131890-S Optimization methods **Winter 2023/2024**

Lecture 1: Introduction to analytical optimization

Daniel Kaszyński *6 October 2023 r.*

Organisation

- **Instructor**: mgr Daniel Kaszyński, dkaszy[@]sgh.waw.pl
- **Consultation**: Friday 12:00-13:00, G-213.
- **Course grading**: A written exam on the content presented in class or included in the materials

  Example task: Describe method XYZ, show differences between ABC and XYZ, solve an analytical optimization task, carry out two iterations of the XYZ method
- **Required literature**:
  - [KW19] Kochenderfer, M.J. and Wheeler, T.A., 2019. *Algorithms for optimization.* Mit Press.
  - [CZ04] Chong, E.K. and Zak, S.H., 2004. *An introduction to optimization.* John Wiley & Sons.
  - [BI86] Birkholc, A., 1986. *Analiza matematyczna: funkcje wielu zmiennych.* Państwowe Wydawnictwo Naukowe.
  - [SS08] Sydsæter, K., Hammond, P., Seierstad, A. and Strom, A., 2008. *Further mathematics for economic analysis.* Pearson education.
  - [CO14] Cortez, P., 2014. *Modern optimization with R.* New York: Springer.

1.1 Definition of extremum

Extremum is the central concept concerning the analytical optimization problem. The extremum x^* is the solution of the **evaluation function** f for which x^* generates 'the best' solution. Let's consider a function of one variable mapping from real to real values $f : \mathbb{R} \to \mathbb{R}$.

Definition 1: Local extremum

By a local extremum (type minimum) we will call a point x^* for which:

$$\exists\, d > 0 \quad \forall\, 0 < |\delta| < |d| \quad \Rightarrow \quad f(x^* + \delta) \geqslant f(x^*) \tag{1.1}$$

Definition 2: Global extremum

By a global extremum (type minimum) we will call a point x^* for which:

$$\forall\, d > 0 \quad \forall\, |\delta| > 0 \quad \Rightarrow \quad f(x^* + \delta) > f(x^*) \tag{1.2}$$

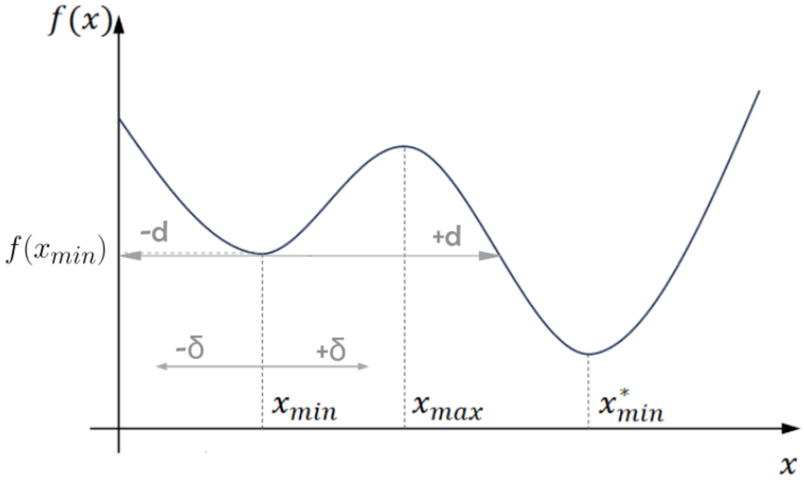

Figure 1.1: Local extremum of a function $f : \mathbb{R} \to \mathbb{R}$

Notice that the difference between a *local* and *global* extremum is the area in which we obtain better solution (from the point of view of the evaluation function). In case of a local extrema, we can point a neighborhood around x_{min} in which we obtain worse solutions. The neighborhood for x_{min} might be very small, where as for a global extremum x^*_{min} it's any neighborhood around an extremum.

1.2 Non-linear optimization without constraints in 1D. $f : \mathbb{R} \to \mathbb{R}$

The basic definition related to the non-linear optimization are related to the definition of derivative of a function.

