

Nichtlineare Gleichungen

1. Der Fixpunktsatz von Banach

Ist $T : X \rightarrow X$ eine Abbildung von einer nicht leeren Menge X in sich, so nennen wir $x_0 \in X$ einen *Fixpunkt* von T , falls

$$T(x_0) = x_0$$

gilt.

10.1. BEISPIEL. In den Übungen zur Analysis 1 zeigt man mit Hilfe des Zwischenwertsatzes, daß jede stetige Abbildung $f : [0, 1] \rightarrow [0, 1]$ wenigstens einen Fixpunkt besitzt.

10.2. DEFINITION. Sei (X, d) ein metrischer Raum. Eine Abbildung $T : X \rightarrow X$ heißt *stark kontrahierend*, falls es eine Konstante $K \in [0, 1)$ gibt mit

$$\forall x, y \in X : \quad d(T(x), T(y)) \leq K d(x, y).$$

Wir nennen K dann auch eine *Kontraktionskonstante* der starken Kontraktion T .

Eine Abbildung $T : X \rightarrow X$ ist also genau dann stark kontrahierend, wenn sie Lipschitz-stetig mit einer Lipschitz-Konstanten $K \in [0, 1)$ ist.

10.3. BEISPIEL. Die Abbildung

$$f : (0, 1] \rightarrow (0, 1], \quad x \mapsto \frac{x}{2},$$

ist stark kontrahierend mit Kontraktionskonstante $1/2$ und besitzt keinen Fixpunkt in $(0, 1]$.

Ein wichtiges konstruktives Kriterium für die Existenz eines Fixpunktes ist der folgende Fixpunktsatz von Banach.

10.4. BANACHSCHER FIXPUNKTSATZ. Sei (X, d) ein vollständiger metrischer Raum und sei $T : X \rightarrow X$ eine stark kontrahierende Abbildung mit Kontraktionskonstante $K \in [0, 1)$. Dann besitzt T genau einen Fixpunkt $a \in X$. Für diesen gilt: Ist $x_0 \in X$ beliebig, so ist

$$(10.1) \quad a = \lim_{n \rightarrow \infty} x_n \quad \text{mit } x_{n+1} := T(x_n) \text{ für alle } n \in \mathbb{N}_0.$$

Man hat die *a priori Fehlerabschätzung*

$$(10.2) \quad \forall n \in \mathbb{N} : \quad d(x_n, a) \leq \frac{K^n}{1 - K} d(x_1, x_0)$$

und die *a posteriori Fehlerabschätzungen*

$$(10.3) \quad \forall n \in \mathbb{N} : \quad d(x_{n+1}, a) \leq \frac{K}{1 - K} d(x_{n+1}, x_n)$$

und

$$(10.4) \quad \forall x \in X : \quad d(T(x), a) \leq \frac{K}{1 - K} d(T(x), x).$$

BEWEIS. (a) *Zur Eindeutigkeit*: Sind $a, b \in X$ mit $T(a) = a$ und $T(b) = b$, so folgt nach Voraussetzung

$$d(a, b) = d(T(a), T(b)) \leq Kd(a, b).$$

Wegen $0 \leq K < 1$ ist dies nur möglich, wenn $d(a, b) = 0$ und damit $a = b$ gilt. Zu zeigen bleibt also noch die Existenz eines Fixpunktes.

(b) Sei $x_0 \in X$ beliebig und im folgenden fest und sei $(x_n)_{n=0}^\infty$ die gemäß (10.1) induktiv konstruierte Folge aus X . Durch vollständige Induktion zeigen wir:

$$(10.5) \quad \forall n \in \mathbb{N}_0 : \quad d(x_{n+1}, x_n) \leq K^n d(x_1, x_0).$$

Für $n = 0$ ist dies trivial. Ist $n \in \mathbb{N}_0$ mit $d(x_{n+1}, x_n) \leq K^n d(x_1, x_0)$, so folgt nach Voraussetzung

$$d(x_{n+2}, x_{n+1}) = d(T(x_{n+1}), T(x_n)) \leq Kd(x_{n+1}, x_n) \leq K^{n+1}d(x_1, x_0).$$

(c) Unter Verwendung von (10.5) und der Dreiecksungleichung erhalten wir für alle $n, p \in \mathbb{N}_0$:

$$(10.6) \quad d(x_{n+p}, x_n) \leq \sum_{k=0}^{p-1} d(x_{n+k+1}, x_{n+k}) \leq \sum_{k=0}^{p-1} K^{n+k} d(x_1, x_0) \leq \frac{K^n}{1-K} d(x_1, x_0).$$

Die Folge $(x_n)_{n=0}^\infty$ ist also eine Cauchy-Folge in (X, d) . Da (X, d) ein vollständiger metrischer Raum ist, existiert der Grenzwert

$$a := \lim_{n \rightarrow \infty} x_n \quad \text{in } (X, d).$$

Wegen der Stetigkeit der Abbildung $d(\cdot, x_n) : X \rightarrow \mathbb{R}$ folgt aus (10.6) für $p \rightarrow \infty$ die a priori Abschätzung (10.2).

(d) Wegen $T(x_n) = x_{n+1}$ gilt nach Voraussetzung

$$0 \leq d(T(a), a) \leq d(T(a), x_{n+1}) + d(x_{n+1}, a) \leq d(a, x_n) + d(x_{n+1}, a) \rightarrow 0 \quad \text{für } n \rightarrow \infty$$

und damit $d(T(a), a) = 0$, d.h. $T(a) = a$. Der Punkt a ist also der nach (a) einzige Fixpunkt von T .

(e) Nimmt man x_n als neuen Startwert für die Iteration, so erhält man aus der a priori Abschätzung (10.2)

$$\forall p \in \mathbb{N}_0 : \quad d(x_{n+p}, a) \leq \frac{K^p}{1-K} d(x_{n+1}, x_n).$$

Der Spezialfall $p = 1$ liefert die a posteriori Abschätzung (10.3).

Sei nun $x \in X$ beliebig. Nimmt man als neuen Startwert der Iteration $x_0 := x$, so folgt die Abschätzung (10.4) aus (10.2) mit speziell $n = 1$. $\square$

10.5. BEISPIELE. (a) Die Tangensfunktion besitzt in dem Intervall $(\pi, \frac{3}{2}\pi)$ offensichtlich einen Fixpunkt x_0 . Für alle $x, y \in \mathbb{R}$ gilt jedoch nach dem Mittelwertsatz 5.20 der Differentialrechnung mit einem ξ zwischen x und y :

$$|\tan(x) - \tan(y)| = |\tan'(\xi)| \cdot |x - y| = \frac{|x - y|}{\cos(\xi)^2} \geq |x - y|.$$

Die Tangensfunktion ist daher auf keiner Teilmenge von $\mathbb{R}$ mit mehr als einem Punkt stark kontrahierend. Es gilt jedoch für $x \in [\pi, \frac{3}{2}\pi]$:

$$\tan(x) = x \iff x = \tan(x - \pi) \iff \arctan(x) = x - \pi \iff g(x) = x$$

mit $g(t) := \arctan(t) + \pi$. Wegen $\arctan([0, \infty)) \subset [0, \frac{1}{2}\pi]$ gilt

