Directional Derivative

The **directional derivative** $D_d f(\mathbf{x}) = f'(\mathbf{x}; \mathbf{d})$ of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^n$ in direction $\mathbf{d} \in \mathbb{R}^n$ is defined as

$$0 > \nabla f(x)^T d \simeq \quad D_d f(x) := \lim_{t \downarrow 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}.$$

$$\overset{\mathbf{d} = e_i}{=} \lim_{t \downarrow 0} \frac{f(x + t e_i) - f(x)}{t} = \frac{\partial f}{\partial x_i}(x)$$

Properties:

- If $\mathbf{d} = \mathbf{e_i}$, the unit vector of coordinate i, then $D_d f = \frac{\partial f}{\partial x_i}(\mathbf{x})$.
- If f is continuously differentiable, then $D_d f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{d}$.
- If $D_d f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \mathbf{d} < 0$, then $\mathbf{d}$ is a **direction of descent**, i.e.

$$f(\mathbf{x} + \alpha \mathbf{d}) < f(\mathbf{x}) \quad \text{for all } \alpha \in (0, \overline{\alpha}].$$

Taylor Expansion I

Taylor expansion

- If f differentiable,
  then f may be approximated in a neighbourhood of $\hat{\mathbf{x}}$ by an affine-linear function as

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \overbrace{\nabla f(\hat{\mathbf{x}})^\top \underbrace{(\mathbf{x} - \hat{\mathbf{x}})}_{= \mathbf{d}t}}^{\text{directional derivative in direction } t\mathbf{d}} + R(\|\mathbf{x} - \hat{\mathbf{x}}\|),$$

  where

$$\lim_{\mathbf{x} \to \hat{\mathbf{x}}} \frac{R(\|\mathbf{x} - \hat{\mathbf{x}}\|)}{\|\mathbf{x} - \hat{\mathbf{x}}\|} = 0$$

  or with $t \in \mathbb{R}$, $\mathbf{d} \in \mathbb{R}^n$, $\xi_t \in (0, t)$, resp.,

$$f(\hat{\mathbf{x}} + t\mathbf{d}) = f(\hat{\mathbf{x}}) + t\,\nabla f(\hat{\mathbf{x}} + \xi_t \mathbf{d})^\top \mathbf{d}.$$

Taylor Expansion II

- If f 2times differentiable, then f may be approximated in a neighbourhood of $\hat{\mathbf{x}}$ by a quadratic function as

$$f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top(\mathbf{x}-\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^\top H_f(\hat{\mathbf{x}})\underbrace{(\mathbf{x}-\hat{\mathbf{x}})}_{t\mathbf{d}}$$
$$+ R(\|\mathbf{x}-\hat{\mathbf{x}}\|^2)$$

or, resp.,

$$f(\hat{\mathbf{x}}+t\mathbf{d}) = f(\hat{\mathbf{x}}) + t\,\nabla f(\hat{\mathbf{x}})^\top\mathbf{d} + \frac{t^2}{2}\mathbf{d}^\top H_f(\hat{\mathbf{x}}+\xi_t\mathbf{d})\mathbf{d}.$$

Taylor Expansion - Example

Approximate around $(\hat{x}_1, \hat{x}_2) = (0, 0)$ the function

$$f(x_1, x_2) = \exp(x_1)\ln(1 + x_2). \overset{x=\hat{x}}{=} 1 \cdot 0 = 0$$

$$\nabla f(x_1, x_2) = \begin{pmatrix} \exp(x_1)\ln(1+x_2) \\ \exp(x_1)\dfrac{1}{1+x_2} \end{pmatrix} \overset{x=\hat{x}}{=} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$H_f(x_1, x_2) = \begin{pmatrix} \exp(x_1)\ln(1+x_2) & \dfrac{\exp(x_1)}{1+x_2} \\ \dots & -\dfrac{\exp(x_1)}{(1+x_2)^2} \end{pmatrix} \overset{x=\hat{x}}{=} \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$$

$$f(x) = f(\hat{x}) + \nabla f(\hat{x})^T (x - \hat{x}) + \frac{1}{2}(x - \hat{x}) H_f(\hat{x})(x - \hat{x}) + \dots$$

$$= 0 + (0, 1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2}(x_1, x_2)\underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{= \begin{pmatrix} x_2 \\ x_1 - x_2 \end{pmatrix}} + \dots$$

$$= x_2 + \cancel{\tfrac{2}{2}}\, x_1 x_2 - \frac{x_2^2}{2} + \dots$$

Error Analysis

Instead of the (usually unknown) real values

$$\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)^\top$$

we measure approximate values

$$\mathbf{x} = (x_1, \dots, x_n)^\top.$$

The measurement errors (uncertainty, observational deviations)

$$\Delta x_i := x_i - \hat{x}_i, 1, \dots, n$$

yield an error for the value of a function $f : \mathbb{R}^n \to \mathbb{R}$

$$\Delta f(\mathbf{x}) := f(\mathbf{x}) - f(\hat{\mathbf{x}}) \approx \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}})$$