Definition 3: Derivative of a function

By a derivative of a function $f : \mathbb{R} \to \mathbb{R}$ we call a function:

$$f'(x) = \frac{df}{dx}(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{f(x) - f(x-h)}{h} \tag{1.1}$$

Continuity of a function f is a necessary condition for a function to be differentiable!

In an analogous way we can describe the second derivative of a function:

$$f''(x) = \frac{d^2 f}{dx^2}(x) = (f'(x))' = \lim_{h\to 0} \frac{f'(x+h) - f'(x)}{h} = \lim_{h\to 0} \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} \tag{1.2}$$

1.2.1 First Order Conditions $f : \mathbb{R} \to \mathbb{R}$

Theorem 1 (First Order Conditions $f : \mathbb{R} \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R} \to \mathbb{R}$, $f \in C^1$. If a function f has an extremum at the point $x \in \mathbb{D}$, then $f'(x) = 0$.

Proof. If a function f has a minimum extremum at the point x, then there must exist $|d| > 0$, such that for all $0 < |\delta| < |d|$, we have $f(x+\delta) > f(x)$, or $f(x+\delta) - f(x) > 0$. Dividing both sides by a δ we obtain:

$$\frac{f(x+\delta) - f(x)}{\delta} > 0 \qquad \wedge \qquad \frac{f(x+\delta) - f(x)}{\delta} < 0$$

for $\delta > 0$ i $\delta < 0$. Respectively at the limit $\delta \to 0^+$ i $\delta \to 0^-$ we have:

$$\lim_{\delta \to 0^+} \frac{f(x+\delta) - f(x)}{\delta} = f'_+(x) \geqslant 0 \qquad \wedge \qquad \lim_{\delta \to 0^-} \frac{f(x+\delta) - f(x)}{\delta} = f'_-(x) \leqslant 0$$

If a function f is differentiable then $f'(x) = f'_+(x) = f'_-(x) = 0$. □

First Order Conditions give us a way to filter solutions from the search space, to those where the first derivative of a function is equal to zero (stationary points).

Caution! Just because a derivative of a function at the point x is equal to zero, doesn't mean that it is an extremum. For an example consider functions $f(x) = x^2$ and $f(x) = x^3$.

1.2.2 Taylor's Theorem $f : \mathbb{R} \to \mathbb{R}$

We will now introduce one of the most important theorem of mathematical analysis. The **Fundamental theorem of calculus** states following:

Theorem 2 (Fundamental theorem of calculus)

$$\int_a^b f(x)\, dx = F(b) - F(a) \tag{1.3}$$

where $F(x)$ is an anti-derivative or indefinite integral at point X. It means that the area under the curve of a function f between points a and b we have to calculate: (1) the area under the curve from $-\infty$ to b, (2) the area under the curve from $-\infty$ to a, (3) subtracting these values. Intuition behind equation (1.3) is shown on Figure 1.1.

To simplify symbols let us assume that $\int f'(x)\, dx = f(x)$, then we can rewrite (1.3) as:

$$f(b) - f(a) = \int_a^b f'(x)\, dx \tag{1.4}$$

Moving forward we will use symbols $a = x$ $b = x + h$. We use point a as a starting point (or reference point),

