2 Extremwerte und konvexe Funktionen

In diesem Abschnitt diskutieren wir lokale Extrema von Funktionen mehrerer Variabler, und verallgemeinern die entsprechenden notwendigen und hinreichenden Kriterien aus Analysis 1. Dabei spielt die zweite Ableitung eine entscheidende Rolle. Wir behandeln im Anschluss Grundtatsachen über konvexe Funktionen. Als bekannt setzen wir voraus: auf einer kompakten Teilmenge des $\mathbb{R}^n$ nimmt eine stetige Funktion ihre Extremwerte an.

Definition 2.1 *Die Funktion $f : M \to \mathbb{R}$, $M \subset \mathbb{R}^n$, hat in $x \in M$ ein lokales Minimum, falls es ein $\delta > 0$ gibt mit*

$$f(y) \geq f(x) \quad \text{für alle } y \in B_\delta(x) \cap M.$$

Ist sogar $f(y) > f(x)$ für $y \in B_\delta(x)\backslash\{x\}$, so heißt das Minimum isoliert. Ein (isoliertes) lokales Maximum ist entsprechend definiert.

Satz 2.1 (notwendige Bedingung für Extrema) *Sei $\Omega \subset \mathbb{R}^n$ offen, und $f : \Omega \to \mathbb{R}$ habe in $x \in \Omega$ ein lokales Extremum. Ist f differenzierbar in x, so folgt $Df(x) = 0$.*

Beweis: Für $v \in \mathbb{R}^n$ hat die Funktion $t \mapsto f(x + tv)$ ein lokales Extremum bei $t = 0$, also folgt aus der eindimensionalen Version und Satz 3.1

$$0 = \frac{d}{dt} f(x + tv)|_{t=0} = Df(x)v \quad \text{für alle } v \in \mathbb{R}^n.$$

□

Definition 2.2 (kritischer Punkt) *Ein Punkt $x \in \Omega$ mit $Df(x) = 0$ heißt kritischer Punkt von f.*

Die kritischen Punkte sind mögliche Kandidaten für Extremstellen, allerdings zeigt schon das Beispiel $f(x_1, x_2) = x_1^2 - x_2^2$, dass in kritischen Punkten nicht unbedingt ein lokales Extremum vorliegt. Um das genauer zu analysieren, müssen wir die zweite Ableitung heranziehen.

Definition 2.3 (zweite Ableitung) *Sei $f \in C^2(\Omega)$ mit $\Omega \subset \mathbb{R}^n$ offen. Die zweite Ableitung von f im Punkt $x \in \Omega$ ist die Bilinearform*

$$D^2 f(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ D^2 f(x)(v, w) = \sum_{i,j=1}^{n} \partial_i \partial_j f(x) v_i\, w_j.$$

Die zugehörige Matrix $\big(\partial_i \partial_j f(x)\big)_{1 \leq i,j \leq n}$ heißt Hessematrix *von f an der Stelle x, und die zugehörige quadratische Form*

$$v \mapsto D^2 f(x)(v, v) = \sum_{i,j=1}^{n} \partial_i \partial_j f(x) v_i v_j$$

heißt Hesseform *von f in x.*

Die Hessematrix ist symmetrisch, und $D^2f(x)$ ist symmetrische Bilinearform. Denn nach Satz 2.2 gilt $\partial_i\partial_j f = \partial_j\partial_i f$ für $f \in C^2(\Omega)$, und daraus folgt

$$D^2f(x)(v,w) = \sum_{i,j=1}^{n} \partial_i\partial_j f(x) v_i w_j = \sum_{i,j=1}^{n} \partial_j\partial_i f(x) w_j v_i = D^2f(x)(w,v).$$

Als erstes wollen wir dir Formel für die zweite Ableitung längs Kurven herleiten.

Lemma 2.1 *Sei $\Omega \subset \mathbb{R}^n$ offen, $f \in C^2(\Omega)$ und $\gamma \in C^2(I,\Omega)$. Dann gilt*

$$(f\circ\gamma)''(t) = D^2f(\gamma(t))\big(\gamma'(t),\gamma'(t)\big) + Df(\gamma(t))\gamma''(t). \tag{2.1}$$

BEWEIS: Nach Kettenregel ist $(f\circ\gamma)'(t) = \sum_{j=1}^n \partial_j f(\gamma(t))\gamma_j'(t)$, und weiter

$$(f\circ\gamma)''(t) = \sum_{i,j=1}^{n} \partial_i\partial_j f(\gamma(t))\gamma_i'(t)\gamma_j'(t) + \sum_{j=1}^{n} \partial_j f(\gamma(t))\gamma_j''(t).$$

□

Wir benötigen nun eine lokale Entwicklung, die die zweite Ableitung mit einbezieht.