Cauchy-Schwarz ineq.:

or in absolute values $|\Delta f(\mathbf{x})| \approx |\nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}})| \leq |\nabla f(\hat{\mathbf{x}})| \cdot |\mathbf{x} - \hat{\mathbf{x}}|$

Propagation of Uncertainty

Suppose

$$|\Delta f(\mathbf{x})| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i(\hat{\mathbf{x}})} \right| |\Delta x_i| \,.$$

If the "scattering" of errors is known as $|x_i| \leq S_i$, then the absolute maximal error S reads

$$|\Delta f(\mathbf{x})| \leq S := \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i(\hat{\mathbf{x}})} \right| S_i .$$

Propagation of Uncertainty - Example

Given: measurement of 2 time intervals
$x = (20 \pm 0.1)[ms]$ and $y = (15 \pm 0.1)[ms]$

Searched for: maximal error of time difference $x - y$

Solution: Set $f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) = x - y$.

$$\nabla f(x, y) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Absolute maximal error:

$$S = \left|\frac{\partial f}{\partial x}(20, 15)\right| \cdot 0.1 + \left|\frac{\partial f}{\partial y}(20, 15)\right| \cdot 0.1 = 0.2\,[ms]$$

Relative maximal error:

$$\frac{S}{|f(20, 15)|} = \frac{0.2}{5} = 4\%.$$

Error Analysis - Rules of Thumbs

Sum or difference:
Absolute maximal error is sum of maximal errors of inputs

Product or quotient:
Relative maximal error is sum of relative errors of inputs

Remark:
More realistic for an error estimation is a stochastic error analysis (Gaussian propagation of uncertainty)
$\rightsquigarrow$ stochastics lecture

Optimization in Higher Dimensions

As an application we consider again the problem of minimizing a given function

$$f : \mathbb{R}^n \to \mathbb{R}$$

but now with n variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$.

For simplicity, we consider the case without restrictions (constraints) on $\mathbf{x}$.
The (local) conditions derived in this subsection hold in the interior of a domain, but not on boundaries.

Note that maximization problems for $g : \mathbb{R}^n \to \mathbb{R}$ are equivalent to minimization problems where $f = -g$.

Unrestricted Optimization for any $n \in \mathbb{N}$

We consider

Problem (Unrestricted optimization problem (UOP))

Minimize $f(x)$ subject to the constraint $x \in \mathbb{R}^n$.

Thereby let $f : \mathbb{R}^n \to \mathbb{R}$ an at least 1x (better 2x) differentiable function.

Idea:
Search at first for local minima (from which one obtains global extrema under certain conditions).

Aims:
Necessary conditions
Sufficient conditions
Wish: necessary & sufficient conditions

Remark:
Numerical methods are mostly based on necessary conditions.

Revision/Outlook: Definite Matrices

Matrices do not have signs, but:

Definition (Definiteness of a Matrix)

Let $H \in \mathbb{R}^{n \times n}$ and $v \in \mathbb{R}^n \setminus \mathbf{0}$.

$$\text{If } v^\top H v \left\{ \begin{array}{c} > \\ < \\ \geq \\ \leq \end{array} \right\} 0 \text{ for all } v, \text{ then } H \text{ is } \left\{ \begin{array}{c} \textbf{positive definite} \\ \textbf{negative definite} \\ \textbf{positive semi-definite} \\ \textbf{negative semi-definite} \end{array} \right. .$$

If $v^\top H v > 0$ for some v and $v^\top H v < 0$ for another v (i.e. H is neither positive semi-definite nor negative semi-definite), then H is **indefinite**.
If $v^\top H v = 0$ for some v, then H is **singular**.

For functions in several variables, the Hessian plays the role of the 2nd derivative.
The definiteness of the Hessian generalizes the sign of the 2nd derivative of a function in 1 variable.

