Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(11) Nichtlineare Optimierung

Motivation

- Viele Verfahren der Prädiktiven Modellierung oder, wie heutzutage auch oft gesagt wird, der "Künstlichen Intelligenz", z.B. Neuronale Netze als generalisierte logistische Regression oder Support-Vektor-Maschinen, basieren auf mathematisch recht überschaubaren Modellen.
- Ihre momentane breite Verwendung verdanken diese Verfahren im Wesentlichen den verbesserten Computer-Hardware-Komponenten der letzten 15 Jahre, die zum Lernen ihrer Parameter ("Trainieren") genutzt werden, weniger wesentlichen neuen theoretischen Einsichten. Für einen aktuellen Überblick, siehe zum Beispiel Prince (2023) und Murphy (2023).
- Das Lernen von Parametern von Modellen der Prädiktiven Modellierung enstpricht der Optimierung von Funktionen, wie wir in (12) Logistische Regression und (13) Neuronale Netze sehen werden, insbesondere der Minimierung sogenannter *Loss Functions*.
- Zur Optimierung von Funktionen werden in diesem Bereich insbesondere und im einfachsten Fall sogenannten *Gradientenverfahren* eingesetzt. In diesem Abschnitt wollen wir deshalb zunächst ein Grundverständnis von Gradientenverfahren erarbeiten.

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Definition (Multivariate reellwertige Funktion)

Eine Funktion der Form

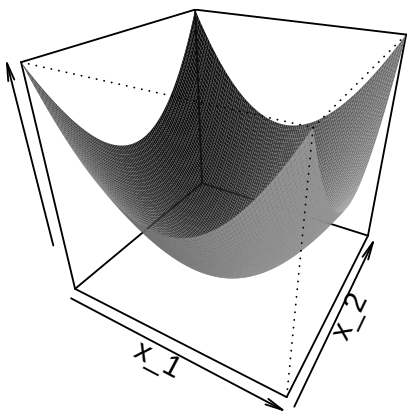
$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n) \tag{1}$$

heißt *multivariate reellwertige* Funktion.

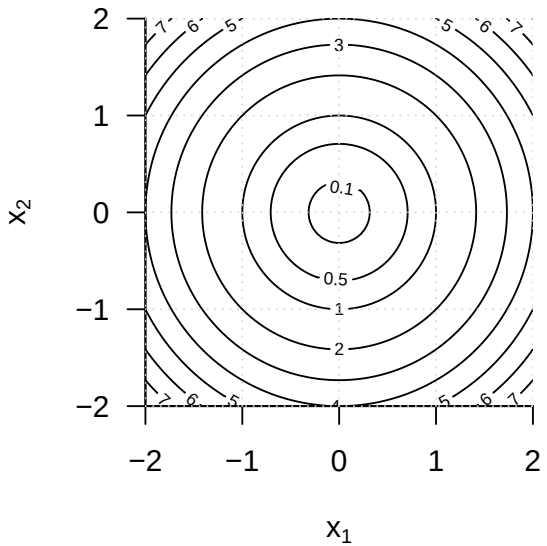
Beispiel für $n := 2$

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2 \tag{2}$$

f(x): = $x_1^2 + x_2^2$



f(x): = $x_1^2 + x_2^2$



Definition (Partielle Ableitung)

Es sei $D \subseteq \mathbb{R}^n$ eine Menge und

$$f : D \to \mathbb{R}, x \mapsto f(x) \tag{3}$$

eine multivariate reellwertige Funktion. f heißt in $x \in D$ *nach* x_i *partiell differenzierbar*, wenn der Grenzwert

$$\frac{\partial}{\partial x_i} f(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{4}$$

existiert. $\frac{\partial}{\partial x_i} f(x)$ heißt dann die *partielle Ableitung von* f *nach* x_i *an der Stelle* x. Wenn f für alle $x \in D$, nach x_i partiell differenzierbar ist, dann heißt f *nach* x_i *partiell differenzierbar* und die Funktion

$$\frac{\partial}{\partial x_i} f : D \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_i} f(x) \tag{5}$$

heißt *partielle Ableitung von* f *nach* x_i.

f heißt *partiell differenzierbar in* $x \in D$, wenn f für alle $i = 1, \ldots, n$ in $x \in D$ nach x_i partiell differenzierbar ist, und f heißt partiell differenzierbar, wenn f für alle $i = 1, \ldots, n$ in allen $x \in D$ nach x_i partiell differenzierbar ist.

