 *Let f be a function bounded from below and let its gradient satisfy the Lipschitz condition with a constant L in a sublevel set $S = W_{f(x_1)}(f)$ for some $\boldsymbol{x}_1$. Let $(\boldsymbol{x}_k)_k$ be a sequence obtained using the steepest descent method with the exact rule (which is correct if the set S is compact), limited rule with $A > \frac{1}{2L}$ or the Armijo rule. Then, either there exists a number K such that $Df(\boldsymbol{x}_K) = \mathbf{0}^T$, or the sequence $\big(Df(\boldsymbol{x}_k)\big)_k$ converges to $\mathbf{0}^T$.*

Proof. The proof is done by showing that the assumptions of the theorem imply the assumptions of Lemma 24.

At first we consider the Armijo rule. Let $\alpha_k = s\beta^{m_k}$ be the step chosen by this rule in the k-th step. Then,

$$f(\boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + \sigma\alpha_k Df(\boldsymbol{x}_k)\boldsymbol{d}_k. \qquad (\oplus)$$

The number α_k has been chosen in the optimal way, i.e., with the lowest power of β to satisfy the last inequalty. Therefore, for an even lower power of β we have

$$f(\mathbf{x}_k + \beta^{-1}\alpha_k\mathbf{d}_k) > f(\mathbf{x}_k) + \sigma\alpha_k Df(\mathbf{x}_k)\mathbf{d}_k. \qquad (\otimes)$$

Due to $Df(\mathbf{x}_k)\mathbf{d}_k < 0$ and $\mathbf{x}_k \in S$, using $(\oplus)$ we obtain $\mathbf{x}_{k+1} \in S$. By the intermediate value theorem, we obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = f(\mathbf{x}_k + \alpha_k\mathbf{d}_k) - f(\mathbf{x}_k) = \alpha_k Df(\tilde{\mathbf{x}}_k)\mathbf{d}_k,$$

where $\tilde{\mathbf{x}}_k$ is a point of the line segment $\overline{\mathbf{x}_k\mathbf{x}_{k+1}}$. This line segment is a subset of S; in particular, $\tilde{\mathbf{x}}_k \in S$. Taking into account that $\mathbf{d}_k = -\big(Df(\mathbf{x}_k)\big)^T$, we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &= \alpha Df(\tilde{\mathbf{x}}_k)\mathbf{d}_k = -\alpha_k\big(Df(\mathbf{x}_k) - Df(\mathbf{x}_k) + Df(\tilde{\mathbf{x}}_k)\big)\big(Df(\tilde{\mathbf{x}}_k)\big)^T \\ &= -\alpha_k\|Df(\mathbf{x}_k)\|^2 + \alpha_k\big(Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}}_k)\big)\big(Df(\mathbf{x}_k)\big)^T \\ &\leqslant -\alpha_k\|Df(\mathbf{x}_k)\|^2 + \alpha_k\|Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}}_k)\|\|Df(\mathbf{x}_k)\|. \end{aligned}$$

There is $\mathbf{x}_k \in S$ and also $\tilde{\mathbf{x}}_k \in S$ for all k. From the Lipschitz condition satisfied by the gradient of f it follows that

$$\|Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}}_k)\| \leqslant L\|\mathbf{x}_k - \tilde{\mathbf{x}}_k\| \leqslant L\|\mathbf{x}_k - \mathbf{x}_{k+1}\| = L\alpha_k\|Df(\mathbf{x}_k)\|.$$

After substituting this estimation to the previous inequality, we obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\alpha_k\|Df(\mathbf{x}_k)\|^2(1 - \alpha_k L).$$

It follows from the above that $(\oplus)$ is satisfied if $1 - \alpha_k L \geqslant \sigma$. Let m_σ be such a number that $1 - s\beta^{m_\sigma}L \geqslant \sigma > 1 - s\beta^{m_\sigma - 1}$. It follows that

$$\sigma s\beta^{m_\sigma} > \frac{\sigma(1-\sigma)\beta}{L}.$$

On the other hand, m_σ does not have to be optimal, satisfying $(\oplus)$ and $(\otimes)$. However, the Armijo rule guarantees that $m_k \leqslant m_\sigma$, because m_k is the smallest number such that $(\oplus)$ is satisfied. Hence,

$$\sigma s\beta^{m_k} \geqslant \sigma s\beta^{m_\sigma} > \frac{\sigma(1-\sigma)\beta}{L}.$$

After substituting the above to $(\oplus)$ and using the fact that $\mathbf{d}_k$ has the steepest descent direction, we obtain the following inequality:

$$\begin{aligned} f(\mathbf{x}_k - \alpha_k\mathbf{d}_k) - f(\mathbf{x}_k) &\leqslant -\sigma s\beta^{m_k}\|Df(\mathbf{x}_k)\|^2 \leqslant -\sigma s\beta^{m_\sigma}\|Df(\mathbf{x}_k)\|^2 \\ &< -\frac{\sigma(1-\sigma)\beta}{L}\|Df(\mathbf{x}_k)\|^2, \end{aligned}$$

i.e., the inequality $(\substack{**\\*})$ of Lemma 24.

With the exact minimization rule, we assume that it is well-defined in each step, i.e., there exists a finite number $\alpha_k \geqslant 0$ to minimise $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ for $\alpha \geqslant 0$.
A calculation similar to that in the proof for the Armijo rule yields the estimation

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\alpha_k\|Df(\mathbf{x}_k)\|^2(1 - \alpha_k L). \qquad (\substack{**\\**})$$

The right-hand side of this inequality is minimal for $\alpha_k = \frac{1}{2L}$. From that we obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\frac{1}{4L}\|Df(\mathbf{x}_k)\|^2,$$

i.e., the inequality $(\substack{**\\*})$ from Lemma 24.

The proof for the restricted minimization rule is similar to the above. We need to assume that the interval $[0, A]$ in which we search for α_k contains the point $\frac{1}{2L}$ to minimise $(\substack{**\\**})$. □

Theorem 65 justifies correctness of the steepest descent methods, but it does not refer to the rate of their convergence. After further strengthtening its assumptions, we can prove that the convergence is at least linear and the stop criterion is correct.

We define the set $S = \{\mathbf{x} \in \mathbb{R}^n\colon f(\mathbf{x}) \leqslant f(\mathbf{x}_1)\}$, where $\mathbf{x}_1$ is the initial point. By $m(\mathbf{x})$ and $M(\mathbf{x})$ we denote respectively the smallest and the greatest eigenvalue of the Hessian matrix of f at $\mathbf{x}$.

Lemma 25 *Suppose that the set S is convex and compact and the function f is twice continuously differentiable in S. Let $m = \inf_{\mathbf{x}\in S} m(\mathbf{x}) > 0$. Then the point $\overline{\mathbf{x}}$ being the limit of the sequence $(\mathbf{x}_k)_k$ obtained by a steepest descent method is in S. It is the minimum of f in S and for all $\mathbf{x} \in S$ there is*

$$\|\mathbf{x} - \overline{\mathbf{x}}\| \leqslant \frac{1}{m}\|Df(\mathbf{x})\|, \quad f(\mathbf{x}) - f(\overline{\mathbf{x}}) \leqslant \frac{1}{m}\|Df(\mathbf{x})\|^2.$$

Proof. Due to $m = \inf_{\mathbf{x}\in S} m(\mathbf{x}) > 0$, the Hessian $D^2f(\mathbf{x})$ is positive-definite at all points of S and it follows that f is a convex function in S.

The set S is compact; the function f is bounded from below and its gradient satisfies the Lipschitz condition. By Theorem 65, $Df(\mathbf{x}_k) \to \mathbf{0}$, i.e., $\overline{\mathbf{x}} \in S$. The critical point of a convex function is its minimum.

