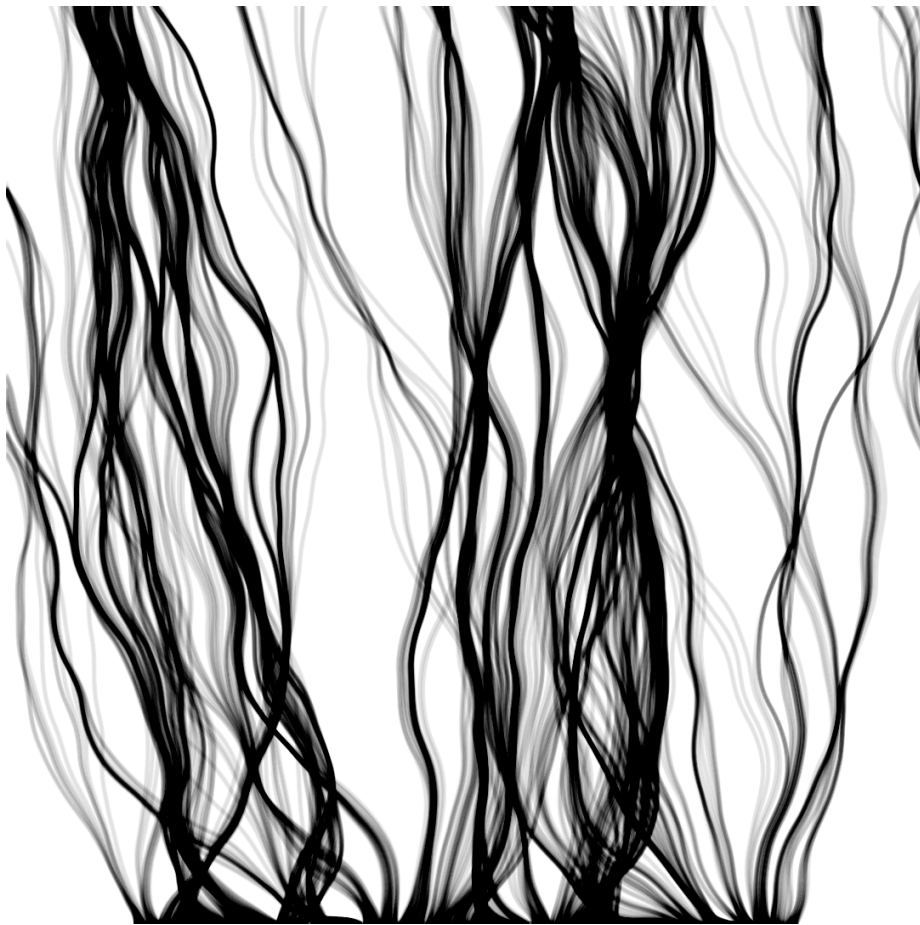


Angewandte Mathematik

Dr. Johannes Riesterer

18. Oktober 2020



©Johannes Riesterer

Vorwort

Kann jeder Mathematik lernen? Als Antwort auf diese Frage möchte ich auf den interessanten Lebenslauf einer der bedeutendsten Mathematikerinnen aller Zeiten eingehen (Auszug aus Wikipedia):

Emmy Noether war eine deutsche Mathematikerin, die grundlegende Beiträge zur abstrakten Algebra und zur theoretischen Physik lieferte. Insbesondere hat Noether die Theorie der Ringe, Körper und Algebren revolutioniert. Das nach ihr benannte Noether-Theorem gibt die Verbindung zwischen Symmetrien von physikalischen Naturgesetzen und Erhaltungsgrößen an.

Sie zeigte in mathematischer Richtung keine besondere Frühreife, sondern hatte in ihrer Jugend Interesse an Musik und Tanzen. Sie besuchte die Städtische Höhere Töchterschule – das heutige Marie-Therese-Gymnasium – in der Schillerstraße in Erlangen. Mathematik wurde dort nicht intensiv gelehrt. Im April 1900 legte sie die Staatsprüfung zur Lehrerin der englischen und französischen Sprache an Mädchenschulen in Ansbach ab. 1903 holte sie in Nürnberg die externe Abiturprüfung am Königlichen Realgymnasium – dem heutigen Willstätter-Gymnasium – nach.

Inhaltsverzeichnis

1	Vorwissen und Konventionen	6
2	Mehrdimensionale Differentialrechnung	7
2.1	Richtungsableitung und Gradient reellwertiger Funktionen	9
2.2	Extrema	17
2.3	Gradient einer mehrdimensionalen Funktion	22
3	Mehrdimensionale Integralrechnung	24
3.1	Lebesgue Maß	24
3.2	Lebesgue Integral	25
	Tabellenverzeichnis	26
	Abbildungsverzeichnis	27

1 Vorwissen und Konventionen

Differenzierbarkeit reeller Funktionen

Eine reelle Funktion $f : (a, b) \rightarrow \mathbb{R}$ heißt differenzierbar in $x \in (a, b)$, falls der Grenzwert $\lim_{h \rightarrow 0} \frac{f(x+h)-f(x)}{h}$ existiert. In diesem Fall heißt dieser Grenzwert die Ableitung (Steigung) von f in x und wird mit $f'(x)$ bezeichnet.

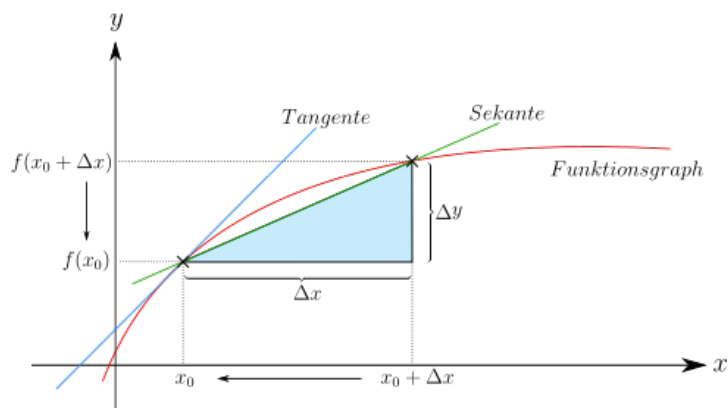


Abbildung 1: Quelle: Wikipedia: https://de.wikipedia.org/wiki/Datei:Differential_quotient_of_a_function.s

Mittelwertsatz einer Veränderlichen

Sei $f : [a, b] \rightarrow \mathbb{R}$ stetig und differenzierbar für alle $x \in (a, b)$. Dann gibt es $\xi \in (a, b)$ mit $f'(\xi) = \frac{f(b)-f(a)}{b-a}$.

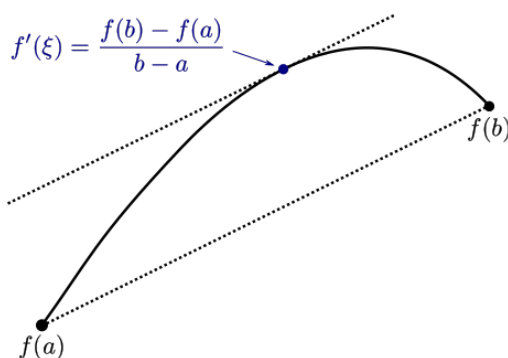


Abbildung 2: Quelle: Wikipedia: <https://commons.wikimedia.org/wiki/File:Mittelwertsatz3.svg>

Taylorapproximation einer Veränderlichen

Jede reelle Funktion f , deren $p+1$ -ten Ableitungen existieren und stetig sind lässt sich mit Hilfe der Taylorreihe

$$f(x) = f(a) + \frac{1}{2!}f'(x-a) + \frac{1}{3!}f''(a)(x-a)^2 + \dots + \frac{1}{p!}f^{(p)}(a)(x-a)^{p-1} + R_{p+1}(x, a)$$

und dem Restglied $R_p(x, a) := \frac{1}{(p+1)!}f^{(p+1)}(\xi)(x-a)^p$ mit einem $\xi \in (x, a)$ darstellen.