$$g([\pi, \frac{3}{2}\pi]) \subseteq [\pi, \frac{3}{2}\pi].$$

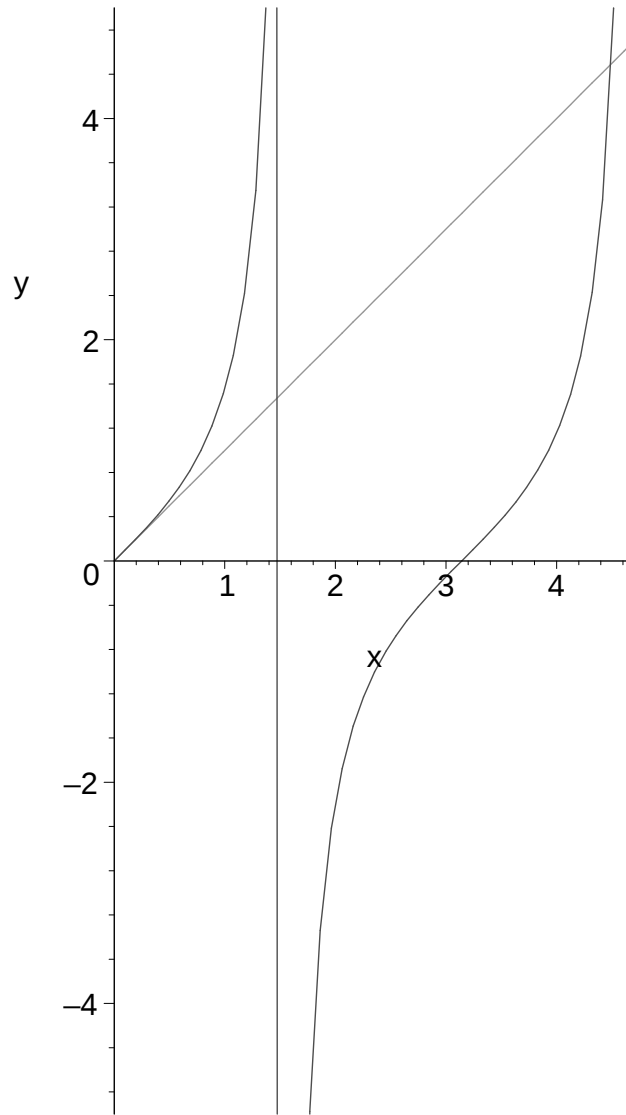


ABBILDUNG 1. Die Tangensfunktion im Bereich $[0, 3\pi/2]$.

Da $[\pi, \frac{3}{2}\pi]$ eine abgeschlossene Teilmenge des vollständigen metrischen Raums $(\mathbb{R}, d_{|\cdot|})$ ist, ist $([\pi, \frac{3}{2}\pi], d)$ ein vollständiger metrischer Raum, wenn wir mit d die von der euklidischen Metrik $d_{|\cdot|}$ auf $[\pi, \frac{3}{2}\pi] \times [\pi, \frac{3}{2}\pi]$ eingeschränkte Metrik bezeichnen. Für alle $x, y \in [\pi, \frac{3}{2}\pi]$ gibt es nach dem Mittelwertsatz ein ξ zwischen x und y mit

$$|g(x) - g(y)| = |g'(\xi)| \cdot |x - y| = \frac{1}{1 + \xi^2} |x - y| \leq \frac{1}{1 + \pi^2} |x - y|.$$

g ist also eine stark kontrahierende Abbildung von $[\pi, \frac{3}{2}\pi]$ in sich mit einer Kontraktionskonstante $K_0 \leq \frac{1}{1 + \pi^2} < K := 0.092 < 1$ und hat nach dem Banachschen Fixpunktsatz genau einen Fixpunkt a in $[\pi, \frac{3}{2}\pi]$, der dann auch der einzige Fixpunkt der Tangensfunktion in $[\pi, \frac{3}{2}\pi]$ sein muß. Mit dem im Banachschen Fixpunktsatz angegebenen iterativen Verfahren können wir a näherungsweise berechnen, wobei wir für die Näherungen Fehlerabschranken angeben können. Wir verwenden als Startwert den Wert $x_0 := 4$ und erhalten in den ersten Iterationsschritten folgende Näherungen und Fehlerabschätzungen:

Fixpunktiteration zur Berechnung des Fixpunktes a von g			
n	x_n	$\frac{K^n}{1-K} x_1 - x_0 $	$\frac{K}{1-K} x_n - x_{n-1} $
1	4.467410318	0.04735875469	0.04735875469
2	4.492175746	0.004357005431	0.002509272440
3	4.493351224	0.0004008444997	0.0001191012952
4	4.493406710	0.00003687769397	0.000005621929515

(b) Sei $I := [0, 1]$. Dann ist $(C(I, \mathbb{R}), d_I)$ mit

$$d_I(f, g) := \|f - g\|_I = \sup_{t \in I} |f(t) - g(t)| \quad \text{für alle } f, g \in C(I, \mathbb{R})$$

nach Satz 9.8 ein vollständiger metrischer Raum. Für alle $f \in C(I, \mathbb{R})$ definieren wir

$$(T(f))(t) := \int_0^1 \cos(f(s)st) ds \quad \text{für } 0 \leq t \leq 1.$$

Nach Satz 7.9 ist $T(f)$ wieder eine stetige Funktion auf I . Seien nun $f, g \in C(I, \mathbb{R})$ beliebig. Für alle $s, t \in I$ gibt es nach dem Mittelwertsatz der Differentialrechnung ein ξ zwischen $f(s)$ und $g(s)$ mit

$$|\cos(f(s)st) - \cos(g(s)st)| = |st \sin(\xi st)| \cdot |f(s) - g(s)| \leq s d_I(f, g).$$

Es folgt also

$$d_I(T(f), T(g)) = \sup_{t \in I} |(T(f))(t) - (T(g))(t)| \leq \int_0^1 s d_I(f, g) ds = \frac{1}{2} d_I(f, g).$$

T ist also eine stark kontrahierende Abbildung von $(C(I, \mathbb{R}), d_I)$ in sich und besitzt daher nach dem Banachschen Fixpunktsatz genau einen Fixpunkt, den man näherungsweise mit Hilfe des im Banachschen Fixpunktsatz angegebenen Iterationsverfahrens berechnen kann.

Wir geben noch einige Varianten des Banachschen Fixpunktsatzes an:

10.6. FOLGERUNG. Sei (X, d) ein vollständiger metrischer Raum und sei $T : X \rightarrow X$ eine Abbildung, so daß für ein $m \in \mathbb{N}$ die m -malige Hintereinanderausführung

$$T^m := \underbrace{T \circ T \circ \cdots \circ T}_{m\text{-mal}}$$

eine stark kontrahierende Abbildung von X in sich ist. Dann besitzt T genau einen Fixpunkt.

BEWEIS. Nach dem Banachschen Fixpunktsatz hat T^m genau einen Fixpunkt a in X . Da jeder Fixpunkt von T auch einer von T^m ist, kann T höchstens einen Fixpunkt besitzen. Wegen $T^m(T(a)) = T^{m+1}(a) = T(T^m(a)) = T(a)$ ist auch $T(a)$ ein Fixpunkt von T^m . Wegen der Eindeutigkeitsaussage im Banachschen Fixpunktsatz muß $a = T(a)$ gelten. $\square$

10.7. SATZ. Sei (X, d) ein vollständiger metrischer Raum, sei $\emptyset \neq A \subseteq X$ und sei $T : A \rightarrow X$ eine Abbildung, für die eine Konstante $K \in [0, 1)$ existiert mit

$$\forall x, y \in A : \quad d(T(x), T(y)) \leq Kd(x, y).$$

Gibt es ein $x_0 \in A$ und ein $r > 0$ mit $B_r(x_0) = \{x \in X ; d(x, x_0) \leq r\} \subseteq A$ und

$$(10.7) \quad d(T(x_0), x_0) \leq (1 - K)r,$$

so gilt: Die durch $x_{n+1} := T(x_n)$ für alle $n \in \mathbb{N}_0$ induktiv definierte Folge konvergiert gegen einen Fixpunkt $a = T(a) \in B_r(x_0)$. a ist der einzige Fixpunkt von T in $B_r(x_0)$ und es gelten die Fehlerabschätzungen (10.2) und (10.3) aus dem Banachschen Fixpunktsatz.

BEWEIS. (a) Die Eindeutigkeitsaussage folgt wie im Beweis zum Banachschen Fixpunktsatz.