Lemma 2.2 *Sei $\Omega \subset \mathbb{R}^n$ offen und $f \in C^2(\Omega)$. Dann gilt*

$$\frac{f(x+h) - \big(f(x) + Df(x)h + \frac{1}{2}D^2f(x)(h,h)\big)}{|h|^2} \to 0 \quad \text{mit } h \to 0.$$

BEWEIS: Wir betrachten die Funktion $\varphi(t) = f(x+th)$. Mit partieller Integration gilt

$$\varphi(1) - \big(\varphi(0) + \varphi'(0)\big) = \int_0^1 (\varphi'(t) - \varphi'(0))\,dt = \int_0^1 (1-t)\varphi''(t)\,dt,$$

also folgt aus Lemma 2.1

$$f(x+h) - \big(f(x) + Df(x)h\big) = \int_0^1 (1-t)D^2f(x+th)(h,h)\,dt. \tag{2.2}$$

Daraus ergibt sich die Darstellung

$$f(x+h) - \big(f(x) + Df(x)h + \frac{1}{2}D^2f(x)(h,h)\big) = \int_0^1 (1-t)\big(D^2f(x+th)(h,h) - D^2f(x)(h,h)\big)\,dt.$$

Zu $\varepsilon > 0$ gibt es ein $\delta > 0$ mit $|\partial_i\partial_j f(y) - \partial_i\partial_j f(x)| < \varepsilon$ für $|y-x| < \delta$. Für $|h| < \delta$ folgt

$$\sum_{i,j=1}^{n} \big|\big(\partial_i\partial_j f(x+th) - \partial_i\partial_j f(x)\big)h_i h_j\big| < n^2|h|^2\varepsilon,$$

woraus sich die Behauptung ergibt. □

Als zweites Hilfsmittel brauchen wir folgende Tatsache über quadratische Formen.

Lemma 2.3 *Sei $b : V \times V \to \mathbb{R}$ eine symmetrische Bilinearform auf dem n-dimensionalen Euklidischen Vektorraum V mit Norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. Setze*

$$\lambda = \inf\{b(x,x) : x \in V,\ \|x\| = 1\}.$$

Dann gibt es ein $v \in V$ mit $\|v\| = 1$ und $b(v,v) = \lambda$.

BEWEIS: Sei zunächst $V = \mathbb{R}^n$ und $\langle\cdot,\cdot\rangle$ das Standardskalarprodukt. Die Funktion

$$f : \mathbb{R}^n \to \mathbb{R},\ f(x) = b(x,x) = \sum_{i,j=1}^{n} b_{ij}x_ix_j \quad \text{wobei } b_{ij} = b(e_i,e_j),$$

ist stetig auf $\mathbb{R}^n$, und $\{x \in \mathbb{R}^n : |x| = 1\} = \mathbb{S}^{n-1}$ ist kompakt. Die Existenz von v folgt also aus Satz 2.1 in Kapitel 4. Für V und $\langle\cdot,\cdot\rangle$ beliebig wählen wir eine Orthonormalbasis $v_1,\dots,v_n$ von V. Mit $T : \mathbb{R}^n \to V$, $T(x) = \sum_{i=1}^n x_iv_i$, gilt dann $\|T(x)\| = |x|$. Wir können also das Bewiesene anwenden auf die Bilinearform $\tilde{b} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $\tilde{b}(x,y) = b(T(x),T(y))$. □

Definition 2.4 *Sei V ein $\mathbb{R}$-Vektorraum. Eine symmetrische Bilinearform $b : V \times V \to \mathbb{R}$ heißt positiv definit (bzw. positiv semidefinit), falls gilt:*

$$b(v,v) > 0 \quad \textit{(bzw. } b(v,v) \geq 0\textit{)} \quad \textit{für alle } v \in V\backslash\{0\}.$$

Notation: $b > 0$ bzw. $b \geq 0$. Entsprechend für negativ (semi-)definit.