Cauchy-Schwarzsche Ungleichung

Für zwei Vektoren $v, w \in \mathbb{R}^n$ gilt:

$$\frac{\langle v, w \rangle}{\|v\| \cdot \|w\|} = \cos(\varphi)$$

wobei φ der Innenwinkel zwischen v und w ist.

Äquivalenz von Normen

Die Normen $\|v\| := \sqrt{\sum_{i=1}^n v_i^2}$ und $\|v\|_\infty := \max\{v_1, \dots, v_n\}$ sind Äquivalent. Sie lassen sich mit Konstanten $k_1\|v\| < \|v\|_\infty k_2\|v\|$ gegeneinander abschätzen.

Symmetrische Matrizen

Für eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ ist Äquivalent:

- A hat positive Eigenwerte.
- $v^T A v > 0$ für alle $v \neq 0$.
- alle Unterdeterminanten sind positiv. Speziell für $n=2$ und $A = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$ bedeutet dies $a > 0$ und $ad - b^2 > 0$.

Definition 1 (Konventionen). In diesem Abschnitt ist $U \subset \mathbb{R}^n$ stets eine offene

Teilmenge des \mathbb{R}^n . $e_i := \begin{pmatrix} 0 \\ \vdots \\ 1 \text{ (i-te Zeile)} \\ \vdots \\ 0 \end{pmatrix}$ bezeichnet den i -ten Basisvektor des

\mathbb{R}^n .

2 Mehrdimensionale Differentialrechnung

Definition 2 (Konvergenz). Eine Folge (a_n) in \mathbb{R}^n heißt konvergent gegen den Grenzwert $a \in \mathbb{R}^n$, wenn gilt:

$$\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n > N : d(a, a_n) < \varepsilon$$

in Worten: Es gibt für jedes beliebige (noch so kleine) ε einen Index N derart, dass für alle Indizes $n > N$, alle weiteren Folgenglieder, gilt: der Abstand $d(a, a_n)$ ist kleiner als ε .

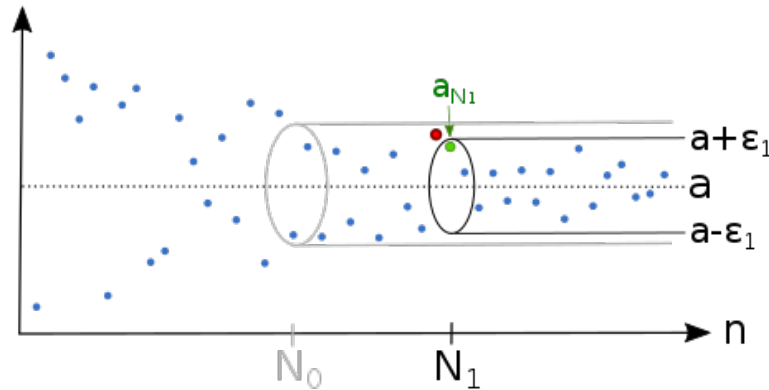


Abbildung 3: Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Epsilonschlauch_klein.svg

Definition 3 (Grenzwert). Sei $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ eine Funktion und $a \in X$. Wir nennen $L_a \in \mathbb{R}^m$ Grenzwert von f bezüglich der Annäherung von x an a , falls für jede konvergente Folge $x_n \rightarrow a$ die Folge $f(x_n)$ nach L_a konvergiert. In diesem Fall bezeichnen wir

$$\lim_{x \rightarrow a} f(x) = L_a .$$

Dies ist gleichbedeutend damit, dass für jedes $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass $d(f(x), L_a) < \varepsilon$ gilt für jedes x mit $d(x, a) < \delta$.

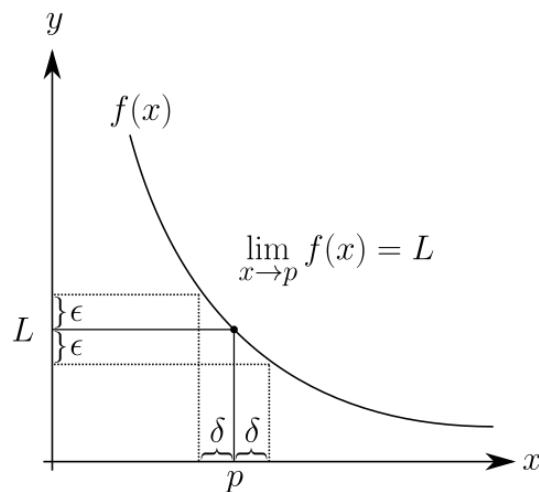


Abbildung 4: Quelle: Wikipedia: https://de.wikipedia.org/wiki/Datei:Limes_Definition_Vektorgrafik.svg

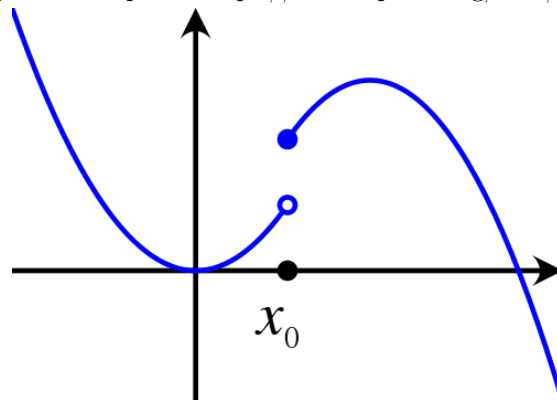


Abbildung 5: Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Upper_semi.svg

Definition 4 (Stetige Funktion). Eine reellwertige Funktion $f : U \rightarrow \mathbb{R}$ heißt stetig, wenn für alle $y \in U$ der Grenzwert $\lim_{x \rightarrow y} f(x) = L_y$ existiert.

2.1 Richtungsableitung und Gradient reellwertiger Funktionen

Ableitungen beschreiben bildlich gesprochen das Verhalten einer Funktion bezüglich beliebig kleiner Änderungen der Eingabewerte.

Definition 5 (Richtungsableitung). Sei $f : U \rightarrow \mathbb{R}$ eine Funktion. Für einen Vektor $h \in \mathbb{R}^n$ und einen Punkt $a \in U$ heißt der Grenzwert (falls er existiert)

$$\partial_h f(a) := \lim_{t \rightarrow 0} \frac{f(a + th) - f(a)}{t}$$

Richtungsableitung von f am Punkt a in Richtung h . Sie misst die Änderung der Funktion in Richtung h .

Speziell nennen wir für die Standard Basisvektoren e_i

$$\frac{\partial f(a)}{\partial x_n} := \partial_{e_i} f(a) := \lim_{t \rightarrow 0} \frac{f(a + te_i) - f(a)}{t}$$

die partielle Ableitung von f in a nach x_i .

Definition 6 (Partielle Differenzierbarkeit). Eine Funktion $f : U \rightarrow \mathbb{R}$ heißt partiell differenzierbar im Punkt $a \in U$, falls alle partiellen Ableitungen

$$\frac{\partial f(a)}{\partial x_1}, \dots, \frac{\partial f(a)}{\partial x_n}$$

existieren.