The Taylor formula for $x \in S$ gives us

$$f(x) = f(\bar{x}) + \frac{1}{2}(x-\bar{x})^T D^2(\tilde{x})(x-\bar{x}),$$
$$f(\bar{x}) = f(x) + Df(x)(\bar{x}-x) + \frac{1}{2}(\bar{x}-x)^T D^2(\hat{x})(\bar{x}-x).$$

After substituting the first equality to the second, we obtain

$$Df(x)(\bar{x}-x) + (x-\bar{x})^T \frac{D^2f(\hat{x}) - D^2f(\tilde{x})}{2}(x-\bar{x}) = 0.$$

We can, therefore, estimate

$$Df(x)(x-\bar{x}) = (x-\bar{x})^T \frac{D^2f(\hat{x}) - D^2f(\tilde{x})}{2}(x-\bar{x}) \geqslant m\|x-\bar{x}\|^2,$$

and then

$$\|Df(x)\|\|x-\bar{x}\| \geqslant (x-\bar{x})^T \frac{D^2f(\hat{x}) - D^2f(\tilde{x})}{2}(x-\bar{x}) \geqslant m\|x-\bar{x}\|^2.$$

After dividing the sides of the last inequality by $m\|x-\bar{x}\|$, we obtain

$$\|x-\bar{x}\| \leqslant \frac{1}{m}\|Df(x)\|.$$

Using the Taylor formula again, due to the convexity of f in S, we obtain

$$f(\bar{x}) = f(x) + Df(x)(\bar{x}-x) + \frac{1}{2}(\bar{x}-x)^T D^2f(\hat{x})(\bar{x}-x) \geqslant f(x) + Df(x)(\bar{x}-x).$$

Hence,

$$f(x) - f(\bar{x}) \leqslant Df(x)(x-\bar{x}),$$

i.e.,

$$f(x) - f(\bar{x}) = |f(x) - f(\bar{x})| \leqslant \|Df(x)\|\|x-\bar{x}\| \leqslant \frac{1}{m}\|Df(x)\|^2. \quad \square$$

Lemma 26 *Let S be convex and compact. Let f be of class C^2 in s and $m = \inf_{x\in S} m(x)$. Denote $M = \sup_{x \in S} M(x)$. Then, $M < +\infty$ and the sequence $(x_k)_k$ obtained with the steepest descent method with the exact minimization rule there is*

$$f(x_{k+1}) - f(\bar{x}) \leqslant \Bigl(1 - \frac{m}{2M}\Bigr)\bigl(f(x_k) - f(\bar{x})\bigr).$$

With the restricted minimization rule there is

$$f(x_{k+1}) - f(\bar{x}) \leqslant \Bigl(1 - m\gamma + \frac{mM\gamma^2}{2}\Bigr)\bigl(f(x_k) - f(\bar{x})\bigr),$$

where $\gamma = \min\{\frac{1}{M}, A\}$.

Proof. Consider the exact rule. By the Taylor formula, for $\delta \geqslant 0$, we have

$$\begin{aligned} f\Bigl(x_k - \delta\bigl(Df(x_k)\bigr)^T\Bigr) &\leqslant f(x_k) + Df(x_k)\Bigl(-\delta\bigl(Df(x_k)\bigr)^T\Bigr) + \delta^2\frac{M}{2}\|Df(x_k)\|^2 \\ &= f(x_k) - \delta\|Df(x_k)\|^2 + \delta^2\frac{M}{2}\|Df(x_k)\|^2. \end{aligned} \qquad (\odot)$$

The minimum of the right-hand side is taken at $\delta = \frac{1}{M}$. Recall that x_{k+1} is the minimum for $\alpha \geqslant 0$:

$$f(x_{k+1}) = \inf_{\alpha \geqslant 0} f\Bigl(x_k - \alpha\bigl(Df(x_k)\bigr)^T\Bigr).$$

Therefore,

$$f(x_{k+1}) \leqslant f\Bigl(x_k - \frac{1}{M}\bigl(Df(x_k)\bigr)^T\Bigr) \leqslant f(x_k) - \frac{1}{2M}\|Df(x_k)\|^2.$$

Now we subtract $f(\bar{x})$ from both sides and we apply the inequality $\|Df(x_k)\|^2 \geqslant m\bigl(f(x_k) - f(\bar{x})\bigr)$, being a consequence of Lemma 25:

$$f(x_{k+1}) - f(\bar{x}) \leqslant f(x_k) - f(\bar{x}) - \frac{m}{2M}\bigl(f(x_k) - f(\bar{x})\bigr).$$

A simple calculation yields the claim.

Now we consider the restricted rule. The point x_{k+1} is the minimum for $\alpha \in [0, A)$. The minimum on the right-hand side of $(\odot)$ is obtained with $\delta = \gamma$. The rest of the proof is identical to the proof for the exact rule. □

The above considerations may be concluded as follows: by Lemma 25, the stop condition based on the norm of the gradient of f is correct; it gives us estimations of the approximation error of $\bar{x}$ by x_{k+1} as well as the approximation error of the minimal function value $f(\bar{x})$ by $f(x_{k+1})$. Note that "the more convex" is the function f in a neighbourhood of $\bar{x}$, i.e., the greater is m, the sharper is the dependency between the norm of the gradient and the distance between x_{k+1} and $\bar{x}$. Lemma 26 suggests that the convergence is fastest if the function f behaves similarly in all directions, i.e., the eigenvalues of the Hessian matrix are close to each other. Then, the quotient $\frac{m}{M}$ is great (close to 1), which decreases the contraction factor $1 - \frac{m}{2M}$. Thus, the steepest descent algorithm works best for functions such that their corresponding numbers m and M are of the same order of magnitude.

The correctness of the stop criterion for the Armijo rule still holds, as Lemma 25 does not depend on the step length. Lemma 26 for this case has to be modified; however, the (at least) linear rate of convergence remains.

What is the advantage of the Armijo rule? It is simple to implement, as it does not need methods of searching minima of functions of one variable. The actual rate of convergence depends on the parameters s, β, δ. Alas, there is no general rule of choosing these parameters—user's experience and intuition must be employed.

Newton's method

The descent methods described above choose the directions for the next point based on the Taylor expansion up to the first-order term:

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + Df(\mathbf{x})\mathbf{d}.$$

The Newton method uses the expansion up to the second-order term of the function $f\colon \mathbb{R}^n \to \mathbb{R}$; assume that it is twice differentiable. Then, we use the approximation

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + Df(\mathbf{x})\mathbf{d} + \frac{1}{2}\mathbf{d}^T D^2 f(\mathbf{x})\mathbf{d}.$$

Instead of finding a minimum of f, we minimise the expression on the right-hand side above. For this to make sense we have to assume that the Hessian matrix $D^2(\mathbf{x})$ is positive-definite. With $f(\mathbf{x})$ fixed, the problem reduces to the following:

$$\begin{cases} h(\mathbf{d}) = \frac{1}{2}\mathbf{d}^T D^2 f(\mathbf{x})\mathbf{d} + Df(\mathbf{x})\mathbf{d} \to \min, \\ \mathbf{d} \in \mathbb{R}^n. \end{cases}$$

As the Hessian is, by assumption, positive-definite, the problem above has the unique solution

$$\mathbf{d} = -\big(D^2 f(\mathbf{x})\big)^{-1}\big(Df(\mathbf{x})\big)^T.$$

Note that if the gradient of f at $\mathbf{x}$ is zero, then $\mathbf{d} = \mathbf{0}$ and we shall not leave the critical point. The Newton algorithm is

Preparation: Choose the initial point $\mathbf{x}_1$ and the parameter $\varepsilon > 0$.

k-th step: Take $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$, where $D^2 f(\mathbf{x}_k)\mathbf{d}_k = -\big(Df(\mathbf{x}_k)\big)^T$.

Stop condition: $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

There are a number of doubts about this algorithm. It is not clear whether the sequence $(\mathbf{x}_k)_k$ is convergent; moreover, it is easy to find a regular function, having a minimum, for which the method generates a divergent (unbounded) sequence. The theorem below describes a sufficient condition for the convergence.