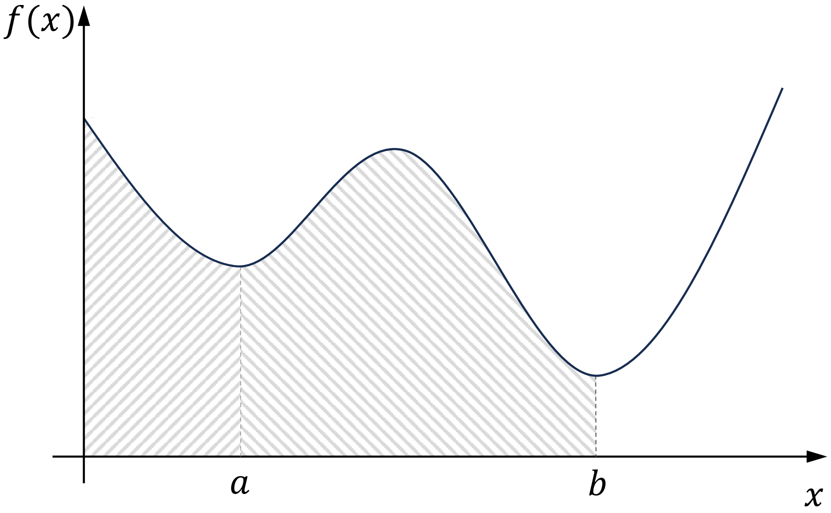

Figure 1.1: Definite integral

and point b as an offset by h from a. Notice that if we set $h = 0$ and start increasing it, then equation (1.4) answers the question by how much the area under the function f increases.

$$f(x+h) - f(x) = \int_x^{x+h} f'(a)\; da$$

By rearranging this equation and bringing the indefinite integral to the beginning of a coordinate system (as for now it was attached at the point x) we obtain:

$$f(x+h) = f(x) + \int_0^h f'(x+a)\; da \tag{1.5}$$

The equation (1.5) is important from the perspective of further transformations. We can see that x is interpreted as a constant value (as the integral is on a). Also, the left side of the equation is in the similar form of a function inside the integral. Let's try to express the integral in terms of equation (1.5):

$$f(x+h) = f(x) + \int_0^h \left[f'(x) + \int_0^a f''(x+b) db \right] da$$

Using the addition property of an integral:

$$f(x+h) = f(x) + \int_0^h f'(x) da + \int_0^h \left[\int_0^a f''(x+b) db \right] da$$

We can try to express the first integral in the following form:

$$\int_0^h f'(x) da = [f'(x)a]_0^h = f'(x)h$$

Which gives us:

$$f(x+h) = f(x) + f'(x)h + \int_0^h \int_0^a f''(x+b)db\ da$$

Caution: The expression inside the double integral we can also express using the previously noticed property:

$$f(x+h) = f(x) + f'(x)h + \int_0^h \int_0^a f''(x) + \left[\int_0^b f''(x+c)dc\right] db\ da$$

The inside of the double integral can be written as:

$$\int_0^h \int_0^a f''(x)db\ da = \int_0^h \left[f''(x)b\right]_0^a da = \int_0^h f''(x)a\ da = \int_0^h \left[\frac{1}{2}f''(x)a^2\right]_0^h = f''(x)h^2$$

Introducing this equation we obtain:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \int_0^h \int_0^a \int_0^b f''(x+c)dc\ db\ da$$

The equation inside the integral is always worked out of equation 1.5. Such procedure can be performed indefinitely (technically as many times as the function f is differentiable). In general this equation is given as:

$$f(x+h) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}h^n \tag{1.6}$$

The qquation (1.6) is called as a *Taylor's equation*. Notice that in practice we usually won't differentiate f infinitely many times. We will do it only few times that satisfies us with its accuracy. Thus we can write out N-th expansions of a function using Taylor's equation:

$$f(x+h) = f(x) + \sum_{n=1}^{N-1} \frac{1}{n!}f^{(n)}(x)h^n + \frac{f^{(N)}(x+\theta h)}{N!}h^N \tag{1.7}$$

where $R_N(x,h) = \frac{f^{(N)}(x+\theta h)}{N!}h^N$ is the Lagrange remainder. We can show that, this remainder has the following property:

$$\lim_{h\to 0} \frac{R_N(x,h)}{h^N} = 0$$

Which means that the remainder $R_N(x,h)$ of approximation using Taylor's equation decreases to 0 at a rate faster than the polynomial of N-th order.

Theorem 3 (Taylor's equation $f : \mathbb{R} \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R} \to \mathbb{R}$ and $f \in C^N$ at every point of the segment $[x, x+h]$. Then for some θ we have:

$$f(x+h) = f(x) + \sum_{n=1}^{N-1} \frac{1}{n!}f^{(n)}(x)h^n + \frac{f^{(N)}(x+\theta h)}{(N)!}h^N$$

Taylor’s theorem is an important result that is often used in practice!