Definition 7 (Differenzierbarkeit). Eine Funktion $f : U \rightarrow \mathbb{R}$ heißt differenzierbar im Punkt $a \in U$, falls alle partiellen Ableitungen

$$\frac{\partial f(a)}{\partial x_1}, \dots, \frac{\partial f(a)}{\partial x_n}$$

existieren und stetig sind. Mann nennt in diesem Fall die $1 \times n$ -Matrix

$$df(a) := \left(\frac{\partial f(a)}{\partial x_1}, \dots, \frac{\partial f(a)}{\partial x_n} \right)$$

das Differential von f im Punkt a . Der Vektor

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

wird als Gradient bezeichnet. Es ist $df(a) \cdot h = \langle \nabla f(a), h \rangle$.

★★★ Der Unterschied zwischen Differenzierbarkeit und partieller Differenzierbarkeit ist also, dass die partiellen Ableitungen zusätzlich zur Existenz auch stetig sein müssen.

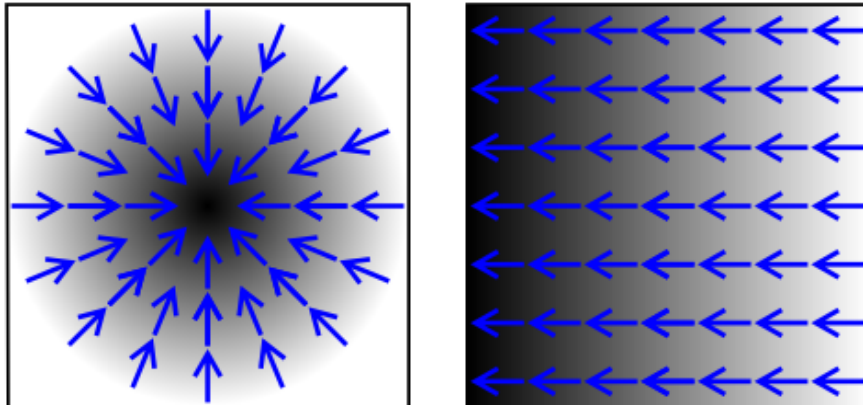


Abbildung 6: Quelle: Wikipedia: <https://commons.wikimedia.org/wiki/File:Gradient2.svg>

Bemerkung 1. Für das Differential einer differenzierbaren Funktion $f : U \rightarrow \mathbb{R}$ gilt für alle $a \in U$:

- $df(a)(h) := df(a) \cdot h$ ist eine lineare Abbildung von \mathbb{R}^n nach \mathbb{R} .
- $df(a) \cdot h = \partial_h f(a)$.
- $d(f \cdot g) = g(a)df + f(a)dg$
- $d(f + g) = df + dg$

Beweis. • Multiplikation mit einer Matrix ist eine lineare Abbildung.

- Für die Basisvektoren ist per Definition $df(a) \cdot e_i = \partial_{e_i} f(a)$. Da jeder Vektor h eine Linearkombination der Basisvektoren ist und df linear ist, folgt die Behauptung.
- Folgt direkt aus der entsprechenden Eigenschaft reeller Funktionen.
- Folgt direkt aus der entsprechenden Eigenschaft reeller Funktionen.

□

Satz 1 (Steilste Anstiegsrichtung). Sei $f : U \rightarrow \mathbb{R}$ differenzierbare Funktion, $a \in U$ und $v := \operatorname{argmax}_{h \in S^n} \partial_h f(a)$. Dann gilt

$$\|\nabla f(a)\|v = \nabla f(a).$$

Beweis. Mit der CSU Ungleichung folgt für beliebiges h

$$\partial_h f(a) = df(a)h = \langle \nabla f(a), h \rangle = \|\nabla f(a)\| \cdot \|h\| \cdot \cos(\varphi)$$

wobei φ den Innenwinkel zwischen $\nabla f(a)$ und h bezeichnet. Für $\|h\| = 1$ wird somit $\partial_h f(a)$ maximal, wenn $\varphi = 0$ und somit $h = \frac{\nabla f(a)}{\|\nabla f(a)\|}$ ist. □

Satz 2 (Lokale Linearisierung). Ist $f : U \rightarrow \mathbb{R}$ differenzierbar, so gibt es ein Restglied $R(h)$ mit $\lim_{h \rightarrow 0} \frac{R(h)}{\|h\|} = 0$ so dass für alle $a \in U$ und $h \in \mathbb{R}$

$$f(a+h) = f(a) + df(a) \cdot h + R(h)$$

gilt.

★★★ Eine differenzierbare Funktion kann auf hinreichend kleinen Umgebungen beliebig genau durch eine lineare Funktion approximiert werden.

◊◊ Der Beweis beruht im Wesentlichen auf dem Mittelwertsatz einer Veränderlichen.

Beweis. Wir wählen einen offenen, achsenparallelen Quader $Q \subset U$, so dass er vollständig in U enthalten und $a \in Q$ ist. Jeder Punkt $a+h \in Q$ lässt sich damit durch einen achsenparallelen Streckenzug durch die Punkte

$$\begin{aligned} a_0 &:= a \\ a_i &:= a_{i-1} + h_i e_i; \quad i = 1, \dots, n \end{aligned}$$

mit a verbinden.

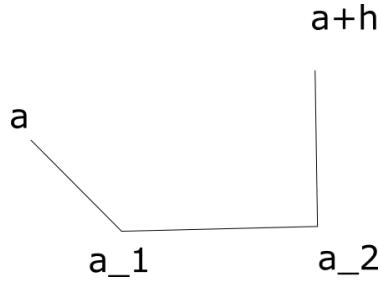


Abbildung 7: Kantenzug mit achsenparallelen Kanten

Damit ist $f(a+h) - f(a) = \sum_{i=1}^n (f(a_i) - f(a_{i-1}))$ und mit $\varphi_i(t) := f(a_i + te_i)$ gilt $f(a_i) - f(a_{i-1}) = \varphi_i(h_i) - \varphi_i(0)$. Wegen dem Mittelwertsatz einer Veränderlichen gibt es τ_i mit

$$\varphi_i(h_i) - \varphi_i(0) = h_i \varphi'_i(\tau_i).$$

Da $\varphi'_i(t) = \frac{\partial f(a_{i-1} + te_i)}{\partial x_i}$ folgt mit $\xi_i := a_i + \tau_i e_i$

$$f(a+h) - f(a) - df(a) \cdot h = \sum_{i=1}^n \left(\frac{\partial f(\xi_i)}{\partial x_i} - \frac{\partial f(a)}{\partial x_i} \right) h_i$$

und damit

$$|f(a+h) - f(a) - df(a) \cdot h| \leq \|h\|_\infty \sum_{i=1}^n \left| \frac{\partial f(\xi_i)}{\partial x_i} - \frac{\partial f(a)}{\partial x_i} \right|.$$

Für $h \rightarrow 0$ gilt $\xi_i \rightarrow a$ und da die partiellen Ableitung stetig sind nach Voraussetzung und alle Normen äquivalent sind folgt

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - df(a) \cdot h}{\|h\|} = 0$$

und damit die Behauptung. \square

Bemerkung 2. Umformuliert bedeutet Satz 2.3, dass für das Differential einer differenzierbare Funktion

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - df(a)h}{\|h\|} = 0 \quad (1)$$

Ist L eine weiter lineare Abbildung mit $\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - L(a)h}{\|h\|} = 0$, so ist $L = df$. Das Differential ist somit eindeutig durch die Eigenschaft der lokalen Linearisierung bestimmt.