(b) Durch vollständige Induktion nach n zeigen wir zunächst, daß für alle $n \in \mathbb{N}_0$ gilt:

$$(10.8) \quad x_n \in B_r(x_0) \quad \text{und} \quad d(x_{n+1}, x_n) \leq K^n d(x_1, x_0).$$

Im Fall $n = 0$ ist dies offensichtlich erfüllt. Sei nun $k \in \mathbb{N}_0$, so daß (10.8) für alle $k = 0, \dots, n$ erfüllt ist. Mit dieser Induktionsvoraussetzung und der Dreiecksungleichung erhalten wir unter Verwendung von (10.7):

$$d(x_{n+1}, x_0) \leq \sum_{j=0}^n d(x_{j+1}, x_j) \leq \sum_{j=0}^n K^j d(x_1, x_0) \leq \frac{d(x_1, x_0)}{1 - K} = \frac{d(T(x_0), x_0)}{1 - K} \leq r$$

Also ist $x_{n+1} \in B_r(x_0) \subseteq A$ und es folgt

$$d(x_{n+2}, x_{n+1}) = d(T(x_{n+1}), T(x_n)) \leq Kd(x_{n+1}, x_n) \leq K^{n+1}d(x_1, x_0).$$

(c) Wie in Teil (c) des Beweises zum Banachschen Fixpunktsatz erhält man die Konvergenz der Folge gegen einen Grenzwert $a \in B_r(x_0)$ (man beachte, daß $B_r(x_0)$ eine vollständige Teilmenge von (X, d) ist und in A liegt) und die a priori Abschätzung (10.2). Auch die Tatsache, daß a ein Fixpunkt von T ist, zeigt man analog wie im Beweis zum Banachschen Fixpunktsatz.

(d) Durch vollständige Induktion nach j zeigt man

$$\forall n, j \in \mathbb{N}_0 : \quad d(x_{n+2+j}, x_{n+1+j}) \leq K^{j+1}d(x_{n+1}, x_n).$$

Für alle $n, p \in \mathbb{N}_0$ erhalten wir also nach (b) und der Dreiecksungleichung

$$d(x_{n+2+p}, x_{n+1}) \leq \sum_{j=0}^p d(x_{n+2+j}, x_{n+1+j}) \leq \sum_{j=0}^p K^{j+1}d(x_{n+1}, x_n) \leq \frac{K}{1 - K}d(x_{n+1}, x_n).$$

Für $k \rightarrow \infty$ folgt hieraus die a posteriori Abschätzung (10.3). □

10.8. SATZ (Lokaler Konvergenzsatz). Sei (X, d) ein vollständiger metrischer Raum und sei $T : X \rightarrow X$ eine Abbildung, die einen Fixpunkt $a \in X$ besitze. Es gebe ein $\varepsilon > 0$ und eine Konstante $K \in [0, 1)$ mit

$$\forall x, y \in U_\varepsilon(a) : \quad d(T(x), T(y)) \leq Kd(x, y).$$

Dann gilt für $0 < r < \varepsilon$: Es ist $T(B_r(a)) \subseteq B_r(a)$ und für jedes $x_0 \in B_r(a)$ konvergiert die durch $x_{n+1} := T(x_n)$ für alle $n \in \mathbb{N}_0$ induktiv definierte Folge gegen a und es gelten die Fehlerabschätzungen (10.2) und (10.3) aus dem Banachschen Fixpunktsatz.

BEWEIS. Für $0 < r < \varepsilon$ gilt $B_r(a) \subset U_\varepsilon(a)$ und für alle $x \in B_r(a)$ ist auch $T(x) \in B_r(a)$ wegen

$$d(T(x), a) = d(T(x), T(a)) \leq Kd(x, a) \leq Kr < r.$$

Als abgeschlossene Teilmenge des vollständigen metrischen Raums (X, d) ist $B_r(a)$ vollständig. Der Banachsche Fixpunktsatz ist also auf $T|_{B_r(a)} : B_r(a) \rightarrow B_r(a)$ und den vollständigen metrischen Raum $(B_r(a), d|_{B_r(a) \times B_r(a)})$ anwendbar und es folgen die Behauptungen $\square$

10.9. DEFINITION. Sei $(x_n)_{n=0}^\infty$ eine gegen einen Punkt $a \in X$ konvergente Folge in einem metrischen Raum (X, d) und sei $p \geq 1$. Die Folge $(x_n)_{n=0}^\infty$ heißt *konvergent von der Ordnung p* , falls eine Konstante $C > 0$ existiert mit

$$\forall n \in \mathbb{N}_0 : \quad d(x_{n+1}, a) \leq Cd(x_n, a)^p.$$

Im Fall $p = 1$ fordern wir zusätzlich $C < 1$ und sprechen von *linearer Konvergenz*.

Die Konvergenz der Iteration des Banachschen Fixpunktsatzes ist also zumindest linear.

2. Das Newton-Verfahren

Wir betrachten die folgende *Problemstellung*: Gegeben sei eine auf einer offenen Menge $\emptyset \neq \Omega \subseteq \mathbb{R}^N$ definierte Abbildung $f : \Omega \rightarrow \Omega$. Es gebe einen Punkt $a \in \Omega$ mit $f(a) = 0$. Wir suchen nach einem geeigneten Verfahren zur näherungsweisen Berechnung von a unter geeigneten, noch zu präzisierenden Voraussetzungen an f .

Zur Motivation des Ansatzes im folgenden Satz betrachten wir zunächst den Fall $N = 1$ und $\Omega = I$, $I \neq \emptyset$ ein offenes Intervall in $\mathbb{R}$. Ist f auf I stetig differenzierbar, so wird sich für x_0 in der "Nähe" von a die Tangente t an den Graphen von f im Punkt $(x_0, f(x_0))$ nicht "wesentlich" von f unterscheiden. Die Nullstelle x_1 von t sollte also eine verbesserte Näherung für a sein. Die Tangente

$$t : \mathbb{R} \rightarrow \mathbb{R}, \quad x \mapsto t(x) = f(x_0) + f'(x_0)(x - x_0)$$

schneidet die x -Achse natürlich nur, wenn $f'(x_0) \neq 0$ ist. Dies trifft wegen der Stetigkeit von f' in einer Umgebung von a zu, falls $f'(a) \neq 0$ erfüllt ist. Man erhält für x_1 :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Man fährt auf diese Weise fort, definiert also induktiv:

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{für alle } n \in \mathbb{N}_0.$$

a ist Fixpunkt von g mit

$$g(x) := x - \frac{f(x)}{f'(x)} \quad \text{für alle } x \in I \text{ mit } f'(x) \neq 0.$$

Unter der zusätzlichen Voraussetzung, daß f' Lipschitz-stetig ist, kann man mit Hilfe des lokalen Konvergenzsatzes 10.8 lineare Konvergenz dieses Verfahrens gegen a zeigen. Setzen wir $f \in C^2(I, \mathbb{R})$ voraus, so können wir sogar quadratische Konvergenz (also Konvergenz der Ordnung $p = 2$) beweisen.