Bei der Monotonie ist es so, dass wir ausdrücklich den Begriff streng monoton verwenden, wenn wir Gleichheit ausschließen wollen. Hier ist der Sprachgebrauch ausnahmsweise anders, definit bedeutet automatisch die strikte Ungleichung. Eine Bilinearform, für die es $v, w \in V$ gibt mit $b(v,v) > 0$ und $b(w,w) < 0$ heißt *indefinit*. Ein Beispiel ist die Bilinearform $b(x,y) = x_1y_1 - x_2y_2$ auf $\mathbb{R}^2$. Anders als bei reellen Zahlen muss also nicht eine der Ungleichungen $b > 0$, $b = 0$ oder $b < 0$ eintreten.

Satz 2.2 (Lokale Extrema) *Sei $f \in C^2(\Omega)$, $\Omega \subset \mathbb{R}^n$ offen und $x \in \Omega$.*

(a) *Wenn f in x ein lokales Minimum hat, so ist $D^2f(x)$ positiv semidefinit.*

(b) *Ist $Df(x) = 0$ und $D^2f(x)$ positiv definit, so hat f in x ein isoliertes lokales Minimum.*

BEWEIS: In (a) gilt $Df(x) = 0$ nach Satz 2.1. Für $v \in \mathbb{R}^n$ beliebig hat $t \mapsto f(x+tv)$ bei $t = 0$ ein lokales Minimum, also folgt aus dem eindimensionalen Fall und (2.1)

$$0 \leq \frac{d^2}{dt^2} f(x+tv)|_{t=0} = D^2f(x)(v,v).$$

Für (b) setzen wir $\lambda = \inf_{|v|=1} D^2f(x)(v,v)$. Nach Lemma 2.3 gibt es ein $v \in \mathbb{R}^n$, $|v| = 1$, mit $D^2f(x)(v,v) = \lambda$. Da $D^2f(x)$ positiv definit ist, folgt insbesondere $\lambda > 0$. Sei nun für $h \neq 0$

$$R(h) = \frac{f(x+h) - \big(f(x) + \frac{1}{2}D^2f(x)(h,h)\big)}{|h|^2}.$$

Nach Lemma 2.2 gibt es zu $\varepsilon > 0$ ein $\delta > 0$ mit $|R(h)| < \varepsilon$ für $|h| < \delta$. Es folgt

$$\frac{f(x+h) - f(x)}{|h|^2} = \frac{1}{2} D^2 f(x)\Big(\frac{h}{|h|}, \frac{h}{|h|}\Big) + R(h) > \frac{\lambda}{2} - \varepsilon.$$

Behauptung (b) folgt, wenn wir zum Beispiel $\varepsilon = \lambda/4$ wählen. □

Um die Funktion in der Nähe eines kritischen Punkts zu verstehen, ist der folgende Satz aus der Linearen Algebra nützlich.

Satz 2.3 (Hauptachsentransformation) *Sei $b : V \times V \to \mathbb{R}$ symmetrische Bilinearform auf dem n-dimensionalen, Euklidischen Vektorraum V. Dann existiert eine Orthonormalbasis $v_1, \dots, v_n$, so dass gilt:*

$$b(v_i, v_j) = \lambda_i \, \delta_{ij} \quad \text{für alle } i, j = 1, \dots, n.$$

Beweis: Wir zeigen: für $v \in V$ mit $\|v\| = 1$ und $b(v, v) = \inf\{b(x, x) : \|x\| = 1\} = \lambda$ gilt

$$b(v, w) = \lambda \langle v, w \rangle \quad \text{für alle } w \in V. \tag{2.3}$$

Die Menge der $w \in V$ mit $b(v, w) = \lambda \langle v, w \rangle$ ist ein Unterraum, der v enthält; deshalb können wir $\langle v, w \rangle = 0$ und $\|w\| = 1$ annehmen. Dann ist $\|(\cos t)v + (\sin t)w\|^2 = 1$ für alle t, und (2.3) folgt aus der Minimumeigenschaft:

$$b(v, w) = \frac{1}{2} \frac{d}{dt} \, b\big((\cos t)v + (\sin t)w, (\cos t)v + (\sin t)w\big)|_{t=0} = 0 = \lambda \langle v, w \rangle.$$