Beweis. Für $v \in \mathbb{R}^n$ mit $\|v\| = 1$ gilt

$$(L(a) - df(a))(v) = \lim_{t \rightarrow 0} (L(a) - df(a)) \left(\frac{tv}{\|tv\|} \right) = \lim_{t \rightarrow 0} \frac{(L(a) - df(a))(tv)}{\|tv\|} = 0$$

\square

Definition 8 (Differenzierbarer Weg). *Ein Weg ist eine Abbildung*

$$\gamma : [a, b] \rightarrow \mathbb{R}^n$$

$$\gamma(t) := \begin{pmatrix} \gamma_1(t) \\ \vdots \\ \gamma_n(t) \end{pmatrix}$$

mit reellen, stetigen Funktionen $\gamma_i : [a, b] \rightarrow \mathbb{R}$ (Damit ist auch γ stetig). Der Weg heißt differenzierbar, falls alle Ableitungen $\gamma'_i(t)$ existieren. In diesem Fall bezeichnen wir

$$\gamma'(t) := \begin{pmatrix} \gamma'_1(t) \\ \vdots \\ \gamma'_n(t) \end{pmatrix}$$

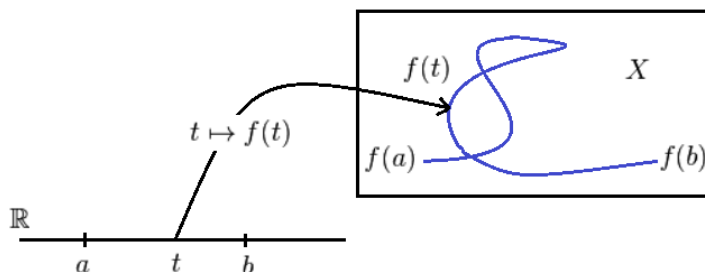


Abbildung 8: Quelle: Wikipedia: <https://de.wikipedia.org/wiki/Datei:EbeneKurve.png>

Satz 3 ((Baby) Kettenregel). *Sei $\gamma : I \rightarrow U$ ein differenzierbarer Weg und $f : U \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Dann ist $f \circ \gamma : I \rightarrow \mathbb{R}$ differenzierbar und hat die Ableitung*

$$\frac{d(f \circ \gamma)}{dt}(t) = df(\gamma(t))\gamma'(t) = \sum_{i=1}^n \frac{\partial f(\gamma(t))}{\partial x_i} \gamma'_i(t)$$

★★★ Das Differential einer differenzierbaren Funktion kann als eine Abbildung von Tangenten interpretiert werden.

Beweis. Wegen der Differenzierbarkeit des Weges und der Funktion gilt für hinreichend kleine $k \in \mathbb{R}$ und $h \in \mathbb{R}^n$

$$\gamma(t+k) = \gamma(t) + k\gamma'(t) + r_1(k)|k|, \text{ mit } \lim_{k \rightarrow 0} r_1(k) = 0$$

$$f(\gamma(t) + h) = f(\gamma(t)) + df(\gamma(t)) \cdot \gamma'(t)h + r_2(h)||h||, \text{ mit } \lim_{h \rightarrow 0} r_2(h) = 0$$

Mit $h := \gamma(t+k) - \gamma(t)$ folgt

$$f(\gamma(t+k)) = f(\gamma(t)) + df(\gamma(t)) \cdot \gamma'(t)k + R(k)$$

mit dem Restglied

$$R(k) := df(\gamma(t))r_1(k)|k| + r_2(\gamma(t+k) - \gamma(t))||\gamma'(t)k + r_1(k)|k||$$

Da $\lim_{k \rightarrow 0} R(k) = 0$ folgt die Behauptung. □

Satz 4 (Mittelwertsatz). Sei $f : U \rightarrow \mathbb{R}$ eine differenzierbare Funktion und $a, b \in U$, so dass die Verbindungsstrecke von a nach b vollständig in U liegt. Dann gibt es einen Punkt $\xi \in [a, b]$ mit

$$f(b) - f(a) = df(\xi)(b - a).$$

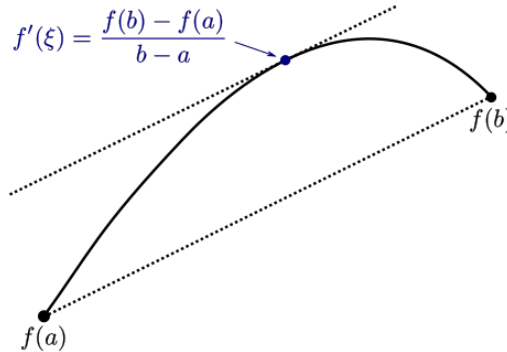


Abbildung 9: Quelle: Wikipedia

Beweis. Setze $\gamma(t) := a + t(b - a)$, $t \in [0, 1]$ (Verbindungsgerade zwischen a und b) und $F := f \circ \gamma : [0, 1] \rightarrow \mathbb{R}$. Damit ist $F(1) - F(0) = f(1) - f(0)$ und nach der Kettenregel ist F differenzierbar. Mit dem Mittelwertsatz einer Veränderlichen gibt es $\tau \in (0, 1)$ mit $F(1) - F(0) = F'(\tau) = df(\gamma(\tau))(b - a)$. Somit ist $\xi := \gamma(\tau)$ der gesuchte Punkt. \square

Satz 5 (Satz von Schwarz). Wenn für eine Funktion $f : U \rightarrow \mathbb{R}$ die Ableitungen $\frac{\partial}{\partial x_i} f(a)$, $\frac{\partial}{\partial x_j} f(a)$ und $\frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(a)$ existieren und letztere stetig ist, dann existiert auch $\frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f(a)$ und es gilt

$$\frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(a) = \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f(a)$$

★★★ Bei wiederholten Ableiten spielt die Reihenfolge keine Rolle, wenn eine der Ableitungen existiert und stetig ist.

◊◊ Der Beweis geht in zwei Schritten: Man wendet den Mittelwertsatz einer Veränderlichen auf eine Hilfsfunktion an. Damit kann man die Differenz der Ableitungen abschätzen und zeigen, dass sie beliebig klein wird und damit gleich sind.

Beweis. Setze $\varphi(x, y) := f(a + xe_i + ye_j)$ mit $(x, y) \in V \subset \mathbb{R}^2$. Bei hinreichend kleiner Wahl von V existieren die Ableitungen $\frac{\partial}{\partial y} \varphi$, $\frac{\partial}{\partial x} \varphi$ und $\frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi$ existiert und ist stetig nach Voraussetzung an f . Es reicht nun zu zeigen, dass $\frac{\partial}{\partial x} \frac{\partial}{\partial y} \varphi(0, 0)$ existiert und

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} \varphi(0, 0) = \frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi(0, 0)$$

gilt. Sei dazu $\epsilon > 0$ und $V' \subset V$ so gewählt, dass $|\frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi(x, y) - \frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi(0, 0)| < \epsilon$ gilt für $(x, y) \in V'$. Innerhalb eines achsenparallelen Quaders $Q \subset V'$ mit Ecken $(0, 0)$ und (h, k) setzen wir $u(x) := \varphi(x, k) - \varphi(x, 0)$. Zweimaliges Anwenden des Mittelwertsatzes einer Veränderlichen liefert

$$\begin{aligned} u(h) - u(0) &= hu'(\xi) \\ &= h\left(\frac{\partial}{\partial x} \varphi(\xi, k) - \frac{\partial}{\partial x} \varphi(\xi, 0)\right) = hk \frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi(\xi, \eta). \end{aligned}$$

Damit erhalten wir die Abschätzung

$$\left| \frac{u(h) - u(0)}{hk} - \frac{\partial}{\partial y} \frac{\partial}{\partial x} \varphi(0, 0) \right| < \epsilon.$$

Da

$$\begin{aligned} \lim_{k \rightarrow 0} \frac{u(h) - u(0)}{hk} &= \lim_{k \rightarrow 0} \frac{1}{h} \left(\frac{\varphi(h, k) - \varphi(h, 0)}{k} - \frac{\varphi(0, k) - \varphi(0, 0)}{k} \right) \\ &= \left(\frac{\frac{\partial}{\partial y} \varphi(h, 0) - \frac{\partial}{\partial y} \varphi(0, 0)}{k} \right) \end{aligned}$$

und

$$\lim_{h \rightarrow 0} \left(\frac{\frac{\partial}{\partial y} \varphi(h, 0) - \frac{\partial}{\partial y} \varphi(0, 0)}{k} \right) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} \varphi(0, 0)$$

folgt die Behauptung. \square

Anwendung: Taylorreihe

Definition 9 (C^k -Funktion). Eine Funktion $f : U \rightarrow \mathbb{R}$ für die alle partiellen Ableitungen

$$\frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_k}} f(a)$$

mit $i_1 + \cdots + i_k \leq k$ existieren und stetig sind heißt C^k -Funktion oder k -mal stetig differenzierbar. Eine C^1 -Funktion ist also eine differenzierbare Funktion.