10.10. SATZ (Newton¹-Verfahren). Sei $\Omega \subseteq \mathbb{R}^N$ offen und $f \in C^2(\Omega, \mathbb{R}^N)$. Es gebe ein $a \in \Omega$ mit $f(a) = 0$. Ferner sei die Jacobi-Matrix $J_f(a)$ von f in a invertierbar. Dann gibt es Konstanten $C > 0$ und $r \in (0, 1/C)$, so daß für alle $x_0 \in \Omega$ mit $|x_0 - a| \leq r$ gilt: Die durch

$$(10.9) \quad x_{n+1} := x_n - J_f(x_n)^{-1} f(x_n)$$

induktiv definierte Folge $(x_n)_{n=0}^\infty$ konvergiert gegen a und erfüllt

$$\forall n \in \mathbb{N}_0 : \quad |x_{n+1} - a| \leq C|x_n - a|^2.$$

BEWEIS. Wie in den Übungen gezeigt wurde ist die Menge G der invertierbaren Matrizen aus $M_{N \times N}(\mathbb{R})$ offen in $M_{N \times N}(\mathbb{R})$ und die Matrixinversion ist als Abbildung von G in sich stetig. Es gibt daher Konstanten $\rho, m_1, m_2 > 0$ mit

$$B_\rho(a) = \{x \in \mathbb{R}^N; |x - a| \leq \rho\} \subset \Omega \quad \text{und} \quad J_f(B_\rho(a)) \subset G$$

sowie

$$\sup_{x \in B_\rho(a)} \|J_f(x)\| \leq m_1 \quad \text{und} \quad \sup_{x \in B_\rho(a)} \|J_f(x)^{-1}\| \leq m_2.$$

Man beachte hierzu, daß $B_\rho(a)$ eine kompakte Teilmenge des $\mathbb{R}^N$ ist und stetige Funktionen auf kompakten Teilmengen des $\mathbb{R}^N$ beschränkt sind. Da f zweimal stetig partiell differenzierbar ist, ist $J_f \in C^1(\Omega, M_{N \times N}(\mathbb{R}))$ und daher nach dem Mittelwertsatz der Differentialrechnung Lipschitz-stetig auf $B_\rho(a)$. Es gibt also ein $L > 0$ mit

$$\forall x, y \in B_\rho(a) : \quad \|J_f(x) - J_f(y)\| \leq \|J_f(x) - J_f(y)\|_2 \leq L|x - y|.$$

Nach dem Satz von Taylor 7.38 gilt für alle $x \in B_\rho(a)$:

$$f(x) = f(a) + J_f(a)(x - a) + R_1(x, a) = J_f(a)(x - a) + R_1(x, a)$$

mit

$$\begin{aligned} |R_1(x, a)| &= \left| 2 \sum_{|\alpha|=2} \frac{(x-a)^\alpha}{\alpha!} \int_0^1 (1-s)(D^\alpha f)(a + s(x-a)) ds \right| \\ &\leq 2 \sum_{|\alpha|=2} \frac{|x-a|^2}{\alpha!} \int_0^1 (1-s) |(D^\alpha f)(a + s(x-a))| ds \\ (10.10) \quad &\leq \max_{|\alpha|=2} \max_{u \in B_\rho(a)} |(D^\alpha f)(u)| |x-a|^2 \sum_{|\alpha|=2} \frac{2!}{\alpha!} \int_0^1 (1-s) ds \\ &\leq C|x-a|^2, \end{aligned}$$

wobei

$$C := \frac{N^2}{2} \max_{|\alpha|=2} \max_{u \in B_\rho(a)} |(D^\alpha f)(u)| < \infty,$$

da die zweiten partiellen Ableitungen von f stetig und somit auf $B_\rho(a)$ beschränkt sind. Offensichtlich ist a Fixpunkt der in einer Umgebung U von a definierten Abbildung

$$g : U \rightarrow \mathbb{R}^N, \quad x \mapsto g(x) := x - J_f(x)^{-1} f(x).$$

Wir setzen $K := m_2 C + m_1 m_2^2 L$ und fixieren $r < \min\{\rho, 1/K\}$. Für alle $x \in B_r(a) \subset B_\rho(a)$ können wir nun abschätzen:

$$\begin{aligned} |g(x) - a| &= |g(x) - g(a)| = |x - J_f(x)^{-1} f(x) - a + J_f(a)^{-1} f(a)| \\ &\leq \underbrace{|J_f(a)^{-1}(J_f(a)(x-a) + f(a) - f(x))|}_{=: A(x)} + \underbrace{|(J_f(a)^{-1} - J_f(x)^{-1})f(x)|}_{=: B(x)} \end{aligned}$$

¹SIR ISAAC NEWTON (4.1.1643–31.3.1727).

Für den ersten Summanden $A(x)$ erhalten wir nach (10.10):

$$A(x) \leq \|J_f(a)^{-1}\| \cdot |J_f(a)(x-a) + f(a) - f(x)| \leq m_2 C |x-a|^2$$

und für den zweiten Summanden folgt mit Hilfe des Mittelwertsatzes 7.30 der Differentialrechnung in mehreren Veränderlichen wegen $f(a) = 0$ und der Lipschitz-Stetigkeit von $\xi \mapsto J_f(\xi)$ auf $B_\rho(a)$:

$$\begin{aligned} B(x) &\leq \|J_f(a)^{-1} - J_f(x)^{-1}\| \cdot |f(x)| = \\ &= \|J_f(x)^{-1} J_f(x) J_f(a)^{-1} - J_f(x)^{-1} J_f(a) J_f(a)^{-1}\| \cdot |f(x) - f(a)| \leq \\ &\leq \|J_f(x)^{-1}\| \cdot \|J_f(x) - J_f(a)\| \cdot \|J_f(a)^{-1}\| \cdot |x-a| \sup_{0 \leq t \leq 1} \|J_f(a + t(x-a))\| \leq \\ &\leq m_2^2 L m_1 |x-a|^2. \end{aligned}$$

Also gilt:

$$|g(x) - a| \leq K|x-a|^2 \leq Kr|x-a| \leq qr$$

mit $q := rK < 1$. Insbesondere ist $g(x) \in B_r(a)$ für alle $x \in B_r(a)$. Durch vollständige Induktion sieht man nun: Die durch (10.9) induktiv definierte Folge nimmt für $x_0 \in B_r(a)$ ihre Werte in $B_r(a)$ an und erfüllt

$$(10.11) \quad |x_{n+1} - a| = |g(x_n) - a| \leq K|x_n - a|^2 \leq q|x_n - a|.$$

Durch vollständige Induktion folgt

$$|x_n - a| \leq q^n |x_0 - a| \rightarrow 0 \quad \text{für } n \rightarrow \infty.$$

Wegen (10.11) ist die Konvergenz quadratisch. $\square$

10.11. BEMERKUNG. Damit man nicht in jedem Schritt eine aufwendige Matrixinversion durchführen muß, geht man in der Praxis wie folgt vor: Multiplizieren wir (10.9) von links mit $J_f(x)$, so erhalten wir

$$J_f(x_n)x_{n+1} = J_f(x_n)x_n - f(x_n)$$

also

$$(10.12) \quad J_f(x_n)z_n = -f(x_n)$$

mit $z_n = x_{n+1} - x_n$. Man berechnet nun z_n als Lösung des linearen Gleichungssystems (10.12) und erhält x_{n+1} als $x_{n+1} = z_n + x_n$.

3. Implizite Funktionen und Umkehrfunktionen

10.12. BEISPIELE. (a) Lösen wir die Gleichung

$$(10.13) \quad 2x^2 + 3y = 0$$

nach y auf, so erhalten wir $y = -\frac{2}{3}x^2$ für alle $x \in \mathbb{R}$. Man sagt, die hierdurch gegebene Funktion

$$g: \mathbb{R} \rightarrow \mathbb{R}, \quad x \mapsto -\frac{2}{3}x^2$$

ist implizit durch (10.13) definiert. Wollen wir hingegen (10.13) nach x auflösen, so erhalten wir nur dann reelle Lösungen, wenn $y \leq 0$ ist. In diesem Fall erhalten wir zwei Lösungen

$$x_1(y) = \sqrt{\frac{3}{2}y} \quad \text{und} \quad x_2(y) = -\sqrt{\frac{3}{2}y}.$$

(b) Die Gleichung $x^2 + y^2 + 1 = 0$ hat für keine Werte von x eine reelle Lösung $y(x)$ und für keine Werte von y eine reelle Lösung $x(y)$.

Wir wollen in diesem Abschnitt das folgende Problem behandeln: Gegeben sei eine Funktion f in zwei Veränderlichen x und y . Unter welchen Bedingungen läßt sich die Gleichung

$$(10.14) \quad f(x, y) = 0$$

nach y auflösen, d.h. läßt sich eine Funktion $x \mapsto g(x)$ finden mit

$$f(x, g(x)) \equiv 0?$$

Im Falle der Auflösbarkeit nennt man g die durch (10.14) bestimmte *implizite Funktion*. Was läßt sich hierbei über die Differenzierbarkeitseigenschaften von g sagen, wenn f stetig partiell differenzierbar ist?