Bemerkungen

- $e_i \in \mathbb{R}^n$ bezeichnet den $i$ten Einheitsvektor.
- $\frac{f(x+he_i)-f(x)}{h}$ misst die Änderung $f(x + he_i) - f(x)$ von f pro Strecke h in Richtung e_i.
- Für $h \to 0$ misst der Differenzquotient die Änderungsrate von f in x in Richtung e_i.
- $\frac{\partial}{\partial x_i} f(x)$ ist eine Zahl, $\frac{\partial}{\partial x_i} f$ ist eine Funktion.
- Praktisch berechnet man $\frac{\partial}{\partial x_i} f$ als die (einfache) Ableitung

$$\frac{d}{dx_i} \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) \tag{6}$$

der univariaten reellwertigen Funktion

$$\tilde{f} : \mathbb{R} \to \mathbb{R}, x_i \mapsto \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) := f(x_1, ..., x_i, ..., x_n). \tag{7}$$

- Man betrachtet alle x_j mit $j \neq i$ also als Konstanten.

Beispiel (1)

Wir betrachten die Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{8}$$

Weil die Definitionsmenge dieser Funktion zweidimensional ist, kann man zwei partielle Ableitungen berechnen

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) \text{ und } \frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x). \tag{9}$$

Um die erste dieser partiellen Ableitungen zu berechnen, betrachtet man die Funktion

$$f_{x_2} : \mathbb{R} \to \mathbb{R}, x_1 \mapsto f_{x_2}(x_1) := x_1^2 + x_2^2, \tag{10}$$

wobei x_2 hier die Rolle einer Konstanten einnimmt. Um explizit zu machen, dass x_2 kein Argument der Funktion ist, die Funktion aber weiterhin von x_2 abhängt haben wir die Subskriptnotation $f_{x_2}(x_1)$ verwendet. Um nun die partielle Ableitung zu berechnen, berechnen wir die (einfache) Ableitung von f_{x_2},

$$f'_{x_2}(x) = 2x_1\,. \tag{11}$$

Es ergibt sich also

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) = \frac{\partial}{\partial x_1}(x_1^2 + x_2^2) = f'_{x_2}(x) = 2x_1\,. \tag{12}$$

Analog gilt mit der entsprechenden Formulierung von f_{x_1}, dass

$$\frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x) = \frac{\partial}{\partial x_2}(x_1^2 + x_2^2) = f'_{x_1}(x) = 2x_2\,. \tag{13}$$

Definition (Zweite partielle Ableitungen)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion und $\frac{\partial}{\partial x_i} f$ sei die partielle Ableitung von f nach x_i. Dann ist die zweite partielle Ableitung von f nach x_i und x_j definiert als

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial}{\partial x_i} f \right) \tag{14}$$

Bemerkungen

- Wie die zweite Ableitung ist auch die zweite partielle Ableitung rekursiv definiert.
- Zu jeder partiellen Ableitung $\frac{\partial}{\partial x_i} f$ gibt es n zweite partiellen Ableitungen $\frac{\partial^2}{\partial x_j \partial x_i} f, j = 1, \ldots, n$.

Theorem (Satz von Schwarz)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine partiell differenzierbare multivariate reellwertige Funktion. Dann gilt

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \text{ für alle } 1 \leq i, j \leq n. \tag{15}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem von Schwarz besagt, dass die Reihenfolge des partiellen Ableitens irrelevant ist
- Das Theorem erleichtert die Berechnung von zweiten partiellen Ableitungen.
- Das Theorem hilft, Fehler bei der Berechnung zweiter partieller Ableitungen aufzudecken.

Beispiel (1) (fortgeführt)

Wir wollen die partiellen Ableitungen zweiter Ordnung der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{16}$$

berechnen. Mit den Ergebnissen für die partiellen Ableitungen erster Ordnung dieser Funktion ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_1} (2x_1) = 2 \\
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_1} (2x_2) = 0 \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_2} (2x_1) = 0 \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_2} (2x_2) = 2
\end{aligned} \tag{17}$$

Offenbar gilt

$$\frac{\partial^2}{\partial x_1 \partial x_2} f(x) = \frac{\partial^2}{\partial x_2 \partial x_1} f(x). \tag{18}$$

Beispiel (2)

Wir wollen die partiellen Ableitungen erster und zweiter Ordnung der Funktion

$$f : \mathbb{R}^3 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_1 x_2 + x_2\sqrt{x_3}. \tag{19}$$

berechnen.