Definition 10 (Höhere Ableitungen). Für eine Funktion $f : U \rightarrow \mathbb{R}$, $a \in U$ und Vektoren $v^1, \dots, v^p \in \mathbb{R}^n$ heißt

$$d^p f(a)(v^1, \dots, v^p) := \partial_{v^1} \dots \partial_{v^p} f(a)$$

die p -te Richtungsableitung von f . Sie ist wegen dem Satz von Schwarz wohldefiniert. Mit Bemerkung 1 ist

$$d^p f(a)(v^1, \dots, v^p) = \sum_{i_1=1}^n \cdots \sum_{i_p=1}^n \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_p}} f(a) \cdot v_{i_1}^1 \cdots v_{i_p}^p.$$

Für einen Vektor $z \in \mathbb{R}^n$ definieren wir

$$d^p f(a)z^p := d^p f(a) \underbrace{(z, \dots, z)}_{p\text{-mal}}.$$

Bemerkung 3. Für $p = 2$ und $u, v \in \mathbb{R}^n$ ist

$$d^2f(a)(u, v) = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(a) v_i u_j$$

und mit

$$f''(a) := \begin{pmatrix} \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_1} f(a) & \cdots & \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_n} f(a) \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_1} f(a) & \cdots & \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_n} f(a) \end{pmatrix}$$

ist $d^2f(a)(u, v) = u^T \cdot f''(a) \cdot v$. Die Matrix $f''(a)$ wird auch Hesse-Matrix genannt. Nach dem Satz von Schwarz ist sie symmetrisch.

Satz 6 (Taylorapproximation). Sei $f : U \rightarrow \mathbb{R}$ eine C^{p+1} -Funktion und $x, a \in U$, so dass die Verbindung zwischen x und a in U liegt. Dann gilt

$$f(x) = f(a) + \sum_{k=1}^p \frac{d^k f(a)}{k!} (x - a)^k + R_{p+1}(x; a)$$

mit dem Restglied $R_{p+1}(x; a) := \frac{1}{(p+1)!} d^{p+1}f(\xi)(x - a)^{p+1}$ für ein $\xi \in [a, x]$.

Beweis. Sei $h := (x - a)$ und $F(t) := f(a + th)$ mit $t \in [0, 1]$. Wiederholte Anwendung der (Baby) Kettenregel mit $\gamma(t) := a + th$ ergibt

$$\begin{aligned} F'(t) &= \sum_{i=1}^n \frac{\partial}{\partial x_i} f(a + th) h_i \\ F''(t) &= \sum_{j=1}^n \sum_{i=1}^n \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_i} f(a + th) h_i h_j \\ &\vdots \\ F^p(t) &= \sum_{i_1=1}^n \cdots \sum_{i_p=1}^n \frac{\partial}{\partial x_{i_1}} \cdots \frac{\partial}{\partial x_{i_p}} f(a + th) h_{i_1} \cdots h_{i_p}. \end{aligned}$$

Mit der Taylorapproximation für Funktionen einer Veränderlichen folgt

$$F(1) = F(0) + F'(0) + \frac{1}{2!} F''(0) + \cdots + \frac{1}{p!} F^p(0) + R_{p+1}$$

mit dem Restglied $R_{p+1} := \frac{1}{(p+1)!} F^{p+1}(\tau)$ mit $\tau \in [0, 1]$. Da nach Konstruktion $F(0) = f(a)$ und $F(1) = f(x)$ folgt insgesamt die Behauptung. \square

Satz 7 (Qualitative Taylorformel). Sei $T_p(x, a) = f(a) + \sum_{k=1}^p \frac{d^k f(a)}{k!} (x - a)^k$ die Taylorapproximation einer C^p -Funktion. Dann gilt:

$$\lim_{x \rightarrow a} \frac{f(x) - T_p(x; a)}{\|x - a\|^p} = 0.$$

Beweis. Da die partiellen Ableitungen stetig sind, gibt es für jedes $\epsilon > 0$ ein Radius $r > 0$, dass für alle $y \in K_r(a)$

$$\frac{1}{p!}(d^p f(y) - d^p f(a))h^p \leq \epsilon \|h\|_\infty^p.$$

Mit der Taylorapproximation ist

$$f(x) = T_{p-1}(x, a) + \frac{1}{p!}d^p f(\xi)(x - a)^p = T_p(x, a) + \frac{1}{p!}(d^p f(\xi) - d^p f(a))h^p(x - a)^p$$

Mit obiger Abschätzung folgt die Behauptung. \square

2.2 Extrema

Definition 11 (Extremum). Sei $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$ eine reelle Funktion. Ein Punkt $a \in X$ heißt *lokales Maximum* bzw. *Minimum*, falls eine Umgebung U von a existiert, so dass $f(x) \leq f(a)$ bzw. $f(x) \geq f(a)$ für alle $x \in U$ gilt. Liegt einer der beiden Fälle vor, so spricht man von einem *lokalen Extremum*. Gilt strikt $f(x) < f(a)$ bzw. $f(x) > f(a)$, so nennt man das *Extremum isoliert*. Ist $U = X$ so nennt man es auch *globales Maximum* bzw. *Minimum*.

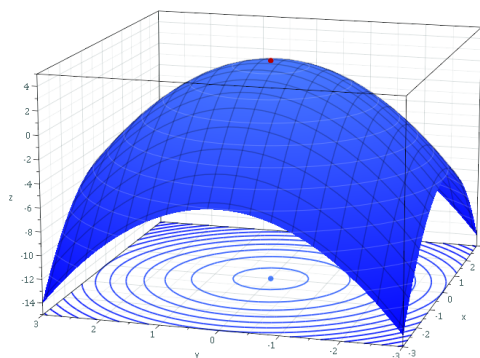


Abbildung 10: Quelle: Wikipedia: <https://en.wikipedia.org/wiki/File:MaximumParaboloid.png>

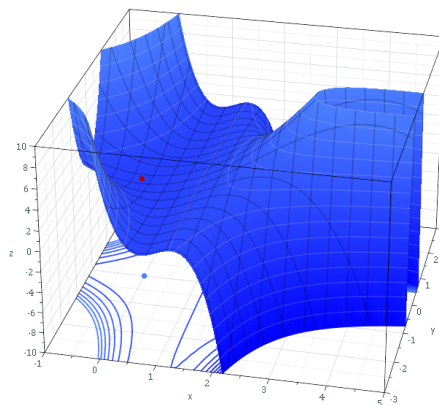


Abbildung 11: Quelle: Wikipedia: <https://en.wikipedia.org/wiki/File:MaximumCounterexample.png>

Satz 8 (Notwendiges Kriterium). Ist $f : U \rightarrow \mathbb{R}$ differenzierbar und hat f in $a \in U$ ein lokales Extremum, so gilt

$$\frac{\partial}{\partial x_1} f(a) = \dots = \frac{\partial}{\partial x_n} f(a) = 0.$$

Dies ist gleichbedeutend mit $df(a) = 0$.