Allgemeiner werden wir $\mathbb{R}^N$ -wertige Funktionen in Veränderlichen $x \in \mathbb{R}^M$, $y \in \mathbb{R}^N$ betrachten.

10.13. SATZ (über implizite Funktionen). Sei $\Omega \subseteq \mathbb{R}^{M+N} = \mathbb{R}^M \times \mathbb{R}^N$ offen und sei $f \in C^1(\Omega, \mathbb{R}^N)$. Für einen Punkt $(x_0, y_0) \in \Omega$ gelte

$$f(x_0, y_0) = 0.$$

Ferner sei die Matrix

$$A := \frac{\partial f}{\partial y}(x_0, y_0) := \left(\frac{\partial f_j}{\partial y_k}(x_0, y_0) \right)_{\substack{j=1, \dots, N \\ k=1, \dots, N}}$$

invertierbar. Dann gilt:

- (a) Es gibt ein $\delta_1 > 0$ und ein $\delta_2 > 0$, so daß zu jedem $x \in U_1 := U_{\delta_1}(x_0)$ genau ein $g(x) \in U_2 := U_{\delta_2}(y_0)$ existiert mit

$$f(x, g(x)) = 0.$$

- (b) Die durch (a) definierte Funktion $g : U_1 \rightarrow U_2 \subset \mathbb{R}^N$ ist stetig partiell differenzierbar und für alle $x \in U_1$ gilt

$$(10.15) \quad J_g(x) = - \left(\frac{\partial f}{\partial y}(x, y) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x)).$$

- (c) Ist $f \in C^k(\Omega, \mathbb{R}^N)$ für ein $k \in \mathbb{N} \cup \{\infty\}$, so gilt auch $g \in C^k(U_1, \mathbb{R}^N)$.

BEWEIS. Da f nach Voraussetzung stetig partiell differenzierbar ist, gibt es $\delta'_1, \delta_2 > 0$ mit $B_{\delta'_1}(x_0) \times B_{\delta_2}(y_0) \subset \Omega$ und

$$(10.16) \quad \sup \left\{ \left\| \frac{\partial f}{\partial y}(x_0, y_0) - \frac{\partial f}{\partial y}(x, y) \right\| ; |x - x_0| \leq \delta'_1, |y - y_0| \leq \delta_2 \right\} < \frac{1}{2\|A^{-1}\|}.$$

Da $A = \frac{\partial f}{\partial y}(x_0, y_0)$ invertierbar ist und die Menge der invertierbaren Matrizen in $M_{N \times N}(\mathbb{R})$ offen ist, können wir wegen der Stetigkeit von $\frac{\partial f}{\partial y}$ durch eventuelles Verkleinern von δ'_1 und δ_2 erreichen, daß gilt:

$$(10.17) \quad \forall u \in B_{\delta'_1}(x_0), v \in B_{\delta_2}(y_0) : \quad \frac{\partial f}{\partial y}(u, v) \text{ ist in } M_{N \times N}(\mathbb{R}) \text{ invertierbar.}$$

Wegen der Stetigkeit von f in (x_0, y_0) und wegen $f(x_0, y_0) = 0$ gibt es ein $\delta_1 \in (0, \delta'_1)$ mit

$$(10.18) \quad \forall x \in B_{\delta_1}(x_0) : \quad |f(x, y_0)| < \frac{\delta_2}{3\|A^{-1}\|}.$$

Angeregt durch das Newtonverfahren betrachten wir für alle $x \in B_{\delta_1}(x_0)$ die Abbildung

$$T_x : B_{\delta_2}(y_0) \rightarrow \mathbb{R}^N, \quad y \mapsto T_x(y) := y - A^{-1}f(x, y).$$

Für die Jacobi-Matrix $J_{T_x}(y)$ von T_x in $y \in B_{\delta_2}(y_0)$ erhält man unter Verwendung von (10.16) die Abschätzung (wobei E_N die Einheitsmatrix in $M_{N \times N}(\mathbb{R})$ bezeichne):

$$\begin{aligned} \|J_{T_x}(y)\| &= \left\| E_N - A^{-1} \frac{\partial f}{\partial y}(x, y) \right\| = \left\| A^{-1} \left(\frac{\partial f}{\partial y}(x_0, y_0) - \frac{\partial f}{\partial y}(x, y) \right) \right\| \\ &\leq \|A^{-1}\| \cdot \left\| \frac{\partial f}{\partial y}(x_0, y_0) - \frac{\partial f}{\partial y}(x, y) \right\| < \frac{1}{2} \end{aligned}$$

Unter Verwendung dieser Ungleichung folgt für alle $x \in B_{\delta_1}(x_0)$ und alle $y_1, y_2 \in B_{\delta_2}(y_0)$ folgt nach dem Mittelwertsatz 7.30 der Differentialrechnung:

$$(10.19) \quad |T_x(y_1) - T_x(y_2)| \leq \sup_{t \in [0,1]} \|J_{T_x}(y_1 + t(y_2 - y_1))\| \cdot |y_1 - y_2| \leq \frac{1}{2} |y_1 - y_2|.$$

Für alle $y \in B_{\delta_2}(y_0)$ und alle $x \in B_{\delta_1}(x_0)$ erhalten wir mit (10.18) und (10.19)

$$\begin{aligned} (10.20) \quad |T_x(y) - y_0| &= |T_x(y) - T_x(y_0) - A^{-1}f(x, y_0)| \\ &\leq |T_x(y) - T_x(y_0)| + \|A^{-1}\| \cdot |f(x, y_0)| \\ &\leq \frac{1}{2} |y - y_0| + \frac{\delta_2}{3} \leq \frac{5}{6} \delta_2. \end{aligned}$$

Wir halten fest: Für alle $x \in B_{\delta_1}(x_0)$ ist T_x eine stark kontrahierende Abbildung der abgeschlossenen und daher vollständigen Teilmenge $B_{\delta_2}(y_0)$ des $\mathbb{R}^N$ in sich mit Kontraktionskonstante $K \leq 1/2$. Nach dem Banachschen Fixpunktsatz 10.4 gibt es daher genau ein $g(x) \in B_{\delta_2}(y_0)$ mit $T_x(g(x)) = g(x)$, d.h. mit $A^{-1}f(x, g(x)) = 0$, d.h. mit $f(x, g(x)) = 0$. Insbesondere ist $g(x_0) = y_0$. Insbesondere folgt wegen $T_x(g(x)) = g(x)$ aus (10.20) auch $g(x) \in B_{\frac{5}{6}\delta_2}(y_0) \subset U_{\delta_2}(y_0)$ für alle $x \in B_{\delta_1}(x_0)$. Damit ist (a) bewiesen.