Mit den Rechenregeln für Ableitungen ergibt sich für die partiellen Ableitungen erster Ordnung

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = 2x_1 + x_2, \\
\frac{\partial}{\partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = x_1 + \sqrt{x_3}, \\
\frac{\partial}{\partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = \frac{x_2}{2\sqrt{x_3}}.
\end{aligned} \tag{20}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_1 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_1}(2x_1 + x_2) = 2, \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_2}(2x_1 + x_2) = 1, \\
\frac{\partial^2}{\partial x_3 \partial x_1} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_3}(2x_1 + x_2) = 0.
\end{aligned} \tag{21}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_2 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_1}(x_1 + \sqrt{x_3}) = 1, \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_2}(x_1 + \sqrt{x_3}) = 0, \\
\frac{\partial^2}{\partial x_3 \partial x_2} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_3}(x_1 + \sqrt{x_3}) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{22}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_3 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_1}\left(\frac{x_2}{2}\sqrt{x_3}\right) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_2}\left(\frac{x_2}{2\sqrt{x_3}}\right) = \frac{1}{2\sqrt{x_3}}, \\
\frac{\partial^2}{\partial x_3 \partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_3}\left(x_2 \frac{1}{2} x_3^{-\frac{1}{2}}\right) = -\frac{1}{4} x_2 x_3^{-\frac{3}{2}}.
\end{aligned} \tag{23}$$

Weiterhin erkennt man, dass die Reihenfolge der partiellen Ableitungen irrelevant ist, denn es gilt

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(x) = 1, \\
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_1} f(x) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_2} f(x) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{24}$$

Definition (Gradient)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist der *Gradient* $\nabla f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^n. \tag{25}$$

Bemerkung

- $\nabla f(x)$ fasst die partiellen Ableitungen von f an der Stelle $x \in \mathbb{R}^n$ in einem Vektor zusammen.
- Gradienten sind multivariate vektorwertige Abbildungen der Form $\nabla f : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto \nabla f(x)$.
- Wir zeigen später, dass $-\nabla f(x)$ die Richtung des steilsten Abstiegs von f in $\mathbb{R}^n$ anzeigt.
- Für $n = 1$ gilt $\nabla f(x) = f'(x)$.

Beispiele

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \in \mathbb{R}^2. \tag{26}$$

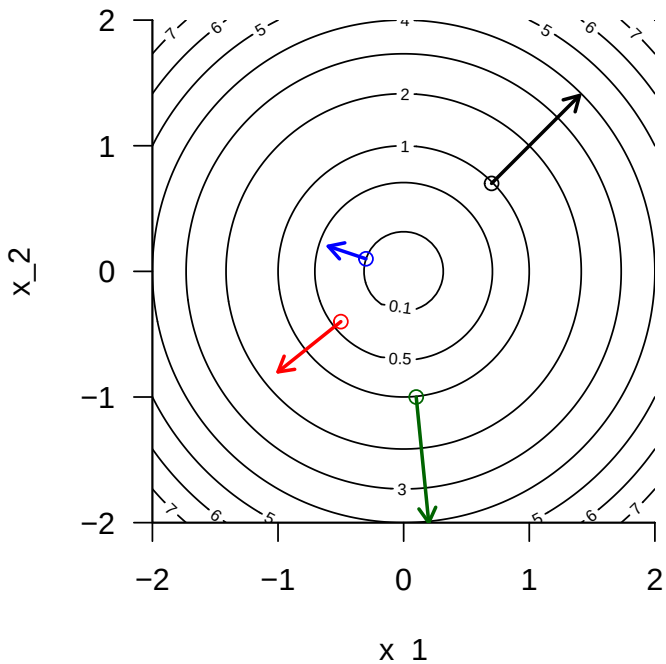
Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + \sqrt{x_3} \\ \frac{x_2}{2\sqrt{x_3}} \end{pmatrix} \in \mathbb{R}^3. \tag{27}$$

Beispiel (1) (fortgeführt)

Gradienten von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$ bei

$$x = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} \qquad x = \begin{pmatrix} -0.3 \\ 0.1 \end{pmatrix} \qquad x = \begin{pmatrix} -0.5 \\ -0.4 \end{pmatrix} \qquad x = \begin{pmatrix} 0.1 \\ -1.0 \end{pmatrix}$$