Beweis. Setze $F_k(t) := f(a + te_k)$. Da f ein Extremum in a hat, hat F_k in einer hinreichend kleinen Umgebung um 0 ein Extremum. Da F_k eine Funktion einer Veränderlichen ist, gilt $F'_k(0) = 0$. Da $\frac{\partial}{\partial x_k} f(a) = F'_k(0)$ folgt die Behauptung. \square

Satz 9 (Hinreichendes Kriterium). Ist $f : U \rightarrow \mathbb{R}$ eine \mathcal{C}^2 -Funktion und ist $f'(a) = 0$ für ein $a \in U$. Dann gilt:

- $f''(a) > 0 \Rightarrow f$ hat in a ein isoliertes lokales Minimum.
- $f''(a) < 0 \Rightarrow f$ hat in a ein isoliertes lokales Maximum.
- $f''(a) \geq 0 \Rightarrow f$ hat in a kein lokales Extremum.

Beweis. Sei $f'(a) = 0$ und $f''(a) > 0$. Mit der Taylorformel gilt für hinreichend kurze Vektoren $h \in \mathbb{R}^n$

$$f(a+h) = f(a) + \frac{1}{2} h^T f''(a) h + R(h)$$

mit $\lim_{h \rightarrow 0} \frac{R(h)}{\|h\|^2} = 0$. Für $\|h\| \leq 1$ hat $h^T f''(a) h$ sein Maximum m auf dem Einheitskreis $\{h \in \mathbb{R}^n \mid \|h\| = 1\}$ da $f''(a) > 0$.

$$h^T f''(a) h = \|h\| \frac{1}{\|h\|} h^T f''(a) \|h\| \frac{1}{\|h\|} h \geq m \|h\|^2.$$

Wir wählen ϵ so klein, dass $R(h) \leq \frac{m}{2} \|h\|^2$ gilt für $\|h\| < \epsilon$ (was geht wegen Taylorformel). Damit erhalten wir

$$f(a+h) \geq f(a) + m \|h\|^2.$$

und damit hat f ein lokales Minimum in a .

Der Fall $f''(a) < 0$ wird mit Betrachtung von $-f$ durch den vorigen Fall bewiesen.

Es sei nun $f''(a) \geq 0$ und v mit $v^T f''(a) v > 0$ und w mit $w^T f''(a) w < 0$. Betrachten wir die Funktionen

$$\begin{aligned} F_v(t) &:= f(a + tv) \\ F_w(t) &:= f(a + tw) \end{aligned}$$

dann ist

$$\begin{aligned} F'_v(t) &= 0; \quad F''_v(0) = v^T f''(a) v > 0 \\ F'_w(t) &= 0; \quad F''_w(0) = w^T f''(a) w < 0 \end{aligned}$$

und somit hat F_v ein isoliertes lokales Maximum und F_w ein isoliertes lokales Minimum und damit f kein lokales Extremum in a . \square

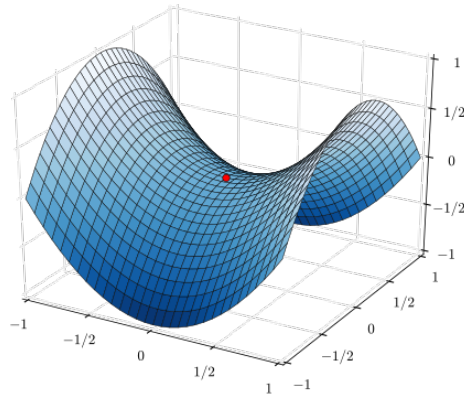


Abbildung 12: Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Saddle_point.svg

Anwendung: Gradientenverfahren

Wie kann man Minima einer differenzierbaren Abbildung $f : \mathbb{R}^n \rightarrow \mathbb{R}$ finden? Wir wissen, dass an jedem Punkt $\mathbf{x}_k \in \mathbb{R}^n$ der negative Gradient $\mathbf{d}_k := -\nabla f(\mathbf{x}_k)$ in die steilste Abstiegsrichtung zeigt. Die Idee des Gradientenabstieg ist, ein bestimmtes Stück in diese Richtung abzustiegen, damit der Funktionswert kleiner wird, also $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ zu setzen. Für hinreichend kleines α_k folgt mit Satz 2.3 über die lokale Linearisierung $f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = f(\mathbf{x}_k) + \alpha_k df(\mathbf{x}_k) \mathbf{d}_k + R(\alpha_k \mathbf{d}_k)$. Somit gilt $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$, falls $\nabla f(\mathbf{x}_k) \neq 0$ und falls die Folge $f(\mathbf{x}_k)$ beschränkt ist, so ist dieser Fixpunkt \mathbf{x}^* ein Minimum, da $\nabla f(\mathbf{x}^*) = 0$ gelten muss.

Algorithm 1 Gradientenabstieg

- 1: Initialisiere $k := 0$ und zufälligen Startwert \mathbf{x}_0 .
 - 2: Initialisiere Genauigkeit $\epsilon > 0$.
 - 3: **while** $\|\nabla f(\mathbf{x}_k)\| > \epsilon$ **do** ▷ So lange kein Extrema vorliegt
 - 4: Bestimme α_k mit $f(\mathbf{x}_k + \alpha \mathbf{d}_k) = f(\mathbf{x}_k) + \alpha_k df(\mathbf{x}_k) \mathbf{d}_k + R(\alpha_k \mathbf{d}_k)$ ▷
 Schrittweite bestimmen
 - 5: Setze $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k \mathbf{d}_k$. ▷ Absteigen
 - 6: $k \leftarrow k + 1$ ▷ Wiederholen
-

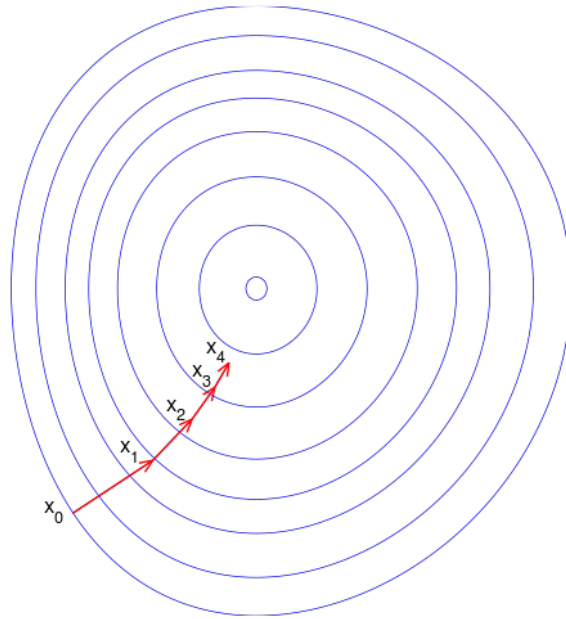


Abbildung 13: Quelle: Wikipedia

Definition 12. Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Eine Kurve $\gamma : I \rightarrow \mathbb{R}^n$, auf der f konstant ist, also $f(\gamma(t)) = c$ für ein festes $c \in \mathbb{R}$ gilt, heißt Höhenlinie.

Bemerkung 4. Der Gradient steht senkrecht auf Höhenlinien. Dies bedeutet, dass

$$\langle \nabla f(\gamma(t)), \gamma'(t) \rangle = 0$$

gilt.