Expansion of a function using a Taylor’s series of 1st and 2nd order:

$$f(x+h) \approx f(x) + f'(x)h + \frac{1}{2}f''(x)h^2$$

Using programming language **R** we can expand an example function into a Taylor’s series of 1st and 2nd order: $f(x) = \frac{x^2}{e^x}$:

```r
# Dane wejsciowe
f  <- function(x) x^2/exp(x)
x0 <- 2.5
h_seq <- seq(0, 10, length = 100)

# Pochodne numeryczne
d1f <- function(f, x, h = 10^-6) (f(x+h)-f(x))/h
d2f <- function(f, x, h = 10^-6) (f(x+2*h)-2*f(x+h)+f(x))/h^2

# Aproksymacja Taylora funkcji f wokol x0
taylor_1 <- function(f, x, h) f(x)+d1f(f, x)*(h-x)
taylor_2 <- function(f, x, h) f(x)+d1f(f, x)*(h-x)+1/2*d2f(f, x)*(h-x)^2

# Wykresy
plot(h_seq, f(h_seq), type=’l’, col=’black’, xlab = ’x’, ylab = ’y’)
lines(h_seq, taylor_1(f, x0, h_seq), col=’red’)
lines(h_seq, taylor_2(f, x0, h_seq), col=’blue’)
legend(7.8, 0.55, legend=c(’f(x)’, ’taylor_1’, ’taylor_2’),
       col=c(’black’, ’red’, ’blue’), lty=1, cex=1)
```

Listing 1: Example: expanding function into a Taylor’s series

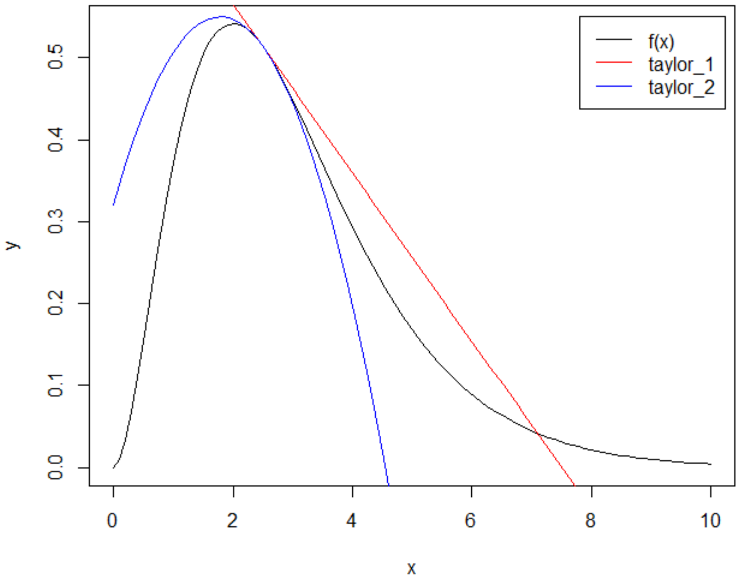

Figure 1.2: Example: expanding function into a Taylor’s series

1.2.3 Second Order Conditions $f : \mathbb{R} \to \mathbb{R}$

We showed earlier that First Order Conditions are the required (each extremum has such property), but not enough (points that are not extremas, but have such property). To if a stationary point has an extremum we will use *sufficient conditions* – Second Order Conditions.

Theorem 4 (Second Order Conditions $f : \mathbb{R} \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R} \to \mathbb{R}$, $f \in C^n$. If for some $x \in \mathbb{D}$ we get: $f'(x) = 0, f''(x) = 0, \ldots, f^{(n-1)}(x) = 0$, but for $f^{(n)}(x) \neq 0$, then:

1. *If n is even, then function f has an extremum at the point x; If $f^{(n)}(x) > 0$ then it is a minimum, if $f^{(n)}(x) < 0$ then it is a maximum.*
2. *If n is odd, then function f doesn't have an extremum at the point x.*

