Optimization and Data Science

Lecture 6: Second Order Optimality Conditions

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

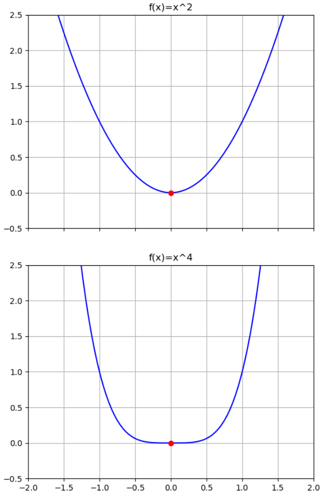
Contents

Optimality conditions for differentiable functions in 1-D, $X_{ad} = X = \mathbb{R}$

- First order necessary condition $f'(x^*) = 0$
- Example: $f(x) = x^2$ local minimum at $x^* = 0$, there: $f'(x^*) = 2x^* = 0$.
- Second order sufficient condition $f'(x^*) = 0, f''(x^*) > 0$.
- $f(x) = x^2$: local minimum at $x^* = 0$, $f''(x^*) = 2 > 0$.
- $f(x) = x^4$: local minimum at $x^* = 0$, $f'(x^*) = 4x^{*3} = 0$. $f''(x^*) = 12x^{*2} = 0$, but $f''(x^*) = 12x^{*2} \not> 0$.
- If $f'(x^*) = 0$ and $f''(x^*) < 0$, then maximum.

$\rightsquigarrow$ Second order necessary condition for minimum: $f'(x^*) = 0$ and $f''(x^*) \geq 0$.

- $X_{ad} \neq X$ (e.g., interval) $\rightsquigarrow$ check boundary points.



Contents

First and second order derivatives in $\mathbb{R}^n$

- **Gradient** of $f : \mathbb{R}^n \to \mathbb{R}$:

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_k}(x) \right)_{k=1}^{n} \in \mathbb{R}^n$$

  with **partial derivatives**:

$$\frac{\partial f}{\partial x_k}(x) := \lim_{h \to 0} \frac{f(x + he_k) - f(x)}{h}, \quad k = 1, \ldots, n.$$

- Matrix of second derivatives: **Hessian matrix**:

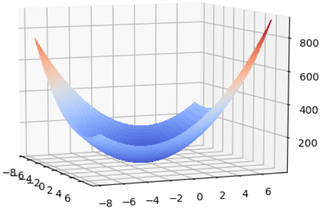
$$\nabla^2 f(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\ldots,n} \in \mathbb{R}^{n \times n}$$

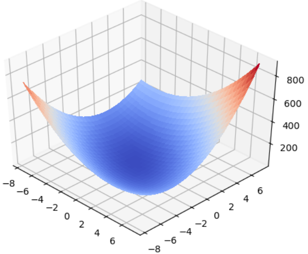
- If all second derivatives are continuous, the Hessian matrix is symmetric.

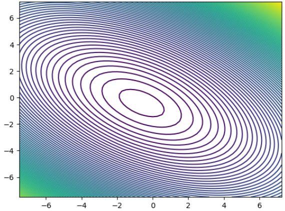
Example: Hessian matrix

- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x) = f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$







- Partial derivatives:

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 8x_1 + 5x_2 + 7, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = 5x_1 + 12x_2 + 8.$$

- Compute the second partial derivatives.

Example: Hessian matrix

- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x) = f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

- Partial derivatives:

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 8x_1 + 5x_2 + 7, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = 5x_1 + 12x_2 + 8.$$

- Second partial derivatives:

$$\frac{\partial^2 f}{\partial x_1^2}(x_1, x_2) = 8, \quad \frac{\partial^2 f}{\partial x_2 \partial x_1}(x_1, x_2) = 5$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(x_1, x_2) = 5, \quad \frac{\partial^2 f}{\partial x_2^2}(x_1, x_2) = 12.$$

- Hessian matrix:

$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix}$$

Contents

Example: Hessian matrix

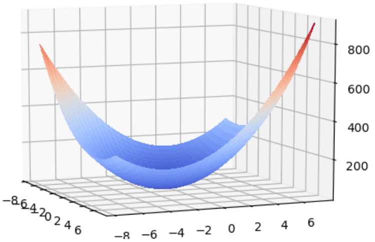
- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

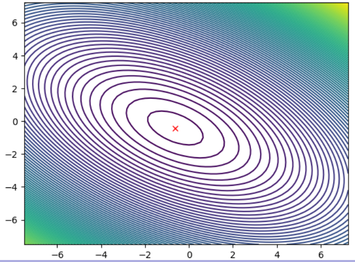
$$f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

- Already computed: $x^* \in \mathbb{R}^2$ with $\nabla f(x^*) = 0$ (1st order necessary condition).
- Obviously the function has a minimum in this point.
- Hessian matrix (is constant):

$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix} \quad \text{for all } x \in \mathbb{R}^2.$$

- What is the multi-dimensional equivalent to the condition $f''(x^*) > 0$ (in $\mathbb{R}$) ...
- ... since $\nabla^2 f(x)$ is a matrix?