Beweis. Aus $f(\gamma(t)) = c$ folgt $\frac{d}{dt}f(\gamma(t)) = 0$. Mit der Kettenregel folgt $\frac{d}{dt}f(\gamma(t)) = \nabla f(\gamma(t)) \cdot \gamma'(t) = 0$ und damit $\langle \nabla f(\gamma(t)), \gamma'(t) \rangle = 0$. \square

Anwendung: Backpropagation

Ein Neuronales Netz ist eine Funktion $f : \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}^m$, die einem Input (auch Feature genannt) in Abhängigkeit der Gewichte Ω einen Ausgabewert zuordnet. Die Funktion ist dabei aus einfachen Bausteinen zusammengesetzt.

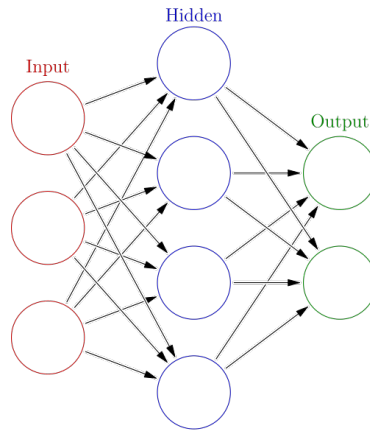


Abbildung 14: Quelle: Wikipedia

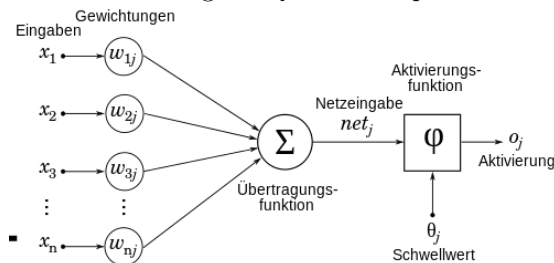


Abbildung 15: Quelle: Wikipedia

Gegeben ist ein Datensatz $D := \{(x_i, y_i)\}$. Finde Gewichte Omega, so dass Lossfunktion

$$L_D : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$$

minimal wird. Zum Beispiel $L_D(\omega) := \sum_{(x_i, y_i) \in D} (f(\omega, x_i) - y_i)^2$.
Wende Gradientenverfahren an:

Algorithm 2 Gradientabstieg

- 1: Initialisiere $k := 0$ und zufällige Gewichte ω_0 .
 - 2: Initialisiere Genauigkeit $\epsilon > 0$.
 - 3: **while** $\|\nabla L_D(\omega)\| > \epsilon$ **do** ► So lange kein Extrema vorliegt
 - 4: Bestimme α_k mit $L_D(\omega_k + \alpha d_k) = L_D(\omega_k) + \alpha_k dL_D(\omega_k) d_k + R(\alpha_k d_k)$
► Schrittweite bestimmen
 - 5: Setze $\omega_{k+1} := \omega_k + \alpha_k d_k$. ► Absteigen
 - 6: $k \leftarrow k + 1$ ► Wiederholen
-

Wenn der Datensatz D sehr groß ist (Big Data), ist die Berechnung des Gradient der Lossfunktion entsprechend aufwendig. Um diesen Aufwand zu reduzieren, kann man das Gradientenverfahren modifizieren, so dass man Gradienten nur auf Teilräumen berechnet. Man erhält somit das sogenannte Mini-Batch Gradientenverfahren und das stochastische Gradientenverfahren.

Algorithm 3 Gradientabstieg

- 1: Initialisiere $k := 0$ und zufällige Gewichte w_0 .
 - 2: Initialisiere Genauigkeit $\epsilon > 0$.
 - 3: Wähle Teilmenge $D'_0 \subset D$
 - 4: **while** $\|\nabla L_{D'_0}(\omega)\| > \epsilon$ **do** ▶ So lange kein Extrema vorliegt
 - 5: Bestimme α_k mit $L_{D'_0}(\omega_k + \alpha d_k) = L_{D'_0}(\omega_k) + \alpha_k dL_{D'_0}(\omega_k)d_k + R(\alpha_k d_k)$ ▶ Schrittweite bestimmen
 - 6: Setze $\omega_{k+1} := \omega_k + \alpha_k d_k$. ▶ Absteigen
 - 7: Wähle neue Teilmenge $D'_{k+1} \subset D$.
 - 8: $k \leftarrow k + 1$ ▶ Wiederholen
-

Wenn man als Teilmenge immer nur eine einelementige Menge wählt, so spricht man vom stochastischen Gradientenverfahren.

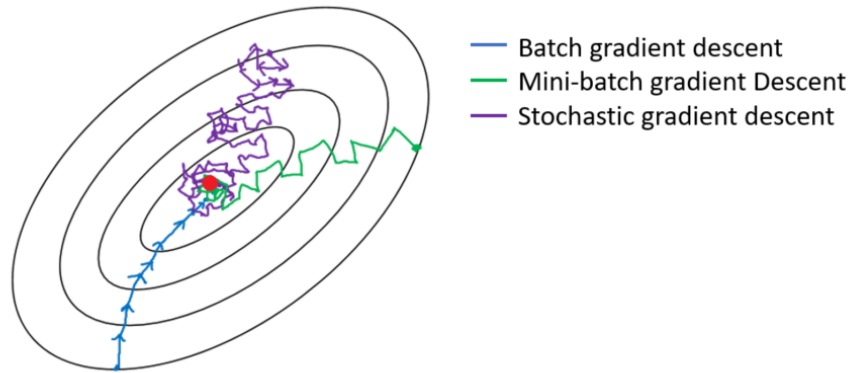


Abbildung 16: Quelle: <https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a>

2.3 Gradient einer mehrdimensionalen Funktion

Definition 13. Eine Funktion $F : U \rightarrow \mathbb{R}^m$ heißt differenzierbar, wenn es eine lineare Abbildung dF gibt, so dass

$$F(a + h) = F(a) + dF(a)h + R(h)$$

mit $\lim_{h \rightarrow 0} \frac{R(h)}{\|h\|} = 0$ gilt für alle $a \in U$ und $h \in \mathbb{R}^n$.

Bemerkung 5. Im Fall $n = 1$ stimmt diese Definition mit der Alten Definition 7 überein.

Beweis. Nach Satz gilt für eine differenzierbare Funktion $f(a + th) = f(a) + dfth + R(th)$ mit $\lim_{t \rightarrow 0} \frac{R(th)}{\|h\|} = 0$. Umstellen ergibt

$$df(a)h = \lim_{t \rightarrow 0} \frac{f(a + th) - f(a)}{t}$$

und somit existieren alle linearen Abbildungen und wegen der Linearität von df sind diese auch stetig. \square

Beispiel 1 (Affine Abbildung). Für $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ ist die Abbildung $F(x) := Ax + b$ differenzierbar, da $F(a+h) = A(a+h) + b = Aa + Ah + b = Aa + b + Ah = F(a) + Ah$ und damit für $dF(a) := A$ und $R(h) = 0$ die Definition 13 erfüllt ist.