Jetzt konstruieren wir induktiv $v_1, \dots, v_k$ orthonormal und $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ mit

$$b(v_i, w) = \lambda_i \langle v_i, w \rangle \quad \text{für alle } w \in V.$$

Für $k = n$ folgt die Behauptung des Satzes, indem wir $w = v_j$ einsetzen. Seien $v_1, \dots, v_{k-1}$ und $\lambda_1, \dots, \lambda_{k-1}$ schon gefunden, wobei $1 \le k < n$. Setze $V_k = \{v_1, \dots, v_{k-1}\}^\perp$ und

$$\lambda_k = \inf\{b(x, x) : x \in V_k, \, \|x\| = 1\}.$$

Nach Lemma 2.3 gibt es ein $v_k \in V_k$ mit $\|v_k\| = 1$ und $b(v_k, v_k) = \lambda_k$, also gilt

$$b(v_k, w) = \lambda_k \langle v_k, w \rangle \quad \text{für alle } w \in V_k \quad \text{nach (2.3)}.$$

Aber für $1 \le i < k$ liefert die Symmetrie von b und die Induktionsannahme

$$b(v_k, v_i) = b(v_i, v_k) = \lambda_i \langle v_i, v_k \rangle = 0 = \lambda_k \langle v_k, v_i \rangle.$$

Es folgt $b(v_k, w) = \lambda_k \langle v_k, w \rangle$ für alle $w \in V$, womit der Induktionsschluss gezeigt ist. □

Ein Endomorphismus $B : V \to V$ eines Euklidischen Vektorraums V heißt symmetrisch (oder hermitesch), wenn die zugeordnete Bilinearform $b(v, w) = \langle Bv, w \rangle$ symmetrisch ist. Dann ist (2.3) äquivalent zu der Gleichung

$$Bv = \lambda v,$$

das heißt die v_k aus dem Satz sind Eigenvektoren von B zu den Eigenwerten λ_k. Wir haben also bewiesen: symmetrische Endomorphismen eines Euklidischen Vektorraums sind diagonalisierbar, und zwar in einer Orthonormalbasis. Für symmetrische Matrizen $B \in \mathbb{R}^{n \times n}$

bedeutet das äquivalent: es gibt eine Matrix $T \in \mathbb{O}(n)$, so dass $T^{-1}BT$ eine Diagonalmatrix ist; dabei sind die Spaltenvektoren von T gerade die Eigenvektoren $v_k \in \mathbb{R}^n$ von B. Der Beweis durch Minimieren der quadratischen Form verwendet nicht das charakteristische Polynom, und kann auf gewisse symmetrische Operatoren in unendlichdimensionalen Hilberträumen verallgemeinert werden.

In den Koordinaten $x = x_1v_1 + \ldots x_nv_n$ bezüglich der Eigenvektorbasis hat die quadratische Form die einfache Darstellung

$$b(x,x) = \sum_{i=1}^{n} \lambda_i x_i^2 \quad \text{mit } \lambda_1 \leq \ldots \leq \lambda_n.$$

Für $n = 2$ wollen wir die Mengen $M_c = \{x \in \mathbb{R}^2 : b(x,x) = c\}$ beschreiben; dabei können wir nach Übergang zu $-b$ annehmen, dass $\lambda_2 > 0$ ist. Es ergeben sich drei Fälle:

$\lambda_1 > 0$: Im Nullpunkt hat $b(x,x)$ ein globales, isoliertes Minimum und für $c > 0$ ist M_c eine achsensymmetrische Ellipse mit Scheiteln in $(\pm\sqrt{c/\lambda_1}, 0)$ und $(0, \pm\sqrt{c/\lambda_2})$.

$\lambda_1 = 0$: Auch hier ist im Nullpunkt ein globales Minimum, allerdings ist M_0 die gesamte x_1-Achse; für $c > 0$ besteht M_c aus den parallelen Geraden $x_2 = \pm\sqrt{c/\lambda_2}$.

$\lambda_1 < 0$: M_0 ist Vereinigung der beiden Ursprungsgeraden $x_2 = \pm\sqrt{-\lambda_1/\lambda_2}\, x_1$. Für $c > 0$ ist M_c eine nach oben und unten geöffnete Hyperbel mit Scheiteln $(0, \pm\sqrt{c/\lambda_2})$ und M_0 als Asymptotenlinien. Für $c < 0$ erhalten wir ebenfalls eine Hyperbel mit Asymptoten M_0, die aber nach links und rechts geöffnet ist und die Scheitel $(\pm\sqrt{c/\lambda_1}, 0)$ hat.