Theorem 66 *Let f be a function of class C^3 in a neighbourhood of a local minimum $\bar{\mathbf{x}}$ and let the Hessian $D^2 f(\bar{\mathbf{x}})$ be positive-definite. Then, there exists $\delta > 0$ and $c > 0$ such that for any $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \delta)$ there is*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}\| \leqslant c\|\mathbf{x}_k - \bar{\mathbf{x}}\|^2.$$

Proof. The continuity of $D^2 f$ implies the existence of $\delta > 0$ such that for $\mathbf{x} \in B(\bar{\mathbf{x}}, \delta)$ the norms of $D^2 f(\mathbf{x})$ and $\big(D^2 f(\mathbf{x})\big)^{-1}$ are bounded and greater than some $r > 0$.

Expanding the gradient of f at $\mathbf{x}_k$, we get

$$\big(Df(\mathbf{x}_k + \mathbf{h})\big)^T = \big(Df(\mathbf{x}_k)\big)^T + D^2 f(\mathbf{x}_k)\mathbf{h} + O(\|\mathbf{h}\|^2).$$

Substituting to the above $\mathbf{h} = -\mathbf{h}_k = -(\mathbf{x}_k - \bar{\mathbf{x}})$, we obtain

$$\big(Df(\mathbf{x}_k)\big)^T - D^2 f(\mathbf{x}_k)\mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = \big(Df(\bar{\mathbf{x}})\big)^T = \mathbf{0}.$$

Let $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \delta)$. Multiplying the sides of the above by $\big(D^2 f(\mathbf{x}_k)\big)^{-1}$, we obtain

$$\mathbf{0} = \big(D^2 f(\mathbf{x})\big)^{-1}\big(Df(\mathbf{x}_k)\big)^T - \mathbf{h}_k + O(\|\mathbf{h}_k\|^2 = -\mathbf{d}_k - \mathbf{h}_k + O(\|\mathbf{h}_k\|^2) \qquad (*)$$

There is

$$-\mathbf{d}_k - \mathbf{h}_k = \mathbf{x}_k - \mathbf{x}_{k+1} - (\mathbf{x}_k - \bar{\mathbf{x}}) = -(\mathbf{x}_{k+1} - \bar{\mathbf{x}}) = -\mathbf{h}_{k+1}.$$

Therefore, by (*), there exists a constant $c > 0$ such that

$$\|\mathbf{h}_{k+1}\| \leqslant c\|\mathbf{h}_k\|^2;$$

hence, the rate of convergence is quadratic.

If $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \frac{\alpha}{c})$, where $\alpha \in (0, 1)$ and $\frac{\alpha}{c} \leqslant \delta$, then the last inequality gives us also

$$\|\mathbf{h}_{k+1}\| \leqslant c\frac{\alpha}{c}\|\mathbf{h}_k\| = \alpha\|\mathbf{h}_k\|,$$

which means that the sequence of vectors $(\mathbf{h}_k)_k$, being differences of the points $\mathbf{x}_k$ and the solution $\bar{\mathbf{x}}$ tends to $\mathbf{0}$. □

Remark. (1) If the point $\mathbf{x}_k$ is distant from the solution $\bar{\mathbf{x}}$, then the Hessian $D^2 f(\mathbf{x}_k)$ does not have to be positive-definite.
(2) If the Hessian is positive-definite at $\mathbf{x}_k$, then the vector $\mathbf{d}_k$ constructed in the k-th step has a descent direction for the function f:

$$Df(\mathbf{x}_k)\mathbf{d}_k = -Df(\mathbf{x}_k)\big(D^2 f(\mathbf{x}_k)^{-1})\big)^{-1}\big(Df(\mathbf{x}_k)\big)^T < 0.$$

(3) Even if the Hessian is positive-definite at x_k, there is no guarantee that $f(x_{k+1}) < f(x_k)$, because there is no minimization along the line having the descent direction. The length of the vector $\mathbf{d}_k$ may simply be too big.
(4) A drawback of the Newton method is that it may converge to a critical point being a local maximum or a saddle point (this may be the case if the Hessian is not positive-definite). The assumption of pseudoconvexity of f guarantees that the critical point is a minimum.

A remedy to the above may be minimization along the Levenberg–Marquardt trajectory. The system of linear equations $D^2f(x_k)\mathbf{d} = -\big(Df(x_k)\big)^T$ may be replaced by the following:

$$\big(D^2f(x_k) + \nu I\big)\mathbf{d} = -\big(Df(x_k)\big)^T,$$

where ν is a parameter and I is the identity matrix. We can define a function of one variable,

$$g(\nu) \overset{\text{def}}{=} f\Big(x_k - \big(D^2f(x_k) + \nu I\big)^{-1}\big(Df(x_k)\big)^T\Big).$$

Then, we can chose ν so as to minimise the function g; the point x_{k+1} is then the argument of f corresponding to that ν. Note that if $\nu = 0$, then the point x_{k+1} is the one obtained with the Newton method. The eigenvalues of the matrix $\big(D^2f(x_k) + \nu I\big)$ are the eigenvalues of $D^2f(x_k)$ increased by ν; hence, there exists ν_0 such that the matrix $\big(D^2f(x_k) + \nu I\big)$ is positive-definite for all $\nu > \nu_0$.

The Levenberg–Marquardt trajectory is the parametric curve made of the points $x_k - \big(D^2f(x_k) + \nu I\big)^{-1}\big(Df(x_k)\big)^T$, where $\nu > \nu_0$. Note that by increasing ν we obtain the vectors $-\big(D^2f(x_k) + \nu I\big)^{-1}\big(Df(x_k)\big)^T$ whose directions tend to the direction of $-Df(x_k)$, i.e., to the steepest descent direction, and their lengths tend to 0. Therefore, if the gradient of f at x_k is nonzero, we have a guarantee of finding on the Levenberg–Marquardt trajectory a point x_{k+1} such that $f(x_{k+1}) < f(x_k)$. Minimization along Levenberg–Marquardt trajectories may also produce a divergent sequence of points or a sequence convergent to a saddle point.

Another drawback of the Newton method are the difficulty of finding a good initial point x_1 and the parameter ε for the stop condition. The Levenberg–Marquardt approach makes the choice of initial points considerably easier.

If the dimension n of the problem is large, each iteration, involving the computation of n^2 coefficients of the Hessian and solving a system of n equations, is costly.

Conjugate directions and conjugate gradient methods

Definition 38 *Let H be a symmetric and positive-definite $n \times n$ matrix. Nonzero vectors $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are called conjugate with respect to the matrix H if*

$$\mathbf{d}_i^T H \mathbf{d}_j = 0, \quad i, j \in \{1, \ldots, n\}, \quad i \neq j.$$

Note that if the vectors $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are conjugate with respect to a matrix, then they are linearly independent.

Consider a quadratic function $f(x) = \frac{1}{2}x^T Hx + \mathbf{b}^T x + c$ with a symmetric and positive-definite matrix H. The minimum of this function may be found using a conjugate directions method: subsequent points approximating the minimum are searched in the directions of vectors conjugate with respect to the matrix H. The algorithm is the following:

Preparation: Choose the initial point x_1.

k-th step:

1. Choose the vector $\mathbf{d}_k$.
2. Take $x_{k+1} = x_k + t_k\mathbf{d}_k$, where t_k is chosen by the exact minimization rule, $t_k = \arg\min\{\, f(x_k + t\mathbf{d}_k) : t \geqslant 0 \,\}$

Stop condition: $\|Df(x_{k+1})\| = 0$.

Theorem 67 *A conjugate directions method with the exact minimization rule finds the minimum of a quadratic function $f(x) = \frac{1}{2}x^T Hx + \mathbf{b}^T x + c$ with a positive-definite matrix H after at most n steps.*

Proof. Denote $\mathbf{g}_k = Df(x_k)$. It is easy to notice that

$$\mathbf{g}_{k+1}\mathbf{d}_k = 0, \quad k = 1, \ldots, n-1.$$

To this end, consider the function $h(t) = f(x_k + t\mathbf{d}_k)$; it has the minimum at t_k, therefore $h'(t_k) = 0$. On the other hand, $h'(t) = D(x_k + t\mathbf{d}_k)\mathbf{d}_k$ and $h'(t_k) = \mathbf{g}_{k+1}\mathbf{d}_k$.