Definition (Hesse-Matrix)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist die *Hesse-Matrix* $\nabla^2 f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \frac{\partial^2}{\partial x_n \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n \partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{28}$$

Bemerkung

- $\nabla^2 f(x)$ fasst die partiellen Ableitungen zweiter Ordnung von f in einer Matrix zusammen.
- Hesse-Matrizen sind multivariate matrixwertige Abbildungen der Form $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}, x \mapsto \nabla^2 f(x)$.
- Für $n = 1$ gilt $\nabla^2 f(x) = f''(x)$.
- Mit $\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x)$ für $1 \leq i, j \leq n$ folgt, dass $\left(\nabla^2 f(x)\right)^T = \nabla^2 f(x)$.

Beispiel

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \frac{\partial^2}{\partial x_2 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_3 \partial x_1} f(x) & \frac{\partial^2}{\partial x_3 \partial x_2} f(x) & \frac{\partial^2}{\partial x_3 \partial x_3} f(x) \end{pmatrix}$$

$$:= \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{4} x_2 x_3^{-3/2} \end{pmatrix}$$

Definition (Glatte multivariate reellwertige Funktion)

Eine multivariate reellwertige Funktion

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) \tag{29}$$

heißt *glatt*, wenn ihr Gradient und ihre Hesse-Matrix existieren und für alle $x \in \mathbb{R}^n$ stetig sind.

Bemerkungen

- Der Gradient und die Hesse-Matrix einer glatten Funktion könnten überall in $\mathbb{R}^n$ berechnet werden.

Theorem (Multivariater Mittelwertsatz erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

$$f(x + p) = f(x) + \nabla f(x + tp)^T p. \tag{30}$$

Theorem (Multivariater Mittelwertsatz zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p. \tag{31}$$

Bemerkung

- Wir verzichten auf Beweise.
- Nocedal and Wright (2006) bezeichnen die Theoreme als “Taylor’s Theorem”, das ist ein wenig misleading.
- ∇f und $\nabla^2 f$ werden an einer Stelle zwischen x und $x + p$ evaluiert.

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Definition (Optimierungsproblem)

Ein *Optimierungsproblem* hat die allgemeine Form

$$\min_x f(x), \tag{32}$$

wobei $x \in \mathbb{R}^n$ und $f : \mathbb{R}^n \to \mathbb{R}$ eine glatte multivariate reellwertige Funktion ist. Die Lösung x^* eines Optimierungsproblems wird bezeichnet mit

$$x^* = \arg\min_x f(x). \tag{33}$$

Bemerkungen

- Weil gilt, dass $\max_x f(x) = \min_x -f(x)$ genügt es, sich mit Minimierungsproblemen zu befassen.
- Im Allgemeinen ist die Lösung x^* eines Optimierungsproblems eine Menge.
- Man denkt bei $\arg\min_x f(x)$ alllerdings auch einfach an Elemente dieser Menge.

Definition (Globale und lokale Minimalstellen/Minima)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion.

- $x^* \in \mathbb{R}^n$ heißt *globale Minimalstelle* von f, wenn $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$ gilt. $f(x^*) \in \mathbb{R}$ heißt dann das *globale Minimum* von f.

- $x^* \in \mathbb{R}^n$ heißt *lokale Minimalstelle* von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) \leq f(x)$ für alle $x \in N \subset \mathbb{R}^n$. $f(x^*) \in \mathbb{R}$ wird dann ein *lokales Minimum* von f genannt.

- $x^* \in \mathbb{R}^n$ heißt *strikte lokale Minimalstelle* von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) < f(x)$ für alle $x \in N \subset \mathbb{R}^n$. $f(x^*) \in \mathbb{R}$ wird dann ein *striktes lokales Minimum* von f genannt.

Bemerkung

- Eine Umgebung von $x \in \mathbb{R}^n$ ist eine offene Menge, die x enthält.