Positive definite matrices

Definition

A matrix $A \in \mathbb{R}^{n\times n}$ with

$$A = A^\top, \text{ i.e., } A_{ij} = A_{ji} \text{ for all } i,j = 1,\dots,n,$$

is called **symmetric**.

Definition

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called

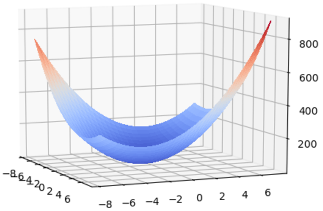
- **positive semi-definite** if

$$x^\top Ax = \sum_{i,j=1}^{n} x_i A_{ij} x_j \geq 0 \quad \text{for all } x \in \mathbb{R}^n,$$

- **positive definite** if additionally

$$x^\top Ax = 0 \Rightarrow x = 0.$$

Example: Hessian matrix



- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

- Hessian matrix:

$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix} \quad \text{for all } x \in \mathbb{R}^2$$

- ... is positive definite, since:

$$\begin{aligned}(x_1, x_2)\begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= (x_1, x_2)\begin{pmatrix} 8x_1 + 5x_2 \\ 5x_1 + 12x_2 \end{pmatrix} = 8x_1^2 + 10x_1x_2 + 12x_2^2 \\ &= 8x_1^2 + 2\sqrt{5}\sqrt{5}x_1x_2 + 12x_2^2 \\ &= 5x_1^2 + 2\sqrt{5}\sqrt{5}x_1x_2 + 5x_2^2 + 3x_1^2 + 7x_2^2 \\ &= (\sqrt{5}x_1 + \sqrt{5}x_2)^2 + 3x_1^2 + 7x_2^2 \geq 0 \quad \text{for all } x \in \mathbb{R}^2\end{aligned}$$

- ... and: $(\sqrt{5}x_1 + \sqrt{5}x_2)^2 + 3x_1^2 + 7x_2^2 = 0 \Rightarrow x_1 = 0, x_2 = 0.$

More characterization of positive definite matrices

A characterization of positive (semi-) definiteness can be given by the eigenvalues:

Definition

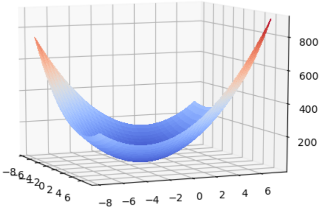
- $\lambda \in \mathbb{C}$ is called **eigenvalue** of $A \in \mathbb{C}^{n\times n}$, if there exists $x \in \mathbb{C}^n \setminus \{0\}$ with $Ax = \lambda x$.
- x is called the corresponding **eigenvector**.

- Eigenvalues can be computed from

$$Ax = \lambda x, x \neq 0 \;\Leftrightarrow\; (A - \lambda I)x = 0, x \neq 0 \;\Leftrightarrow\; (A - \lambda I) \text{ is singular } \;\Leftrightarrow\; \det(A - \lambda I) = 0.$$

- A matrix in $\mathbb{C}^{n\times n}$ has n eigenvalues (that do not have to be different).
- A real matrix in $\mathbb{R}^{n\times n}$ may have complex eigenvalues $\lambda \in \mathbb{C}$.
- Symmetric matrices have only real eigenvalues $\lambda \in \mathbb{R}$.
- The eigenvalues of a positive definite matrix are > 0.
- The eigenvalues of a positive semi-definite matrix are ≥ 0.

Example: Hessian matrix



- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

- Hessian matrix:

$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix} \quad \text{for all } x \in \mathbb{R}^2$$

- Eigenvalues:

$$\det \begin{pmatrix} 8-\lambda & 5 \\ 5 & 12-\lambda \end{pmatrix} = (8-\lambda)(12-\lambda) - 25 = \lambda^2 - 20\lambda + 71 = 0$$

$$\iff \quad \lambda_{1,2} = 10 \pm \underbrace{\sqrt{100-71}}_{<10} > 0.$$

$\rightsquigarrow$ Hessian is positive definite.

Contents

Second order sufficient optimality condition

- Here, we study unconstrained problems only.
- For constrained problems we need special conditions that we study later!

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. If for $x^* \in \mathbb{R}^n$

- $\nabla f(x^*) = 0$ *and*
- *the Hessian matrix $\nabla^2 f(x)$ is* **positive semi-definite** *for all $x \in B_\epsilon(x^*)$ and some $\epsilon > 0$,*

then x^* is a local minimizer.