For the quadratic function f we obtain

$$\mathbf{g}_{k+1} - \mathbf{g}_k = Df(\mathbf{x}_{k+1}) - Df(\mathbf{x}_k) = \big(H\mathbf{x}_{k+1}\big)^T - \big(H\mathbf{x}_k\big)^T = t_k\mathbf{d}_k^T H.$$

Using this equality we prove by induction that

$$\mathbf{g}_{k+1}\mathbf{d}_j = 0, \quad k = 1, \dots, n-1,\ j = 1 \dots, k. \qquad (\square)$$

For $k = 1$ it is already proved. Suppose that $\mathbf{g}_k\mathbf{d}_j = 0$ for $j = 1, \dots, k-1$. Then, for $j = 1, \dots, k-1$

$$\mathbf{g}_{k+1}\mathbf{d}_j = (\mathbf{g}_k + t_k\mathbf{d}_k^T H)\mathbf{d}_j = \mathbf{g}_k\mathbf{d}_j + t_k\mathbf{d}_k^T H\mathbf{d}_j = 0$$

because $\mathbf{g}_k\mathbf{d}_j = 0$ due to the inductive assumption and $\mathbf{d}_k^T H\mathbf{d}_j = 0$ because the vectors $\mathbf{d}_k$ and $\mathbf{d}_j$ are conjugate with respect to H. Also, the equality $\mathbf{g}_{k+1}\mathbf{d}_k$ is already proved, which completes the proof of $(\square)$.

By $(\square)$, the derivatives of the function f at $\mathbf{x}_{k+1}$ in the directions of $\mathbf{d}_1, \dots, \mathbf{d}_k$ are zero; hence,

$$Df(\mathbf{x}_{k+1})\mathbf{d} = 0, \quad \mathbf{d} \in \operatorname{lin}\{\mathbf{d}_1, \dots, \mathbf{d}_k\}.$$

Let

$$K_k = \mathbf{x}_{k+1} + \operatorname{lin}\{\mathbf{d}_1, \dots, \mathbf{d}_k\} = \mathbf{x}_1 + \operatorname{lin}\{\mathbf{d}_1, \dots, \mathbf{d}_k\}.$$

Let $F = f|_{K_k}$. The function F is convex, because f is convex and K_k is a convex set. All directional derivatives equal to 0 is a sufficient condition for a minimum of a convex function. Therefore,

$$\mathbf{x}_{k+1} = \arg\min\{\, F(\mathbf{x})\colon \mathbf{x} \in K_k \,\} = \arg\min\big\{\, f(\mathbf{x})\colon \mathbf{x} \in \mathbf{x}_1 + \operatorname{lin}\{\mathbf{d}_1, \dots, \mathbf{d}_k\} \,\big\}.$$

To complete the proof we notice that $K_n = \mathbb{R}^n$. □

The method analysed in the last theorem is very effective, as we can find the minimum in at most n steps. Its drawback is the necessity of finding the entire set of conjugate vectors of the matrix H (these may be the eigenvectors of H). The conujgate gradient method described below finds conjugate vectors in consecutive steps of the algorithm. This method is also known as the Fletcher–Reeves algorithm.

Preparation: Choose the initial point $\mathbf{x}_1$.

1-st step: Take $\mathbf{d}_1 = -\mathbf{g}_1^T$ (the vector with the steepest descent direction) and

$$\mathbf{x}_2 = \mathbf{x}_1 + t_1\mathbf{d}_1, \quad t_1 = \arg\min\{f(\mathbf{x}_1 + t\mathbf{d}_1)\}.$$

k-th step, $k > 1$: (we already know the vectors $\mathbf{d}_1, \dots, \mathbf{d}_{k-1}$)

1. Take

$$\beta_{k-1} = \frac{\mathbf{g}_k\mathbf{g}_k^T}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^T}, \quad \mathbf{d}_k = -\mathbf{g}_k^T + \beta_k\mathbf{d}_{k-1}.$$

2. Take $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$, where $t_k = \arg\min\{f(\mathbf{x}_k + t\mathbf{d}_k)\}$.

Stop condition: $\|Df(\mathbf{x}_{k+1})\| = 0$.

Theorem 68 *The Fletcher–Reeves algorithm with the exact minimization rule applied to a quadratic function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *with a positive Hessian matrix* H *constructs vectors* $\mathbf{d}_1, \mathbf{d}_2, \dots$ *conjugate with respect to the matrix* H. *Moreover, for* $i \leqslant m = \max\{\, i\colon \mathbf{g}_i \neq 0^T \,\}$ *there is*

$$\mathbf{g}_i\mathbf{g}_j^T = 0, \quad j = 1, \dots, i-1, \qquad (\oplus)$$

and

$$\mathbf{g}_i\mathbf{d}_i = -\mathbf{g}_i\mathbf{g}_i^T. \qquad (\otimes)$$

Proof. If $m = 0$, then the initial point is the solution and there is nothing to prove. Let $m \geqslant 1$. The proof is done by the induction with respect to i.

If $i = 1$, then we only need to prove that $\mathbf{g}_i\mathbf{d}_i = -\mathbf{g}_i\mathbf{g}_i^T$, which is obvious due to $\mathbf{d}_1 = -\mathbf{g}_1^T$.

Suppose that the vectors $\mathbf{d}_1, \dots, \mathbf{d}_i$ are conjugate with respect to H and the equalities $(\oplus)$ and $(\otimes)$ are satisfied for some $i < m$. From the proof of Theorem 67 we know that for a quadratic function f

$$\mathbf{g}_{i+1} - \mathbf{g}_i = t_i\mathbf{d}_i^T H. \qquad (\ominus)$$

Using this and $\mathbf{g}_{i+1}\mathbf{d}_i = 0$, we obtain

$$0 = \mathbf{g}_{i+1}\mathbf{d}_i = \mathbf{g}_i\mathbf{d}_i + t_i\mathbf{d}_i^T H\mathbf{d}_i.$$

From the above we obtain

$$t_i = -\frac{\mathbf{g}_i\mathbf{d}_i}{\mathbf{d}_i^\mathsf{T} H \mathbf{d}_i} = \frac{\mathbf{g}_i\mathbf{g}_i^\mathsf{T}}{\mathbf{d}_i^\mathsf{T} H \mathbf{d}_i}; \qquad (\odot)$$

the last equality is based on the inductive assumption.

For $j < i$, by the inductive assumption and Step 2 of the algorithm we have

$$\begin{aligned}\mathbf{g}_{i+1}\mathbf{g}_j^\mathsf{T} &= \mathbf{g}_i\mathbf{g}_j^\mathsf{T} + t_i\mathbf{d}_i^\mathsf{T} H\mathbf{g}_j = \mathbf{g}_i\mathbf{g}_j^\mathsf{T} + t_i\mathbf{d}_i^\mathsf{T} H(\mathbf{d}_j - \beta_{j-1}\mathbf{d}_{j-1}) \\ &= \mathbf{g}_i\mathbf{g}_j^\mathsf{T} - t_i\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j + t_i\beta_{j-1}\mathbf{d}_i^\mathsf{T} H\mathbf{d}_{j-1} = 0,\end{aligned}$$

because the first term of the last sum is 0 by the inductive assumption (equality $(\oplus)$) and the other two terms are 0 because the vectors $\mathbf{d}_i$, $\mathbf{d}_j$ and $\mathbf{d}_{j-1}$ are conjugate with respect to H.