Theorem (Notwendige Bedingung erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann gilt

$$\nabla f(x^*) = 0_n. \tag{34}$$

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Dazu nehmen wir an, dass x^* zwar eine lokale Minimalstelle von f ist, aber $\nabla f(x^*) \neq 0_n$ ist. Dazu definieren wir zunächst $p := -\nabla f(x^*)$. Dann gilt, dass

$$p^T \nabla f(x^*) = -\nabla f(x^*)^T \nabla f(x^*) = -||\nabla f(x^*)|| < 0. \tag{35}$$

Weil ∇f in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass auch

$$p^T \nabla f(x^* + tp) < 0 \text{ für alle } t \in [0, T]. \tag{36}$$

gilt. Nun gilt für $\tilde{t} \in]0, T]$ aber mit dem Mittelwertssatz erster Ordnung, dass

$$f(x^* + \tilde{t}p) = f(x^*) + \nabla f(x^* + tp)^T \tilde{t}p = f(x^*) + \tilde{t}p^T \nabla f(x^* + tp) \text{ für ein } t \in]0, \tilde{t}[. \tag{37}$$

Also folgt $f(x^* + \tilde{t}p) < f(x^*)$ für alle $\tilde{t} \in]0, T]$. Wir haben also eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla f(x^*) \neq 0_n$ gilt. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla f(x^*) \neq 0_n$ gilt. Also muss $\nabla f(x^*) = 0_n$ gelten, wenn x^* eine lokale Minimalstelle ist.

Theorem (Notwendige Bedingung zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann ist $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ ist positiv semidefinit.

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Wir haben schon gesehen, dass $\nabla f(x^*) = 0_n$ ist, wenn x^* eine lokale Minimalstelle von f. ist. Für einen Widerspruchsbeweis nehmen wir nun an, dass x^* zwar eine lokale Minimalstelle von f ist, aber dass $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dann ist es möglichen einen Vektor p zu finden, so dass gilt

$$p^T \nabla^2 f(x^*) p < 0. \tag{38}$$

Weil $\nabla^2 f(x^*)$ in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass

$$p^T \nabla^2 f(x^* + tp) p < 0 \text{ für alle } t \in [0, T]. \tag{39}$$

gilt. Mithilfe des Mittelwertsatzes zweiter Ordnung gilt dann für alle $\bar{t} \in]0, T]$ und ein $t \in]0, \bar{t}[$, dass

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x^*) + \frac{1}{2} \bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*). \tag{40}$$

Wir haben also wieder eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Also muss $\nabla^2 f(x^*)$ positiv semidefinit sein, wenn x^* eine lokale Minimalstelle ist.

Theorem (Hinreichende Bedingungen zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es seien $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ positiv definit. Dann ist x^* eine strikte Minimalstelle von f.

Beweis

Wir halten zunächst fest, dass weil die Hesse-Matrix stetig und positiv definit in x^* ist, wir ein $r > 0$ wählen können, so dass $\nabla^2 f(x)$ positiv definit für alle x in

$$D = \{x | ||x - x^*|| < r\} \tag{41}$$

ist. Für einen Vektor p mit $||p|| > 0$ und $||p|| < r$ gilt $x^* + p \in D$. Für ein $t \in]0, 1[$ gilt dann mit dem Mittelwertssatz zweiter Ordnung, dass

$$\begin{aligned} f(x^* + p) &= f(x^*) + \nabla f(x^*) p^T + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p. \end{aligned} \tag{42}$$

Weil aber $x^* + tp \in D$ ist, gilt, dass $p^T \nabla^2 f(x^* + tp) p > 0$ ist und somit $f(x^* + p) > f(x^*)$. In jeder Richtung p von x^* weg erhöht sich also der Wert von f und damit ist x^* eine strikte Minimalstelle.

Zusammenfassung

Optimierungsproblem

$$\min_x f(x) = \max_x -f(x) \text{ für } f : \mathbb{R}^n \to \mathbb{R}$$

Lokale Minimalstelle

$$x^* = \arg\min_x f(x), x^* \in \mathbb{R}^n \Leftrightarrow f(x^*) \leq f(x) \text{ für alle } x \in N \subset \mathbb{R}^n$$

Notwendige Bedingung erster Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n$$

Notwendige Bedingung zweiter Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n \text{ und } \nabla^2 f(x^*) \text{ positiv semidefinit}$$

Hinreichende Bedingung zweiter Ordnung

$$\nabla f(x^*) = 0_n \text{ und } \nabla^2 f(x^*) \text{ positiv definit } \Rightarrow x^* = \arg\min_x f(x)$$

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Allgemeine Form von Optimierungsalgorithmen

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Berechnung von x_{k+1} basierend auf Information über f an der Stelle x_k.
2. STOP, wenn Minimalstelle gefunden ist oder kein Fortschritt mehr erzielt wird.