Our example

- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

- Hessian matrix was constant:

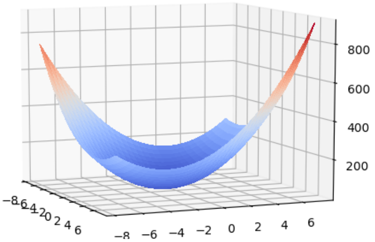
$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix} \quad \text{for all } x \in \mathbb{R}^2.$$

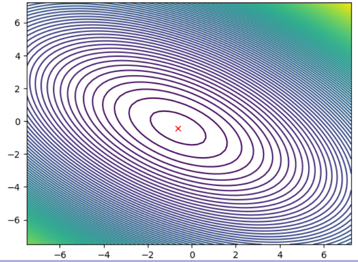
- ... and positive definite (thus also positive semi-definite) **for all** $x \in \mathbb{R}^2$.

$\rightsquigarrow$ x^* with

$$\nabla f(x^*) = 0$$

satisfies 2nd order sufficient condition and is local minimizer.





Second order sufficient optimality condition (for a strict minimizer)

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be twice continuously differentiable. If for* $x^* \in \mathbb{R}^n$

- $\nabla f(x^*) = 0$ *and*
- *the Hessian matrix* $\nabla^2 f(x^*)$ *is* **positive definite (in** x^***)**,

then x^* *is a strict local minimizer. Moreover, it holds*

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|_2^2 \text{ for all } x \in B_\varepsilon(x^*)$$

for some $\alpha, \epsilon > 0$.

Proof.

Luenberger: Linear and Nonlinear Programming §7.3, proposition 3. □

Our example

- Function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x_1, x_2) = 4x_1^2 + 5x_1x_2 + 6x_2^2 + 7x_1 + 8x_2 + 9.$$

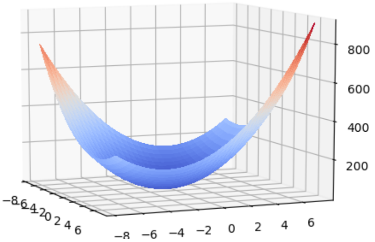
- Hessian matrix was constant

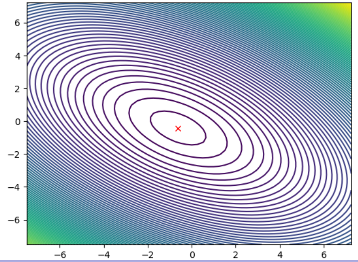
$$\nabla^2 f(x) = \begin{pmatrix} 8 & 5 \\ 5 & 12 \end{pmatrix}$$

- ... and positive definite **for all** $x \in \mathbb{R}^2$.
- Thus, it is also positive definite for x^* with

$$\nabla f(x^*) = 0$$

$\rightsquigarrow$ This point satisfies 2nd order sufficient condition and is even a **strict** local minimizer.





Second order necessary optimality condition

- Recall: we found out in $\mathbb{R}$:
- Second order necessary condition for minimum: $f'(x^*) = 0$ and $f''(x^*) \geq 0$.
- What is the generalization for $\mathbb{R}^n$?

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. If $x^* \in \mathbb{R}^n$ is a local minimizer, then

- $\nabla f(x^*) = 0$ *and*
- *the Hessian matrix $\nabla^2 f(x^*)$ is* **positive semi-definite (in x^* only)**.

Proof.

Luenberger: Linear and Nonlinear Programming §7.3, proposition 2. □

- In our example, this condition was satisfied, since the Hessian was positive semi-definite for all $x \in \mathbb{R}^2$.

Contents

Useful tool: Taylor expansion

- Approximation of differentiable function by polynomial.
- Using function and derivative values at one fixed point x ...
- ... to approximate function value in the vicinity of x.
- Consider $n = 1$, i.e., $f : \mathbb{R} \to \mathbb{R}$. Taylor expansion around x:

$$f(x+h) = \underbrace{f(x)}_{\text{constant w.r.t. } h} + \underbrace{f'(x)h}_{\text{linear in } h} + \underbrace{\frac{1}{2}f''(x)h^2}_{\text{quadratic in } h} + \underbrace{\frac{1}{6}f'''(x)h^3}_{\text{3rd order in } h} + \dots$$

$$= \underbrace{\sum_{k=0}^{N} \frac{f^{(k)}(x)}{k!}h^k}_{\text{expansion of order } N} + \underbrace{\frac{f^{(N+1)}(x+th)}{(N+1)!}h^{N+1}}_{\text{remainder term}} \text{ with some } t \in [0,1].$$

- Assumption: all derivatives exist.