Revision/Outlook: Definite Matrices - Examples

Identity matrix (unit matrix) $\mathbb{1}_n = E_n$

$$v^T E_n v = \sum_{i=1}^{n} v_i^2 > 0 \Rightarrow E_n > 0 \quad \text{(pos. def.)} \qquad = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & 1 \end{pmatrix}$$

Null matrix $\quad 0_{n\times m} \overset{m=n}{=} 0_n = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$

$v^T 0_n v = 0 \quad \Rightarrow \quad 0_n$ pos. semi-def.
neg. semi-def.
singular

$$H = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \qquad v = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad v^T H v = (-1, 1)\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$= -1 - 1 = -2 < 0$$

$\Rightarrow H$ pos. (semi-) def.

Necessary Condition of 1st Order

Theorem (Necessary Condition of 1st Order)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $\hat{\mathbf{x}} \in \mathbb{R}^n$ a local minimum of f.

Then there holds

$$\nabla f(\hat{\mathbf{x}}) = \mathbf{0}.$$

Definition (Stationary Point)

Any point $\mathbf{x} \in \mathbb{R}^n$ with $\nabla f(\mathbf{x}) = \mathbf{0}$ is called a **stationary point** of $f : \mathbb{R}^n \to \mathbb{R}$.

Stationary points are not automatically minima, but candidates for (local) minima!

Most numerical methods try to approximate stationary points.

Necessary Condition of 2nd Order

Theorem (Necessary Condition of 2nd Order)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be 2x continuously differentiable and $\hat{\mathbf{x}} \in \mathbb{R}^n$ a local minimum of f.

Then the Hessian

$$H_f(\hat{\mathbf{x}})$$

is positive semi-definite, i.e. $H_f(\hat{x}) \geq 0$.

In case $n = 1$ we have

$$f''(\hat{x}) \geq 0.$$

Likewise, this condition only yields potential candidates for a (local) minimum.

Sufficient Condition of 2nd Order

Theorem (Sufficient Condition of 2nd Order)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be 2x continuously differentiable and $\hat{\mathbf{x}} \in \mathbb{R}^n$ a stationary point of f with positive definite Hessian, i.e. $H_f(\hat{x}) > 0$.

Then $\hat{\mathbf{x}}$ is a strict local minimum of f.

In case $n = 1$ we assume that

$$f''(\hat{x}) > 0.$$

This allows us to decide whether a candidate $\hat{\mathbf{x}}$ that fulfills necessary conditions (1st or 2nd order) is indeed a local minimum.

Revision: Gradient

- The **gradient** of a function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ is defined as a column vector

$$\nabla f(x_1, \ldots, x_n) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_1, \ldots, x_n) \end{pmatrix}.$$

In case $n = 1$ the gradient is the 1st derivative of f at point x, shortly $f'(x)$.

Illustratively, the gradient describes the steepest ascent.

The gradient is orthogonal on level curves.

Revision: Hessian

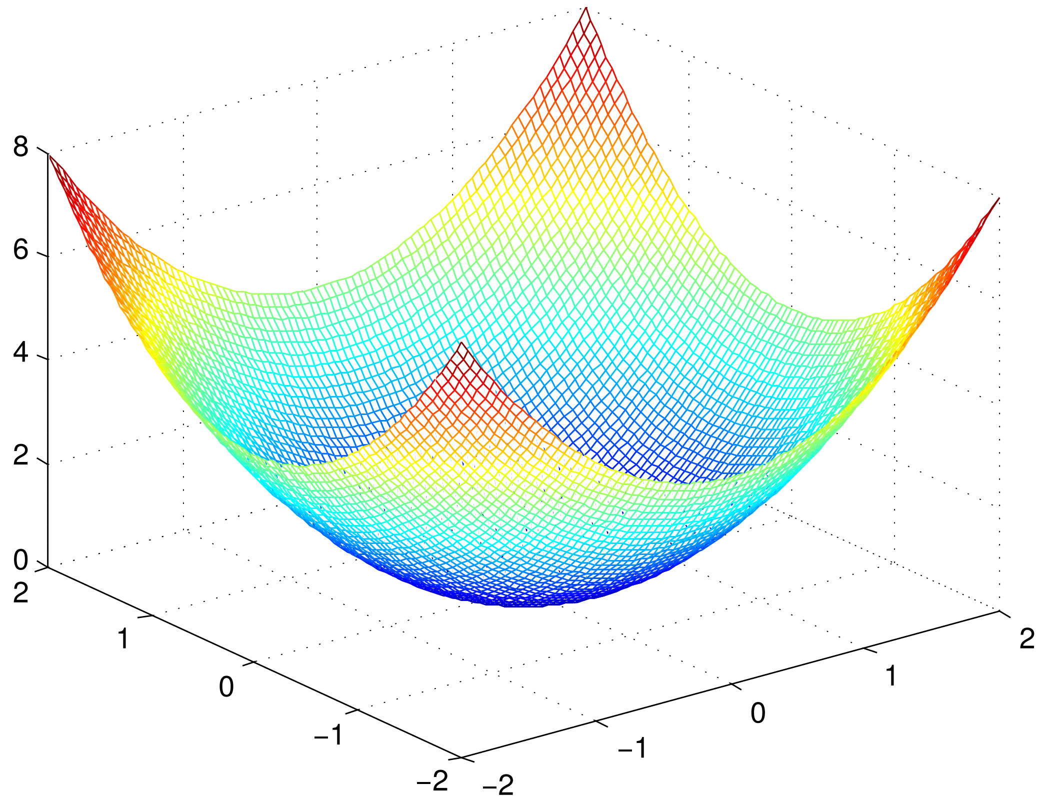
- The **Hessian** of a function f at a point $\mathbf{x}$ is defined as the quadratic matrix