A similar calculation with $i = j$, using the inductive assumption and $(\odot)$ gives us

$$\mathbf{g}_{i+1}\mathbf{g}_i^\mathsf{T} = \mathbf{g}_i\mathbf{g}_i^\mathsf{T} - t_i\mathbf{d}_i^\mathsf{T} H\mathbf{d}_i + t_i\beta_{i-1}\mathbf{d}_i^\mathsf{T} H\mathbf{d}_{i-1} = \mathbf{g}_i\mathbf{g}_i^\mathsf{T} - \frac{\mathbf{g}_i\mathbf{g}_i^\mathsf{T}}{\mathbf{d}_i^\mathsf{T} H\mathbf{d}_i}\mathbf{d}_i^\mathsf{T} H\mathbf{d}_i = 0.$$

We proved that $\mathbf{g}_{i+1}\mathbf{g}_j^\mathsf{T} = 0$ for $j = 1, \ldots, i$, which is the inductive step of the proof of $(\oplus)$.

It remains to be proved that the vectors $\mathbf{d}_1, \ldots, \mathbf{d}_m$ are conjugate with respect to H. Using $(\ominus)$ and the formula for $\mathbf{d}_{i+1}$ we obtain

$$\begin{aligned}\mathbf{d}_{i+1}^\mathsf{T} H\mathbf{d}_j &= -\mathbf{g}_{i+1}H\mathbf{d}_j + \beta_i\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j = -\frac{1}{t_j}\mathbf{g}_{i+1}(\mathbf{g}_j - \mathbf{g}_{j+1}) + \beta_i\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j \\ &= -\frac{1}{t_j}\mathbf{g}_{i+1}\mathbf{g}_{j+1}^\mathsf{T} + \beta_i\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j.\end{aligned}$$

For $j < i$ we obtain at once $\mathbf{d}_{i+1}^\mathsf{T} H\mathbf{d}_j = 0$, because, by inductive assumption, $\mathbf{g}_{i+1}\mathbf{g}_{j+1}^\mathsf{T} = 0$ and $\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j = 0$.

For $j = i$, using the formulae for t_i and β_i, we obtain

$$\begin{aligned}\mathbf{d}_{i+1}^\mathsf{T} H\mathbf{d}_j &= -\frac{1}{t_j}\mathbf{g}_{i+1}\mathbf{g}_{j+1}^\mathsf{T} + \frac{\mathbf{g}_{i+1}\mathbf{g}_{i+1}^\mathsf{T}}{\mathbf{g}_i\mathbf{g}_i^\mathsf{T}}\mathbf{d}_i^\mathsf{T} H\mathbf{d}_j \\ &= -\frac{1}{t_j}\mathbf{g}_{i+1}\mathbf{g}_{j+1}^\mathsf{T} + \frac{\mathbf{g}_{i+1}\mathbf{g}_{i+1}^\mathsf{T}}{\mathbf{g}_i\mathbf{g}_i^\mathsf{T}}\,\frac{\mathbf{g}_i\mathbf{g}_i^\mathsf{T}}{t_i} = 0.\end{aligned}$$

We proved that

$$\mathbf{d}_{i+1}^\mathsf{T} H\mathbf{d}_j = 0, \quad j = 1, \ldots, i.$$

The last thing to be proved is the equality $(\otimes)$. Using the formula for $\mathbf{d}_{i+1}$ and $(\Box)$, we obtain

$$\mathbf{g}_{i+1}\mathbf{d}_{i+1} = \mathbf{g}_{i+1}(-\mathbf{g}_{i+1}^\mathsf{T} + \beta_i\mathbf{d}_i) = -\mathbf{g}_{i+1}\mathbf{g}_{i+1}^\mathsf{T} + \beta_i\mathbf{g}_{i+1}\mathbf{d}_i = -\mathbf{g}_{i+1}\mathbf{g}_{i+1}^\mathsf{T}.$$

This completes the entire inductive step. □

The conjugate gradient method is a powerful method of solving systems of linear equations with a symmetric and positive-definite $n \times n$ matrix, where n is large. It is most often used to solve systems with millions of unknown variables obtained by discretization of partial differential equations, which is beyond the scope of this lecture. The idea is to find the minimum of the quadratic function $\frac{1}{2}\mathbf{x}^\mathsf{T} A\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x}$, whose gradient is $\mathbf{x}^\mathsf{T} A - \mathbf{b}^\mathsf{T}$; clearly, at the minimum $\bar{\mathbf{x}}$ the gradient is zero, i.e., $A\bar{\mathbf{x}} = \mathbf{b}$. It turns out that the points obtained in consecutive iterations initially approach the solution, but the rounding errors (always present if floating point arithmetic is used) destroy the convergence; from a certain step the distance between the consecutive points and the solution may (rapidly) increase. Therefore, for huge systems of equations the conjugate gradient method is used as an iterative method: the computations are broken after m iterations, where m is much smaller than n.

If f is not a quadratic function, then the Fletcher–Reeves conjugate gradient method has to be modified. Even without rounding errors we cannot expect the method to find the minimal point of a function of n variables after n iterations. Therefore the stop condition has to be based on a test, e.g., comparing the length of the gradient with some tolerance threshold:

Preparation: Choose the initial point $\mathbf{x}_1$ and the parameter $\varepsilon > 0$.

1-st step: Take $\mathbf{d}_1 = -\mathbf{g}_1^\mathsf{T}$ (the vector with the steepest descent direction) and

$$\mathbf{x}_2 = \mathbf{x}_1 + t_1\mathbf{d}_1, \quad t_1 = \arg\min\{f(\mathbf{x}_1 + t\mathbf{d}_1)\}.$$

k-th step, $k > 1$: (we already know the vectors $\mathbf{d}_1, \ldots, \mathbf{d}_{k-1}$)

1. Take

$$\beta_{k-1} = \frac{\mathbf{g}_k\mathbf{g}_k^\mathsf{T}}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^\mathsf{T}}, \quad \mathbf{d}_k = -\mathbf{g}_k^\mathsf{T} + \beta_k\mathbf{d}_{k-1}.$$

2. Take $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$, where $t_k = \arg\min\{\, f(\mathbf{x}_k + t\mathbf{d}_k) \colon t \geqslant 0\,\}$.

Stop condition: $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

Note that making more than n iterations makes no sense for a quadratic function f. Therefore many implementations of the method make a "reset" every m iterations, where $m \leqslant n$; the vector $\mathbf{d}_k$ after the reset is $-\mathbf{g}_k^T$, i.e., it has the steepest descent direction.

In general, the choice of β_{k-1} as above does not guarantee the vector $\mathbf{d}_k$ to have a descent direction, which is why the formula for β_{k-1} is subject to various modifications. A modification often giving better results than the original Fletcher–Reeves method is the following:

$$\beta_{k-1} = \frac{\mathbf{g}_k(\mathbf{g}_k - \mathbf{g}_{k-1})^T}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^T}.$$

If f is a quadratic function, then $\mathbf{g}_k\mathbf{g}_{k-1}^T = 0$, and so this formula is equivalent to the original one.

Example. Figure 11 shows level sets of the Rosenbrock "banana valley" function,

$$f(x, y) = (x-1)^2 + 100(x^2 - y)^2,$$

whose unique minimum is $\overline{x} = (1, 1)$. This function is known as troublesome in numerical optimization, which is why it is one of popular, or even classical tests for various algorithms. The initial point $\mathbf{x}_1 = (-0.5, 0.5)$ was taken for three algorithms discussed above.

The steepest descent method with restricted minimization rule with $A = 0.5$ yields the sequence of points shown in figure (a); note the very slow convergence to the minimum; the first 150 points have been plotted. One can see that improving the accuracy of computing minima in the steepest descent directions does not accelerate the convergence.

Figure (b) shows the sequence constructed with the Newton method augmented with minimization along Levenberg–Marquardt trajectories; three such trajectories were necessary, and the method has dealt with the problem pretty well.

In Figure (c) we can see the sequence constructed using the modified Fletcher–Reeves method, with reset every 2 iterations. After the initial quick progress the algorithm got stuck at a considerable distance from the minimal point.