Definition (Gradientenverfahren)

Es sei $f : \mathbb{R}^n \to \mathbb{R}$ eine multivariate reellwertige Funktion. Dann hat das *Gradientenverfahren* zur Minimierung von f folgende allgemeine Form:

Initialisierung

0. Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$, eine Lernrate $\alpha > 0$ und ein Konvergenzkriterium $\delta > 0$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Setze $x_{k+1} := x_k - \alpha \nabla f(x_k)$.
2. STOP, wenn $||\nabla f(x_{k+1})|| < \delta$, ansonsten gehe zu 1.

Bemerkungen

- Die Lernrate α bestimmt, wie weit ein Schritt in Richtung des Gradienten erfolgt.
- Das Konvergenzkriterium bestimmt, wie klein der Gradient sein muss, damit das Verfahren endet.

Theorem (Gradientenverfahren)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $x_k \in \mathbb{R}^n$. Dann ist die Gradientenrichtung

$$p_k^G := -\nabla f(x_k) \tag{43}$$

die Richtung des steilsten Abstiegs von f in x_k.

Bemerkungen

- Es gibt unendliche viele mögliche Richtungen p in x_k.
- $\nabla f(x) \in \mathbb{R}^n$ ist eine Richtung in der Definitionsmenge von f (Parameterraum).
- Die Gradientenrichtung ist davon die Richtung, in der die Zielfunktion f am schnellsten abnimmt.
- Zum Vergleich von Richtungen genügt es, Richtungen der Länge $||p|| = 1$ zu vergleichen.

Beweis

Mit dem Mittelwertsatz zweiter Ordnung gilt für jede Richtung p und Schrittlängenparameter α, dass

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f(x_k) + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p \text{ für ein } t \in]0, \alpha[. \tag{44}$$

Die Änderungsrate von f in Richtung p in x_k ist also der Koeffizient von α, also $p^T \nabla f(x_k)$ (man denke an $x = tv$ für einen Ort x, eine Geschwindigkeit v und eine Zeit t). Also gilt, dass die Richtung des steilsten Abstiegs p in x_k mit Länge 1 die Lösung des Optimierungsproblems

$$\min_p p^T \nabla f(x_k) \text{ mit der Nebenbedingung } ||p|| = 1. \tag{45}$$

ist. Wir erinnern nun zunächst daran, dass für $x, y \in \mathbb{R}^n$ gilt der Kosinus des Winkel zwischen x und y durch

$$\cos\alpha = \frac{\langle x, y\rangle}{||x||||y||} = \frac{x^T y}{||x||||y||} \tag{46}$$

gegeben ist. Damit aber gilt, dass

$$p^T \nabla f(x_k) = ||p|| \cdot ||\nabla f(x_k)|| \cos\theta = 1 \cdot ||\nabla f(x_k)|| \cos\theta = ||\nabla f(x_k)|| \cos\theta \tag{47}$$

und somit liegt hier bei $\cos\theta = -1$ eine Minimalstelle vor. Dies bedeutet aber, dass die minimierende Länge p exakt antiparallel zu $\nabla f(x_k)$ und von Länge 1 sein muss. Also ist die Minimalstelle des Optimierungsproblems

$$p = \frac{-\nabla f(x_k)}{||\nabla f(x_k)||}. \tag{48}$$

Damit ist $p_k^G := -\nabla f(x_k)$ aber der Richtungsvektor beliebiger Länge in der die Abnahme von f maximal ist.

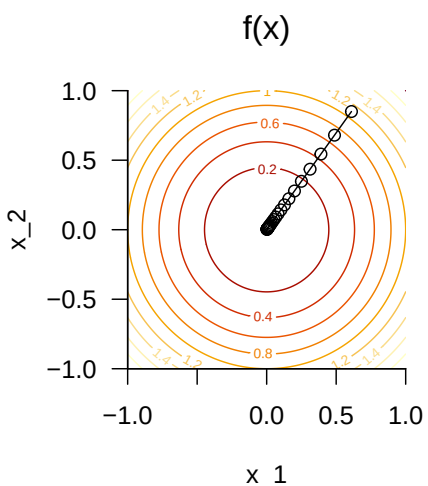
Beispiel

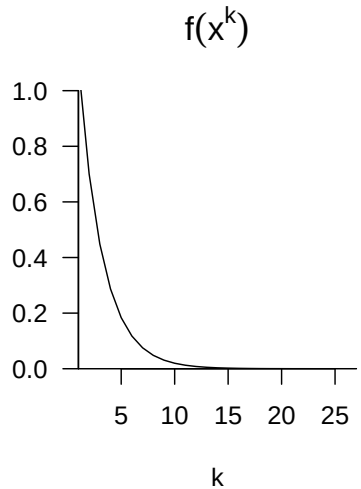
Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$