$$H_f(x_1,\ldots,x_n) = H_f(\mathbf{x}) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}) \end{pmatrix}.$$

  In case $n = 1$ the Hessian ist the 2nd derivative of f at a point x, shortly $f''(x)$.

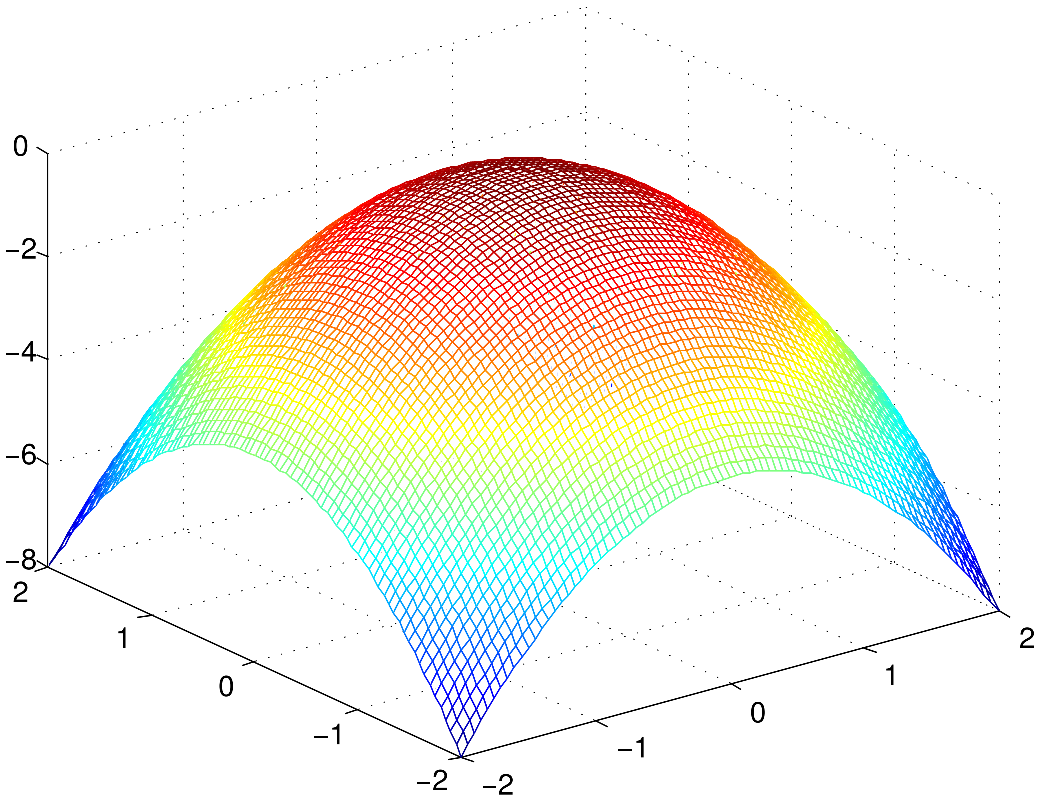
  Illustratively, the Hessian describes the local curvature of a function.