Definieren wir induktiv die Funktionen $g_n : B_{\delta_1}(x_0) \rightarrow \mathbb{R}^N$, $n \in \mathbb{N}_0$, durch

$$g_0 \equiv y_0 \quad \text{und} \quad g_{n+1}(x) := T_x(g_n(x)) = g_n(x) + A^{-1}f(x, g_n(x)) \quad \text{für alle } x \in B_{\delta_1}(x_0), n \in \mathbb{N}_0,$$

so folgt durch vollständige Induktion die Stetigkeit von g_n für alle $n \in \mathbb{N}_0$. Aus der a priori Abschätzung (10.2) im Banachschen Fixpunktsatz folgt

$$\forall x \in B_{\delta_1}(x_0) : \quad |g_n(x) - g(x)| \leq \frac{\left(\frac{1}{2}\right)^n}{1 - \frac{1}{2}} |g_1(x) - g_0(x)|$$

und somit

$$\|g_n - g\|_{B_{\delta_1}(x_0)} \leq 2^{1-n} \|g_1 - g_0\|_{B_{\delta_1}(x_0)} \rightarrow 0 \quad \text{für } n \rightarrow \infty.$$

Die Folge $(g_n)_{n=0}^\infty$ ist also eine auf $B_{\delta_1}(x_0)$ gleichmäßig gegen g konvergente Folge von auf $B_{\delta_1}(x_0)$ stetigen Funktionen. Nach Satz 9.4 ist also auch die Funktion g auf $B_{\delta_1}(x_0)$ stetig. Indem wir δ_1 notfalls verkleinern, können wir wegen der Stetigkeit von g in x_0 erreichen, daß gilt

$$(10.21) \quad \forall x \in U_{\delta_1}(x_0) : \quad g(x) \in U_{\delta_2}(g(x_0)) = U_{\delta_2}(y_0).$$

Für alle $(x, y) \in \Omega$ und alle $\xi \in \mathbb{R}^M, \eta \in \mathbb{R}^N$ gilt

$$J_f(x, y) \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x, y) & \dots & \frac{\partial f_1}{\partial x_M}(x, y) & \frac{\partial f_1}{\partial y_1}(x, y) & \dots & \frac{\partial f_1}{\partial y_N}(x, y) \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1}(x, y) & \dots & \frac{\partial f_N}{\partial x_M}(x, y) & \frac{\partial f_N}{\partial y_1}(x, y) & \dots & \frac{\partial f_N}{\partial y_N}(x, y) \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_M \\ \eta_1 \\ \vdots \\ \eta_N \end{pmatrix}$$

$$= \frac{\partial f}{\partial x}(x, y)\xi + \frac{\partial f}{\partial y}(x, y)\eta.$$

Da f auf Ω stetig partiell und somit auch total differenzierbar ist, folgt für alle $x \in U_{\delta_1}(x_0)$, $y \in U_{\delta_2}(y_0)$ und alle $\xi \in \mathbb{R}^M, \eta \in \mathbb{R}^N$ mit $x + \xi \in U_{\delta_1}(x_0)$ $y + \eta \in U_{\delta_2}(y_0)$:

$$(10.22) \quad f(x + \xi, y + \eta) = f(x, y) + \frac{\partial f}{\partial x}(x, y)\xi + \frac{\partial f}{\partial y}(x, y)\eta + \varphi(\xi, \eta)$$

mit

$$(10.23) \quad \frac{|\varphi(\xi, \eta)|}{|(\xi, \eta)|} \rightarrow 0 \quad \text{für } |(\xi, \eta)| \rightarrow 0.$$

Speziell mit $y = g(x), \eta = g(x + \xi) - g(x)$ erhalten wir für alle $x \in B_{\delta_1}(x_0)$ wegen $f(x, g(x)) = 0 = f(x + \xi, g(x + \xi))$

$$0 = \frac{\partial f}{\partial x}(x, g(x))\xi + \frac{\partial f}{\partial y}(x, g(x))(g(x + \xi) - g(x)) + \varphi(\xi, g(x + \xi) - g(x)).$$

Da $\frac{\partial f}{\partial y}(x, g(x))$ invertierbar ist, folgt hieraus

$$(10.24) \quad g(x + \xi) = g(x) - \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x))\xi + \psi(\xi)$$

mit

$$(10.25) \quad \psi(\xi) = - \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \varphi(\xi, g(x + \xi) - g(x)).$$

Für alle $0 \neq \xi \in \mathbb{R}^M$ mit $x + \xi \in U_{\delta_1}(x_0)$ hat man dann

$$(10.26) \quad \frac{|\psi(\xi)|}{|\xi|} \leq \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \right\| \cdot \frac{|\varphi(\xi, g(x + \xi) - g(x))|}{|(\xi, g(x + \xi) - g(x))|} \cdot \frac{|(\xi, g(x + \xi) - g(x))|}{|\xi|}$$

Wegen der Stetigkeit von g gilt $g(x + \xi) \rightarrow g(x)$ für $|\xi| \rightarrow 0$ und somit auch (nach (10.23))

$$\frac{|\varphi(\xi, g(x + \xi) - g(x))|}{|(\xi, g(x + \xi) - g(x))|} \rightarrow 0 \quad \text{für } |\xi| \rightarrow 0.$$

Da der erste Faktor in (10.26) von ξ unabhängig ist, folgt

$$(10.27) \quad \frac{|\psi(\xi)|}{|\xi|} \rightarrow 0 \quad \text{für } |\xi| \rightarrow 0,$$

wenn wir zeigen können, daß es Konstanten $\varepsilon > 0$ und $C > 0$ gibt, so daß $U_\varepsilon(x) \subset U_{\delta_1}(x_0)$ gilt und

$$(10.28) \quad \forall \xi \in U_\varepsilon(x) \setminus \{0\} : \frac{|(\xi, g(x+\xi) - g(x))|}{|\xi|} \leq \frac{|\xi|}{|\xi|} + \frac{|g(x+\xi) - g(x)|}{|\xi|} = \\ = 1 + \frac{|g(x+\xi) - g(x)|}{|\xi|} \leq 1 + C.$$

Wegen (10.23) gibt es zu $K := \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \right\|^{-1}$ ein $\varepsilon_1 > 0$ mit $U_{\varepsilon_1}(x) \subset U_{\delta_1}(x_0)$, $U_{\varepsilon_1}(g(x)) \subset U_{\delta_2}(y_0)$ und

$$(10.29) \quad \forall \xi \in U_{\varepsilon_1}(0), \eta \in U_{\varepsilon_1}(0) : |\varphi(\xi, \eta)| \leq \frac{K}{2}|(\xi, \eta)| \leq \frac{K}{2}(|\xi| + |\eta|).$$

Wegen der Stetigkeit von g in x gibt es ein $\varepsilon > 0$ mit $\varepsilon < \varepsilon_1$, so daß für alle $\xi \in U_\varepsilon(0)$ gilt $|g(x+\xi) - g(x)| < \varepsilon_1$. Mit (10.24), (10.25) und (10.29) folgt für $0 < |\xi| < \varepsilon$:

$$\begin{aligned} |g(x+\xi) - g(x)| &\leq \left| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x)) \xi \right| + \\ &\quad + \left| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \varphi(\xi, g(x+\xi) - g(x)) \right| \\ &\leq \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x)) \right\| \cdot |\xi| + \\ &\quad + \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \right\| \cdot \frac{K}{2} \cdot (|\xi| + |g(x+\xi) - g(x)|) \\ &\leq \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x)) \right\| \cdot |\xi| + \frac{|\xi|}{2} + \frac{1}{2}|g(x+\xi) - g(x)| \end{aligned}$$

Hieraus erhalten wir für $0 < |\xi| < \varepsilon$:

$$\frac{|g(x+\xi) - g(x)|}{|\xi|} \leq C := 2 \left\| \left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \frac{\partial f}{\partial x}(x, g(x)) \right\| + 1.$$

Also gilt (10.28) und somit auch (10.27). Zusammen mit (10.24) folgt die totale Differenzierbarkeit von g in allen $x \in U_{\delta_1}(x_0)$ und (10.15). Mit Hilfe der Cramerschen Regel folgt hieraus wegen der Stetigkeit von g die Stetigkeit von J_g auf $U_{\delta_1}(x_0)$. Also ist g auf $U_{\delta_1}(x_0)$ stetig partiell differenzierbar. Durch vollständige Induktion nach k schließt man mit Hilfe der Cramerschen Regel und (10.15): Ist $f \in C^k(\Omega, \mathbb{R}^N)$, so ist g auf $U_{\delta_1}(x_0)$ ebenfalls k mal stetig partiell differenzierbar. $\square$