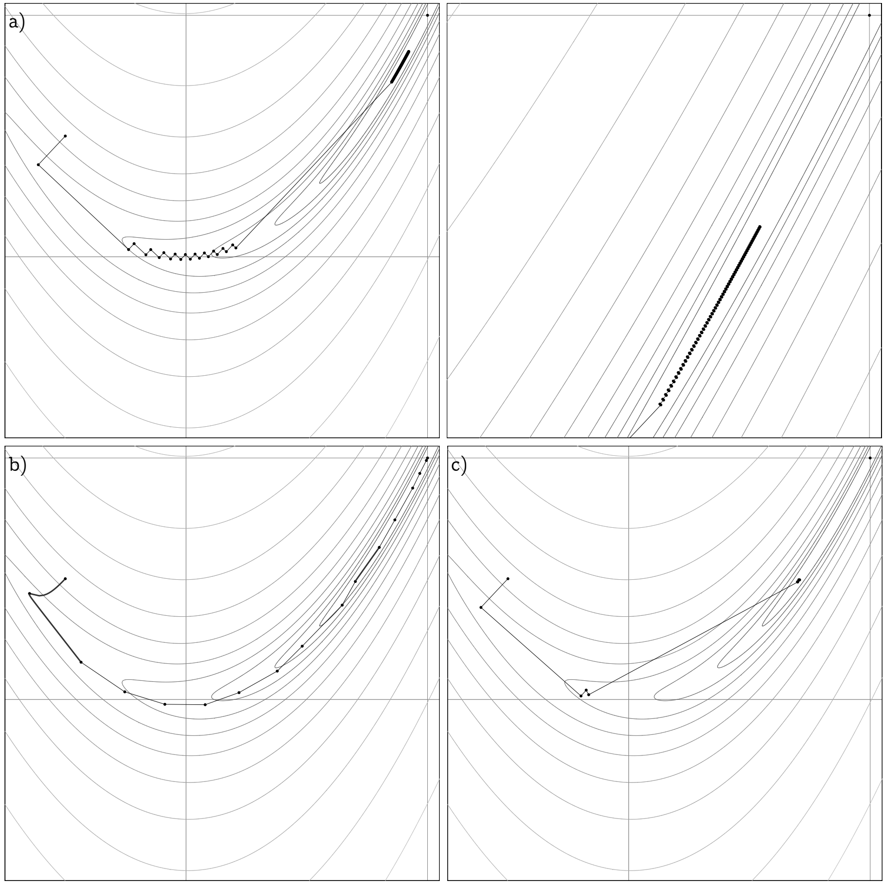

Figure 11: Searching the minimum of the Rosenbrock function

14. Algorithms for optimization with constraints

Now we focus on numerical methods for optimization problems with inequality constraints. We shal see problems arising when the steepest descent methods and their naive modifications are used and we shall see a more efective, though more complicated approach.

The problem considered here is

$$\begin{cases} f(x) \to \min, \\ g_i(x) \leqslant 0, \quad i = 1, \dots, m, \\ x \in \mathbb{R}^n, \end{cases}$$

where $f, g_1, \dots, g_m \colon \mathbb{R}^n \to \mathbb{R}$. Thus, the feasible set is

$$W = \{ x \in \mathbb{R}^n \colon g_1(x) \leqslant 0, \dots, g_m(x) \leqslant 0 \}.$$

In our considerations we use the notion of feasible directions, defined as follows:

$$F(x) = \{ \mathbf{d} \in \mathbb{R}^n \colon \mathbf{d} \neq \mathbf{0} \text{ and there exists } \lambda^* > 0 \\ \text{such that } x + \lambda\mathbf{d} \in W \text{ for all } \lambda \in [0, \lambda^*] \}.$$

Zoutendijk algorithm for affine constraints

Consider a simple modification of the steepest descent method. If a point x is in the interior of W, then we can move in the steepest descent direction until we hit the boundary. If the point is already at the boundary, then it is natural to choose a direction of the possibly steep descent being feasible. Such a direction is called a feasible descent direction at the point x; a vector $\mathbf{d}$ having such a direction satisfies the inequality $Df(x)\mathbf{d} < 0$.

It turns out that if the constraint functions are affine, then this idea works quite well; it is named the Zoutendijk algorithm. Recall that problems with affine constraints are simpler than the general case. This simplicity makes it possible to extend the analysis to problems with mixed affine constraints, i.e.,

$$\begin{cases} f(x) \to \min, \\ Ax \leqslant \mathbf{b}, \\ Qx = \mathbf{a}, \\ x \in \mathbb{R}^n. \end{cases} \tag{*}$$

Here A is an $m \times n$ matrix, Q is $l \times n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{a} \in \mathbb{R}^l$. The following lemma characterises the set of feasible descent directions. Its proof is left as an exercise.

Lemma 27 *Let x be a feasible point for the problem (*). Assume that the matrix A and the vector $\mathbf{b}$ may be divided into blocks A_1, A_2 and $\mathbf{b}_1$, $\mathbf{b}_2$ such that $Ax \leqslant \mathbf{b}$ is the conjunction of $A_1x = \mathbf{b}_1$, $A_2x < \mathbf{b}_2$ (depending on the set of active constraints at x, this may require renumbering of the constraints). The vector $\mathbf{d} \in \mathbb{R}^n$ has a feasible direction at x if $A_1\mathbf{d} \leqslant \mathbf{0}$ and $Q\mathbf{d} = \mathbf{0}$. If in addition $Df(x)\mathbf{d} < 0$, then $\mathbf{d}$ has a feasible descent direction.*

How to choose the best descent direction at x? It would be the simplest to solve the problem

$$Df(x)\mathbf{d} \to \min, \quad \mathbf{d} \in F(x), \quad \|\mathbf{d}\| \leqslant 1. \tag{**}$$

The restriction for the norm of $\mathbf{d}$ is indispensable. Without it, for any vector $\mathbf{d}$ having a feasible descent direction there is $\lim_{\lambda\to\infty} Df(x)\lambda\mathbf{d} = -\infty$ and the problem above has no solution.

Using the block A_1 of the matrix A as in Lemma 27, we can rewrite the problem (**) in the form

$$\begin{cases} Df(x)\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ \mathbf{d}^T\mathbf{d} \leqslant 1. \end{cases}$$

Note that the only nonlinear part above is the norm restriction. In practice, without loss of the algorithm quality it is replaced by linear restrictions, which make it possible to use fast methods of linear optimization (e.g. the simplex algorithm). The most popular replacements for the Euclidean norm $\|\mathbf{d}\|_2$ are

- the maximum norm, $\|\mathbf{d}\|_\infty = max_j|d_j|$, which gives us

  $$-1 \leqslant \mathbf{d}_j \leqslant 1, \quad j = 1, \dots, n,$$

- the first norm, $\|\mathbf{d}\|_1 = \sum_j |d_j|$; in this case we have the inequalities

  $$\begin{cases} \sum_{j=1}^n \eta_j \leqslant 1, \\ -\eta_j \leqslant d_j \leqslant \eta_j, \quad j = 1, \dots, n, \end{cases}$$

  where $\eta_1, \dots, \eta_n$ are new, auxiliary variables.

Below we consider the algorithm with the maximum norm restriction:

$$\begin{cases} Df(\mathbf{x})\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases}$$

The algorithm is the following:

Preparation: Choose the initial point $\mathbf{x}_1$.

k-th step:

1. Given a point $\mathbf{x}_k$, find the blocks A_1, A_2 of the matrix A and the blocks $\mathbf{b}_1$, $\mathbf{b}_2$ of $\mathbf{b}$ so as to obtain $A_1\mathbf{x} = \mathbf{b}_1$ and $A_2\mathbf{x} < \mathbf{b}_2$ (like in Lemma 27).
2. Choose the vector $\mathbf{d}_k$ by solving the problem

$$\begin{cases} Df(\mathbf{x}_k)\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases} \qquad \left(\substack{**\\ *}\right)$$

3. If $Df(\mathbf{x}_k)\mathbf{d}_k = 0$, then stop, as the point $\mathbf{x}_k$ satisfies the necessary first-order condition. Else continue.
4. Take $\alpha_k = \arg\min_{\alpha \in [0, A_k]} f(\mathbf{x}_k + \alpha\mathbf{d}_k)$, where A_k is the greatest number such that the line segment $\overline{\mathbf{x}_k, \mathbf{x}_k + A_k\mathbf{d}_k}$ is contained in the feasible set W.
5. Take $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.

