Higher Order Derivatives

**Definition 1 (Higher - order derivatives)** refer to the derivatives of derivatives, taking higher-order derivatives involves repeatedly finding the derivative of a function. Example: the second derivative is the derivative of the first derivative, the third derivative is the derivative of the second derivative, and so on.

Consider \( f : \mathbb{R}^n \to \mathbb{R} \), assume it is differentiable, so all partial derivatives \( \frac{\partial f}{\partial x_i} : \mathbb{R}^n \to \mathbb{R} \) exist. If this function is differentiable, we can take its derivative:

\[
\frac{\partial}{\partial x_i} \left( \frac{\partial f}{\partial x_j} \right) = \frac{\partial^2 f}{\partial x_i \partial x_j}
\]

These are called second order partial derivatives.

⚠️ In general, we cannot change the order of derivatives:

\[
\frac{\partial^2 f}{\partial x_i \partial x_j} \neq \frac{\partial^2 f}{\partial x_j \partial x_i}
\]

Example:

\[
f(x, y) = \frac{x \cdot y^3}{x^2 + y^2}
\]

\[
\nabla f(x, y) = \left( \frac{y^3(y^2 - x^2) - xy^2(3x^2 + y^2)}{(x^2 + y^2)^2}, \frac{3x^2y^2 - x^4 + y^4}{(x^2 + y^2)^2} \right)
\]

Have:

\[
\frac{\partial f}{\partial x}(0, y) = y \quad \forall y, \quad \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} \right) = 0
\]

\[
\frac{\partial f}{\partial y}(x, 0) = 0 \quad \forall x, \quad \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right) = 0
\]

we can see that \( 1 \neq 0 \).

**Definition 2 (Continuously Differentiable.)** We say that \( f : \mathbb{R}^n \to \mathbb{R} \) is continuously differentiable if all partial derivatives exist and are continuous.

We say that \( f \) is twice continuously differentiable if \( f \) is continuously differentiable and all its partial derivatives \( \frac{\partial^2 f}{\partial x_i} \) are again continuously differentiable.

Analogously: \( k \) times continuously differentiable
Notation:
\[ C^k(\mathbb{R}^n, \mathbb{R}^m) = \{ f : \mathbb{R}^n \to \mathbb{R}^m | k \text{ times continuously differentiable} \} \]
\[ C^\infty(\mathbb{R}^n, \mathbb{R}^m) = \{ f : \mathbb{R}^n \to \mathbb{R}^m | \text{\infty often continuously differentiable} \} \]

**Theorem 3 (Schwartz)** Assume that \( f \) is twice continuously differentiable. Then we can exchange the order in which we take partial derivatives:
\[
\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}
\]

Analogously: \( k \) times continuously differentiable \( \implies \) can exchange order of first \( k \) partial derivatives.

⚠️ Caution about derivatives:
- \( f : \mathbb{R}^n \to \mathbb{R} \) ← function
- \( \nabla f : \mathbb{R}^n \to \mathbb{R}^n \) ← first derivatives(\( \frac{\partial f}{\partial x_i} \)): \( n \) partial derivatives
- \( Hf : \mathbb{R}^n \to \mathbb{R}^{n \times n} \) ← second derivatives(\( \frac{\partial^2 f}{\partial x_i \partial x_j} \)): \( n^2 \) partial derivatives

**Definition 4 (Hessian Matrix)** \( f : \mathbb{R}^n \to \mathbb{R}, \) then we define the Hessian of \( f \) at point \( x \) by,
\[
(Hf)_{ij}(x) := \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad i, j = 1, 2, 3, ..., n
\]
Minima/Maxima

Definition 5 (Critical Point) \( f : \mathbb{R}^n \to \mathbb{R} \) differentiable. If \( \nabla f(x) = 0 \) then we call \( x \) a critical point.

- \( f \) has a local minimum at \( x_0 \) if there exists \( \epsilon > 0 \), such that \( \forall x \in B_\epsilon(x_0) : f(x) \geq f(x_0) \)
- \( f \) has a strict local minimum at \( x_0 \) if there exists \( \epsilon > 0 \), such that \( \forall x \in B_\epsilon(x_0) : f(x) > f(x_0) \)

- \( f \) has a local maximum at \( x_0 \) if there exists \( \epsilon > 0 \), such that \( \forall x \in B_\epsilon(x_0) : f(x) \leq f(x_0) \)
- \( f \) has a strict local maximum at \( x_0 \) if there exists \( \epsilon > 0 \), such that \( \forall x \in B_\epsilon(x_0) : f(x) < f(x_0) \)
- If \( f \) is differentiable and \( x_0 \) is a critical point that is neither a local minima nor a local maximum. We call it a saddle point.

- \( f \) has a global minimum at \( x_0 \) if \( \forall x : f(x) \geq f(x_0) \)

- \( f \) has a global maximum at \( x_0 \) if \( \forall x : f(x) \leq f(x_0) \)
How can we identity which type of point we have?