Betrachten wir in den drei Fällen die zugehörigen Graphen im $\mathbb{R}^3$, so haben wir für $\lambda_1 > 0$ anschaulich eine Mulde, für $\lambda_1 = 0$ einen Hohlweg und für $\lambda_1 < 0$ einen Sattel. Aus der Mulde wird für $-b$ eine Kuppe. Es ist instruktiv, diese Beschreibung an expliziten Beispielen zu verfizieren und auch ein Bild der zugehörigen Graphen in $\mathbb{R}^3$ anzufertigen. Man nennt einen kritischen Punkt x von f nicht degeneriert, wenn die Bilinearform $D^2f(x)$ nicht entartet ist beziehungsweise äquivalent wenn die Eigenwerte λ_i von $D^2f(x)$ alle ungleich Null sind. In diesem Fall bezeichnet man die Anzahl der negativen Eigenwerte als den Index des kritischen Punkts. Die drei Fälle oben sind also der Reihe nach Index Null, degeneriert und Index Eins.

Definition 2.5 (konvexe Funktion) *Sei $K \subset \mathbb{R}^n$ konvex. Dann heißt $f : K \to \mathbb{R}$ konvex, falls für alle $x, y \in K$ gilt:*

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \quad \textit{für alle } t \in [0,1].$$

f heißt strikt konvex, falls die strikte Ungleichung gilt für $x \neq y$ und $t \in (0,1)$.

Wie man leicht sieht, ist Konvexität von f äquivalent dazu, dass der Epigraph

$$G^+(f) = \{(x,z) \in K \times \mathbb{R} : z \geq f(x)\}$$

eine konvexe Menge im $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$ ist.

Satz 2.4 (Konvexitätskriterien) *Sei $\Omega \subset \mathbb{R}^n$ offen und konvex, und $f \in C^1(\Omega)$. Die folgenden Aussagen sind äquivalent:*

(a) *f ist konvex.*

(b) *$f(y) \geq f(x) + Df(x)(y-x)$ für alle $x, y \in \Omega$.*

(c) *$\big(Df(y) - Df(x)\big)(y-x) \geq 0$ für alle $x, y \in \Omega$.*

Ist sogar $f \in C^2(\Omega)$, so ist außerdem äquivalent:

(d) *$D^2 f(x) \geq 0$ für alle $x \in \Omega$.*

Beweis: Die Aussage wird jeweils auf den eindimensionalen Fall reduziert, indem wir für $x, y \in \Omega$ die Funktion $\varphi(t) = (1-t)f(x) + tf(y) - f\big((1-t)x + ty\big)$ betrachten. Es ist klar, daß f genau dann konvex ist, wenn, für alle $x, y \in \Omega$, $\phi(t) \geq 0$ für alle $t \in (0,1)$ gilt. Die Ableitung von ϕ ist

$$\phi'(t) = f(y) - f(x) - Df\big((1-t)x + ty\big)(y-x).$$

Unter Voraussetzung (a) hat φ in $t = 0$ ein Minimum, daraus folgt (b):

$$0 \leq \varphi'(0) = f(y) - f(x) - Df(x)(y-x).$$

Aussage (c) folgt aus (b) durch Vertauschen von x und y und Addition. Die Implikation (c) $\Rightarrow$ (a) zeigen wir durch Widerspruch. Angenommen $\varphi(t)$ hat in $\tau \in (0,1)$ ein Minimum $\varphi(\tau) < 0$. Für $t_1 < t_2$ gilt nach (c) mit $x(t) = (1-t)x + ty$

$$\varphi'(t_1) - \varphi'(t_2) = \frac{1}{t_2 - t_1}\big(Df(x(t_2)) - Df(x(t_1))\big)\big(x(t_2) - x(t_1)\big) \geq 0.$$

Für $t < \tau$ folgt $\varphi'(t) \geq \varphi'(\tau) = 0$, und hieraus $\varphi(0) \leq \varphi(\tau) < 0$, ein Widerspruch.

Sei nun $f \in C^2(\Omega)$. Mit $h = y - x$ folgt aus (2.2)

$$f(y) = f(x) + Df(x)(y-x) + \int_0^1 (1-t)D^2 f(\varphi(t))(y-x, y-x)\, dt.$$

Dies zeigt die Implikation (d) $\Rightarrow$ (b). Umgekehrt folgt (d) aus (b) mit Satz 2.2, denn die Funktion $g(y) = f(y) - \big(f(x) + Df(x)(y-x)\big)$ hat in x ein Minimum. □

Eine Funktion f mit $f((1-t)x + ty) \geq (1-t)f(x) + tf(y)$ für alle $x, y \in \Omega$, $t \in [0,1]$, heißt [*konkav und es gelten entsprechende Aussagen mit umgekehrten Ungleichungen.*