Higher order derivatives

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:

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

These are called second order partial derivatives. \( \frac{\partial^2 f}{\partial x_i \partial x_j} \neq \frac{\partial^2 f}{\partial x_j \partial x_i} \)

⚠️ in general, we cannot change the order of derivatives.

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

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

Have: \( \frac{\partial f}{\partial x}(0, y) = y \quad \forall \ y \), \( \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} \right) = 1 \)

\( \frac{\partial f}{\partial y}(x, 0) = 0 \quad \forall \ x \), \( \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right) = 0 \)
**Def:** 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 \( \frac{\partial f}{\partial x_i} \) are again continuously differentiable.

**Notation:** \( C^k(\mathbb{R}^n, \mathbb{R}^m) = \{ f: \mathbb{R}^n \to \mathbb{R}^m \mid k \text{ times cont. differentiable} \} \)

\( C^\infty(\mathbb{R}^n, \mathbb{R}^m) = \{ f: \mathbb{R}^n \to \mathbb{R}^m \mid \infty \text{ often cont. diff.} \} \)

**Theorem (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 cont. diff. \( \Rightarrow \) can exchange order of first \( k \) partial derivatives.
Caution about derivatives

\( f : \mathbb{R}^n \rightarrow \mathbb{R} \) \quad \text{← function}

\( \nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) \quad \text{← first derivative: } n \text{ partial derivatives}

\( H f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n} \) \quad \text{← second derivative: } \begin{align*}
&n^2 \text{ "partial derivatives"} \\
&\frac{\partial f}{\partial x_i; \partial x_j}
\end{align*}

\underline{Def:} \quad \text{Hessian matrix}

\( f : \mathbb{R}^n \rightarrow \mathbb{R} \), then we define the Hessian of \( f \) at point \( x \) by,

\[
(H f)_{ij} (x) := \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad i, j = 1, 2, \ldots, n
\]
**Minima/Maxima**

**Def:** $f: \mathbb{R}^n \to \mathbb{R}$ differentiable. If $\forall f(x) = 0$ then we call $x$ a **critical point**.

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

$f$ has a **strict local minimum** at $x_0$ if $\exists \varepsilon > 0$ such that $\forall x \in B_\varepsilon(x_0): f(x) > f(x_0)$

$f$ has a **local maximum** (resp, a strict local maximum) $\forall x \in B_\varepsilon(x_0): f(x) \leq f(x_0)$.

If $f$ is diff. and $x_0$ is a critical point that is neither a local min./local max. we call it a **saddle point**.
f has a global minimum at $x_0$ if
\[ \forall x: f(x) \geq f(x_0) \]

How can we identify which type of point we have?

Intuition in $\mathbb{R}$:

- **Local Minimum**: $f'(x) = 0$, $f''(x) > 0$

- **Local Maximum**: $f'(x) = 0$, $f''(x) < 0$

- **Saddle Point**: $f'(x) = 0$, $f''(x) = 0$
Theorem: \( f: \mathbb{R}^n \to \mathbb{R}, \ f \in C^2(\mathbb{R}^n) \). 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 \( Hf(x_0) \) is positive semi definite (negative semi definite).

(ii) If \( Hf(x_0) \) is positive definite (negative definite), then \( x_0 \) is a strict local min (max). If \( Hf(x_0) \) is indefinite then \( x_0 \) is a saddle point.
Example: Linear least squares

\[ f: \mathbb{R}^n \rightarrow \mathbb{R} \]

\[ \text{pred } \hat{y}(w) = A w \]

(weight vector)

(input data)

(\text{params we want to find})

\[ f(w) = \| y - \hat{y}(w) \|_2^2 = \| y - Aw \|_2^2 \]

how good pred. is with param \( w \).

Want to minimize \( f(w) \). Need to look at \( \nabla f: \mathbb{R}^n \rightarrow \mathbb{R}^n \).

Compute gradient:

\[ f(\mathbf{w}) = \sum_{j=1}^{n} (y_j - \sum_{k=1}^{n} a_{jk} w_k)^2 \]

\[ \frac{df}{d\omega_i} = \sum_{j=1}^{n} (-a_{ji}) \cdot 2 \cdot (y_j - \sum_{k=1}^{n} a_{jk} w_k) \]

\[ = -2 \cdot \sum_{j=1}^{n} a_{ij} \cdot (\text{At}(y-Aw))_i \]
\( \nabla f(\omega) = -2A^T(y-A\omega) \)

**Intuition:** "syntax" close to 1-dim case:

\[
\begin{align*}
    f(\omega) &= (y-a\omega)^2 \\
    f'(\omega) &= -a(y-a\omega) \quad 2 = -2a(y-a\omega)
\end{align*}
\]

Matrix-vector calculus: lookup table ("matrix cookbook") for gradients of many important functions:

**\( f: \mathbb{R}^n \rightarrow \mathbb{R} \)**

- \( f(x) = a^T x \quad (a \in \mathbb{R}^n) \)
  \[
  = \langle a, x \rangle \\
  \frac{\partial f}{\partial x} = a \in \mathbb{R}^n
  \]

- \( f(x) = x^T A x \Rightarrow \frac{\partial f}{\partial x} = (A + A^T)x \in \mathbb{R}^n \)

**\( f: \mathbb{R}^{n \times m} \rightarrow \mathbb{R} \)**

- \( f(x) = \begin{pmatrix} a^T & b \end{pmatrix} \Rightarrow \frac{\partial f}{\partial x} = a \cdot b^T \in \mathbb{R}^{n \times m} \)
\[
f(x) = \frac{a^T x^T C x b}{x^m X^m x^m x^m m^x x^1}
\]
\[
\frac{\partial f}{\partial x} = C^T x a b + C x b a^T
\]
\[f(x) = \text{tr}(x) \rightarrow \text{Trace}\]
\[
\frac{\partial f}{\partial x} = I
\]
\[f(x) = \text{tr}(A x) \Rightarrow \frac{\partial f}{\partial x} = A\]
\[f(x) = \text{tr}(x^T A x) \Rightarrow \frac{\partial f}{\partial x} = (A + A^T)x\]
\[f(x) = \det(x) \rightarrow \text{Determinant}\]
\[
\frac{\partial f}{\partial x} = \det(x) \cdot (x^T)^{-1}
\]
\[
\frac{\partial \det}{\partial x_{sr}} = \det(x) \cdot (x^{-1})_{rs}
\]
$f: \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}$

$\text{inverse.}$

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

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