**Intuition in \( \mathbb{R} \):**

![Graphs showing local minima, local maxima, and saddle points](image)

**Theorem 6** \( f : \mathbb{R}^n \to \mathbb{R}, \ f \in C^2(\mathbb{R}^n) \). Assume that \( x_0 \) is a critical point, i.e \( \nabla f(x_0) = 0 \). Then:

(i) If \( x_0 \) is a local minimum (maximum), then the Hessian \( H_f(x_0) \) is positive semi definite (negative semi definite).

(ii) If \( H_f(x_0) \) is positive definite (negative definite), then \( x_0 \) is a strict local minimum (maximum). If \( H_f(x_0) \) is indefinite then \( (x_0) \) is a saddle point.
Matrix/Vector Calculus

Example: Linear Least Squares

\[ f : \mathbb{R}^n \to \mathbb{R} \]
\[ \text{pred } \hat{y}(w) = Aw \text{ where,} \]
\[ \hat{y} \text{ is prediction, } A \text{ - input data}\]
\[ w \text{ - weight vector (parameters we want to find).} \]

\[ f(w) = \| y - \hat{y}(w) \|_2^2 = \| y - Aw \|_2^2 \]
\[ f(w) \text{ - how good pred. is with parameter } w. \]

We want to minimize \( f(w) \). Thus, we need to look at \( \nabla f : \mathbb{R}^n \to \mathbb{R}^n \).

Compute Gradient:

\[
\begin{align*}
\frac{\partial f}{\partial w_i} &= \sum_{j=1}^{n} 2(-a_{ji})(y_j - \sum_{k=1}^{n} a_{jk}w_k) \\
&= \sum_{j=1}^{n} 2(-a_{ji})(y_j - \sum_{k=1}^{n} a_{jk}w_k) \\
&\quad \text{where } \sum_{k=1}^{n} a_{jk}w_k = (Aw)_j, \\
&\quad \text{and} \\
&\quad (y_j - \sum_{k=1}^{n} a_{jk}w_k) = y - (Aw)_j, \
&\quad -2 \sum_{j=1}^{n} 2(-a_{ji})(y_j - \sum_{k=1}^{n} a_{jk}w_k) = (A^T(y - Aw))_i \\
\n\n\n\end{align*}
\]

\[ \nabla f(w) = -2A^T(y - Aw) \]

Intuition: "syntax" close to 1-dim case:
\[ f(w) = (y - aw)^2 \]
\[ f'(w) = -a(y - aw) \cdot 2 = -2a(y - aw) \]

Matrix-Vector Calculus: Lookup table ("matrix cookbook") for gradients of many important functions:

\[ f : \mathbb{R}^n \to \mathbb{R}. \]

- \( f(x) = a^T x \quad (a \in \mathbb{R}^n) \)
  \[ f(x) = \langle a, x \rangle \]
  \[ \frac{\partial f}{\partial x} = a \in \mathbb{R}^n \]
- \( f(x) = x^T Ax \implies \frac{\partial f}{\partial x} = (A + A^T)x \in \mathbb{R}^n \)
\[ f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}. \]

- \( f(x) = a^T X b \implies \frac{\partial f}{\partial x} = ab^T \in \mathbb{R}^{n \times m} \)
  where \( X \) is \( \mathbb{R}^{n \times m} \), \( a^T \) is \( 1 \times n \) and \( b \) is \( m \times 1 \) dimensions.

- \( f(x) = a^T X^T C X b \implies \frac{\partial f}{\partial x} = C^T X a b + C X b a^T \)
  where \( a^T \) is \( a \times m \), \( X^T \) is \( m \times n \), \( C \) is \( n \times n \), \( X \) \( n \times m \), and \( b \) is \( m \times 1 \) dimensions.

- \( f(X) = \text{tr}(X) \implies \frac{\partial f}{\partial X} = I \)
  where \( \text{tr}(X) \) is the trace and \( I \) is the identity matrix.

- \( f(X) = \text{tr}(AX) \implies \frac{\partial f}{\partial X} = A \)

  \( f(X) = \text{tr}(X^T A X) \implies \frac{\partial f}{\partial X} = (A + A^T) X \)

- \( f(X) = \det(X) \rightarrow \text{Determinant} \)

  \[ \frac{\partial \det}{\partial X} = \det(X)(X^{-1}) \]

  \[ \frac{\partial \det}{\partial x_{sr}} = \det(X)(X^{-1})_{rs} \]

\[ f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m} \text{ Inverse}. \]

- \( f(A) = A^{-1}, \ f_{ij} := (A^{-1})_{ij} \)

  \[ \frac{\partial f_{ij}}{\partial a_{uv}} = -(a_{iu})^{-1}(a_{uj})^{-1} \]