Let's take a look at the choice of the vector $\mathbf{d}_k$. The vector $\mathbf{d}_k = \mathbf{0}$ satisfies all restrictions, therefore the minimal value of the function $Df(\mathbf{x}_k)\mathbf{d}_k$ is less than or equal to 0.

Lemma 28 *The necessary first-order condition is satisfied at* $\mathbf{x}_k$ *if and only if the solution* $\mathbf{d}_k$ *of* $\left(\substack{**\\ *}\right)$ *satisfies the equality* $Df(\mathbf{x}_k)\mathbf{d}_k = 0$

Proof. Recall that the necessary first-order condition for the problem (*) is satisfied at $\mathbf{x}_k$ if and only if there exist vectors $\boldsymbol{\mu} \in [0, \infty)^{m_1}$ and $\boldsymbol{\lambda} \in \mathbb{R}^l$ such that

$$Df(\mathbf{x}_k) + \boldsymbol{\mu}^T A_1 + \boldsymbol{\lambda}^T Q = \mathbf{0}^T,$$

where the $m_1 \times n$ matrix A_1 corresponds to the constraints active at $\mathbf{x}_k$.

By the Farkas lemma (Lemma 15), if this system has a solution, then the system

$$\begin{cases} Df(\mathbf{x}_k)\mathbf{d} < 0, \\ A_1\mathbf{d} \leqslant 0, \\ Q\mathbf{d} = \mathbf{0} \end{cases} \qquad \left(\substack{**\\ **}\right)$$

is inconsistent. Due to the observation that $\mathbf{d} = \mathbf{0}$ is a solution of the system above with the inequality $Df(\mathbf{x}_k)\mathbf{d} < 0$ replaced by the equality $Df(\mathbf{x}_k)\mathbf{d} = 0$, the right implication is proved.

To prove the left implication, we notice that if $Df(\mathbf{x}_k)\mathbf{d} = 0$, then $\mathbf{d}_k$ does not satisy $\left(\substack{**\\ **}\right)$. Using again the Farkas lemma we notice that the necessary first-order condition is then satisfied. □

Zoutendijk algorithm for nonlinear constraints

One can wonder if the Zoutendijk algorithm works just as well for nonlinear constraints:

Preparation: Choose the initial point $\mathbf{x}_1$.

k-th step:

1. Given a point $\mathbf{x}_k$, choose the vector $\mathbf{d}_k$ by solving the problem

$$Df(\mathbf{x}_k)\mathbf{d} \to \min, \quad \mathbf{d} \in F(\mathbf{x}_k), \quad \|\mathbf{d}\| \leqslant 1.$$

2. If $Df(\mathbf{x}_k)\mathbf{d}_k = 0$, then stop, as the point $\mathbf{x}_k$ satisfies the necessary first-order condition. Else continue.
3. Take $\alpha_k = \arg\min_{\alpha \in [0, A_k]} f(\mathbf{x}_k + \alpha\mathbf{d}_k)$, where A_k is the greatest number such that the line segment $\overline{\mathbf{x}_k, \mathbf{x}_k + A_k\mathbf{d}_k}$ is contained in the feasible set W.
4. Take $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.

Figure 12 shows the steepest descent methods used to find the minimum of the function $f(x_1, x_2) = -2x_1 - x_2$ in two different sets; the minimal point is denoted by $\bar{\mathbf{x}}$. In the case (a) the minimum is found in the second step. If the set W is not convex, as in the case (b), the algorithm may go to a dead end. This is, alas,

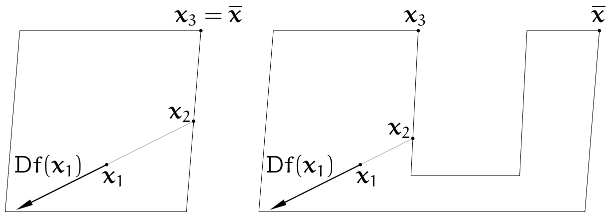

Figure 12: Steepest descent method for two different feasible sets

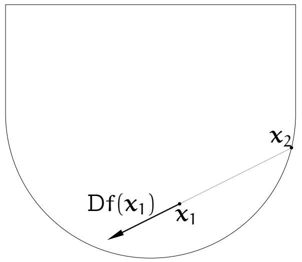

Figure 13: Failure of the descent method in a convex feasible set

a property of all descent algorithms. Therefore, we have to require that the feasible set be convex. Does it suffice? No. In Figure 13 we can see that the algorithm may fail even with a convex feasible set. The steepest descent direction cannot be found, because the set of feasible descent directions $F(x_2)$ is not closed. It is easy to find a way out: we need to choose such a vector $\mathbf{d}_k$ that the descent is steep and also a long part (line segment) of the halfline having the direction of $\mathbf{d}_k$ be contained in the set W. We still pursue the simplicity, i.e., linearity, of the optimization problem posed in order to choose $\mathbf{d}_k$. The solution is prompted by the following lemma:

Lemma 29 *Let x be a feasible point. If the functions f and g_i for $i \in I(x)$ are differentiable at x and the functions g_i for $i \notin I(x)$ are continuous, then any vector $\mathbf{d}$ such that $Df(x)\mathbf{d} < 0$ and $Dg_i(x)\mathbf{d} < 0$ for $i \in I(x)$ has a feasible descent direction.*

Proof. First we prove that $\mathbf{d}$ has a feasible direction, i.e., for a sufficiently small $\lambda > 0$ there is $x + \lambda\mathbf{d} \in W$. For $i \notin I(x)$, due to the continuity of g_i we have

$g_i(x + \lambda\mathbf{d}) < 0$. For $i \in I(x)$ we have

$$g_i(x + \lambda\mathbf{d}) = g_i(x) + \lambda Dg_i(x)\mathbf{d} + o(\lambda\mathbf{d}).$$

There is $\mathbf{g}_i(x) = 0$, $Dg_i(x) < 0$ and $o(\lambda\mathbf{d}) \to 0$ for $\lambda \to 0$; hence, $g_i(x + \lambda\mathbf{d}) < 0$ for all sufficiently small positive λ.

The vector $\mathbf{d}$ has a feasible direction. By the asumptions of the lemma it has also a descent direction. There is

$$f(x + \lambda\mathbf{d}) = f(x) + \lambda Df(x)\mathbf{d} + o(\lambda\mathbf{d}).$$

Hence, for λ sufficiently small there is $f(x + \lambda\mathbf{d}) < f(x)$. □

Lemma 29 gives us only a *sufficient* condition. There are optimization problems with inequality constraints such that one of feasible descent directions does not satisfy the lemma's assumptions.

Choosing a vector $\mathbf{d} \in \mathbb{R}^n$ such that $Df(x)\mathbf{d} < 0$ and $Dg_i(x)\mathbf{d} < 0$ for $i \in I(x)$ may be done by solving the following problem:

$$\begin{cases} \max\{\, Df(x)\mathbf{d},\ Dg_i(x)\mathbf{d},\ i \in I(x)\,\} \longrightarrow \min, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases}$$

The target function above is tough to implement; we can reduce its minimization to the much simpler linear optimization problem

$$\begin{cases} \eta \to \min, \\ Df(x)\mathbf{d} \leqslant \eta, \\ Dg_i(x)\mathbf{d} \leqslant \eta, \quad i \in I(x), \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases} \qquad (\otimes)$$

Here the optimization is done with respect to *two* variables: $\mathbf{d} \in \mathbb{R}^n$ and $\eta \in \mathbb{R}$. Note that $\eta \leqslant 0$, because the pair $(\mathbf{d}, \eta) = (\mathbf{0}, 0)$ satisfiest the constraints above. If the target function has a negative value, then by Lemma 29 the vector $\mathbf{d}$ has a feasible descent direction. If $\eta = 0$ is a solution and the linear independence condition is satisfied by the constraints, then the necessary first-order condition is satisfied at x. The inverse implication is also true.