10.14. SATZ (von der Umkehrfunktion). Sei $G \subseteq \mathbb{R}^N$ offen, $k \in \mathbb{N} \cup \{\infty\}$ und sei $f \in C^k(G, \mathbb{R}^N)$. Ist $a \in G$ und ist $J_f(a)$ invertierbar, so gibt es offene Mengen $U, V \subseteq \mathbb{R}^N$ mit $a \in U \subseteq G$ und $f(a) \in V$, so daß $f|_U : U \rightarrow V$ bijektiv ist und $(f|_U)^{-1} : V \rightarrow U$ auf V k -mal stetig partiell differenzierbar ist. Es gilt dann

$$(10.30) \quad J_{(f|_U)^{-1}}(f(a)) = (J_f(a))^{-1}.$$

BEWEIS. Die Menge $\Omega := \mathbb{R}^N \times G$ ist offen in $\mathbb{R}^{2N} = \mathbb{R}^N \times \mathbb{R}^N$ und die Abbildung

$$F : \Omega \rightarrow \mathbb{R}^N, \quad (x, y) \mapsto F(x, y) := x - f(y),$$

ist auf Ω k -mal stetig partiell differenzierbar. Es ist $F(f(a), a) = 0$ und $\frac{\partial F}{\partial y}(f(a), a) = -J_f(a)$ nach Voraussetzung invertierbar. Nach dem Satz über implizite Funktionen gibt es also $\delta_1 > 0$ und $\delta_2 > 0$, so daß es zu jedem $x \in U_{\delta_1}(f(a))$ genau ein $g(x) \in U_{\delta_2}(a)$ gibt mit $F(x, g(x)) = 0$, d.h. mit $x = f(g(x))$ und die hierdurch definierte Funktion

$g : U_{\delta_1}(f(a)) \rightarrow \mathbb{R}^N$ ist k -mal stetig partiell differenzierbar. Wegen der Stetigkeit von f ist die Menge $U := U_{\delta_2}(a) \cap f^{-1}(U_{\delta_1}(f(a)))$ eine offene Umgebung von a und $f|_U : U \rightarrow U_{\delta_1}(f(a))$ ist bijektiv mit $(f|_U)^{-1} = g$.

(10.30) folgt mit Hilfe von (10.15) oder auch durch Anwendung der Kettenregel auf $x = f(g(x))$ für alle $x \in U_{\delta_1}(a)$ $\square$

10.15. BEISPIEL. Die Abbildung

$$f : [0, \infty) \times \mathbb{R} \rightarrow \mathbb{R}^2, \quad (r, \varphi) \mapsto f(r, \varphi) := (r \cos(\varphi), r \sin(\varphi))$$

ist surjektiv und beliebig oft stetig partiell differenzierbar mit

$$J_f(r, \varphi) = \begin{pmatrix} \cos(\varphi) & -r \sin(\varphi) \\ \sin(\varphi) & r \cos(\varphi) \end{pmatrix} \quad \text{und} \quad \det(J_f(r, \varphi)) = r.$$

Wegen $f(r, \varphi) = f(r, \varphi + 2\pi)$ ist f nicht injektiv, hat aber nach dem Satz von der Umkehrfunktion für alle $r_0 > 0$ und $\varphi_0 \in \mathbb{R}$ in einer Umgebung V eines jeden Punktes eine Umkehrfunktion $g : V \rightarrow U$ von V auf eine Umgebung U von (r_0, φ_0) . Mit $(x, y) := f(r, \varphi)$ folgt $r = \sqrt{x^2 + y^2}$ und

$$J_g(x, y) = J_f(r, \varphi)^{-1} = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\frac{\sin(\varphi)}{r} & \frac{\cos(\varphi)}{r} \end{pmatrix} = \frac{1}{x^2 + y^2} \begin{pmatrix} x\sqrt{x^2 + y^2} & y\sqrt{x^2 + y^2} \\ -y & x \end{pmatrix}.$$

Im folgenden Abschnitt geben wir eine wichtige Anwendung des Satzes über implizite Funktionen auf die Untersuchung von lokalen Extremstellen bei Vorliegen von Nebenbedingungen.

4. Extrema mit Nebenbedingungen

10.16. SATZ (Multiplikatorenregel von Lagrange). Sei $\Omega \subseteq \mathbb{R}^N$ offen und seien $f \in C^1(\Omega, \mathbb{R})$ und $g = (g_1, \dots, g_L) \in C^1(\Omega, \mathbb{R}^L)$ mit $L < N$. Sei $x_0 \in \Omega$ mit $g(x_0) = 0$ und $\text{rang} F_g(x_0) = L$ (also mit maximal möglichem Rang). Ist x_0 lokale Extremstelle von f auf $\{x \in \Omega; g(x) = 0\}$, so gibt es $\lambda_1, \dots, \lambda_L \in \mathbb{R}$ mit

$$(10.31) \quad \frac{\partial f}{\partial x_j}(x_0) + \sum_{k=1}^L \lambda_k \frac{\partial g_k}{\partial x_j}(x_0) = 0 \quad \text{für } j = 1, \dots, N,$$

d.h. mit

$$(\nabla f)(x_0) + J_g(x_0) \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_L \end{pmatrix} = 0.$$

BEWEIS. Wegen $\text{rang} J_g(x_0) = L$ gibt es L der Variablen $x_1, \dots, x_N$ etwa (nach eventueller Umnummerierung) $x_1, \dots, x_L$, mit

$$(10.32) \quad \det \left(\frac{\partial g_j}{\partial x_k}(x_0) \right)_{\substack{j=1, \dots, L \\ k=1, \dots, L}} \neq 0.$$

Schreiben wir $x = (x_j)_{j=1}^N$ in der Form $x = \begin{pmatrix} u \\ v \end{pmatrix}$ mit $u = (x_j)_{j=1}^L \in \mathbb{R}^L$ und $v = (x_j)_{j=L+1}^N \in \mathbb{R}^{N-L}$ und entsprechend $x_0 = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}$, so gibt es nach dem Satz über implizite Funktionen

Umgebungen U von u_0 und V von v_0 sowie eine Funktion $\varphi \in C^1(V, \mathbb{R}^L)$ mit $\varphi(v_0) = u_0$ und $g(\varphi(v), v) \equiv 0$ auf V sowie mit

$$(10.33) \quad J_\varphi(v) = -\left(\frac{\partial g}{\partial u}(x_0)\right)^{-1} \frac{\partial g}{\partial v}(x_0)$$

Die durch $F(v) := f(\varphi(v), v)$ für alle $v \in V$ definierte Funktion $F : V \rightarrow \mathbb{R}$ hat nach Voraussetzung in v_0 eine lokale Extremstelle. Es folgt daher mit der Kettenregel und (10.33):

$$(10.34) \quad \begin{aligned} 0 &= (\nabla F)(v_0)^t = J_F(v_0) = \frac{\partial f}{\partial u}(u_0, v_0) J_\varphi(v_0) + \frac{\partial f}{\partial v}(u_0, v_0) \\ &= -\frac{\partial f}{\partial u}(x_0) \left(\frac{\partial g}{\partial u}(x_0)\right)^{-1} \frac{\partial g}{\partial v}(x_0) + \frac{\partial f}{\partial v}(x_0) \end{aligned}$$

Da wegen (10.32) die Zeilen von $\frac{\partial g}{\partial u}(x_0)$ linear unabhängig sind, gibt es $\lambda_1, \dots, \lambda_L \in \mathbb{R}$ mit

$$(10.35) \quad -\frac{\partial f}{\partial u}(x_0) = \sum_{k=1}^L \lambda_k \frac{\partial g_k}{\partial u}(x_0) = (\lambda_1, \dots, \lambda_L) \frac{\partial g}{\partial u}(x_0),$$

also mit

$$\frac{\partial f}{\partial u}(x_0) + \sum_{k=1}^L \lambda_k \frac{\partial g_k}{\partial u}(x_0) = 0.$$