Satz 10 (Differenzierbarkeit von Produktfunktionen). Eine Funktion $F := (F_1, F_2) : U \rightarrow \mathbb{R}^m \times \mathbb{R}^k$ ist genau dann differenzierbar, wenn $F_1 : U \rightarrow \mathbb{R}^m$ und $F_2 : U \rightarrow \mathbb{R}^k$ differenzierbar sind. In diesem Fall ist

$$dF(a) = (dF_1(a), dF_2(a)).$$

Beweis. Sind F_1 und F_2 differenzierbar, so gilt für $i = 1, 2$

$$F_i(a+h) = F_i(a) + dF_i h + R_i(h)$$

Dann gilt mit $dF(a) = (dF_1(a), dF_2(a))$ und $R(h) := (R_1(h), R_2(h))$

$$F(a+h) = F(a) + dFh + R(h)$$

mit $\lim_{h \rightarrow 0} \frac{R(h)}{\|h\|} = 0$ und damit ist F differenzierbar. Die Umkehrung folgt analog. \square

Bemerkung 6 (Differenzial von Produktfunktionen). Eine Abbildung $F : U \rightarrow \mathbb{R}^m$ ist genau dann differenzierbar, wenn ihre Koordinaten-Funktionen $F_1 : U \rightarrow \mathbb{R}, \dots, F_m : U \rightarrow \mathbb{R}$ mit $F(a) = \begin{pmatrix} F_1(a) \\ \vdots \\ F_m(a) \end{pmatrix}$ differenzierbar sind. In diesem Fall

gilt

$$dF(a) := \begin{pmatrix} \frac{\partial}{\partial x_1} F_1(a) & \cdots & \frac{\partial}{\partial x_n} F_1(a) \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} F_m(a) & \cdots & \frac{\partial}{\partial x_n} F_m(a) \end{pmatrix}$$

Bemerkung 7 (Differenzierbarkeit von Wegen). Ein Weg $\gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix} : I \rightarrow U$ ist genau dann differenzierbar, wenn γ_i differenzierbar ist für $i = 1, \dots, m$ und

$$\text{dann gilt } \gamma'(t) = \begin{pmatrix} \gamma'_1(t) \\ \vdots \\ \gamma'_m(t) \end{pmatrix}.$$

Satz 11 (Kettenregel). Seien $G : U \subset \mathbb{R}^n \rightarrow V \subset \mathbb{R}^m$ und $F : V \rightarrow Z \subset \mathbb{R}^k$ differenzierbar. Dann ist $F \circ G$ differenzierbar und mit $b := G(a)$ es gilt

$$d(F \circ G)(a) = dF(b) \cdot dG(a)$$

Beweis. Analog zu Baby Kettenregel. \square

Anwendung: Automatisches Ableiten

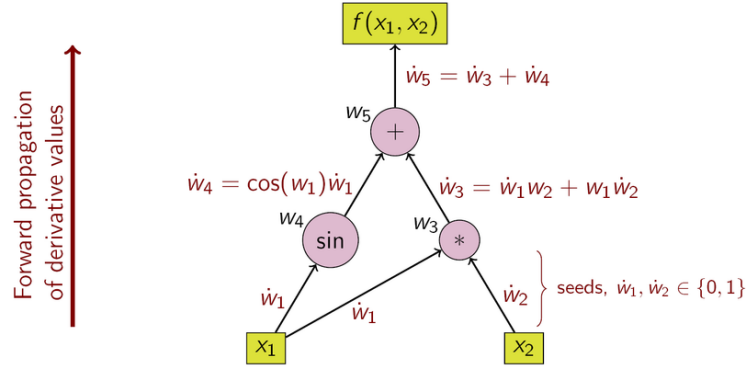


Abbildung 17: Quelle: Wikipedia

Automatisches Ableiten in Pytorch

Mit Hilfe des Automatischen Ableiten kann man den Gradienten der Lossfunktion zu einem Neuroyalen Netz einfach berechnen, da dieser sich aus einfachen Funktionen zusammensetzt.

3 Mehrdimensionale Integralrechnung

3.1 Lebesgue Maß

Für offene Intervalle $(a_i, b_i) \subset \mathbb{R}$ mit $a_i \leq b_i$ nennen wir $I := (a_1, b_1) \times \cdots \times (a_n, b_n)$ einen n -dimensionalen Quader und $\bar{I} := [a_1, b_1] \times \cdots \times [a_n, b_n]$ seinen Abschluss. Wir definieren das Volumen

$$\text{vol}(I) := \prod_{i=1}^n (b_i - a_i).$$

Mit $\mathbb{I}(n) := \{(a_1, b_1) \times \cdots \times (a_n, b_n) \mid (a_i, b_i) \subset \mathbb{R}\}$ bezeichnen wir die Menge aller n -dimensionalen Quader. Für eine Menge $A \subset \mathbb{R}^n$ bezeichnen wir eine Menge von Quadern $\{I_j \mid I_j \in \mathbb{I}(n)\}$ mit $A \subset \bigcup_j I_j$ als Hüllquader für A .

Definition 14 (Lebesguesche äußere Maß). Für eine Menge $A \subset \mathbb{R}^n$ definieren wir das Lebesguesche äußere Maß durch

$$\mu(A) := \inf \left\{ \sum_{j=1}^{\infty} \text{vol}(I_j) \mid I_j \in \mathbb{I}(n); A \subset \bigcup_{j=1}^{\infty} I_j \right\}$$

Definition 15 (Erinnerung Infimum).

Bemerkung 8. Für $A \subset B \subset \mathbb{R}^n$ ist $\mu(A) \leq \mu(B)$

Beweis. Da $A \subset B$ sind Hüllquader von B sind auch Hüllquader von A und damit $\mu(A) \leq \mu(B)$. \square

Satz 12 (σ -additivität). Sei $A_j \subset \mathbb{R}^n$ eine Folge von Mengen. Dann gilt

$$\mu\left(\bigcup_j A_j\right) \leq \sum_{i=1}^{\infty} \mu(A_j)$$

Beweis. Für jedes A_j und $\epsilon > 0$ können wir eine geeignete Überdeckung $A_j \subset \bigcup_k K_{j,k}$ mit Hüllquadrern $K_{j,k}$ finden, so dass $\sum_k \text{vol}(K_{j,k}) \leq \mu(A_j) + \frac{\epsilon}{2^{j+1}}$. Da $\bigcup_j A_j \subset \bigcup_j \bigcup_k K_{j,k}^j$ eine Überdeckung mit Hüllquadrern ist, folgt

$$\mu\left(\bigcup_j A_j\right) \leq \sum_j \sum_k \text{vol}(K_{j,k}) \leq \left(\sum_j \mu(A_j) + \frac{\epsilon}{2^{j+1}}\right) = \left(\sum_j \mu(A_j)\right) + \epsilon$$

(Die letzte Gleichung beruht auf dem Wert der geometrischen Reihe). □

Satz 13. Für $I \in \mathbb{I}(n)$ und $A \subset \mathbb{R}^n$ mit $I \subset A \subset \bar{I}$ gilt $\mu(A) = \text{vol}(I)$.

Beweis. □

3.2 Lebesgue Integral

Anwendung: Fourierreihen, Fouriertransformation und FFT

Tabellenverzeichnis

Abbildungsverzeichnis

1	Quelle: Wikipedia: https://de.wikipedia.org/wiki/Datei:Differential_quotient_of_a_function.svg	
2	Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Mittelwertsatz3.svg	6
3	Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Epsilonschlauch_klein.svg	8
4	Quelle: Wikipedia: https://de.wikipedia.org/wiki/Datei:Limes_Definition_Vektorgrafik.svg	9
5	Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Upper_semi.svg	9
6	Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Gradient2.svg	10
7	Kantenzug mit achesenparallelen Kanten	12
8	Quelle: Wikipedia: https://de.wikipedia.org/wiki/Datei:EbeneKurve.png	13
9	Quelle: Wikipedia	14
10	Quelle: Wikipedia: https://en.wikipedia.org/wiki/File:MaximumParaboloid.png	17
11	Quelle: Wikipedia: https://en.wikipedia.org/wiki/File:MaximumCounterexample.png	17
12	Quelle: Wikipedia: https://commons.wikimedia.org/wiki/File:Saddle_point.svg	19
13	Quelle: Wikipedia	20
14	Quelle: Wikipedia	21
15	Quelle: Wikipedia	21
16	Quelle: https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a	22
17	Quelle: Wikipedia	24