Lemma 30 *If the linear independence condition is satisfied by the constraints at a feasible point x and $\eta = 0$ is the solution of the problem $(\otimes)$, then the necessary first-order condition is satisfied at x. Also, if the first-order condition is satisfied at x, then $\eta = 0$ is the solution of $(\otimes)$ (here the regularity of constraints at x needs not be assumed).*

Proof. If $\eta = 0$ is the solution, then the system $A\mathbf{d} < \mathbf{0}$, where A is the matrix whose rows are gradients of f and g_i, $i \in I(\mathbf{x})$, has no solution. By Gordan's lemma (Lemma 18), there exists $\mathbf{y} \geqslant 0$, $\mathbf{y} \neq \mathbf{0}$ such that $A^T\mathbf{y} = \mathbf{0}$. Let $\mathbf{y} = \big(\hat{\mu}_0,\ \hat{\mu}_i, i \in I(\mathbf{x})\big)$ and let $\hat{\mu}_i = 0$ for $i \notin I(\mathbf{x})$. The equality $A^T\mathbf{y} = \mathbf{0}$ may be rewritten as follows:

$$\hat{\mu}_0 Df(\mathbf{x}) + \sum_{i \in I(\mathbf{x})} \hat{\mu}_i Dg_i(\mathbf{x}) = \mathbf{0}^T.$$

From the assumption of linear independence of gradients of the active constraints we conclude that $\hat{\mu}_0 \neq 0$. Taking $\mu_i = \hat{\mu}_i/\hat{\mu}_0$ for $i = 1, \ldots, m$, we obtain the Lagrange multipliers for the necessary first-order condition.

To prove the inverse implication we notice that if the necessary first-order condition is satisfied at $\mathbf{x}$, then the vector $\mathbf{y} = \big(1,\ \mu_i, i \in I(\mathbf{x})\big)$ satisfies the following: $\mathbf{y} \geqslant 0$, $\mathbf{y} \neq \mathbf{0}$ and $A^T\mathbf{y} = \mathbf{0}$. By Lemma 18, there is no $\mathbf{d} \in \mathbb{R}^n$ such that $A\mathbf{d} < \mathbf{0}$. Then, $\eta = 0$ is the solution of $(\otimes)$. □

The complete Zoutendijk algorithm for nonlinear problems with nonlinear constraints is the following:

Preparation: Choose the initial point $\mathbf{x}_1$.

k-th step:

1. Given a point $\mathbf{x}_k$, choose the vector $\mathbf{d}_k$ by solving the problem
$$\begin{cases} \eta \to \min, \\ Df(\mathbf{x}_k)\mathbf{d} \leqslant \eta, \\ Dg_i(\mathbf{x}_k)\mathbf{d} \leqslant \eta, \quad i \in I(\mathbf{x}_k), \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases}$$
2. If $\eta = 0$, then stop, as the point $\mathbf{x}_k$ satisfies the necessary first-order condition. Else continue.
3. Take $\alpha_k = \arg\min_{\alpha \in [0, A_k]} f(\mathbf{x}_k + \alpha\mathbf{d}_k)$, where A_k is the greatest number such that the line segment $\overline{\mathbf{x}_k, \mathbf{x}_k + A_k\mathbf{d}_k}$ is contained in the set W.
4. Take $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.

Example. Consider the problem

$$\begin{cases} 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \to \min, \\ x_1 + 5x_2 \leqslant 5, \\ 2x_1^2 - x_2 \leqslant 0, \\ x_1 \geqslant 0,\ x_2 \geqslant 0. \end{cases}$$

The Zoutendijk algorithm with the initial point $\mathbf{x}_1 = (0, 0.75)$ generates the following sequence of points:

$$\mathbf{x}_2 = (0.2803, 0.5477),$$
$$\mathbf{x}_3 = (0.5555, 0.8889),$$
$$\mathbf{x}_4 = (0.6479, 0.8397),$$
$$\mathbf{x}_5 = (0.6302, 0.8740).$$

As we can see, this sequence shows considerable oscillations in the feasible set, see Figure 14. This is a typical behaviour of descent methods for problems with constraints.

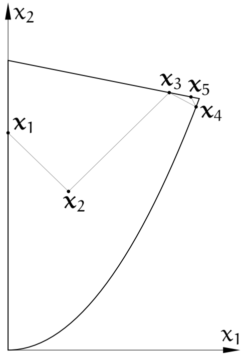

Figure 14: Oscillations of the sequence generated by the Zoutendijk algorithm

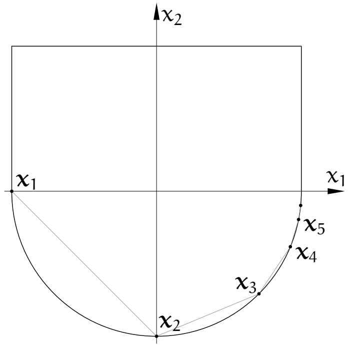

Figure 15: Convergence to a point not being a solution

The Zoutendijk algorithm may fail even for rather simple problems. Consider searching the minimum of the linear function $f(x_1, x_2) = -2x_1 - x_2$ in the set shown in Figure 15; the minimum is at $\overline{x} = (1, 1)$. Beginning at $\mathbf{x}_1 = (-1, 0)$, we obtain a sequence of points converging to $(1, 0)$; thus, the algorithm will not approach the solution $\overline{x}$. Moreover, the value of f at $(1, 0)$ is -2, while $f(\overline{x}) = -3$. However, a slight modification may improve the Zoutendijk algorithm.

Topkis–Veinott modification

In 1967 Topkis and Veinott suggested a modification of the method of choosing $\mathbf{d}_k$ in the Zoutendijk algorithm:

$$\begin{cases} \eta \to \min, \\ Df(\mathbf{x}_k)\mathbf{d} \leqslant \eta, \\ Dg_i(\mathbf{x}_k)\mathbf{d} \leqslant \eta - g_i(\mathbf{x}_k), \quad i = 1, \ldots, m, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases} \qquad (\oplus)$$

The inequalities imposed for gradients of constraints include all constraints; for active constraints, $i \in I(\mathbf{x}_k)$, we have $g_i(\mathbf{x}_k) = 0$ and thus the conditions have not been changed. For inactive constraints, $g_i(\mathbf{x}) < 0$ and the right-hand sides of the inequalities are greater than η. If the value of g_i at $\mathbf{x}_k$ is great, then the inequality is almost irrelevant. If the value of g_i is close to 0, i.e., the constraint is “almost active”, then the corresponding inequality has a significant influence on the choice of $\mathbf{d}_k$. Moreover, in the implementation this modification helps finding the active constraints, as due to the inexact representation of real numbers (the floating point representation), we usually cannot obtain $g_i(\mathbf{x}) = 0$.

The theorem, which characterises the effectiveness of this modification is given without proof:

Theorem 69 *Assume that $f, g_1, \ldots, g_m$ are of class C^1. If the sequence $(\mathbf{x}_k)_k$ constructed by the Zoutendijk algorithm with the Topkis–Veinott modification has a concentration point $\mathbf{x}$, at which the linear independence regularity condition is satisfied, then the necessary first-order condition is satisfied at $\mathbf{x}$.*

Conclusion

The numerical methods described above make it possible to approximate points satisfying the necessary first-order condition. Then, Theorem 47 may be used to guarantee the optimality of those points. In particular, if the constraints are linear, it suffices to assume pseudoconvexity of the function f. Note that similar

assumptions were needed for problems without constraints. The assumption about convexity is a natural and often necessary condition for the numerical methods to work well.