Proof. From the Taylor's equation, for some $0 < \theta < 1$ we have:

$$f(x+h) = \sum_{k=0}^{n-1} \frac{1}{k!} f^{(k)}(x) h^k + \frac{1}{(n)!} f^{(n)}(x + \theta h) h^{(n)}$$

because $f'(x) = 0, f''(x) = 0, \ldots, f^{(n-1)}(x) = 0$ then:

$$f(x+h) = f(x) + \frac{1}{n!} f^{(n)}(x + \theta h) h^n$$

$$f(x+h) - f(x) = \frac{1}{n!} f^{(n)}(x + \theta h) h^n$$

When n is even and $f^{(n)}(x) > 0$ then due to parity of n we have $h^n > 0$. Due to continuity of the function $f^{(n)}$ at the point x we know that for some $\delta > 0$ such, that for each $h : 0 < |h| < \delta$ we have $f^{(n)}(x+h) > 0$, due to that $f^{(n)}(x + \theta h) > 0$. What it means is that a function f has in this point x a minimum. We can show analogously the maximum case. □

1.3 Non-linear optimization without constraints, *multivariate* case, $f : \mathbb{R}^n \to \mathbb{R}$

In the previous section we showed the optimization conditions in a one dimensional case – one decision variable. However, the nature of optimization problems is more complicated and is multivariate.

We first have to introduce some basic terms from the area of differential calculus related to multiple variables. We recommended to get familiar with early chapters of a textbook [BI86].

Definition 4: Directional derivative

Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{D}$ and $h \in \mathbb{R}^n : x + h \in \mathbb{D}$. Directional derivative of a function f at the point x in direction h we call the function:

$$\frac{df}{dh}(x) = \lim_{t \to 0} \frac{f(x + th) - f(x)}{t}$$

Definition 5: Partial derivative

Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{D}$ and $h \in \mathbb{R}^n : x + h \in \mathbb{D}$. Partial derivative of f at the point x with respect to variable x_i, $i = 1, 2, \ldots n$ we call the function:

$$\frac{\partial f}{\partial x_i}(x) = \frac{df}{de_i}(x)$$

where e_i is the i-th versor of space $\mathbb{R}^n$. Partial derivative of f with respect to x_i is then a directional derivative of f in direction of i-th versor, meaning that $h = e_i$.

Definition 6: Gradient of a function

Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{D}$. By a gradient of a function f we call function $\nabla_f(x) : \mathbb{R}^n \to \mathbb{R}^n$ at the point f x:

$$\nabla_f(x) = \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right]$$

Relation of directional derivative and gradient? If the gradient of a function $\nabla_f(\mathbf{x})$ exists at the point $\mathbf{x}$ (which means that the function f is differentiable in $\mathbf{x}$)

$$\nabla_f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]$$

then directional derivative of a function f in direction of a vector $\mathbf{h}$ is equal to the dot product of a gradient $\nabla_f(\mathbf{x})$ and vector $\mathbf{h}$:

$$\frac{df}{dh} = \nabla_f(\mathbf{x}|\mathbf{h}) = \nabla_f(\mathbf{x}) \times \mathbf{h}$$

1.3.1 First Order Conditions $f : \mathbb{R}^n \to \mathbb{R}$

Theorem 5 (First Order Conditions $f : \mathbb{R}^n \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$. If a function f has an extremum in point x, then $\nabla_f(x) = \mathbf{0}$

Proof. Let's consider a function $g_x(t) = f(x + th)$ and $h \in \mathbb{R}^n : x + h \in \mathbb{D}$. Because f has an extremum in x, then g has an extremum at $t = 0$, then $g'(t) = 0$, which means that $g'(t)|_{t=0} = 0$. As a result:

$$g'(t)|_{t=0} = \lim_{\Delta \to 0} \frac{g(t+\Delta) - g(t)}{\Delta} = \lim_{\Delta \to 0} \frac{f(x + \Delta h) - f(x)}{\Delta} = \frac{df}{dh}(x) = \nabla_f(x)h = \mathbf{0}$$

□

Example:

Let's consider a function $f(x) = x_1^2 + x_2^2$. Find extremum of $f(x)$.

$$\nabla_f(x) = \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x)\right] = [2x_1, 2x_2] = \mathbf{0} \implies [x_1, x_2] = [0, 0]$$

1.3.2 Second Order Conditions $f : \mathbb{R}^n \to \mathbb{R}$

To derive Second Order Conditions for a multivariate function we have to introduce a generalization of a second derivative of a function, namely Hessian matrix.