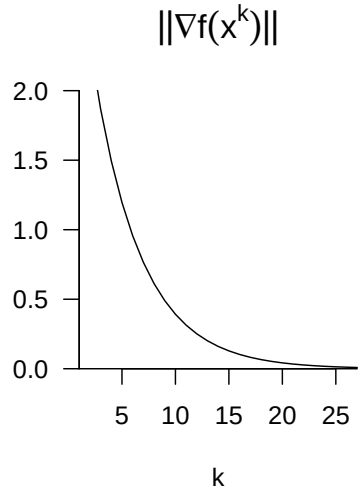
```
# Funktionsdefinitionen
# ------------------------------------------------------------------------------
# Zielfunktion
f = function(x) {
  return(x[1]^2 + x[2]^2)                # f(x) := x_1^2 + x_2^2
}
# Gradient der Zielfunktion
nabla_f = function(x) {
  return(matrix(c(2*x[1],2*x[2]),        # \nabla f(x) := (2x_1, 2x_2)^T
                nrow = 2))
}
# Gradientenverfahren
# ------------------------------------------------------------------------------
# Parameter
n      = 2                             # Dimension
alpha  = 1e-1                          # Lernrate
delta  = 1e-2                          # Konvergenzkriterium

# Initialisierung
x_k    = matrix(c(.61,.85), nrow = 2)  # Zufälliger Startpunkt in [0,1]^2
x      = x_k                           # Initialisierung Iteranden
fx     = f(x_k)                        # Initialisierung Funktionswerte
crt    = norm(nabla_f(x_k))            # Initialisierung Kriterium

# Iterationen
while(norm(nabla_f(x_k)) > delta){

  # Argumentupdate
  x_k    = x_k - alpha*nabla_f(x_k)

  # Dokumentation
  x     = cbind(x,x_k)
  fx    = c(fx,f(x_k))
  crt   = c(crt, norm(nabla_f(x_k)))
}
```

Beispiel

Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$







Liniensuchverfahren als generalisierte Gradientenverfahren

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Wahl einer Abstiegsrichtung p_k
2. Wahl eines Lernparameters $\alpha_k \approx \min_\alpha f(x_k + \alpha p_k)$.
3. Setze $x_{k+1} := x_k + \alpha_k p_k$.
4. Konvergenztest.

$\Rightarrow$ Die Wahl sinnvoller Lernraten α_k ist für eine gute Performanz entscheidend!

(vgl. Ostwald and Starke (2016))

Selbstkontrollfragen

1. Geben Sie die Definition einer multivariaten reellwertigen Funktion wieder.
2. Geben Sie die Definition der partiellen Ableitung wieder.
3. Geben Sie die Definition der zweiten partiellen Ableitung wieder.
4. Geben Sie den Satz von Schwarz wieder.
5. Geben Sie die Definition eines Gradienten einer multivariaten reellwertigen Funktion wieder.
6. Geben Sie die Definition der Hesse-Matrix einer multivariaten reellwertigen Funktion wieder.
7. Geben Sie die allgemeine Form eines Optimierungsproblems wieder.
8. Geben Sie die notwendige Bedingung erster Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
9. Geben Sie die notwendige Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
10. Geben Sie die hinreichende Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
11. Geben Sie die Definition des Gradientverfahrens wieder.
12. Erläutern Sie die Bedeutung der Lernrate und des Konvergenzkriteriums im Gradientenverfahren.
13. Geben Sie das Theorem zum Gradientenverfahren wieder.

Referenzen

Murphy, Kevin P. 2023. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research. New York: Springer.

Ostwald, Dirk, and Ludger Starke. 2016. "Probabilistic Delay Differential Equation Modeling of Event-Related Potentials." *NeuroImage* 136 (August): 227–57. https://doi.org/10.1016/j.neuroimage.2016.04.025.

Prince, Simon J. D. 2023. *Understanding Deep Learning*. Cambridge, Massachusetts: The MIT Press.