Dies liefert die ersten L Gleichungen in (10.31). Setzen wir nun noch (10.35) in (10.34) ein, so folgt

$$0 = (\lambda_1, \dots, \lambda_L) \frac{\partial g}{\partial u}(x_0) \left(\frac{\partial g}{\partial u}(x_0)\right)^{-1} \frac{\partial g}{\partial v}(x_0) + \frac{\partial f}{\partial v}(x_0)$$

und damit

$$\frac{\partial f}{\partial v}(x_0) + (\lambda_1, \dots, \lambda_L) \frac{\partial g}{\partial v}(x_0) = 0,$$

d.h. es gilt auch für $j = L+1, \dots, N$

$$\frac{\partial f}{\partial x_j}(x_0) + \sum_{k=1}^L \lambda_k \frac{\partial g_k}{\partial x_j}(x_0) = 0,$$

Damit ist gezeigt, daß (10.31) erfüllt ist. $\square$

Es gilt also: Ist $g : \Omega \rightarrow \mathbb{R}^L$ stetig differenzierbar mit $\text{rang } J_g(x) = L$ für alle $x \in \Omega$ mit $g(x) = 0$, so können höchstens solche $x_0 \in \Omega$ als lokale Extremstellen von $f \in C^1(\Omega, \mathbb{R})$ unter der Nebenbedingung $g(x_0) = 0$ d.h. mit

$$(10.36) \quad g_j(x_0) = 0 \quad \text{für } j = 1, \dots, L,$$

aufzutreten, für die $\lambda_1, \dots, \lambda_L \in \mathbb{R}$ existieren mit (10.31). Mit den Gleichungssystemen (10.31) und (10.36) stehen uns $N+L$ Gleichungen zur Berechnung der $N+L$ Unbekannten $\lambda_1, \dots, \lambda_N$ und $x_{0,1}, \dots, x_{0,N}$ zur Verfügung.

10.17. BEISPIEL. Gesucht sind alle Extremstellen der offensichtlich stetig differenzierbaren Funktion

$$f : \mathbb{R}^3 \rightarrow \mathbb{R}, \quad (x, y, z) \mapsto f(x, y, z) := xyz$$

unter den Nebenbedingungen

$$(10.37) \quad g_1(x, y, z) := x + y + z - 5 = 0$$

und

$$(10.38) \quad g_2(x, y, z) := xy + yz + zx - 8 = 0.$$

Mit $g := (g_1, g_2)$ gilt $g \in C^1(\mathbb{R}^3, \mathbb{R}^2)$ und

$$J_g(x) = \begin{pmatrix} 1 & 1 & 1 \\ y+z & x+z & x+y \end{pmatrix}.$$

Es folgt

$$\det \begin{pmatrix} \frac{\partial g_1}{\partial x}(x, y, z) & \frac{\partial g_1}{\partial y}(x, y, z) \\ \frac{\partial g_2}{\partial x}(x, y, z) & \frac{\partial g_2}{\partial y}(x, y, z) \end{pmatrix} = x+z - (y+z) = x-y$$

und

$$\det \begin{pmatrix} \frac{\partial g_1}{\partial x}(x, y, z) & \frac{\partial g_1}{\partial z}(x, y, z) \\ \frac{\partial g_2}{\partial x}(x, y, z) & \frac{\partial g_2}{\partial z}(x, y, z) \end{pmatrix} = x+y - (y+z) = x-z.$$

$J_g(x, y, z)$ hat also höchstens in solchen Punkten (x, y, z) einen Rang < 1 , für die $x = y = z$ gilt. Für solche Punkte gilt aber

$$g_1(x, y, z) = 3x - 5 \neq 0 \quad \text{oder} \quad g_2(x, y, z) = 3x^2 - 8 \neq 0.$$

Für alle (x, y, z) , die die Nebenbedingungen (10.37) und (10.38) erfüllen ist also $\text{rang } J_g(x, y, z) = 2$. Damit ist die Rangbedingung in Satz 10.16 erfüllt. Zu lösen sind nun nach Satz 10.16 die Gleichungen (10.37) und (10.38) und

$$(10.39) \quad 0 = \frac{\partial f}{\partial x}(x, y, z) + \lambda_1 \frac{\partial g_1}{\partial x}(x, y, z) + \lambda_2 \frac{\partial g_2}{\partial x}(x, y, z) = yz + \lambda_1 + \lambda_2(y+z)$$

$$(10.40) \quad 0 = \frac{\partial f}{\partial y}(x, y, z) + \lambda_1 \frac{\partial g_1}{\partial y}(x, y, z) + \lambda_2 \frac{\partial g_2}{\partial y}(x, y, z) = xz + \lambda_1 + \lambda_2(x+z)$$

$$(10.41) \quad 0 = \frac{\partial f}{\partial z}(x, y, z) + \lambda_1 \frac{\partial g_1}{\partial z}(x, y, z) + \lambda_2 \frac{\partial g_2}{\partial z}(x, y, z) = xy + \lambda_1 + \lambda_2(x+y)$$

Addition der drei Gleichungen (10.39)–(10.41) liefert unter Verwendung von (10.37) und (10.38):

$$(10.42) \quad 8 + 3\lambda_1 + 10\lambda_2 = 0$$

Subtrahieren wir (10.40) von (10.39) und (10.41), so erhalten wir

$$(10.43) \quad (y-x)(z+\lambda_2) = 0 \quad \text{und} \quad (y-z)(x+\lambda_2) = 0.$$

Hieraus folgt

Fall (a): $x = y$ oder

Fall (b): $z = -\lambda_2$.

In Fall (a) erhalten wir mit (10.37)

$$z = 5 - 2x$$

und hieraus mit (10.38)

$$x^2 + 2x(5 - 2x) - 8 = 0$$

Es ergeben sich die Lösungen (unter Verwendung der rechten Gleichung in (10.43) und von (10.42))

$$x_1 = y_1 = 2, \quad z_1 = 1, \quad \lambda_2^{(1)} = -x_1 = -2 \neq -z_1, \quad \lambda_1^{(1)} = 4.$$

und

$$x_2 = y_2 = \frac{4}{3}, \quad z_2 = \frac{7}{3}, \quad \lambda_2^{(2)} = -x_2 = -\frac{4}{3} \neq -z_2, \quad \lambda_1^{(2)} = \frac{16}{9}.$$

Insbesondere sieht man hieraus, daß die Fälle (a) und (b) nicht beide gemeinsam eintreffen können.

In Fall (b) muß also $x \neq y$ gelten. Einsetzen von $\lambda_2 = -z$ in der rechten Gleichung von (10.43) liefert

$$(y-z)(x-z) = 0.$$

Im Fall $y = z$ erhält man wie in Fall (a):

$$y_3 = z_3 = 2 \quad \text{und} \quad x_3 = 1$$

und

$$y_4 = z_4 = \frac{4}{3}, \quad \text{und} \quad x_4 = \frac{7}{3}.$$

Im verbleibenden Fall $x = z$ erhält man analog

$$x_5 = z_5 = 2 \quad \text{und} \quad y_5 = 1$$

und

$$x_6 = z_6 = \frac{4}{3}, \quad \text{und} \quad y_6 = \frac{7}{3}.$$

Insbesondere folgt

$$(x_j, y_j, z_j) \in N_g := \{(x, y, z) \in \mathbb{R}^3; g(x, y, z) = 0\}.$$

Subtrahieren wir zweimal die Gleichung (10.38) von der ins Quadrat erhobenen Gleichung (10.37), so folgt für alle $(x, y, z) \in N_g$: $x^2 + y^2 + z^2 = 3$. Die Menge N_g ist als nicht leerer Schnitt einer Ebene mit einer Kugel eine Kreislinie in $\mathbb{R}^3$ und damit kompakt. Die Funktion f nimmt also auf N_g sowohl ihr Maximum als auch ihr Minimum an. Wegen

$$f(x_j, y_j, z_j) = 4 \quad \text{für } j = 1, 3, 5 \quad \text{und}$$

$$f(x_j, y_j, z_j) = \frac{112}{27} = 4 + \frac{4}{27} \quad \text{für } j = 2, 4, 6$$

folgt: x_j ist Maximumstelle für $j = 2, 4, 6$ und Minimumstelle für $j = 1, 3, 5$.