Definition 7: Hessian matrix

Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{D}$. By a Hessian matrix $H_f(x)$ we call a matrix:

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

Caution! Hessian matrix, is a symmetric matrix only, when all second order partial derivatives are continuous (Schwarz's theorem). Which means that: $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$

Also, the previously introduced Taylor's equation for a one dimensional case, can be redefined for the multivariate case:

Theorem 6 (Taylor's theorem $f : \mathbb{R}^n \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$ and $f \in C^2$ in each point of a section $[x, x + h]$. Then:

$$f(x + h) = f(x) + \nabla_f(x)h + \frac{1}{2}h^T H_f(x)h + R_3(x, h)$$

Theorem 7 (Second Order Conditions $f : \mathbb{R}^n \to \mathbb{R}$)
Let $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$. If in x^* we have both:

1. $\nabla_f(x^*) = \mathbf{0}$

2. $H_f(x^*) > 0$

Then w x^* is a local minimum of f.

Proof. From the Taylor's theorem for $\mathbb{R}^n$ we have:

$$f(x+h) = f(x) + \nabla_f(x)h + \frac{1}{2}h^T H_f(x)h + o(|h|^2) = f(x) + \frac{1}{2}h^T H_f(x)h + o(|h|^2)$$

where $f(x+h) - f(x) = \frac{1}{2}h^T H_f(x)h + o(|h|^2)$. From the Rayleigh's theorem, value of a quadratic form $h^T H_f(x)h$ can be bounded from below by a $\lambda_{min}|h|^2$:

$$f(x+h) - f(x) = \frac{1}{2}h^T H_f(x)h + o(|h|^2) \geqslant \frac{1}{2}\lambda_{min}|h|^2 + o(|h|^2)$$

For small enough h we get $f(x+h) - f(x) > 0$ □

What is left is to tell what happens when $H_f(x^*) > 0$?

Theorem 8 (Sylvester's criterion)
Let A be a symmetric real values matrix:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}$$

We can define first minors of a matrci A as:

$$M_1 = a_{1,1} \quad M_2 = det\left(\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}\right) \quad \dots \quad M_n = det\left(\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}\right)$$

Then:

1. *Matrix A is positively defined if and only if all first minors M_i of A are positive.*
2. *Matrix is negatively defined if and only if all even first minors M_i of A are positive, and all odd minors M_i are negative.*

Example:

Let $f(x) = x_1^2 + x_2^2$. Find extrema of $f(x)$:

First Order Conditions:

$$\nabla_f(x) = \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x)\right] = [2x_1, 2x_2] = \mathbf{0} \implies [x_1, x_2] = [0, 0]$$

Second Order Conditions:

$$H_f([0,0]) = \left[\begin{array}{cc} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{array}\right] = \left[\begin{array}{cc} 2 & 0 \\ 0 & 2 \end{array}\right] > 0$$

The $f(x)$ function has only one extremum, which is $[0, 0]$ – minimum.

Bibliography

[KW19] M. Kochenderfer and T. Wheeler, “Algorithms for optimization, “*Mit Press*, 2019.

[BI86] A. Birkholc, “Analiza matematyczna: funkcje wielu zmiennych, “*Państwowe Wydawnictwo Naukowe*, 1986.

[CZ04] E.K. Chong and S.H.. Zak, “An introduction to optimization, “*John Wiley & Sons*, 2004.

[CO14] P. Cortez, “Modern optimization with R, “*Springer*, 2014.

[SS08] K. Sydsæter and P. Hammond and A. Seierstadand A. Strom, “Further mathematics for economic analysis “*Pearson education*, 2008.