Hessian in 2d - Example 1



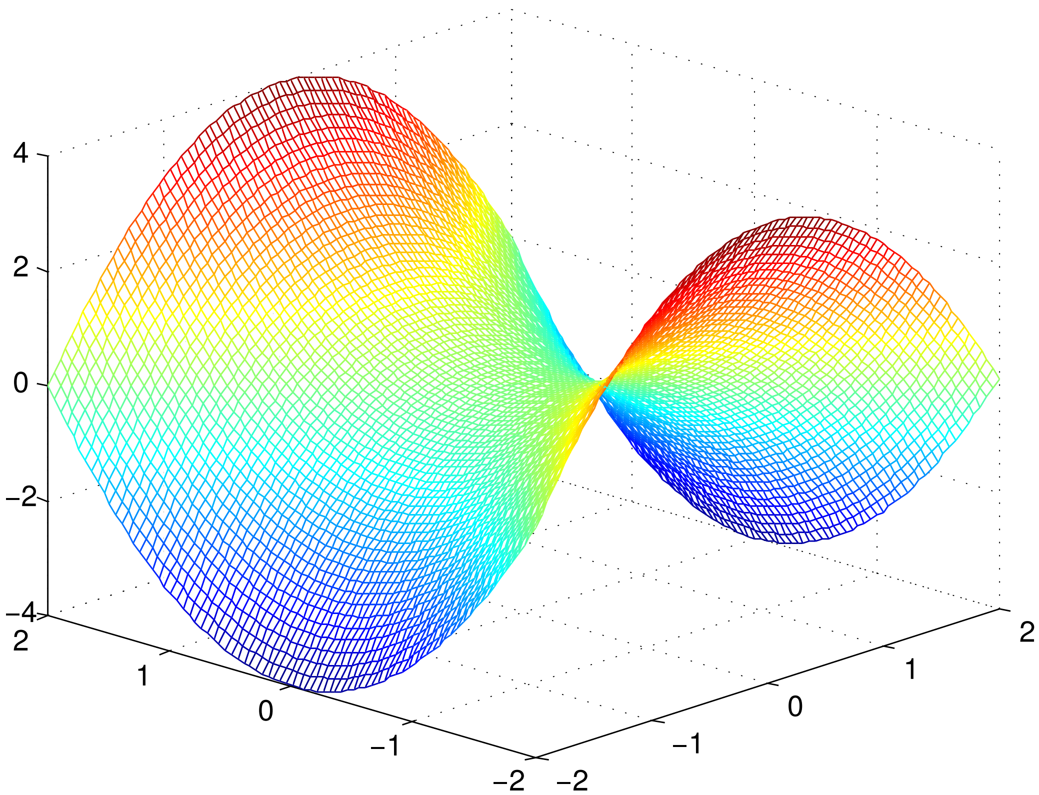
$$f(x_1, x_2) = x_1^2 + x_2^2, \qquad H_f(\hat{x}_1, \hat{x}_2) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Hessian in 2d - Example 2



$$f(x_1, x_2) = -x_1^2 - x_2^2, \quad H_f(\hat{x}_1, \hat{x}_2) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$$

Hessian in 2d - Example 3



$$f(x_1, x_2) = x_1^2 - x_2^2, \quad H_f(\hat{x}_1, \hat{x}_2) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

$x_2^2 - x_1^2$ $\qquad$ $\begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix}$

Hessian in 2d - Examples 4 & 5

Other possibilities for definite matrices in $\mathbb{R}^{2\times 2}$:

$$H_f(\hat{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$ pos. semi-def.

$$H_f(\hat{x}) = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}$$ neg. semi-def.

$$H_f(\hat{x}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

all 3 matrices are singular

Stationary Points

Points $\hat{\mathbf{x}}$ with $\nabla f(\hat{\mathbf{x}}) = (0, 0, \dots, 0)^\top$ are called **stationary points**.

Depending on the properties of the Hessian we find:

- **positive definite:** local minimum
- **negative definite:** local maximum
- **indefinite:** saddle point
- **singular:** everything is possible (e.g. for positive/negative semi-definite Hessian)

Optimization in $\mathbb{R}^n$ - Example

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x,y) = \exp(x) \cdot (2x + y^2)$$

$$\nabla f(x,y) = \begin{pmatrix} \exp(x)\,(2x+y^2+2) \\ \exp(x)\cdot 2y \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \hat{y} = 0$$

$$\Rightarrow \hat{x} = -1$$

The stationary point $\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ is the only candidate.

$$H_f(x,y) = \begin{pmatrix} \exp(x)\,(2x+y^2+2+2) & \exp(x)\cdot 2y \\ \exp(x)\cdot 2y & \exp(x)\cdot 2 \end{pmatrix}$$

$$H_f(-1,0) = \begin{pmatrix} 2/e & 0 \\ 0 & 2/e \end{pmatrix} \text{ pos. def.} \Rightarrow \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

is a strict local min.