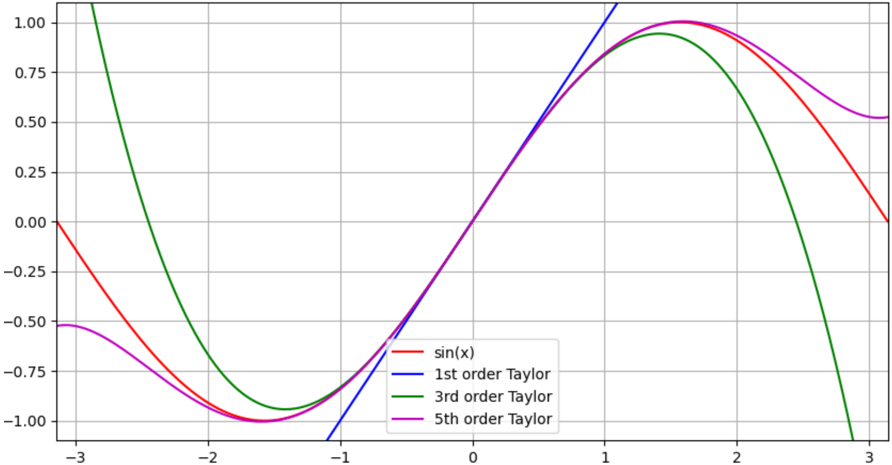
Example: Taylor expansion

Taylor expansion of $f : \mathbb{R} \to \mathbb{R}, f(x) = \sin(x)$ around $x = 0$:

$$\begin{aligned}
\sin(h) &= f(0) + f'(0)h + \frac{1}{2}f''(0)h^2 + \frac{1}{6}f'''(0)h^3 + \frac{1}{24}f^{(4)}(0)h^4 + \frac{1}{120}f^{(5)}(0)h^5 + \dots \\
&= \underbrace{\sin(0)}_{=0} + \underbrace{\cos(0)}_{=1} h - \frac{1}{2}\underbrace{\sin(0)}_{=0} h^2 - \frac{1}{6}\underbrace{\cos(0)}_{=1} h^3 + \frac{1}{24}\underbrace{\sin(0)}_{=0} h^4 + \frac{1}{120}\underbrace{\cos(0)}_{=1} h^5 + \dots \\
&= h - \frac{1}{6}h^3 + \frac{1}{120}h^5 + \dots
\end{aligned}$$

Example: Taylor expansion

Taylor expansion around $x = 0$: $\sin(h) = h - \frac{1}{6}h^3 + \frac{1}{120}h^5 + \ldots$



Taylor expansion for arbitrary dimension

- We only need $N = 0$ (Mean value theorem)

$$f(x+h) = f(x) + \underbrace{\nabla f(x+th)^\top h}_{\text{linear in } h} \text{ with some } t \in [0,1]$$

- ... and $N = 1$:

$$f(x+h) = f(x) + \underbrace{\nabla f(x)^\top h}_{\text{linear in } h} + \underbrace{\frac{1}{2} h^\top \nabla^2 f(x+th) h}_{\text{quadratic in } h} \text{ with some } t \in [0,1].$$

- Now: $x, h, \nabla f(x) \in \mathbb{R}^n$ are **vectors**, $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ is the Hessian **matrix**.

Second order sufficient optimality condition

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. If for $x^* \in \mathbb{R}^n$

- $\nabla f(x^*) = 0$ *and*
- *the Hessian matrix* $\nabla^2 f(x)$ *is* **positive semi-definite** *for all* $x \in B_\epsilon(x^*)$ *and some* $\epsilon > 0$,

then x^* *is a local minimizer.*

Proof.

- Let $h \in \mathbb{R}^n$ with $\|h\| < \epsilon$. Then we have $x^* + h \in B_\epsilon(x^*)$.
- Taylor expansion: There exists $t \in [0, 1]$ such that

$$f(x^* + h) = f(x^*) + \nabla f(x^*)^\top h + \frac{1}{2} h^\top \nabla^2 f(x^* + th)h \geq f(x^*).$$

- $t \in [0, 1] \Rightarrow x^* + th \in B_\epsilon(x^*)$
- $\nabla f(x^*) = 0$ and $h^\top \nabla^2 f(x^* + th)h \geq 0$, since Hessian positive semi-definite in $B_\epsilon(x^*)$.

What is important?

- Second order derivative in n dimensions is the Hessian matrix.
- Second order optimality conditions are based on the concept of positive (semi-) definiteness of the Hessian.
- This can be characterized by the eigenvalues.
- Second order sufficient condition: gradient zero and Hessian positive semi-definite in a neighborhood of a point.
- Second order sufficient condition for a strict minimum: gradient zero and Hessian positive definite in a point.
- Second order necessary condition: gradient zero and Hessian positive semi-definite in the minimizer.
- The proofs are conducted using Taylor expansion ...
- ... which approximates a function in the vicinity of a point using its derivatives at this point.